Statistische Datenanalyse mit SPSS für Windows

4. Auflage

Springer-Verlag Berlin Heidelberg GmbH

Jürgen Janssen · Wilfried Laatz

Statistische Datenanalyse mit SPSS für Windows

Eine anwendungsorientierte Einführung in das Basissystem und das Modul Exakte Tests

Vierte, neubearbeitete und erweiterte Auflage

Mit 385 Abbildungen und 165 Tabellen

Springer

Jürgen Janssen, Dozent für Volkswirtschaftslehre
Dr. Wilfried Laatz, Professor für Soziologie

HWP-Hamburger Universität
für Wirtschaft und Politik
Von-Melle-Park 9
20146 Hamburg
Deutschland
janssenj@hwp-hamburg.de
laatzw@hwp-hamburg.de

ISBN 978-3-540-44002-4

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Janssen, Jürgen: Statistische Datenanalyse mit SPSS für Windows: eine anwendungsorientierte Einführung in das Basissystem und das Modul exakte Tests; mit 165 Tabellen / Jürgen Janssen; Wilfried Laatz. – 4., neubearb. und erw. Aufl.

ISBN 978-3-540-44002-4 ISBN 978-3-662-10038-7 (eBook)
DOI 10.1007/978-3-662-10038-7

http://www.springer.de

Ursprünglich erschienen bei Springer-Verlag Berlin Heidelberg New York 2003

Umschlaggestaltung: Künkel & Lopka, Heidelberg

SPIN 10888311 42/2202-5 4 3 2 1 0 – Gedruckt auf säurefreiem Papier

Vorwort zur vierten Auflage

Zur Anpassung an die Neuerungen von SPSS für Windows war eine Überarbeitung des Buches notwendig.

Alle Neuerungen bis einschließlich der Version 11 sind in das Buch aufgenommen. In Ergänzung der multivariaten statistischen Verfahren ist nun auch die Multidimensionale Skalierung sowie die Reliabilitätsanalyse enthalten. Für die Varianzanalyse werden neue und erweiterte Verfahren angeboten.

Das bewährte Grundkonzept des Buches wurde beibehalten: Dem Anfänger wird ein leichter Einstieg und dem schon erfahrenen Anwender eine detaillierte und umfassende Nachschlagemöglichkeit gegeben. Die Darstellung ist praxisorientiert mit vielen Beispielen. Die Vorgehensweise bei einer statistischen Auswertung wird gezeigt und die Ergebnisse werden ausführlich kommentiert und erklärt. Dabei werden die statististischen Verfahren mit ihren theoretischen Grundlagen und Voraussetzungen in die Darstellung einbezogen. Neben Daten aus dem ALLBUS (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften) werden unter anderen volkswirtschaftliche Daten, Daten aus der Wahlforschung, der Schuldnerberatung, der Qualitätskontrolle und Medizin verwendet.

Der Service zum Zugang zu den verwendeten Datendateien wird fortgeführt. Man kann die Datendateien entweder per Post oder per Internet beziehen (siehe Anhang B). Die von uns zum Buch eingerichtete Website (Http://www.hwp-hamburg.de/JanssenJ/spss.html) bietet nicht nur einen schnellen Zugang zu den Datendateien, sondern enthält weitere Informationsangebote zu unserem Buch. Man kann dort Ergänzungstexte zum Buch und Übungsaufgaben mit Lösungen finden.

Obwohl sich die Version 11 durch Erweiterungen und Verbesserungen auszeichnet, können auch Anwender früherer Versionen dieses Buch sehr gut nutzen.

Die Gliederung des Buches orientiert sich stark an den Elementen und Menüs des Programms damit der Programmbenutzer sich leicht und schnell zurechtfindet. Darüberhinaus besteht folgende Gliederungsstruktur: Kapitel 1 erläutert die Installation des Programms und gibt weitere Hinweise rund um die Installation. Kapitel 2 dient zur Einführung in die Programmbedienung. Ein Anfänger erhält eine sehr leicht nachvollziehbare Anleitung zum Selbststudium für das Umgehen mit SPSS für Windows. Dabei hat er die Möglichkeit, alle gezeigten grundlegenden Anwendungschritte nachzuvollziehen.

Kapitel 3 bis 7 behandelt das Daten- und Dateienmanagement in SPSS. In diesen Kapiteln werden die Menüs "Datei", "Bearbeiten", "Daten" und "Transformieren" behandelt.

Kapitel 8 bis 24 geht auf alle statistischen Verfahren im Menü "Analysieren" ein.

Kapitel 25 bis 27 befassen sich mit der Erzeugung und Überarbeitung von interaktiven und herkömmlichen Grafiken In Kapitel 26.2 wird anhand eines Beispiels in mehreren Schritten die komplette Überarbeitung einer erzeugten herkömmlichen Grafik für Präsentationszwecke aufgezeigt.

In Kapitel 28 werden weitere Programmelemente sowie Programmfunktionen erklärt. In Kapitel 29 wird die Theorie und praktische Anwendung von Exakte Tests erläutert. Exakte Tests erlaubt für die nichtparametrischen Tests sowie für den Chi-Quadrat-Test im Rahmen von Kreuztabellierung genaue Signifikanzprüfungen. Dieses Ergänzungsmodul ist unverzichtbar, wenn nur kleine oder unausgewogene Stichproben vorliegen.

Unser herzlicher Dank geht an SPSS GmbH Software in München für die Überlassung des Programms sowie für die weitere sehr gute Unterstützung und an den Springer-Verlag für die harmonische Zusammenarbeit. Gerne möchten wir erneut unsere Leser ermuntern und bitten: Schreiben Sie uns, wenn Sie Fehler entdecken oder sonstige Verbesserungsvorschläge haben. Insbesondere möchten wir unsere Leser auf unsere E-Mail-Adressen hinweisen.

Hamburg, im Juni 2002.

Jürgen Janssen
Wilfried Laatz

E-Mail: JanssenJ@hwp-hamburg.de
LaatzW@hwp-hamburg.de

Inhaltsverzeichnis

1 Installieren von SPSS

1.1 Anforderungen an die Hard- und Software

Zur Installation und zum Betrieb des Basis-Systems von SPSS für Windows 11 bestehen folgende Systemanforderungen:

- ❒ Pentium-Prozessor oder Prozessor der Pentiumklasse.
- ❒ Für Windows 98 und Windows ME: mindestens 64 MB Arbeitsspeicher
- ❒ Freier Festplattenspeicher von mindestens 80 MB für das Basissystem.
- ❒ CD-Rom-Laufwerk.
- ❒ Grafikkarte mit einer Mindestauflösung von 800*600 (SVGA).
- ❒ Windows 98, Windows NT 4.0, Windows ME, Windows 2000 oder Windows XP. Windows NT Nutzer sollten Service Pack 5 oder 6 nutzen.

1.2 Die Installation durchführen

Falls Sie eine frühere SPSS-Version (Version 7.5 oder jünger) installiert und Menüs oder Symbolleisten ihren Bedürfnissen angepasst haben und Sie diese Anpassungen erhalten wollen, so sollten Sie SPSS 11 in das gleiche Verzeichnis installieren. Ältere Versionen von SPSS sollten vor der Installation von SPSS 11 deinstalliert werden.

Legen Sie die CD-ROM in das Laufwerk. Nach Einlegen der CD-ROM erscheint durch die AutoPlay Funktion ein Menü mit mehreren Optionen. Wählen Sie „SPSS installieren“ zum Starten der Installation. Die Installation erfolgt weitgehend automatisch. Folgen Sie bitte den Anweisungen auf dem Bildschirm.

Sie können das Installationsprogramm aber auch manuell starten. Wählen Sie dazu im Menü „Start“ von Windows die Option „Ausführen“. Geben Sie im Dialogfeld „Ausführen“ den Befehl d:\setup ein. (Wenn das CD-ROM-Laufwerk nicht Laufwerk D: ist, geben Sie den entsprechenden Laufwerksbuchstaben ein.)

Zu den wichtigsten der beim Installationsvorgang erscheinenden Dialogboxen werden im folgenden einige Hinweise gegeben.

In der Dialogbox „Wählen Sie das Zielverzeichnis“ wird "C:\PROGRAM-ME\SPSS" als Programmverzeichnis vorgeschlagen. Sie können dieses mit „Weiter“ bestätigen oder auch ein anderes Verzeichnis für Ihre Programminstallation wählen.

In der Dialogbox „Informationen zum Anwender“ geben Sie Ihren Namen und eventuell Ihren Firmennamen an. In das Eingabefeld „Seriennummer“ geben Sie die Seriennummer ein. Diese befindet sich auf der Hülle der CD-ROM.

In der Dialogbox „Setup-Typ" kann man zwischen den Installationsarten „Standard", „Minimal" und „Benutzer" wählen. Mit der Wahl legt man fest, ob man programmergänzende Komponenten (und welche) installieren möchte. Wenn Sie hinreichenden Plattenspeicher haben, so wählen Sie „Benutzer" und legen selber fest, welche Komponenten Sie übernehmen möchten. Sie sollten die Hilfedateien (11 MB), den Statistik-Assistenten und falls Sie auch mit der Befehlssprache von SPSS arbeiten möchten, eventuell auch das Syntax-Handbuch im PDF-Format (16 MB) installieren. Sie können das Syntax-Handbuch aber bei Bedarf auch direkt von der CD-ROM aufrufen. Über das Menü Hilfe von SPSS können Sie per Acrobat Reader die über 1400 Seiten des Syntax-Handbuchs zur Nutzung der Befehlssprache am Bildschirm einsehen. Zur Nutzung des Syntax-Handbuchs ist Acrobat Reader 3 oder eine Nachfolgeversion erforderlich. Gegebenenfalls müssen Sie Acrobat Reader von der vorliegenden CD-ROM installieren. In den Hilfedateien sind außerdem umfassende Online-Hilfen für Komponenten und Dialogboxen sowie ein Lernprogramm enthalten. Der Statistik-Assistent ist ein interaktiver Ratgeber, der bei der Auswahl der statistischen Analysen und bei der Erstellung von Grafiken um Hilfe gebeten werden kann. Für den Statistik-Assistenten ist Internetexplorer 4 oder 5 erforderlich. Gegebenenfalls können Sie auch den Internetexplorer von der CD-ROM installieren.

Bei einer Standardinstallation werden Beispieldatendateien, Hilfedateien sowie der Statistik-Assistent installiert, nicht aber das Syntax-Handbuch. Bei der Minimalinstallation werden nur die Dateien installiert, die zum Ausführen von SPSS unabdingbar sind.

In der Dialogbox „Installation: Einzelplatz oder Netzwerk" wird gewählt, ob es sich um eine Einzelplatz- oder eine Netzwerkinstallation handelt. Hier beschränken wir uns auf die Einzelplatzinstallation.

In der Dialogbox „Codes für Produktlizenzen" ist der Lizenzcode für SPSS einzugeben. Nur nach Eingabe des Codes (bzw. Eingabe mehrerer Codes für mehrere Module) werden das Basissystem und eventuell erworbene Zusatzmodule von SPSS lauffähig installiert. Die Lizenznummer (wird von SPSS auf einem gesonderten Blatt geliefert) muss einschließlich der Leerzeichen zwischen den Zahlengruppen eingetippt werden. Die Lizenz ist i.d.R. zeitbegrenzt. Das Ablaufdatum kann man sich über den Befehl „Info" des Menüs „Hilfe" anzeigen lassen. Zur Verlängerung der Lizenzperiode siehe unten.

In der Dialogbox „Optionen auswählen" wählt man die Programmmodule aus, für die man eine Lizenz erworben hat.

1.3 Weitere Hinweise

Hardware key. Manche SPSS-Installationen erfordern einen hardware key (25-Pin-Stecker). Er ist dann der Lieferung der Software beigefügt. Bevor man SPSS startet, sollte man den hardware key auf den Paralleldruckerausgang stecken. Muss man zum Starten anderer Software weitere hardware keys nutzen, sollte der für SPSS als erster gesteckt sein.

Komponenten nach der Installation hinzufügen oder löschen. Wenn Sie bislang nicht installierte Komponenten oder Module installieren wollen, so gehen Sie wie folgt vor:

- ▷ Wie bei einer Erstinstallation wählen Sie nach Einlegen der CD-ROM „SPSS installieren" und folgen den Anweisungen auf dem Bildschirm.
- ▷ In der Dialogbox zur Auswahl von zu installierenden Komponenten können Sie die gewünschten Komponenten oder Optionen auswählen. Falls ein neues Modul hinzugefügt werden soll, müssen Sie für dieses auch einen neuen Lizenzcode eingeben.

Zum Entfernen einer Komponente oder eines Moduls gehen Sie analog vor. In der Dialogbox zur Auswahl von zu installierenden Komponenten entfernen Sie die Markierungspunkte von zu entfernenden Komponenten bzw. Modulen. Alle Komponenten, die erhalten bleiben sollen, müssen markiert sein.

Deinstallieren von SPSS. Um SPSS zu deinstallieren gehen Sie wie bei jedem anderen Windows-Programm vor:

- ▷ Wählen Sie über das Start-Menü „Einstellungen" und „Systemsteuerung".
- ▷ Im Fenster „Systemsteuerung" doppelklicken Sie auf „Software".
- ▷ In der Dialogbox „Eigenschaften von Software" markieren Sie in der Softwareliste „SPSS 11 für Windows" und wählen dann die Schalfläche „Hinzufügen/Entfernen".

Verlängern der Lizenzperiode. Ist die Lizenzperiode abgelaufen und haben Sie die Lizenz für eine weitere Periode erworben, so müssen Sie SPSS nicht erneut installieren. Zur Lizenzverlängerung gehen Sie wie folgt vor:

- ▷ Öffnen Sie über das Start-Menü die MS-DOS-Eingabeaufforderung.
- ▷ Wechseln Sie in das Verzeichnis, in das SPSS installiert ist.
- ▷ Doppelklicken Sie auf die Datei „Licrenew.exe".
- ▷ In die sich öffnende Dialogbox zur Eingabe des Lizenzcodes geben Sie den neuen Lizenzcode ein.

Daten aus Datenbanken einlesen. Für den Fall, dass man mit SPSS auf Daten in Datenbanken zugreifen möchte, muss man vorher von der vorliegenden CD-ROM die Menüoption „SPSS Data Access Pack installieren" aufrufen und kann dann die gewünschten ODBC-Treiber (SPSS Data Access Pack) installieren. Zur Installation von Datenbanktreiber für Microsoft wählen Sie „CD-ROM durchsuchen", öffnen das Verzeichnis „Microsoft Data Access Pack" und starten die Installation durch Doppelklicken auf das entsprechende Anwendungsprogramm. (⇨ Kap. 6.1.3.2). Für weitere Informationen sollten Sie die im Verzeichnis „Installationsanweisungen" liegende Instruktionsdokumente auf der vorliegenden CD-ROM lesen.

Installieren mehrerer Versionen von SPSS. Ab der Version SPSS 7.5.2 können mehrere Versionen auf einem PC installiert und auch ausgeführt werden. Es wird aber nicht empfohlen.

Der erste Start von SPSS. Wird SPSS zum erstenmal nach dem Installieren gestartet, so öffnet sich die Dialogbox "SPSS Starteinstellungen" mit Voreinstellungen. Klicken auf "OK" bestätigt diese. Diese Voreinstellungen kann man jederzeit mit der Befehlsfolge "Bearbeiten", "Optionen" ändern (⇨ Kap. 28.5).

Sie starten SPSS für Windows durch die Befehlsfolge „Start", „Programme" und Auswahl von „SPSS 11 für Windows" in der Liste der Programme (oder durch Anklicken des SPSS-Programmsymbols auf dem Desktop). Per Voreinstellung erscheint dann der SPSS Daten-Editor (⇨ Abb. 2.1). Beim ersten Mal ist es überlagert von dem Fenster „SPSS für Windows" (⇨ Abb. 1.2).

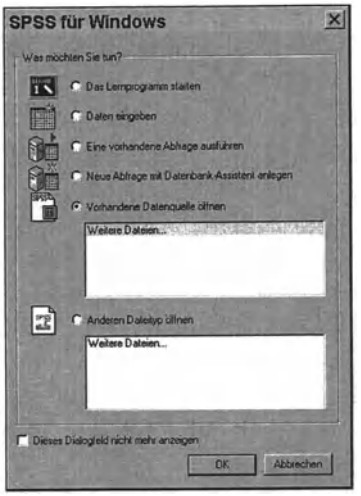


Abb. 1.2. Eröffnungs-Dialogbox „SPSS für Windows"

In ihm können Sie auswählen, was Sie als nächstes tun möchten: „Das Lernprogramm starten", „Daten eingeben", "Eine vorhandene Abfrage ausführen", „Eine neue Abfrage mit dem Datenbank-Assistenten erstellen", „Vorhandene Datenquelle öffnen". Die letzte Option ist voreingestellt. Unter ihr findet sich ein Auswahlfenster mit den zuletzt verwendeten Dateien. Diese Option wird man in der Regel verwenden, um eine Datendatei auszuwählen. Entweder wählt man durch Anklicken ihres Namens eine der zuletzt verwendeten Dateien oder aber man lädt eine andere Datei in der Dialogbox „Datei öffnen", die nach Anklicken von „Weitere Dateien..." erscheint.

Wenn Sie es wünschen, können Sie durch Anklicken des Kontrollkästchens „Dieses Dialogfeld nicht mehr anzeigen" dafür sorgen, dass Sie in Zukunft bei Öffnung von SPSS direkt im Daten-Editorfenster landen. Wir empfehlen dies, denn alle im Eröffnungsfenster angebotenen Aktionen können Sie auch auf andere Weise ausführen.

2 Einführende Übungen mit SPSS

Mit diesem Kapitel werden zwei Ziele angestrebt:

- ❒ Einführen in das Arbeiten mit der Oberfläche von SPSS für Windows.
- ❒ Vermitteln grundlegender Anwendungsschritte für die Erstellung und statistische Auswertung von Datendateien.

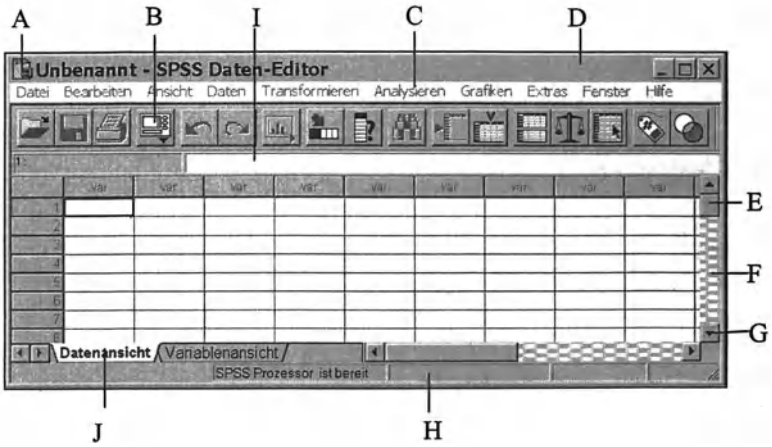
Wir gehen davon aus, dass Sie mit einer Maus arbeiten. Außerdem sollten Sie den Umgang mit der Windows-Oberfläche weitgehend beherrschen. Unter der Windows-Oberfläche kann man die meisten Aktionen auf verschiedene Weise ausführen. Wir werden in der Regel nur eine (die vermutlich gebräuchlichste) benutzen. Bei den ersten Anwendungen werden sie etwas ausführlicher erläutert (z.B. zeigen Sie mit der Maus auf die Option „Datei", und klicken Sie den linken Mauszeiger), später wird nur noch die Kurzform verwendet (*Beispiel:* Wählen Sie die Option „Datei", oder: Wählen Sie „Datei"). Die Maus bestimmt die Position des Zeigers (Cursors) auf dem Bildschirm. Er hat gewöhnlich die Form eines Pfeiles, ändert diese aber bei den verschiedenen Anwendungen. So nimmt er in einem Eingabefeld die Form einer senkrechten Linie an. Durch Verschieben der Maus ändert man die Position. Befindet sich der Cursor an der gewünschten Position (z.B. auf einem Befehl, in einem Feld, auf einer Schaltfläche), kann man entweder durch „Klicken" (einmaliges kurzes Drücken) der linken Taste oder durch „Doppelklicken" (zweimaliges kurzes Drücken der linken Taste) eine entsprechende Aktion auslösen (z.B. einen Befehl starten, eine Dialogbox öffnen oder den Cursor in ein Eingabefeld platzieren). Außerdem ist auch das „Ziehen" des Cursors von Bedeutung (z.B. um ein Fenster zu verschieben oder mehrere Variablen gleichzeitig zu markieren). Hierzu muss der Cursor auf eine festgelegte Stelle platziert werden. Die linke Maustaste wird gedrückt und festgehalten. Dann wird der Cursor durch Bewegen der Maus auf eine gewünschte Stelle gezogen. Ist sie erreicht, wird die Maustaste losgelassen. Von „Markieren" sprechen wir, wenn – entweder durch Anklicken einer Option oder eines Feldes oder durch Ziehen des Cursors über mehrere Felder – Optionen oder größere Textbereiche andersfarbig unterlegt werden.

Wenn in Zukunft angegeben wird, dass ein Menüelement durch Doppelklick gewählt werden soll, ist in der Regel immer auch statt dessen die Auswahl durch Markieren des Menüelements und das Drücken der Eingabetaste möglich.

Außerdem benutzen wir weitestgehend die Voreinstellungen von SPSS. (Änderungsmöglichkeiten ⇨ Kap. 28.5).

2.1 Die Oberfläche von SPSS für Windows

Starten Sie SPSS für Windows (⇨ Kap. 1.3). In der Eröffnungsdialogbox (Abb. 1.2) markieren Sie den Kreis vor der Option „Daten eingeben" und klicken auf „OK". Es öffnet sich das Daten-Editorfenster.



A SPSS-Systemmenüfeld
B Symbolleiste mit Symbolen
C Menüleiste mit Menüs
D Titelleiste
E Bildrollfeld
F Bildlaufleiste
G Bildrollpfeil
H Statusleiste
I Zelleneditorzeile
J Registerblatt

Abb. 2.1. SPSS Daten-Editor

SPSS arbeitet mit fünf Fenstern. Die ersten beiden Fenster wird man bei der Arbeit mit SPSS stets benötigen.

- *Daten-Editor* (mit den Registerblättern „Datenansicht" und „Variablenansicht". Es öffnet sich per Voreinstellung mit dem Registerblatt „Datenansicht" beim Start des Programms (Titelleiste enthält: Name der Datendatei, zuerst „Unbenannt" und den Namen des Fensters „SPSS Daten-Editor"). In diesem Fenster kann man Daten-Dateien erstellen oder öffnen, einsehen und ändern. (Das Registerblatt „Variablenansicht" dient der Datendefinition und wird in 2.3 näher betrachtet.)
- *SPSS Viewer* (Ausgabefenster). (Titelleiste enthält: Name der Ausgabedatei, zuerst „Ausgabe1" und „SPSS Viewer"). In ihm werden Ergebnis (Output) der Arbeit mit SPSS ausgegeben. Interaktive Grafiken können darin direkt bearbeitet werden (⇨ Kap. 25). Es ist zweigeteilt. Links enthält er das Gliederungsfenster, rechts die eigentliche Ausgabe. Man kann diese editieren und für den weiteren Gebrauch in Dateien speichern. Man kann auch weitere Ausgabefenster öffnen (⇨ näher unten und Kap. 4.1.1). (Speziell für Textausgaben existieren auch noch Textviewer und Text-Editor, auf die wir hier nicht eingehen).

Neben diesen beiden Fenstern gibt es drei weitere Fenster:

- *Diagramm-Editor* (Grafikfenster). Es wird benötigt, wenn man die im Ausgabefenster anfallenden herkömmliche Grafiken weiter bearbeiten möchte (andere Farben, Schriftarten etc.). Siehe hierzu ⇨ Kap. 27.1.
- *Pivot-Tabellen-Editor*. In diesem Fenster können Pivot-Tabellen weiter bearbeitet werden.
- *Syntax-Editor*. In dieses Fenster können die in den Dialogboxen ausgewählten Befehle in Form von Befehlstexten übertragen werden. Diese können darin editiert und durch Befehlselemente ergänzt werden, die in den Menüs nicht verfügbar sind. Es ist möglich, eine Befehlsdatei zu erstellen, zu speichern und zu starten.
- *Skript-Editor*. In ihm können SPSS-Skripte in einer speziellen Skriptsprache erstellt, gespeichert und gestartet werden. Diese dienen hauptsächlich zur Gestaltung des Outputs.

Diagramm-Editor und Pivot-Tabellen-Editor öffnen sich durch Doppelklick auf entsprechende Objekte im graphisch orientierten SPSS-Viewer (sie können nicht wie andere Fenster über das Menü „Datei" geöffnet werden). Es stehen dort besondere Bearbeitungsfunktionen zur Verfügung, die an entsprechender Stelle dargestellt werden. Sie unterscheiden sich wie auch der Skript-Editor im Aufbau deutlich von den anderen Fenstern. Die folgenden Ausführungen beziehen sich daher nicht auf sie.

Außer dem Daten-Editor müssen alle anderen Arten von Fenstern erst geöffnet werden. Dies geschieht entweder beim Ausführen entsprechender Befehle automatisch oder über die Menüpunkte „Datei", „Neu" bzw. „Datei", „Öffnen" (nicht bei Grafik- und Pivot-Tabellen-Editor). Das Fenster, in dem jeweils im Vordergrund gearbeitet werden kann, nennt man das *aktive* Fenster. Nach dem Start von SPSS ist dieses der Daten-Editor. Will man in einem anderen Fenster arbeiten, muss es zum aktiven Fenster werden. Das geschieht entweder bei Ausführung eines Befehls automatisch oder indem man dieses Fenster anwählt. Das ist auf unterschiedliche Art möglich. Sie können das Menü „Fenster" anklicken. Es öffnet sich dann eine Drop-Down-Liste, die im unteren Teil alle z.Z. geöffneten Fenster anzeigt. Das aktive Fenster ist durch ein Häkchen vor dem Namen gekennzeichnet. Wenn Sie den Namen des gewünschten Fensters anklicken, wird dieses geöffnet. Alle z.Z. geöffneten Fenster werden auch am unteren Rand des Bildschirms als Registerkarten angezeigt. Das Anklicken der entsprechenden Registerkarte macht das Fenster aktiv. Überlappen sich die Fenster auf dem Desktop (falls sie nicht auf volle Bildschirmgröße eingestellt sind), kann man ein Fenster auch durch Anklicken irgendeiner freien Stelle dieses Fensters öffnen. Schalten Sie auf die verschiedenen Weisen einmal zwischen einem Dateneditor-Fenster und einem Ausgabefenster hin und her. Dafür öffnen Sie zunächst einmal ein Ausgabefenster, indem Sie mit dem Cursor auf das Menü „Datei" zeigen und die linke Maustaste drücken. In der sich dann öffnenden Drop-Down-Liste zeigen Sie zunächst auf „Neu", in der dann sich öffnenden Liste auf „Ausgabe". Hier klicken Sie auf die linke Maustaste. Ein Ausgabefenster „Ausgabe1" ist geöffnet.

Es kann nur ein Dateneditorfenster geöffnet werden. Bei allen anderen Fenstertypen kann man zusätzliche Fenster anfordern, so dass mehrere gleichzeitig geöffnet sind. Dadurch wird es möglich, verschiedene Ausgabeergebnisse (oder eine Folge von Befehlen) einer Sitzung gezielt in unterschiedliche Dateien zu leiten. In welches Fenster z.B. die Ausgabe (Output) gelenkt wird, bestimmt der Nutzer, indem er ein Ausgabefenster zum *Hauptfenster* (designierten Fenster) erklärt. Das geschieht, indem man, während dieses Fenster aktiv ist, das hervorgehobene Symbol anklickt. (Alternativ wählen Sie „Extras" und „Hauptfenster".) Das Symbol wird dann deaktiviert. In der Statusleiste des designierte Fensters erscheint aber gleichzeitig zu dessen Kennzeichnung ein hervorgehobenes Ausrufezeichen.

Im folgenden werden wir uns zunächst einmal im Daten-Editor und Ausgabefenster bewegen und einige Menüs des Dateneditors erkunden.

Die Fenster kann man in der bei Windows-Programmen üblichen Art verkleinern, vergrößern, in Symbole umwandeln und wiederherstellen. Probieren Sie das einmal am „Dateneditorfenster" aus. Zur Veränderung der Größe setzten Sie den Cursor auf eine Seite des Rahmens des Fensters (dass Sie sich an der richtigen Stelle befinden, erkennen Sie daran, dass der Cursor seine Form in einen Doppelpfeil ändert). Dann ziehen Sie den Cursor bei Festhalten der linken Maustaste und beobachten, wie sich das Fenster in der Breite verkleinert oder vergrößert. Die Größe ist fixiert, wenn Sie die Maustaste loslassen. Auf dieselbe Weise können Sie auch die Höhe verändern. Höhe und Breite ändert man gleichzeitig, indem man den Cursor auf eine der Ecken des Rahmens setzt und entsprechend zieht. Eine andere Möglichkeit besteht darin, ein Fenster den ganzen Bildschirm einnehmen zu lassen. Dazu können Sie u.a. das SPSS-Systemmenüfeld (⇨ Abb. 2.1) anklicken und darauf in der Liste die Auswahlmöglichkeit „Maximieren" anklicken. Wiederhergestellt wird die alte Größe durch Anklicken der Auswahlmöglichkeit „Wiederherstellen" im selben Menü. Man kann das Fenster auch zu einer Registerkarte (am unteren Rand des Bildschirms) verkleinern (und damit gleichzeitig deaktivieren), indem man den Menüpunkt „Minimieren" wählt. Durch Doppelklick auf die Registerkarte kann ein Fenster wiederhergestellt werden. Auch die Symbole in der rechten Ecke der Titelleiste dienen diesem Zweck. Anklicken von maximiert das Fenster, gleichzeitig wandelt sich das Symbol in . Anklicken dieses Symbols stellt den alten Zustand wieder her. Anklicken von minimiert das Fenster zur Registerkarte, schließt das Programm.

Nimmt der Inhalt eines Fensters mehr Raum ein, als auf dem Bildschirm angezeigt, kann man den Bildschirminhalt mit Hilfe der Bildlaufleisten verschieben (*scrollen)*. Diese befinden sich am rechten und unteren Rand des Bildschirms. Am oberen und unteren (bzw. linken und rechten) Ende befindet sich jeweils ein Pfeil, der *Bildrollpfeil*. Außerdem enthalten die Bildlaufleisten ein kleines Kästchen, das *Bildrollfeld* (⇨ Abb. 2.1). Klicken Sie einige Male den Pfeil am unteren Ende des Dateneditorfensters an, und beachten Sie die Zahlen am linken Rand dieses Fensters. Sie erkennen, dass mit jedem Klick der Fensterinhalt um eine Zeile nach unten verschoben wird. Halten Sie die Taste dabei gedrückt, läuft das Bild automatisch weiter nach unten. Das Bildrollfeld zeigt an, an welcher Stelle man sich in einer Datei befindet. Es ist bei der bisherigen Übung etwas nach unten gewandert. Außerdem kann man sich mit seiner Hilfe schneller im Fenster bewegen. Man setzt

den Cursor dazu auf das Bildrollfeld und zieht es an die gewünschte Stelle. (*Anmerkung*: Man kann auch durch Drücken der Pfeil-Tasten oder durch Drücken der <Bild auf> bzw. <Bild ab>–Tasten der Tastatur das Bild rollen).

Sollten Sie noch Schwierigkeiten im Umgang mit der Windows-Oberfläche haben, können Sie das Windows-Handbuch zu Rate ziehen.

2.2 Einführen in die Benutzung von Menüs und Symbolleisten

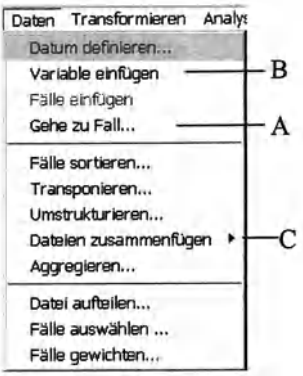
Jedes Fenster enthält eine eigene Menüleiste und eine oder zwei eigene Symbolleisten. In dieser Einführung werden die Menüs und die Symbolleiste des Dateneditorfensters in den Vordergrund gestellt. Im Aufbauprinzip und auch in großen Teilen der Menüs entsprechen sich aber alle Fenster.

Menüs und Dialogboxen des Daten-Editors. In der Menüleiste gibt es folgende Menüs:

- *Datei.* Es dient zum Erstellen, Öffnen, Importieren und Speichern jeder Art von SPSS-Dateien. Daneben ist an Datendateien der Import von Dateien zahlreicher Tabellenkalkulations- oder Datenbankprogrammen, von Dateien anderer Statistikprogamme sowie von ASCII-Dateien möglich. Darüber hinaus dient das Menü der Information über die Datendatei und dem Druck einer Datendatei. Auch andere Dateien (Syntax-, Ausgabe-, Skript-Dateien etc.) können hier erstellt werden.
- *Bearbeiten.* Dient zum Löschen und Kopieren, Einfügen und Suchen von Daten. Der Menüpunkt Optionen führt zu den Dialogboxen für die Grundeinstellung der verschiedenen SPSS-Bereiche.
- *Ansicht.* Ermöglicht es, Status- und Symbolleisten aus- oder einzublenden, die Symbolgröße und das Schriftbild der Daten zu bestimmen, Gitterlinien- ein oder auszublenden, Werte als Labels oder Wert anzeigen zu lassen. Schließlich kann man mit dem letzten Menüpunkt zwischen Daten- und Variablenansicht umschalten.
- *Daten.* Dient der Definition von Datumsvariablen, dem Einfügen von Variablen und Fällen sowie der globalen Änderung von SPSS-Datendateien, z.B. Kombinieren von Dateien, Transponieren und Umstrukturieren der Datenmatrix (von Variablen in Fälle und umgekehrt), Aggregieren sowie Auswahl von Teilgruppen. (Die Änderungen sind temporär, wenn sie nicht ausdrücklich gespeichert werden.)
- *Transformieren.* Veränderung von Variablen und Berechnung neuer. (Die Änderungen sind temporär, wenn sie nicht ausdrücklich gespeichert werden.)
- *Analysieren.* Dient der Auswahl statistischer Verfahren und stellt den eigentlichen Kern des Programms dar.
- *Grafiken.* Dient zur Erzeugung verschiedener Arten von Diagrammen und Grafiken. Diese können im Diagramm-Editor vielfältig gestaltet werden.
- *Extras.* Sammlung verschiedener Optionen. Informationen über SPSS-Datendateien, Arbeiten mit Datensets und Skripten, Erweitern der Menüs durch den Nutzer (im Viewer und Syntax-Editor auch zur Definition des Hauptfensters).

- *Fenster.* Auswahl des aktiven SPSS-Fensters. Minimieren der Fenster.
- *Hilfe.* Bietet ein Hilfefenster. Es ist nach den (nicht ganz glücklichen) Regeln eines Standard-Microsoft-Hilfefensters aufgebaut.

Diese Menüs sind (mit Ausnahme von „Daten" und „Transformieren" in allen Fenstern identisch. (Im Diagramm-Editor fehlt zudem das Menü „Fenster".) Daher können alle Grundfunktionen in allen Fenstern aufgerufen werden. Andere haben dieselbe Bezeichnung und im Grundsatz dieselben Funktionen, sind aber hinsichtlich der verfügbaren Optionen dem jeweiligen Fenster angepasst: „Datei", „Bearbeiten", „Extras". Jedes Fenster hat auch einige, nur in ihm enthaltene, spezielle Menüs. Im Dateneditor sind dies „Daten" und „Transformieren".



A Option, die zu einer Dialogbox führt (mit Pünktchen)
B Direkt ausführbarer Befehl (ohne Pünktchen)
C Option, die zu einem Untermenü führt (mit Pfeil)

Abb. 2.2. Drop-Down-Liste des Menüs „Daten"

Die Menüs in der Menüleiste des Dateneditor-Fensters kann man nutzen oder auch nur erkunden, indem man mit der Maus das gewünschte Menü anklickt. Wir versuchen das zunächst einmal mit dem Menü „Daten". Klicken Sie den Menünamen an. Dann öffnet sich die in Abb. 2.2 dargestellte *Drop-Down-Liste*. Sie zeigt die in diesem Menü verfügbaren Auswahlmöglichkeiten, wir sprechen auch von Optionen oder Befehlen. In diesem Falle sind es 12 Optionen wie „Datum definieren...", „Variable einfügen". Davon ist eine („Fälle einfügen") nur schwach angezeigt. Die fett angezeigten Optionen sind z.Z. aufrufbar, die anderen nicht. Ihr Aufruf setzt bestimmte Bedingungen voraus, die z.Z. noch nicht gegeben sind. Dies gilt auch für einige andere nicht unmittelbar ausführbare Befehle (z.B. „Fälle sortieren"). Wählt man diese an, so wird in einem Drop-Down-Fenster mitgeteilt, dass dieser Befehl nicht ausführbar ist und welche Voraussetzung fehlt. Führen Sie den Cursor auf die Option „Fälle einfügen" und klicken Sie auf die linke Maustaste. Es passiert nichts. Wiederholen Sie das bei der Option „Fälle sortieren...". Es öffnet sich ein Drop-Down–Fenster mit dem Warnhinweis. Unter den fett angezeigten Optionen werden einige nur mit Namen (z.B. „Variable einfügen"), andere mit Namen

und drei Pünktchen (z.B. „Datum definieren...") angezeigt. Im ersten Falle bedeutet das, dass der Befehl direkt ausgeführt wird. Eine Übung möge dies verdeutlichen: Setzen Sie den Cursor auf die Option „Variable einfügen", und drücken Sie die linke Maustaste. Der Befehl wird direkt ausgeführt. Die Drop-Down-Liste verschwindet und über der ersten Spalte des Dateneditorfensters erscheint der Name „VAR00001". Bei Auswahl eines Befehls mit Pünktchen öffnet sich eine *Dialogbox*. Der Befehl „Gehe zu Fall..." öffnet z.B. eine gleichnamige Dialogbox, in der die Fallnummer eingegeben und der entsprechende Fall angesprungen werden kann. Eine Dialogbox enthält meistens folgende grundlegende Bestandteile (⇨ Abb. 2.3)[1]:

- *Quellvariablen-* und *Auswahlvariablenliste* (in allen Dialogboxen, mit denen Prozeduren ausgewählt werden). Die Quellvariablenliste ist die Liste aller Variablen in der Datendatei (bzw. im verwendeten Datenset). Die Auswahlvariablenliste enthält die Variablen, die für eine statistische Auswertung genutzt werden sollen. Sie werden durch Markieren der Variablen in der Quellvariablenliste und anschließendem Klicken auf einen Pfeilschalter [▸] oder durch Doppelklick in dafür vorgesehene Eingabefelder der Auswahlliste übertragen.
- *Informations-, Eingabe- und Auswahlfelder.* Wählen Sie einmal das Menü „Datei", und setzen Sie den Cursor auf die Option „Öffnen. Es erscheint eine Dialogbox (⇨ Abb. 2.5). In ihr befindet sich ein Eingabefeld „Dateiname". In ein solches Eingabefeld ist gewöhnlich etwas einzutragen (hier wäre es ein Name einer zu öffnenden Daei). Mitunter gibt es auch ein damit verbundenes Auswahlfeld (⇨ Erläuterungen zu Abb. 2.5), in dem man aus einer Drop-Down-Liste eine Option auswählen kann. in manchen Dialogboxen findet man auch reine Informationsfelder, die interessierende Informationen, z.B. zur Definition einer Variablen enthalten.
- *Befehlsschaltflächen.* Klickt man diese mit der Maus an, so wird ein Befehl abgeschickt.

Folgende Befehlsschaltflächen (ohne Pünktchen am Ende) führen zur unmittelbaren Befehlsausführung und sind immer vorhanden (⇨ Abb. 2.3):

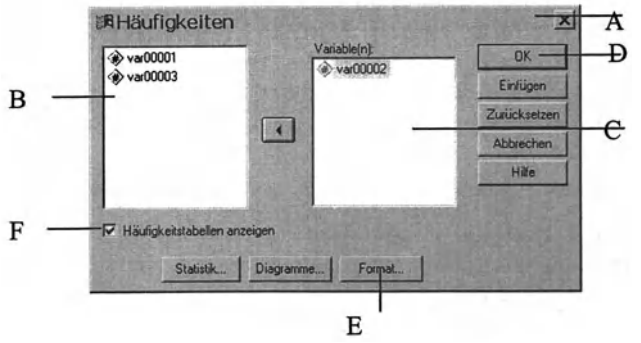
- *OK.* Bestätigt die in der Dialogbox gemachten Angaben und führt die gewünschte Aufgabe aus.
- *Abbrechen.* Damit bricht man die Eingabe in der Dialogbox ab und kehrt zum Ausgangsmenü zurück. Alle Änderungen der Dialogboxeinstellung werden aufgehoben.
- *Hilfe.* Damit fordert man eine kontextbezogene Hilfe im Standardformat von MS Windows an.

In vielen Dialogboxen, insbesondere zur Durchführung von statistischen Auswertungen und zur Erzeugung von Grafiken, gibt es folgende weitere Schaltflächen:

- *Zurücksetzen.* Damit werden schon in der Dialogbox eingegebene Angaben rückgängig gemacht, so dass neue eingegeben werden können, ohne die Dialogbox zu verlassen.

[1] Um die Dialogboxen erkunden zu können, ist es vorteilhaft, wenn Sie durch Eingabe einiger beliebiger Zahlen in mehreren Spalten des Editors eine kleine Datendatei erzeugen.

- *Einfügen*. Nach Anklicken wird der Befehl des Menüs in der Befehlssprache von SPSS ins Syntaxfenster übertragen und dieses aktiviert.



A Dialogbox: Titelleiste
B Quellvariablenliste
C Auswahlvariablenliste
D Schaltfläche, die zu einer sofortigen Ausführung des Befehls führt (ohne Pünktchen)
E Schaltfläche, die zu einer Unterdialogbox führt (mit Pünktchen)
F Kontrollkästchen mit eingeschalteter Option

Abb. 2.3. Dialogbox „Häufigkeiten"

Unterdialogboxen. Neben den genannten Schaltflächen können in Dialogboxen auch Schaltflächen mit Pünktchen vorkommen, z.B. die Schaltflächen „Statistik..." und „Diagramme..." (⇨ Abb. 2.3). Durch Anklicken dieser Schaltflächen werden weitere Dialogboxen (Unterdialogboxen) geöffnet, die zusätzliche Spezifizierungen der gewünschten durchzuführenden Aufgabenstellung erlauben.

Eine aus einer Dialogbox durch Klicken einer Schaltfläche mit Pünktchen (z.B. „Diagramme..." geöffnete (Unter-)Dialogbox hat meistens neben den oben erläuterten Eingabefeldern und Schaltflächen weitere Elemente, mit denen man Spezifizierungen einer Aufgabenstellung vornehmen kann:

❒ *Optionsschalter*. Mit diesen erfolgt eine Auswahl aus einander ausschließenden Optionen. Eine Übung möge diese veranschaulichen[2]: Wählen Sie im Fenster „Häufigkeiten" (sie gelangen dorthin mit „Analysieren", „Deskriptive Statistiken", „Häufigkeiten") die Schaltfläche „Diagramme ...". Es öffnet sich die in Abb. 2.4 dargestellte (Unter-)Dialogbox, in der u.a. in der Gruppe Diagrammtyp verschiedene Optionen mit einem Kreis davor angeführt sind. Einen solchen Kreis bezeichnet man als Optionsschalter. Einer dieser Kreise ist mit einem schwarzen Punkt gekennzeichnet, im Beispiel „Keiner". Damit ist die Option „Keiner" eingestellt (d.h. es wird kein Diagramm erzeugt). Durch Anklicken eines Optionsschalters wählt man die gewünschte Option aus. Es kann nur eine Option gewählt werden.

[2] Setzt voraus, dass Sie einige wenige Daten im Daten-Editor eingegeben haben.

❒ *Kontrollkästchen.* Damit können gleichzeitig mehrere Optionen ausgewählt werden. Ein Kontrollkästchen finden Sie z.B. am unteren Rand der Dialogbox „Häufigkeiten" (⇨ Abb. 2.3). Eine ganze Reihe von Kontrollkästchen finden Sie in der Unter-Dialogbox „Häufigkeiten: Statistiken", in die Sie durch Anklicken der Schaltfläche „Statistiken..." in der Dialogbox „Häufigkeiten" gelangen. Hier können Sie durch Anklicken der Kästchen beliebig viele Maßzahlen zur Berechnung auswählen. Im gewählten Kästchen erscheint jeweils ein Häkchen. Durch erneutes Anklicken können Sie dieses wieder ausschalten.

A Optionsschalter eingeschaltet
B Optionsschalter ausgeschaltet

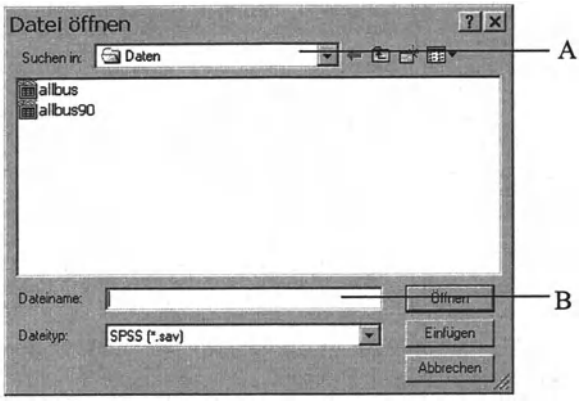
Abb. 2.4. Dialogbox „Häufigkeiten: Diagramme:"

❒ *Weiter.* Neben den bekannten Befehlsschaltflächen „Abbrechen" und „Hilfe" enthalten viele Unterdialogboxen die Schaltfläche „Weiter". Durch Klicken auf diese Schaltfläche (⇨ Abb. 2.4) bestätigt man die ausgewählten Angaben und kehrt zur Ausgangsdialogbox zurück.

❒ *Auswahlfeld.* Die in Abb. 2.5 dargestellte Dialogbox hat ein Auswahlfeld „Suchen in:". Klicken Sie auf den Pfeil neben dem Auswahlfeld. Es öffnet sich dann ein Fenster mit einer Auswahlliste der verfügbaren Verzeichnis. Klicken Sie eines an, erscheint in dem darunter liegenden Auswahlfenster wiederum eine Auswahlliste aller dort verfügbaren Dateien des eingestellten Dateityps. Nach Anklicken einer dieser Dateien, erscheint sie in der Auswahlliste „Dateiname".

Untermenüs. Manche Menüs der Menüleiste enthalten *Untermenüs*. Wenn Sie die schon die Dialogbox „Häufigkeiten" geöffnet haben, kennen Sie das bereits. Öffnen Sie zur Verdeutlichung nun noch einmal das Menü „Analysieren". Sie sehen, dass hier alle Optionen mit einem Pfeil am rechten Rand gekennzeichnet sind. Das bedeutet, dass in den Menüs weitere Untermenüs vorhanden sind. Wählen Sie die Option „Deskriptive Statistiken ▷ ". Es öffnet sich ein weiteres Menü mit mehreren Optionen, u.a. „Häufigkeiten ...". Durch Auswahl von „Abbrechen" gelangen Sie in die Menüleiste zurück.

Gehen Sie nun zur Menüleiste zurück, und öffnen Sie als letztes das Menü „Bearbeiten". Hier ist neu, dass zu den verschiedenen Optionen auch Tastenkombinationen angegeben sind, mit denen die Menüs gewählt werden können. So die Option „Einfügen" mit <Strg> + <V>. Außerdem sind sie durch Querstriche in Gruppen unterteilt. Die erste Gruppe umfasst Optionen zum Ausschneiden, Einsetzen, Kopieren von Texten usw., die zweite Gruppe eine Option zum Suchen von Textstellen, die dritte die Wahlmöglichkeit „Optionen", die zu einer Dialogbox für die Gestaltung der Einstellungen von SPSS führt. Erforschen Sie auf die angegebene Weise ruhig alle Menüs.



A Auswahlfeld mit Drop-Down-Liste (zum Öffnen Pfeil anklicken)
B Eingabefeld

Abb. 2.5. Dialogbox „Datei öffnen"

Hilfe zu Elementen der Dialogbox. Außer der kontextsensitiven Hilfe zu einer Dialogbox, kann man auch Hilfetexte für die einzelnen Elemente einer Dialogbox aufrufen. Dazu klickt man mit der *rechten* Maustaste auf das Element. Es öffnet sich ein Anzeigefenster mit einer Erklärung dieses Elements. In Variablenlisten dagegen öffnet sich dann eine Auswahlliste, in der man außer der wenig informativen „Direkthilfe" auch „Info zu den Variablen" anwählen kann. Bei deren Auswahl erhält man eine Beschreibungen der gerade markierten Variablen.

Symbolleiste benutzen. Alle Fenster verfügen auch über eigene Symbolleisten. Viele häufig benutzte Funktionen lassen sich über die Symbolleiste aufrufen. Man erspart sich dann den Weg über die Menüs. Im Dialogfenster „Ansicht", „Symbolleisten" sind die im Fenster verfügbaren Symbolleisten angeführt. Durch Anklicken des Kontrollkästchen vor dem Namen der Symbolleiste kann man deren Anzeige aus- und einschalten. Klicken Sie das Kontrollkästchen „Große Schaltflächen" an, werden die Symbole in der Leiste größer und damit besser erkennbar angezeigt. Die Symbole erklären ihre Funktion leider nicht hinreichend selbst. Berührt der Cursor aber eines davon, so wird dessen Funktion gleichzeitig sowohl in der Statuszeile als auch in einem Drop-Down-Fenster am Symbol selbst beschrie-

ben. Die Symbolleiste lässt sich auch beliebig verschieben. Klicken Sie dazu an irgendeiner Stelle auf die Leiste (aber nicht auf ein Symbol) und ziehen Sie diese mit gedrückter Taste an die gewünschte Stelle. Mit Loslassen der Taste ist die Symbolleiste fixiert. Um eine Aktion auszuführen, klickt man auf das zuständige Symbol.

Klicken Sie auf ein Symbol, dann werden einige der Aktionen sofort ausgeführt. In vielen Fällen öffnet sich jedoch eine Dialogbox. Sie ist identisch mit der Dialogbox, in die Sie das entsprechende Menü auch führt. Die Dialogbox wird in der üblichen Weise benutzt.

Die folgende Abbildung gibt einen Überblick über die *Symbole* des Dateneditorfensters. Anschließend werden deren Funktionen erläutert.

Datei öffnen. Öffnet eine Dialogbox zur Auswahl einer Datei. Es können nur Dateien des dem derzeit aktiven Fenster entsprechenden Typs geöffnet werden.

Datei speichern. Speichert den Inhalt des derzeit aktiven Fensters. Handelt es sich um eine neue Datei, öffnet sich die Dialogbox „Datei speichern unter".

Drucken. Öffnet eine Dialogbox zum Drucken des Inhalts des aktiven Fensters. Auch eine Auswahl kann gedruckt werden.

Zuletzt verwendete Dialogfelder. Listet die zuletzt geöffneten Dialogboxen zur Auswahl auf. Man kann die gewünschte Dialogbox direkt anspringen. (Die Zahl der Dialogbox kann bis 9 – Voreinstellung – reichen.)

Rückgängig machen. Macht die letzte Dateneingabe rückgängig und springt in die entsprechende Zelle der Datenmatrix zurück.

Wiederholen. Wiederholt eine rückgängig gemachte Dateneingabe.

Gehe zu Diagramm. Ist aktiv, wenn der Diagramm-Editor geöffnet ist. Springt direkt in das Grafikfenster.

Gehe zu Fall. Öffnet eine Dialogbox, aus der man zu einer bestimmten Fallnummer im Dateneditorfenster springen kann. (Fallnummer ist die von SPSS automatisch vergebene Nummer.)

Variablen. Öffnet das Fenster „Variablen" mit einer Variablenliste und Variablenbeschreibung. (Dasselbe bewirkt die Befehlsfolge „Extras", „Variablen...".) Eine ausgewählte Variable kann im Dateneditor direkt angesprungen werden.

Suchen. Öffnet eine Dialogbox, aus der man, ausgehend von einer markierten Zelle, innerhalb der ausgewählten Spalte bestimmte Werte im Dateneditorfenster suchen kann.

Fälle einfügen. Fügt vor einer markierten Zeile einen neuen Fall ein. Dasselbe bewirkt die Befehlsfolge „Daten", „Fälle einfügen".

Variable einfügen. Fügt vor einer markierten Spalte eine neue Variable ein. Dasselbe bewirkt die Befehlsfolge „Daten", „Variable einfügen".

Datei aufteilen. Öffnet eine Dialogbox, mit der eine Datei in Gruppen aufgeteilt werden kann. Dasselbe bewirkt die Befehlsfolge „Daten", „Datei aufteilen...".

Fälle gewichten. Öffnet eine Dialogbox, mit der die Fälle der Datendatei gewichtet werden können. Dasselbe bewirkt die Befehlsfolge „Daten", „Fälle gewichten..."

Fälle auswählen. Öffnet eine Dialogbox, mit der Fälle der Datendatei nach gewissen Bedingungen zur Analyse ausgewählt werden können. Dasselbe bewirkt die Befehlsfolge „Daten", „Fälle auswählen..."

Wertelabels. Durch Anklicken dieses Symbols kann man von Anzeige der Variablenwerte als Wert zur Anzeige als Label umschalten und umgekehrt. Dasselbe bewirkt die Befehlsfolge: „Ansicht", „Wertelabels".

Sets verwenden. Öffnet eine Dialogbox, mit der aus vorher definierten Variablensets derjenige ausgewählt werden kann, der für die Analyse verwendet werden soll. Dasselbe bewirkt die Befehlsfolge: „Extras", „Sets verwenden".

Ein Teil dieser Symbole (Hauptsymbole) findet sich in der Symbolleiste aller Fenster. Es sind dies die ersten sechs Symbole auf der linken Seite (wobei allerdings „Rückgängig machen/Wiederholen" in den anderen Fenster nur durch ein Symbol vertreten sind). Sie dienen zum Laden und Speichern von Dateien, machen die letzte Eingabe rückgängig oder zeigen eine Liste der zuletzt benutzten Dialogboxen. Beachten Sie dabei, dass sich die Funktionen „Öffnen", „Speichern" und „Drucken" nur auf das gerade aktive Fenster beziehen. Weiter sind die Symbole „Gehe zu Fall", „Variablen" und „Sets verwenden" allen Symbolleisten (außer der des Skript-Editors) gemeinsam.

Das Ausgabefenster und das Syntaxfenster verfügen über zwei weitere gemeinsame Symbole:

Gehe zu Daten. Führt direkt in das Dateneditorfenster.

Hauptfenster. Dient dazu, bei mehreren geöffneten Ausgabe- bzw. Syntaxfenstern das Hauptfenster zu bestimmen, in das die Ausgabe bzw. die Syntax geleitet wird.

Skript-Editor und Syntax-Editor teilen mit dem Daten-Editor das Symbol „Suchen". Ansonsten verfügt die Symbolleiste jedes Fensters über einige fensterspezifische Symbole, die an gegebener Stelle besprochen werden.

Hinweis. Das Menüsystem von SPSS lässt sich teilweise auch mit der Tastatur bedienen. Die Hauptmenüs werden dann durch die Kombination <Alt>+<im Menünamen unterstrichener Buchstaben> angewählt. Die Optionen können teilweise über eine Tastenkombination (diese ist dann hinter der Optionsbezeichnung angegeben) ausgewählt werden. Oder man bewegt den Cursor mit der <Auf-> bzw. <Ab->-Steuerungstaste auf die Option und aktiviert sie mit <Enter>. Es stehen viele weitere Steuerungsmöglichkeiten per Taste, insbesondere zum Editieren der Dateien zur Verfügung. Im weiteren wird diese Steuerungsmöglichkeit nicht mehr besprochen. Weitere Einzelheiten können sie im Hilfesystem dem Fenster „Tastatur" und den zugehörigen Unterfenstern entnehmen.

2.3 Daten im Dateneditorfenster eingeben und definieren

2.3.1 Eingeben von Daten

Vor der Auswertung von Daten muss SPSS der zu analysierende Datensatz erst zur Verfügung gestellt werden. Dieses kann auf unterschiedliche Weise geschehen: durch Eintippen der Daten im Dateneditorfenster oder durch Importieren einer mit einem anderen Programm erstellten Datei (eine mit einem Texteditor erstellten ASCII-Datei, eine mit einem Tabellenkalkulations- oder einem Datenbankprogramm oder mit einer anderen SPSS-Version erstellte Datei oder auch einer Datei aus den Statistikprogrammen SAS und Systat). Der Import von Dateien erfolgt mit dem Menü „Datei" der Menüleiste des Daten-Editors (⇨ Kap. 6), Optionen „Öffnen" oder „Datenbank öffnen". Nach dem Datenimport erscheinen dann die Daten im Dateneditorfenster und können darin weiterbearbeitet werden.

Hier soll die Eingabe von Daten im Dateneditorfenster selbst vorgestellt werden. Als Beispieldatensatz werden ausgewählte Variablen für 32 Fälle aus der ALLBUS-Studie (einer allgemeinen Bevölkerungsumfrage) des Jahres 1990 verwendet. Für diese 32 Befragten sind neben einer Fall- und einer Versionsnummer die Variablen Geschlecht, höchster schulischer Bildungsabschluss, Einkommen, politische Einstellung, die Einstellung zur ehelichen Treue sowie vier Fragen, die später zu einem Materialismus-Postmaterialismus-Index zusammengefasst werden, erhoben worden. Der Beispieldatensatz wird mit dem Namen ALLBUS bezeichnet und nur für dieses Kapitel verwendet. Er ist im Anhang A vollständig dokumentiert, damit Sie die folgenden Ausführungen auf dem PC mit SPSS nachvollziehen können. Dazu sollten Sie sich ein Verzeichnis C:\DATEN anlegen.

Ein großer Teil der Beispiele in den späteren Teilen dieses Buches greift ebenfalls auf dieselben Variablen des ALLBUS-Datensatzes zurück (manchmal werden auch weitere Variablen hinzugezogen). Allerdings wird eine größere Stichprobe von ca. 300 Fällen herangezogen, um zu realitätsnäheren Ergebnissen zu kommen. Dieser Datensatz wird als ALLBUS90 bezeichnet. (In diesem Buch werden Dateinamen und Variablennamen zur besseren Lesbarkeit immer groß geschrieben. SPSS für Windows zeigt aber Variablennamen unabhängig von der Schreibweise

immer in Kleinbuchstaben an). Sie können die meisten Anwendungsbeispiele auch mit dem in diesem Kapitel verwendeten und durch Sie einzutippenden Datensatz ALLBUS nachvollziehen. Freilich werden die Ergebnisse zwangsläufig anders ausfallen, als die im Buch dokumentierten, da der Übungsdatensatz ALLBUS von dem Datensatz ALLBUS90 differiert. Wenn Sie aber die Beispiele der späteren Kapitel exakt nachvollziehen wollen, downloaden Sie bitte die Daten von der zum Buch gehörenden Website (⇨ Anhang B) und laden Sie jeweils die dem Beispiel zugehörige Datei.

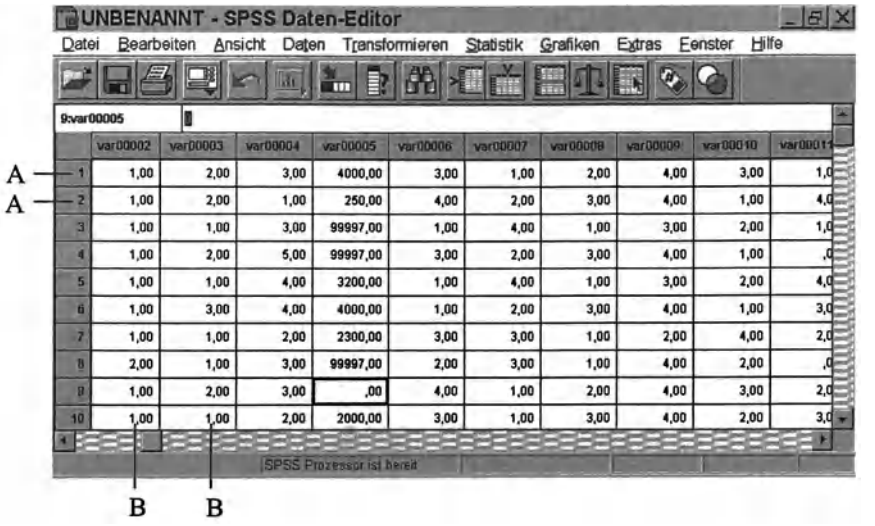
Das SPSS-Dateneditorfenster zeigt sich in Gestalt eines Tabellenkalkulationsblattes. Es hat die Form einer viereckigen Matrix, bestehend aus Zellen, die sich aus Spalten und Zeilen ergeben. Die Zeilen der Matrix sind mit den Ziffern 1, 2 usw. durchnumeriert. Die Spalten sind am Kopf vorerst einheitlich mit VAR beschriftet. Der Wert einer Variablen wird in eine Zelle eingetragen. Die Eingabe muss dabei in bestimmter Weise erfolgen: In einer Zeile der Matrix werden die Werte jeweils eines Befragten (allgemein: eines Falles) eingetippt. In eine Spalte kommen jeweils die Werte für eine Variable. Der Wert ist die verschlüsselte Angabe über die Ausprägung des jeweils untersuchten Falles auf der Variablen. So bedeutet in unserer Übung z.B. bei der Variablen Geschlecht der Wert 1 „männlich" und der Wert 2 „weiblich".

Die auf dem Bildschirm sichtbaren Spalten der Matrix haben eine voreingestellte Breite von acht Zeichen. Voreingestellt ist auch eine rechtsbündige Darstellung der eingegebenen Werte. Das kann nur im Dateneditorfenster in den zugehörigen Definitionsspalten „Spalten" und „Ausrichtung" geändert werden oder durch Markieren der Linie zwischen zwei Spalten, Drücken der linken Maustaste und Ziehen des Cursors. Von diesem Spaltenformat (einem reinen Anzeigeformat) ist das Variablenformat zu unterscheiden, das angibt, wie viel Zeichen ein Variablenwert maximal umfassen kann (dies muss nicht mit der Anzeigebreite korrespondieren). Per Voreinstellung werden die eingetippten Werte der Variablen als numerische Variablen in einem festen, voreingestellten Format mit einer Breite von maximal acht Zeichen und zwei Dezimalstellen aufgenommen (allerdings kann man auch größere Zahlen eintippen. Sie werden aber dann nur mit maximal der angegebenen Zahl von Dezimalstellen angezeigt). Diese Voreinstellungen für das Variablenformat kann mit der Befehlsfolge „Bearbeiten", „Optionen..." im Register „Daten" verändert werden. Für einzelne Variablen ändert man das Format in der Variablenansicht des Dateneditors mit den beiden zugehörigen Definitionsspalten, von der die erste etwas irreführend „Spaltenformat", die zweite „Dezimalstellen" überschrieben ist (⇨ Abb. 2.10)

Abb. 2.6 zeigt das Dateneditorfenster mit den eingetippten Daten für die ersten elf Variablen der ersten zehn Fälle unserer Beispieldatei. Variablen sind:

VAR00001: Fallnummer
VAR00002: Version Nummer
VAR00003: Geschlecht
VAR00004: Allgemeiner Schulabschluss
VAR00005: Monatliches Nettoeinkommen
VAR00006: Politisches Interesse
VAR00007: Wichtigkeit von Ruhe und Ordnung

VAR00008: Wichtigkeit von Bürgereinfluss
VAR00009: Wichtigkeit von Inflationsbekämpfung
VAR00010: Wichtigkeit von freier Meinungsäußerung
VAR00011: Verhaltensbeurteilung: Seitensprung



A Zeilen
B Spalten mit von SPSS automatisch vergebenen Variablennamen als Überschrift

Abb. 2.6. Dateneditorfenster mit Eintragungen

Sie sollten nun SPSS aufrufen und die in der Abbildung sichtbaren Daten und alle anderen im Dateneditorfenster eintippen (alle Fälle sind mit allen Variablen im Anhang A aufgeführt, geben Sie auch offensichtlich falsche Werte in der vorliegenden Form ein). Für die Eingabe gehen Sie mit dem Cursor auf die obere linke Ecke der Matrix (erste Zeile, erste Spalte) und klicken dieses Feld an. Es erscheint jetzt umrandet (bzw. unterlegt). Nun geben Sie den ersten Wert ein. Der Wert erscheint in der Zelleneditorzeile über der Matrix. Wenn Sie die Eingabetaste drücken, wird er in das aktivierte Feld eingetragen und der Cursor rückt eine Zeile nach unten. (Alternativ können Sie auch die Eingabe durch Betätigung der Richtungstasten <Pfeil nach unten> bestätigen.) Soll der Cursor eine Spalte nach rechts rücken, müssen Sie die Eingabe mit der Taste <Pfeil rechts> bestätigen (letzteres dürfte in den meisten Fällen angemessen sein, da man üblicherweise die Daten fallweise eingibt). Der eingegebene Wert erscheint jeweils im markierten Feld, und der Cursor rückt ein Feld in der durch die Richtungstaste festgelegte Richtung weiter. Wenn Sie diese Werte eingeben, wird die per Voreinstellung festgelegte Spaltenbreite größer sein als für die meisten Variablen notwendig. Außerdem werden die Zahlen mit zwei Kommastellen erscheinen, was ebenfalls überflüssig ist, weil unsere Kodierungen nur ganze Zahlen enthalten. Wir werden beides später ändern.

Mit der Eingabe des ersten Wertes vergibt SPSS automatisch einen Variablennamen. Für die erste Variable ist das VAR00001. Dieser steht jetzt über der Spalte. Falls Sie die Werte spaltenweise eingeben, wird Ihnen weiter auffallen, dass SPSS sofort mit dem Eröffnen einer neuen Variablen für sämtliche Fälle vorläufig ein Komma (als systemdefinierter fehlender Wert) einsetzt. Geben Sie auf eine der dargestellten Weisen die Werte für sämtliche Fälle ein. Sie haben nun eine Datenmatrix, die Sie sofort zur statistischen Analyse verwenden können. Weitere Vorbereitungen sind nicht unbedingt nötig, aber meistens nützlich.[3]

2.3.2 Speichern und Laden einer Datendatei

Sicherheitshalber sollten Sie jetzt Ihre Daten speichern. Dafür wählen Sie das Menü:

▷ „Datei" und darin den Befehl „Speichern" oder Klicken auf das Symbol

Beim erstmaligen Speichern erscheint auf dem Bildschirm die in Abb. 2.7 dargestellte Dialogbox (später nur bei Wahl der Option „Speichern unter..."). In dieser Dialogbox wird der Typ der Datei und das Verzeichnis, in dem die Datei gespeichert werden soll, angegeben. Voreingestellt ist als Dateiyp eine „SPSS"-Datei (sav), und der Pfad zeigt auf das Verzeichnis, in dem SPSS liegt, z.B. SPSS11. Ersteres akzeptieren wir so. Als Verzeichnis, in dem die Datei abgespeichert werden soll, sollten Sie aber C:\DATEN wählen (vorausgesetzt, Sie haben dieses Verzeichnis – wie vorgeschlagen – eingerichtet oder richten es jetzt im Dateimanager ein). (Wenn man über die Schaltfläche „Variablen" eine Unterdialogbox öffnet, kann man auch nur einen Teil der Variablen zum Speichern auswählen.)
Um das Verzeichnis C:\DATEN zu wählen, gehen Sie wie folgt vor:

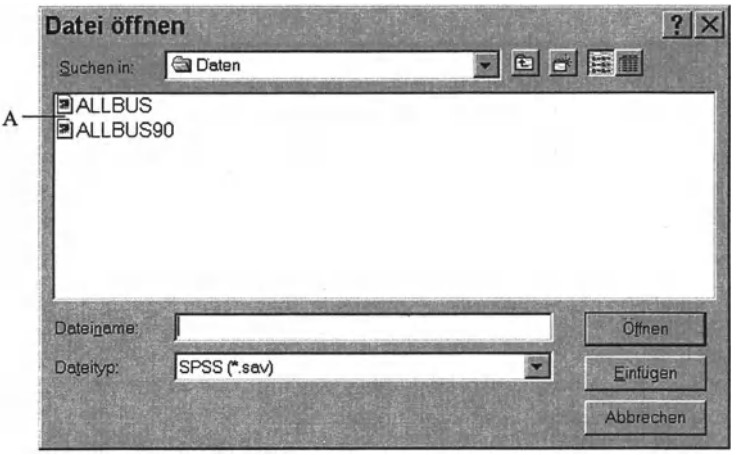
▷ Öffnen Sie durch Klicken auf den Pfeil neben dem Auswahlfeld „Speichern" die Drop-Down-Liste mit den Bezeichnungen der verfügbaren Laufwerke.
▷ Klicken Sie in dieser Liste auf den Namen des gewünschten Laufwerks.
▷ Ist dieser im Auswahlfeld richtig angezeigt, doppelklicken Sie in dem darunterliegenden großen Anzeigefeld auf den Namen des Verzeichnisses „Daten". Der Name des Verzeichnisses erscheint im Anzeigefeld.
▷ Tragen Sie den gewünschten Dateinamen im Feld „Dateiname:" ein. Wir tragen ALLBUS ein. Unter „Dateityp:" könnte ein Dateiformat für die gespeicherte Datendatei ausgewählt werden. Voreingestellt ist das SPSS-Windows-Format. Wir akzeptieren dies und die ebenfalls voreingestellte Namenserweiterung (Ex-

[3] Wir gehen in der Einführung so vor, dass wir sofort in eine leere Matrix Daten eingeben. Das ist möglich, weil SPSS die notwendigsten Variablendefinitionen automatisch vornimmt. Man kann mit den Daten ohne weitere Vorbereitungen sofort arbeiten. Gewünschte Änderungen der Variablendefinitionen können nachträglich durchgeführt werden. Selbstverständlich kann man aber auch zuerst die Variablen definieren. Das empfiehlt sich insbesondere, wenn Daten arbeitsteilig eingegeben und später vereinigt werden sollen. Dann sind vordefinierte Variablen von großem Nutzen. Im folgenden geben wir neben den von uns und in früheren Versionen benutzten Bezeichnungen für Variablenbreite und Spaltenbreite die Bezeichnung der Variablenansicht in Klammern gesetzt an.

tension) SAV und klicken zur Bestätigung auf „Speichern" (oder drücken die Enter-Taste).

Abb. 2.7. Dialogbox „Daten speichern unter"

Nun sollten Sie aber auch gleich das Laden der Datei kennen lernen. (Die Datei darf nicht bereits geöffnet sein.)



A Auswahlliste

Abb. 2.8. Dialogbox „Datei öffnen"

Zum Laden dieser Datendatei wählen Sie die Befehlsfolge:

▷ „Datei", „Öffnen", „Daten...". Alternativ klicken Sie auf .

Es erscheint die in Abb. 2.8 dargestellte Dialogbox „Datei öffnen".

▷ Hier kann zunächst der Typ der Datendatei eingestellt werden. Voreingestellt ist SPSS(*.sav). Diese Voreinstellung wird beibehalten.

▷ Wählen Sie dann auf die soeben beschriebene Weise das gewünschte Verzeichnis (hier: C:\DATEN). Es wird eine Liste der darin enthaltenen Datendateien mit der Extension *.SAV angeführt.
▷ Doppelklicken Sie auf den Namen der gewünschten Datei (hier: ALLBUS). (Oder klicken Sie auf den Namen und drücken die Taste <Enter> bzw. klicken Sie auf die Schaltfläche „Öffnen".)

2.3.3 Variablen definieren

Wir werden im folgenden einige Änderung bei Variablen- und Spaltenformaten vornehmen und einige weitere Eingaben zur Datenbeschreibung durchführen:

- Die von SPSS automatisch vergebenen Variablennamen VAR00001, VAR00002 etc. sollen in „sprechende" Variablennamen geändert werden.
- Das Format der Variablen soll auf die notwendige Zeichenbreite reduziert werden und keine Kommastellen mehr enthalten.
- Den Variablen sollen Labels (Etiketten) zugewiesen werden.
- Den Variablenwerten (soweit sinnvoll) sollen ebenfalls Labels (Etiketten) zugewiesen werden.
- Fehlende Werte sollen als solche deklariert werden.
- Die angezeigte Spaltenbreite soll verringert werden.

Fehlende Werte müssen von statistischen Prozeduren ausgeschlossen werden, wenn deren Einbeziehung das Ergebnis verfälschen würde. SPSS trägt automatisch systemdefinierte fehlende Werte (System-Missings) ein, wenn in Zellen des Eingabebereichs keine Werte eingetragen sind. Nutzt man dies, kann man einige Eingabearbeit sparen. Um verschiedene Arten von fehlenden Werten zu unterscheiden und um das Risiko von Eingabefehlern zu reduzieren, wird aber häufig auch bei fehlenden Werten eine Eingabe vorgenommen. So ist das auch in unserem Beispiel. Um diese ebenfalls bei Bedarf von statistischen Prozeduren ausschließen zu können, muss man sie als (nutzerdefinierte) fehlende Werte deklarieren. Diese Änderung der Variablendefinition ist deshalb unabdingbar. Alle anderen Änderungen dienen dagegen ausschließlich der leichteren Handhabung bei der Datenauswertung, der besseren Lesbarkeit der Variablen in den Auswahllisten sowie der Daten im Dateneditor und der Gestaltung der Ergebnisprotokolle. Sie sind nicht unbedingt notwendig, aber nützlich. Zur einfachen Definition von Variablen und deren Änderungen enthält der Dateneditor das Registerblatt „Variablenansicht". Die gewünschten Änderungen werden auf diesem Registerblatt vollzogen. Um sie zunächst für die erste Variable durchzuführen, gehen Sie wie folgt vor:

▷ Klicken Sie im Daten-Editor auf die Registerkarte „Variablenansicht". Das Registerblatt Variablenansicht öffnet sich (⇨ Abb. 2.10). Es hat die Form eines Tabellenkalkulationsblattes. Je eine Reihe enthält die Datendefinition einer Variablen (d.h. einer Spalte in der Daenansicht). Die Spalten enthalten die einzelnen Elemente der Variablendefinition (beginnend mit „Name", „Typ" und endend mit „Messniveau". In unserem Falle enthält das Blatt bereits Definitionen, denn mit jeder Eingabe eines Datums in irgendeine Spalte des Datenblattes generiert SPSS automatisch eine Variable mit der dazugehörigen

Minimaldefinition (Namen VAR0001 etc., Spaltenformat 8, Dezimalstellen 2 etc.).

- ▷ Gehen Sie mit dem Cursor in die Spalte „Namen". Aktivieren Sie durch Klicken mit der linken Maustaste die Zelle mit dem Namen der ersten Variablen. Überschreiben Sie den bisherigen Namen VAR00001 einfach mit dem neuen Namen NR. (Die in Anhang A enthaltene Variable LFDNR lassen wir aus.)

Als nächstes wird der Variablentyp in verschiedenen Spalten geändert (Voreingestellt ist „Numerisch", Breite 8[4], mit 2 Dezimalstellen). Den „Typ" „numerisch" behalten wir bei.

- ▷ Zur Änderung der Breite: Klicken Sie auf die zur Variablen gehörende Zelle in der Spalte „Spaltenformat". Am Ende dieser Zell erscheinen zwei Pfeile. Mit ihrer Hilfe kann der Wert geändert werden. Klicken Sie auf den unteren Pfeil, bis der Wert von 8 in 4 geändert ist.
- ▷ Zur Änderung der Dezimalstellen: Klicken Sie auf die zur Variablen gehörende Zelle in der Spalte „Dezimalstellen". Am Ende dieser Zelle erscheinen zwei Pfeile. Klicken Sie auf den unteren Pfeil, bis der Wert von 2 in 0 geändert ist.
- ▷ Markieren Sie jetzt die Zelle in der Spalte „Variablenlabel". Die Zelle ist leer. Wir tragen in ihr als Variablenlabel „Fallnummer" ein. (Labels von Variablen können bis zu 120 Zeichen lang sein. Bei den meisten Ergebnisausgaben von statistischen Auswertungen werden aber weniger Zeichen angezeigt.)
- ▷ Abschließend ändern wir die angezeigte Spaltenbreite der Matrix. Dazu aktivieren wir die entsprechende Zelle in der drittletzten Spalte „Spalten" und vermindern mit Hilfe des unteren Pfeils den Wert von 8 auf 5.

Schalten Sie kurz durch Anklicken der Registerkarte „Datenansicht" auf das Datenblatt um. Hier erscheint nun die erste Spalte verändert. Im Kopf steht der neue Name „NR", die Variablenwerte erscheinen ohne Nachkommastellen und die Matrixspalte ist nur noch fünf Stellen breit.

Die anderen Variablendefinitionen sollen in ähnlicher Weise verändert werden. Zur Änderung der Definition der Variablen VAR00002 aktivieren Sie jeweils die entsprechenden Zellen in der zweien Reihe der „Variablenansicht", zur Änderung der Variablen VAR0003 der dritten Reihe etc..

- ▷ Bei VAR00002 ändern Sie den Namen in „VN", die Variablenbreite (Spaltenformat) in 2 und die Zahl der Nachkommastellen in 0. Die Spaltenbreite der Matrix wird auf 2 geändert. Vergeben Sie das „Variablenlabel" „Version Nummer". (ALLBUS wurde in zwei Versionen durchgeführt, die erste bekommt die Versionsnummer 1, die zweite 2.)
- ▷ Bei VAR00003 führen Sie folgende Änderungen durch: Variablennamen „GESCHL", Variablenbreite (Spaltenformat) 1, Nachkommastellen 0, Spaltenbreite (Spalten) 5. Im Eingabefeld der Spalte „Variablenlabel" setzen Sie als Variablenlabel „Geschlecht" ein.

[4] Lassen Sie sich jetzt und im folgenden nicht dadurch verwirren, dass in der Dialogbox „Bearbeiten: Optionen" dies als Breite (wir benutzen diesen Begriff), in der Variablenansicht dagegen als Spaltenformat bezeichnet wird.

▷ Zusätzlich sollen jetzt Wertelabel vergeben werden. Dazu aktivieren Sie zunächst die entsprechende Zelle in der Spalte „Wertelabels". Am rechten Rand der Zelle erscheint ein unterlegtes Quadrat mit drei Pünktchen. Dies zeigt an, dass eine Dialogbox zum Zwecke der weiteren Definition existiert. Klicken Sie mit der linken Maustaste auf dieses Quadrat. Die Dialogbox „Wertelabels definieren" erscheint. Im Eingabefeld „Wert:" tragen Sie den Wert 1 ein, dann in „Wertelabel:" „männlich" und klicken auf die Schaltfläche „Hinzufügen". Es erscheint 1 = „männlich" im großen Informationsfeld für die definierten Wertelabels. Zugleich ist SPSS bereit für die Eingabe eines weiteren Labels. Geben Sie in „Wert:" den Wert 2 ein, dann in „Wertelabel:" „weiblich", und klicken Sie auf die Schaltfläche „Hinzufügen". Jetzt erscheint 2 = „weiblich". Bestätigen Sie mit „OK".

VAR00004 wird wie folgt verändert: Variablenname SCHUL, Variablenbreite (Spaltenformat) 2, Nachkommastellen 0, Spaltenbreite (Spalten) 5, Variablenlabel „Allgemeiner Schulabschluss".

Wertelabels sind:

1 = „Schule beendet ohne Abschluss"
2 = „Volks-/Hauptschulabschluss"
3 = „Mittlere Reife, Realschulabschluss (Fachschulreife)"
4 =„Fachhochschulreife (Abschluss einer Fachoberschule, etc.)"
5 =„Abitur (Hochschulreife)"
6 = „Anderer Schulabschluss"
7 = „Noch Schüler"
97 = „Verweigert"
98 = „Weiß nicht"
99 = „Keine Angabe"

Die Wertelabels können bis zu 60 Zeichen lang sein. Bei den meisten Ergebnisausgaben werden aber weniger Zeichen angezeigt.

Gegenüber den anderen Variablendefinitionen kommt neu hinzu, dass die drei Werte 97, 98 und 99 als „Fehlende Werte" deklariert werden sollen.

▷ Dazu aktivieren Sie zunächst die entsprechende Zelle in der Spalte „Fehlende Werte". Am rechten Rand der Zelle erscheint wieder ein unterlegtes Quadrat mit drei Pünktchen, welches anzeigt, dass eine Dialogbox für die weitere Definition existiert. Klicken Sie mit der linken Maustaste auf dieses Quadrat. Es erscheint die in Abb. 2.9 dargestellte Dialogbox „Fehlende Werte definieren". Hier ist per Voreinstellung „Keine fehlende Werte" angegeben.
▷ Ändern Sie das, indem Sie den Optionsschalter „Einzelne fehlende Werte" anklicken. Geben Sie in die Eingabefelder in der entsprechenden Reihe die Werte 97, 98 und 99 ein (da die Werte unmittelbar nebeneinander liegen, hätte man sie auch als einen Bereich über den Optionsschalter „Bereich und einzelner fehlender Wert" und die dazugehörigen Eingabefelder eingeben können).
▷ Bestätigen Sie mit „OK".

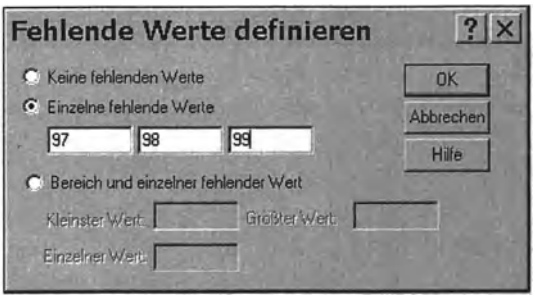


Abb. 2.9. Dialogbox „Fehlende Werte definieren"

VAR00005 bekommt folgende Definitionen: Variablenname EINK, Variablenbreite (Spaltenformat) 5, Nachkommastellen 0, Spaltenbreite 6, Variablenlabel „Monatliches Nettoeinkommen", Wertelabels:

99997 = „Angabe verweigert"
99998 = „Weiß nicht"
99999 = „Keine Angabe"
0 = „Kein eigenes Einkommen"
(99997, 99998, 99999, 0 sind fehlende Werte).

SPSS erlaubt maximal drei diskrete Werte als fehlende Werte zu deklarieren. Da wir hier vier fehlende Werte vorliegen haben, nutzen wir die Möglichkeit, einen Wertebereich kombiniert mit einem diskreten Wert als fehlenden Wert zu deklarieren. Dazu gehen Sie wie folgt vor:

- Aktivieren Sie die Zelle in der Spalte „Fehlende Werte". Klicken Sie mit der linken Maustaste auf das Quadrat auf der rechten Seite. Es erscheint die in Abb. 2.9 dargestellte Dialogbox „Fehlende Werte definieren".
- Klicken Sie auf den Schalter vor der Option „Bereich und einzelner fehlender Wert". Geben Sie 99997 für den niedrigsten Wert in das Eingabefeld „Kleinster Wert:" und 99999 für den höchsten Wert in „Größter Wert:" ein, schließlich in das Kästchen „Einzelner Wert:" den Wert 0.
- Bestätigen Sie mit „OK".

VAR0006 bekommt folgende Definitionen: Variablenname POL, Variablenbreite (Spaltenformat) 1, Nachkommastellen 0, Spaltenbreite (Spalte) 4, Variablenlabel „Politisches Interesse", Wertelabels:

1 = „Sehr stark"	5 = „Überhaupt nicht"
2 = „Stark"	7 = „Verweigert"
3 = „Mittel"	8 = „Weiß nicht"
4 = „Wenig"	9 = „Keine Angabe"

(7, 8 und 9 sind fehlende Werte).

VAR00007 bis VAR00010 unterscheiden sich nur im Variablennamen und den Variablenlabels. Ansonsten ist ihre Definition identisch. Als Namen benutzen wir RUHE, EINFLUSS, INFLATIO, MEINUNG. Die Variablenlabels sind „Wichtigkeit von Ruhe und Ordnung", „Wichtigkeit von Bürgereinfluss", „Wichtigkeit der

Inflationsbekämpfung" und „Wichtigkeit von freier Meinungsäußerung". Diese Angaben geben wir bei jeder Variablen gesondert ein. Die anderen Angaben dagegen sind identisch: Variablenbreite ist (Spaltenformat) 1, Zahl der Nachkommastellen 0, Spaltenbreite (Spalte) 8. Auch die Wertelabels sind für alle vier Variablen identisch:

1 = „Am wichtigsten"	7 = „Verweigert"
2 = „Am zweitwichtigsten"	8 = „Weiß nicht"
3 = „Am drittwichtigsten"	9 = „Keine Angabe"
4 = „Am viertwichtigsten"	0 = Frage nicht gestellt (Version 2)
(7, 8 und 9 sind fehlende Werte).	

Die Dateneingabe kann daher durch Kopieren vereinfacht werden.

Ändern Sie zunächst die Namen der vier Variablen wie oben angegeben. Geben Sie für alle vier Variablen die Variablenlabels ein. Dann ändern Sie alle anderen Definitionen nur für die ehemalige Variable VAR00007, jetzt RUHE. (Sie müssen die Zahl der Nachkommastellen vor dem Spaltenformat ändern, da das Programm sonst moniert, dass die Feldbreite [= Variablenbreite] nicht für die Zahl der Nachkommastellen ausreicht.)

Die identischen Definitionen kopieren Sie anschließend aus den Definitionsfelder der Variablen RUHE in die Definitionsfelder der drei anderen Variablen.

- ▷ Dazu aktivieren Sie zunächst die Zelle zur Spalte „Dezimalstellen" in der Zeile der Variablen RUHE. Wählen Sie im Menü „Bearbeiten", „Kopieren". Setzen Sie den Cursor in die entsprechende Zelle der Zeile EINFLUSS, drücken Sie die linke Maustaste und ziehen Sie nun den Cursor bis zum Namen der letzten Variablen. Wenn Sie die Maustaste loslassen, sind alle drei Variablen markiert. (*Anmerkung:* Nicht nebeneinanderliegende Variablen können nicht gleichzeitig markiert werden.)
- ▷ Wählen Sie im Menü „Bearbeiten", „Einfügen". Die Definition der Nachkommastellen ist auf alle markierten Variablen übertragen. (*Anmerkung:* Die Befehle „Kopieren" und „Einfügen" können auch einfacher über das Kontextmenü, das sich beim Drücken der rechten Maustaste öffnet, gewählt werden.)
- ▷ Wiederholen Sie den Prozess für die anderen identischen Definitionselemente (Spaltenformat, Wertelabels und fehlende Werte).

Anmerkung. Es können nicht mehrere Definitionselemente gleichzeitig kopiert und eingefügt werden, es sei denn, es wird die vollständige Definition einer Variablen übernommen. Dann markieren Sie die ganze Definitionszeile dieser Variablen, indem sie auf die Zeilennummer am linken Rand drücken. Kopieren und Einfügen erfolgt in der angegebenen Weise.

VAR00011 bekommt folgende Definitionen: Variablenname TREUE, Variablenbreite (Spaltenformat) 1, Nachkommastellen 0, Spaltenbreite 5, Variablenlabel „Verhaltensbeurteilung: Seitensprung", Wertelabels:

1= „Sehr schlimm"	7 = „Verweigert"
2 = „Ziemlich schlimm"	8 = „Weiß nicht"
3 = „Weniger schlimm"	9 = „Keine Angabe"
4 = „Überhaupt nicht schlimm"	0 = „Frage nicht gestellt"(Version 2)

(7, 8 und 9 und 0 sind fehlende Werte)

Anmerkung. Wir haben eine ziemlich umfassende Definition der Variablen vorgenommen. Natürlich kann man sich sehr viel Arbeit sparen, wenn man z.B. die von SPSS vergebenen Variablennamen akzeptiert oder mit einem einheitlichen Datentyp arbeitet. Auf Labels kann man verzichten, wenn man den Verschlüsselungsplan (Kodeplan) neben sich liegen hat. Allerdings macht das andererseits auch viele Auswertungen mühsam. Auf Variablenlabels kann man verzichten, wenn man selbsterklärende Variablennamen vergibt. Umgekehrt kann man auf neue Variablennamen verzichten, wenn man Variablenlabels benutzt.

Variableninformationen. Nachdem Sie die Variablen definiert haben, können Sie sich in jetzt jeder Quellvariablenliste oder Auswahlliste diese Definitionen anzeigen lassen. Markieren Sie dazu den Variablennamen und Drücken sie die rechte Maustaste. Es öffnet sich ein lokales Menü. Wählen Sie dort „Info zu Variable", werden neben Namen und Variablentyp die Labels zu dieser Variablen angezeigt. (Um eine vollständige Liste der Wertelabels zu sehen, müssen Sie auf den Pfeil neben dem Fenster „Wertelabels" klicken.) Probieren Sie das in einem der Analysemenüs.

Abb. 2.10 Registerblatt „Variablenansicht" im Dateneditor

Wenn Sie außerdem in den Optionen im Register „Allgemein"(Menü „Bearbeiten", „Optionen") in der Gruppe „Variablenlisten" die Option „Labels anzeigen" gewählt haben (Voreinstellung ⇨ Kap. 28.5), werden in den Quellvariablenlisten

nicht die Variablennamen, sondern die Variablenlabels angezeigt (gefolgt von den in Klammern gesetzten Namen). Sind diese zu lang, öffnet sich sogar eine Zeile, in der das ganze Label zu sehen ist.

Wenn Sie mit dem Cursor auf dem Datenblatt des Datei-Editors auf den Namen einer Variablen im Kopf der Spalte zeigen, wird das Variablenlabel in einer Drop-Down-Zeile angezeigt.

2.4 Daten bereinigen

Daten können aus unterschiedlichen Gründen fehlerhaft sein. Schon bei der Erhebung kommen Mess- und Registrierungsfehler vor, oder es entstehen an irgendeiner Stelle Verschlüsselungs- oder Übertragungsfehler. Bevor man an die Auswertung von Daten geht, sollte man daher zuerst diese Fehler so weit wie möglich beseitigen. Man wird die fehlerhaften Daten suchen und korrigieren. Diesen Prozess nennt man Datenbereinigung. Mit Hilfe der in SPSS verfügbaren Prozeduren wird man Fehler allerdings nur ausfindig machen können, wenn sie durch eines der folgenden Merkmale auffallen:

- Ein Wert liegt außerhalb des zulässigen Bereiches (sind z.B. bei der Variablen Geschlecht 1, 2 und zur Deklaration eines Fehlenden Wertes 0 zugelassen, sind alle anderen Angaben fehlerhafte Werte).
- Logische Inkonsistenzen treten auf (z.B. ist bei einem Alter von fünf Jahren als Familienstand verheiratet angegeben oder bei einer Frage, die gar nicht gestellt werden durfte, ist eine gültige Angabe aufgenommen).
- Außergewöhnliche Werte oder Kombinationen treten auf, die auf einen evtl. Fehler hinweisen (z.B. ein Schüler im Alter von 80 Jahren oder 20 Familienmitglieder).

Welche Fehler auftreten können, hängt u.a. davon ab, welche Vorkehrungen schon bei der Eingabeprozedur getroffen wurden. So kann man mit Datenbankprogrammen oder SPSS-Data-Entry (eine von SPSS angebotene Stand-alone-Software) durch Angabe entsprechender Grenzen die Eingabe nicht zulässiger Werte verhindern. Ebenso können bei Data-Entry Filter eingebaut werden, die beim Auftreten eines bestimmten Variablenwertes die Eingabe von logisch nicht zulässigen Folgewerten verhindern. Die häufigen Formatfehler, die meist Folgefehler nach sich ziehen, werden bei Verwendung von entsprechenden Eingabemasken weitgehend ausgeschlossen. Auch die Verwendung des eben beschriebenen Daten-Editors von SPSS ist hier sehr hilfreich.

Um unzulässige Werte aufzudecken, wird man gewöhnlich eine sogenannte „Grundauszählung" durchführen und deren Ergebnisse inspizieren. Sie ist vor allem bei qualitativen Daten nützlich. Fehler in quantitativen Daten, vor allem wenn sehr viele Ausprägungen auftreten, sind dagegen damit kaum auszumachen. Logische Fehler können auf verschiedene Weise entdeckt werden, z.B. durch Erstellen von Kreuztabellen oder mit Bedingungsbefehlen (If-Befehlen). Außergewöhnliche Fälle und Kombinationen kann man ebenfalls auf verschiedene Weise entdecken.

SPSS hält dafür auch Prozeduren zur Datenexploration zur Verfügung. Wir werden uns hier nur mit den beiden ersten Inspektionsformen beschäftigen.

Für diese Übung sollten Sie die Voreinstellung von SPSS für die Beschriftung der Ausgabe ändern, damit Sie die Kategorienwerte sehen können (per Voreinstellung werden nur die Labels angezeigt ⇨ Kap. 28. 5). Wählen Sie dafür die Befehlsfolge: „Bearbeiten", „Optionen..." und in der dann erscheinenden Dialogbox „Optionen" die Registerkarte „Beschriftung der Ausgabe". Öffnen Sie in der Gruppe „Beschriftung für Pivot-Tabellen" durch Anklicken des Pfeils neben dem Auswahlfeld „Variablen in Beschriftungen anzeigen als:" eine Auswahlliste. Wählen Sie daraus „Namen und Labels". Im Auswahlfeld „Variablenwerte in Beschriftungen anzeigen als:" wählen Sie auf die gleiche Weise „Werte und Labels". Bestätigen Sie mit „OK". (Sie können das nach dieser Übung wieder rückgängig machen.)

Als erstes führen wir eine *Grundauszählung* durch. Das ist eine einfache Häufigkeitsauszählung für alle Variablen. Um eine Grundauszählung zu erstellen:

▷ Wählen Sie „Analysieren", „Deskriptive Statistiken ▷ ", „Häufigkeiten...". Es erscheint die in Abb. 2.11 dargestellte Dialogbox „Häufigkeiten".

Diese enthält auf der linken Seite die Quellvariablenliste mit allen Variablen des Datensatzes. Um daraus die Variablen auszuwählen, für die eine Auszählung vorgenommen werden soll:

▷ Doppelklicken Sie auf den Variablennamen, oder markieren Sie den Variablennamen durch Anklicken mit dem Cursor und klicken Sie auf das Schaltfeld ▸. Dann wird die Variable in das Feld der ausgewählten Variablen verschoben. Gleichzeitig kehrt sich der Pfeil im Schaltfeld um. Klickt man ihn wieder an, wird die Auswahl rückgängig gemacht.

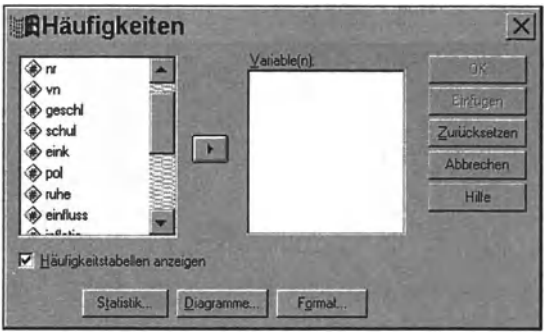


Abb. 2.11. Dialogbox „Häufigkeiten"

Da wir eine Grundauszählung durchführen, sollen alle Variablen ausgewählt werden. Dazu markieren wir alle Variablen der Quellvariablenliste. Wir setzen den Cursor auf die erste Variable, drücken die linke Maustaste und ziehen den Cursor so lange, bis alle Variablen markiert sind. Durch Klicken auf das Schaltfeld ▸ übertragen wir sie alle gleichzeitig in das Auswahlfeld.

▷ Mit „OK" starten wir den Befehl.

Das Ergebnis wird in das Ausgabefenster geleitet. Dieses wird automatisch aktiviert. Die linke Seite, das Gliederungsfenster, lassen wir vorerst außer Acht (⇨ Kap. 4.1.2) und benutzen nur die rechte Seite, das eigentliche Ausgabefenster. Auf dem Bildschirm ist der Anfang des Outputs zu sehen.[5]

Wir scrollen mit der Bildlaufleiste durch die Ausgabe und inspizieren jetzt alle Häufigkeitstabellen auf unzulässige Werte. Bei der Tabelle Geschlecht bemerken wir einen unzulässigen Wert, nämlich eine 3 (⇨ Abb. 2.12).

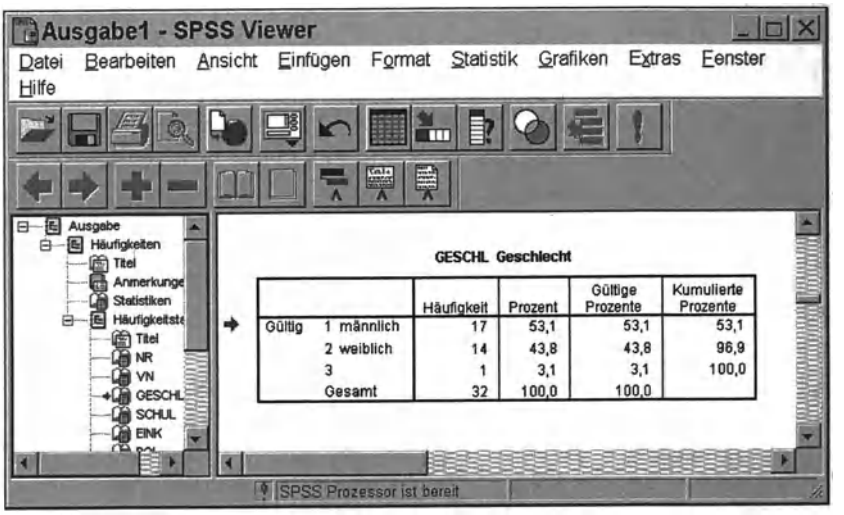


Abb. 2.12. Ausgabefenster mit der Häufigkeitsverteilung der Variablen GESCHL

Ein Beispiel für eine nicht zulässige Kombination: Die Frage nach der Bewertung ehelicher Treue (Variable TREUE) wurde nur den Befragten der Fragebogenversion 1 gestellt, nicht aber denjenigen, die mit der Version 2 befragt wurden. Die Version ist in der Variable VN festgehalten. Entsprechend muss bei allen Befragten, die bei der Variablen VN den Eintrag 2 haben, eine 0 für „Frage nicht gestellt" in der Variablen TREUE stehen. Wo dagegen in VN eine 1 steht, muss bei TREUE einer der anderen zulässigen Werte, das sind die Werte 1 bis 4 und 7 bis 9 eingetragen sein. Ob dies der Fall ist, kann man auf verschiedene Weisen erkunden. Wir untersuchen es jetzt mit Hilfe einer Kreuztabelle.

[5] Bei einigen Prozeduren gibt SPSS zusätzlich zu der eigentlichen Ergebnisausgabe eine mit „Verarbeitete Fälle" überschriebene Tabelle aus, in der die Zahl der gültigen Fälle und der fehlenden Fälle bzw. eingeschlossenen und ausgeschlossenen Fälle für jede Tabelle angegeben wird. Dies ist auch bei Häufigkeitsauszählungen der Fall. Die Zusatztabelle ist allerdings mit „Statistiken" überschrieben, weil in ihr gegebenenfalls auch angeforderte statistische Maßzahlen dargestellt werden. Diese vorangestellte, eher Rahmeninformationen enthaltende, Zusatztabelle besprechen wir durchgängig bei der Interpretation der Ausgabe nicht.

Weil bei der Erstellung von Kreuztabellen prinzipiell die fehlenden Werte nicht berücksichtigt werden, uns aber bei der Variablen TREUE gerade der als fehlend deklarierte Wert 0 interessiert, müssen wir diese Deklaration vorübergehend rückgängig machen.

- ▷ Gehen Sie dazu in das Registerblatt „Variablenansicht“ des Daten-Editors.
- ▷ Aktivieren Sie in der Zeile der Variablen TREUE die Zelle in der Spalte „Fehlende Werte“ und öffnen Sie das Dialogfenster „Fehlende Werte definieren“ durch Anklicken des unterlegten Quadrats auf der rechten Seite der Zelle.
- ▷ Ändern Sie die Eingabe, indem Sie „Bereich und einzelner fehlender Wert” den einzelnen fehlenden Wert 0 löschen.
- ▷ Bestätigen Sie das Ganze mit „OK” (machen Sie das nach der Erstellung der Kreuztabelle wieder rückgängig).

Zur Erstellung der gewünschten Kreuztabelle gehen Sie wie folgt vor:

- ▷ Wählen Sie „Analysieren”, „Deskriptive Statistiken ▷ ” und „Kreuztabellen...”. Es erscheint die Dialogbox zur Erstellung von Kreuztabellen (Abb. 2.13).

Aus der Quellvariablenliste wählen Sie aus, welche Variable in einer Kreuztabelle in die Zeile, welche in die Spalten, d.h. in den Kopf der Tabelle kommen soll. In unserem Fall soll VN in die Zeile, TREUE in den Kopf der Tabelle.

- ▷ Dazu markieren Sie zunächst VN und klicken auf die Schaltfläche ▶ vor dem Auswahlfeld „Zeilen:”. Dann markieren Sie TREUE und klicken auf die Schaltfläche ▶ vor dem Auswahlfeld „Spalten:” Die beiden Variablen sind jetzt in diese Felder übertragen.
- ▷ Bestätigen Sie mit „OK”.

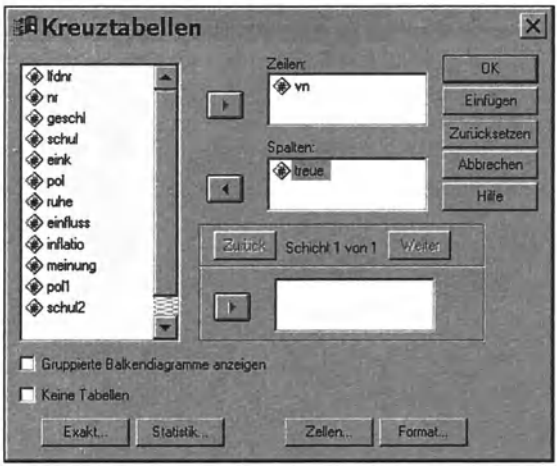


Abb. 2.13. Dialogbox „Kreuztabellen”

Tabelle 2.1. Kreuztabelle für die Variablen TREUE und VN

VN Version Nummer * TREUE Verhaltensbeurteilung: Seitensprung Kreuztabelle

Anzahl

		TREUE Verhaltensbeurteilung: Seitensprung					
		0 Frage nicht gestellt (Fragebogen Version 2)	1 Sehr schlimm	2 Ziemlich schlimm	3 Weniger Schlimm	4 Überhaupt nicht schlimm	Gesamt
VN Version Nummer	1	1	5	5	3	7	21
	2	11					11
Gesamt		12	5	5	3	7	32

Die Durchsicht der Kreuztabelle (⇨ Tabelle 2.1) im Ausgabefenster zeigt, dass eine nicht zulässige Kombination vorliegt, nämlich eine 0 bei der Variablen TREUE, obwohl die Fragebogenversion 1 Verwendung fand.

Wir müssen nun noch herausfinden, bei welchen Fällen die beiden Fehler aufgetreten sind und sie im Dateneditorfenster beseitigen.

Dies würde man in diesem Falle, bei einer solche kleinen Datenmatrix normalerweise wohl direkt bei der betroffenen Variablen in der Datenansicht des Daten-Editors tun. Dabei würde man die Option „Suchen..." im Menü „Bearbeiten" (⇨ Kap. 3.4) verwenden. Um die Verwendung von *Bedingungsausdrücken* zu demonstrieren, wird hier ein umständlicheres Verfahren gewählt.

Zur Identifikation der beiden fehlerhaften Fälle benutzen wir die Kombination eines Datenauswahlbefehls (Datenselektionsbefehl) und eines Statistikbefehls. Zur Identifikation des ersten fehlerhaften Falles suchen wir den Fall heraus, der bei GESCHL den Wert 3 hat und lassen uns seine Fallnummer ausgeben.

- ▷ Wählen Sie im Menü „Daten" die Option „Fälle auswählen...". Es öffnet sich die Dialogbox „Fälle auswählen" (⇨ Abb. 7.10).
- ▷ Klicken Sie auf den Optionsschalter „Falls Bedingung zutrifft".
- ▷ Wählen Sie in der Gruppe „Nicht ausgewählte Fälle" die Option „Filtern" (Voreinstellung). Damit werden nicht ausgewählte Fälle nicht permanent ausgeschlossen und bleiben der Datei für spätere Auswertungen erhalten.
- ▷ Klicken Sie auf die Schaltfläche „Falls ...". Es öffnet sich die in Abb. 2.14 dargestellte Dialogbox, in der wir die Auswahlbedingung angeben müssen.

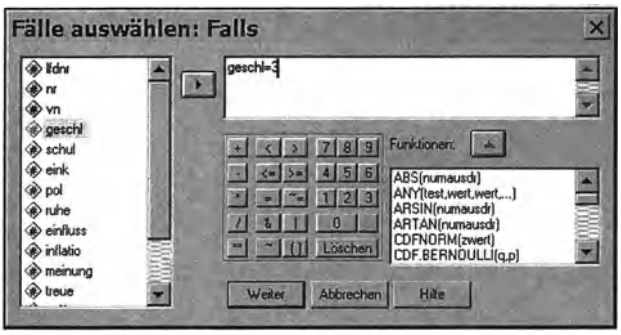


Abb. 2.14. Dialogbox „Fälle auswählen: Falls"

In dem Feld rechts oben wird die Bedingung eingetragen, die eine oder mehrere Variablen erfüllen müssen (⇨ Abb. 2.14).

- ▷ Übertragen Sie GESCHL aus der Quellvariablenliste in das Feld für die Definition der Bedingung.
- ▷ Das Gleichheitszeichen übertragen Sie durch Anklicken von „=" in der Rechnertastatur.
- ▷ Schließlich geben Sie 3 ein, und die Auswahlbedingung ist gebildet .
- ▷ Bestätigen Sie mit „Weiter" und „OK".

Mit der Befehlsfolge „Analysieren", „Berichte ▷ ", „Fälle zusammenfassen..." und Auswahl der Variablen NR bilden Sie eine Tabelle, in der die Fallnummer der so ausgewählten Fälle angezeigt wird. In unserem Beispiel ist es nur der Fall 6.

Parallel verfahren wir bei der Identifikation des zweiten fehlerhaften Falles. Allerdings ist hier die Auswahlbedingung etwas komplizierter, da zwei Bedingungen gleichzeitig gegeben sein müssen: die Versionsnummer VN = 1 und TREUE = 0. Die Auswahlbedingung muss lauten:

vn = 1 & treue = 0

Die anderen Schritte können Sie selbst vollziehen. Die resultierende Liste der Fälle macht deutlich, dass der Fall 4 der gesuchte Fall ist.

Ergebnis der Dateninspektion ist, dass der Fall 6 bei Variable GESCHL anstelle einer 3 eine 1 bekommen muss. Bei Fall 4 ist der Wert der Variablen TREUE falsch. Er muss nun 3 statt 0 lauten.

Wechseln Sie in das Dateneditorfenster, indem Sie es anklicken oder im Hauptmenü „Fenster" den entsprechenden Dateinamen (wenn Sie unserer Empfehlung gefolgt sind, lautet er ALLBUS) anklicken, wechseln Sie gegebenenfalls in die Datenansicht, und ändern Sie die Werte, indem Sie sie einfach durch den richtigen Wert überschreiben. Sichern Sie die bereinigten Daten, indem Sie im Hauptmenü „Datei" die Option „Daten speichern" wählen. (Vergessen Sie nicht, vorher bei TREUE den fehlenden Wert 0 wieder zu deklarieren !)

2.5 Einfache statistische Auswertungen

2.5.1 Häufigkeitstabellen

Die meisten Auswertungen beginnen mit einfachen Häufigkeitsauszählungen. Mit dem Menü „Häufigkeiten" kann man absolute und relative Häufigkeiten sowie vielfältige deskriptive statistische Maßzahlen ermitteln und die Ergebnisse grafisch aufbereiten.

Eine solche Auszählung soll für die Variable „Politisches Interesse" (POL) erstellt werden. Bevor das möglich ist, muss aber zunächst die Auswahl von Fällen aus der vorigen Übung rückgängig gemacht werden. Benutzen Sie dazu die Befehlsfolge:

- ▷ „Daten", „Fälle auswählen...".

▷ Klicken Sie in der Dialogbox auf die Schaltfläche „Zurücksetzen", und bestätigen Sie mit „OK". Jetzt ist die Auswahl aufgehoben.

Zur Erstellung der Häufigkeitstabelle gehen Sie wie folgt vor:

▷ Wählen Sie die Befehlsfolge „Analysieren", „Deskriptive Statistiken ▷", „Häufigkeiten...". Die bekannte Dialogbox öffnet sich (sollte noch eine Variable ausgewählt sein, klicken Sie auf „Zurücksetzen").
▷ Wählen Sie jetzt die Variable POL aus.
▷ Zusätzlich öffnen Sie durch Anklicken der Schaltfläche „Statistik..." die Dialogbox „Häufigkeiten: Statistik" und wählen dort in der Gruppe „Lagemaße" die Option „Modalwert" (häufigster Wert) sowie in der Gruppe „Streuung" die Optionen „Minimum" und „Maximum" durch Anklicken der zugehörigen Kontrollkästchen aus. Die ausgewählten Optionen werden durch ein Häkchen gekennzeichnet.
▷ Bestätigen Sie mit „Weiter".
▷ Öffnen Sie eine neue Dialogbox durch Anklicken der Schaltfläche „Diagramme...".
▷ Hier wählen Sie durch Anklicken eines Optionsschalters in der Gruppe „Diagrammtyp" den Diagrammtyp „Balkendiagramme" aus und legen durch Anklicken von „Prozente" in der Gruppe „Diagrammwerte" fest, dass die Höhe der Balken die Prozentwerte ausdrückt.
▷ Bestätigen Sie durch „Weiter", und starten Sie den Befehl mit „OK".

Tabelle 2.2. Häufigkeitstabelle für die Variable „Politisches Interesse"

POL Politisches Interesse

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1 Sehr stark	6	18,8	18,8	18,8
	2 Stark	8	25,0	25,0	43,8
	3 Mittel	12	37,5	37,5	81,3
	4 Wenig	6	18,8	18,8	100,0
	Gesamt	32	100,0	100,0	

Im Ausgabefenster erscheint Tabelle 2.2. In der ersten Spalte finden wir (wegen unserer Voreinstellung für die Ausgabe) Werte und Wertelabels für die Ausprägungen der Variablen sowie eine Zeile „Gesamt", welche die Angaben alle Werte enthält. Die nächste Spalte enthält die absoluten Häufigkeiten („Häufigkeit") für die Ausprägungen sowie insgesamt. So erfährt man etwa, dass von 32 Befragten 6 „sehr stark", 8 „stark" usw. politisch interessiert sind. Daneben werden die Daten in Prozentwerten, berechnet auf der Basis aller Fälle („Prozent"), angegeben. So sind 18,8 % „sehr stark", 25 % „stark" usw. interessiert. Dahinter sind die Daten ein weiteres Mal prozentuiert. Diesmal unter Ausschluss der fehlenden Werte („Gültige Prozente"). Da bei dieser Variablen keine fehlenden Werte auftreten, sind die beiden Prozentwerte identisch. Schließlich finden sich in der letzten Spalte kumulierte Prozentwerte auf der Basis der gültigen Werte („Kumulierte Prozente").

Die ausgewählten Statistiken werden in einer vorangestellten weiteren, mit „Statistiken" überschriebenen, Tabelle ausgegeben (⇨ Tabelle 2.3, sie ist gegenüber der voreingestellten Darstellung durch Pivotierung geändert ⇨ Kap. 4.1.4). Der häufigste Wert („Modus") beträgt danach 3, der niedrigste („Minimum") 1, der höchste („Maximum") 4. Diese Tabelle enthält auch Angaben über die Zahl der gültigen und fehlenden Fälle. Werden mehrere Häufigkeitsauszählungen in einem Lauf abgerufen, sind diese Angaben in einer einzigen, den Häufigkeitstabellen vorangestellten, Tabelle zusammengefasst.

Auch das Diagramm wird im Ausgabefenster angezeigt. Will man es weiter bearbeiten, muss man durch Doppelklick auf die Grafik den Diagramm-Editor öffnen (⇨ Kap. 27.1)

Tabelle 2.3. Statistiken zur Häufigkeitsauszählung

Statistiken

	N		Modus	Minimum	Maximum
	Gültig	Fehlend			
POL Politisches Interesse	32	0	3	1	4

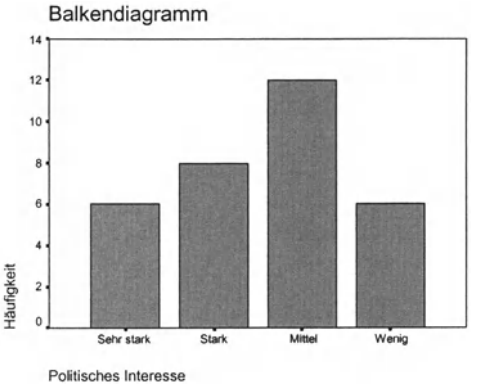


Abb. 2.15. Balkendiagramm für die Variable „Politisches Interesse"

Nun ein Beispiel für metrische Daten. Wir wollen die Verteilung der Einkommen in der Untersuchungsgruppe betrachten. Das Einkommen ist in der Variablen EINK auf eine DM genau erfasst. Es kann also sehr viele verschiedene Werte geben, die zumeist auch noch mit wenigen Fällen besetzt sind. Würde man hier einfach eine Häufigkeitstabelle erstellen, ergäbe sich ein sehr unübersichtliches Bild. Zur Verbesserung der Übersichtlichkeit wollen wir daher Einkommensklassen bilden. Die Klassen sollen mit Ausnahme der ersten eine Spannweite von 1000 DM besitzen. Die erste geht nur von 1 bis unter 500 DM. Der Sinn dieser Festlegung

ist, dass die Klassenmitten der anderen Klassen immer bei den häufig angegebenen ganzen Tausenderwerten liegen. Dadurch wird die Verzerrung aufgrund der Klassenbildung geringer. Die nächsten Klassen reichen also von 500 bis unter 1500, 1500 bis unter 2500 etc.. Den neuen Klassen soll die Klassenmitte als Wert zugeordnet werden. Das ist notwendig, damit statistische Kennwerte richtig berechnet werden. Wir kodieren also die Variable um. Wir wollen aber die Ausgangswerte nicht verlieren, denn aus diesen lassen sich statistische Kennwerte wie das arithmetische Mittel und die Standardabweichung genauer ermitteln. Deshalb erfasst die neue Variable EINK2 die umkodierten Daten. Zum Umkodieren gehen Sie wie folgt vor:

- ▷ Wählen Sie (im Daten-Editor) im Menü „Transformieren“ die Option „Umkodieren ▷“. Es öffnet sich ein Untermenü mit den Optionen „In dieselben Variablen...“ und „In andere Variablen...“.
- ▷ Wählen Sie „In andere Variablen ...“. Es öffnet sich die in Abb. 2.16 dargestellte Dialogbox, in der die Umkodierung vorgenommen wird. Dafür gehen Sie wie folgt vor:
- ▷ Übertragen Sie EINK aus der Quellvariablenliste in das Auswahlfeld „Numerische Var. → Ausgabevar.:“[6]. Anstelle des Namens der Ausgabevariablen steht noch ein Fragezeichen. In den zwei Feldern der Gruppe „Ausgabevariable“ können wir jetzt einen neuen Variablennamen und zugleich ein Variablen-Label vergeben.
- ▷ Tragen Sie in das Feld „Name:“ EINK2 ein, und bestätigen Sie die Eingabe durch Anklicken der Schaltfläche „Ändern“. In das Feld „Label:“ geben Sie „monatliches Nettoeinkommen (klassifiziert)“ ein. Damit ist eine neue Variable definiert.

Abb. 2.16. Dialogbox „Umkodieren in andere Variablen“

- ▷ Klicken Sie auf die Schaltfläche „Alte und neue Werte...“. Es öffnet sich die in Abb. 2.17 dargestellte Dialogbox. Links ist eine Gruppe zum Eintragen der al-

[6] Die Bezeichnung variiert nach Variablentyp.

ten Werte („Alter Wert"), rechts eine, in die die neuen Werte eingetragen werden („Neuer Wert").

- ▷ Da wir ganze Bereiche zu einem neuen Wert zusammenfassen wollen, klicken Sie zunächst den Optionsschalter vor „Bereich" an. Die Bereiche dürfen sich nicht überschneiden. Weil eine DM die kleinste Maßeinheit ist, geben wir daher für die erste Klasse in das linke Feld 1 und in das rechte Feld hinter „bis" den Wert 499 ein, um einen Bereich von 1 bis 499 (= unter 500 DM) festzulegen.
- ▷ Geben Sie dann in der Gruppe „Neuer Wert" in das Feld „Wert" 250 ein. Das ist der Klassenmittelwert, den wir als neuen Wert benutzen wollen.
- ▷ Durch Anklicken der Schaltfläche „Hinzufügen" übertragen Sie diese Definition in das Feld „Alt → Neu:"
- ▷ Wiederholen Sie dasselbe für die Bereiche: 500 bis 1499, 1500 bis 2499, 2500 bis 3499, 3500 bis 4499 und 4500 bis 5499. Alle anderen Werte werden dann automatisch systemdefinierte fehlende Werte. 0 und 99997 dürfen bei dieser Umkodierung nicht mit eingeschlossen werden, weil sie weiterhin als fehlende Werte deklariert bleiben sollen, allerdings werden sie zu systemdefinierten fehlenden Werten. (Denselben Effekt hätte man, würde man die Kombination „alle anderen Werte" und „Systemdefiniert fehlend" anwählen. Wollte man dagegen beides als nutzerdefinierte fehlende Werte behalten, müsste man die Kombination „Alle anderen Werte" und „Alte Werte kopieren" auswählen und nachträglich in der neuen Variablen diese als fehlende Werte deklarieren.)

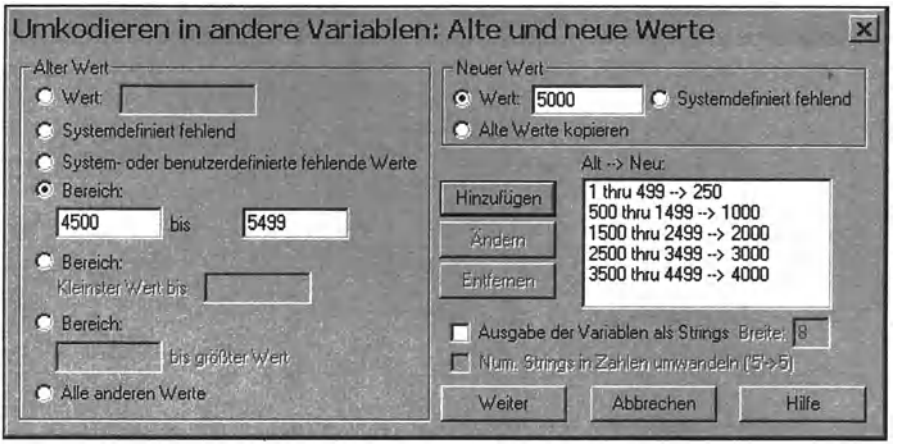


Abb. 2.17. Dialogbox „Umkodieren in andere Variablen: Alte und neue Werte"

- ▷ Bestätigen Sie mit „Weiter" und „OK". Das Dateneditorfenster öffnet sich, und Sie sehen, wie die neue Variable und ihre Werte eingetragen werden.

Jetzt können wir für diese neu gebildete Variable eine Häufigkeitsauszählung vornehmen.

- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Deskriptive Statistiken ▷", „Häufigkeiten...", und wählen Sie die Variable EINK2 aus.

▷ Klicken Sie auf die Schaltfläche „Statistik...“, und wählen Sie in der nun geöffneten Dialogbox in der Gruppe „Lagemaße“ die Option „Mittelwert“ („arithmetisches Mittel“) und in der Gruppe „Streuung“ „Std.abweichung“ („Standardabweichung“) aus. Bestätigen Sie mit „Weiter“.

Tabelle 2.4. Häufigkeitstabelle für die Variable EINK2

Statistiken

	N			
	Gültig	Fehlend	Mittelwert	Standardabweichung
EINK2	19	13	2131,5789	1329,0776

EINK2

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	250,00	2	6,3	10,5	10,5
	1000,00	5	15,6	26,3	36,8
	2000,00	5	15,6	26,3	63,2
	3000,00	4	12,5	21,1	84,2
	4000,00	2	6,3	10,5	94,7
	5000,00	1	3,1	5,3	100,0
	Gesamt	19	59,4	100,0	
Fehlend	System	13	40,6		
Gesamt		32	100,0		

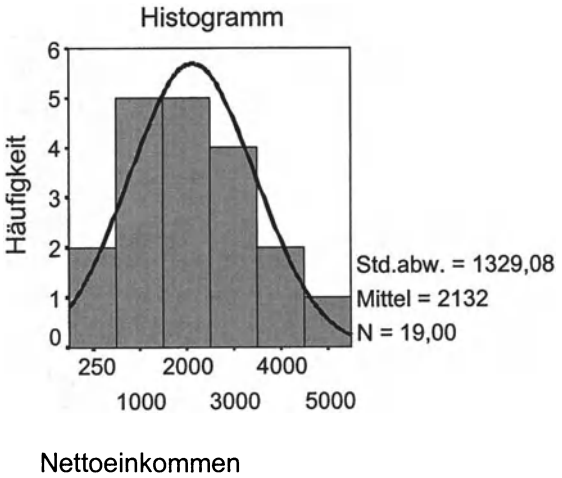


Abb. 2.18. Histogramm mit überlagerter Normalverteilung für die Variable EINK2

▷ Klicken Sie das Schaltfeld „Diagramme..“ an, und wählen Sie in der darauf erscheinenden Dialogbox die Option „Histogramme“ und zusätzlich „Mit Normalverteilungskurve“. (Die Optionen für die Diagrammwerte stehen bei Wahl

des Histogramms nicht zur Verfügung. Auf der senkrechten Achse werden immer absolute Häufigkeiten eingetragen.) Bestätigen Sie die Wahl mit „Weiter".

▷ Mit „OK" in der Dialogbox „Häufigkeiten" führen Sie den Befehl aus. Als Ergebnis erscheint die in Tabelle 2.4 dargestellte Ausgabe, eine Doppeltabelle (Die erste Tabelle ist in unserer Darstellung durch Pivotierung geändert ⇨ Kap. 4.1.4).

Die untere Tabelle zeigt für die einzelnen Klassen die absoluten und die Prozentwerte. Diesmal ist die Prozentuierung der gültigen Werte („Gültige Prozente") interessant, weil immerhin bei 13 Fällen keine gültigen Werte vorliegen. In der Tabelle darüber finden wir u.a. das arithmetische Mittel („Mittelwert") mit 2131,5789 angegeben und die Standardabweichung mit 1329,0776. Sie sollten zum Vergleich einmal dieselben statistischen Kennwerte für die nicht klassifizierte Variable „EINK" ermitteln. Sie werden dann sehen, dass diese nicht identisch sind. Das liegt daran, dass die klassifizierten Werte ungenauer sind als die Ausgangswerte.
Im Ausgabefenster finden Sie auch ein Histogramm der klassifizierten Einkommensverteilung. Überlagert ist diese durch eine Kurve, die anzeigt, wie die Daten verteilt sein müssten, läge eine Normalverteilung vor. Die Beschriftung der ersten Säule des Histogramms ist nicht richtig. Sie beträgt 0 statt 250. Das liegt daran, dass SPSS von gleichen Klassenbreiten ausgeht. Innerhalb von SPSS lässt sich das nicht ändern. (Wenn Sie die Grafik z.B. nach WORD exportieren und dort als Grafik bearbeiten, können Sie die falsche Beschriftung korrigieren.) (Auch bei gleichen Klassenbreiten kann es zu einer falschen Beschriftung kommen. Dies kann man dann aber in SPSS selbst im Grafikfenster in der Dialogbox „Intervallachse: Benutzerdefinierte..." durch Angabe passender Minimum- bzw. Maximumwerte und/oder einer passenden Intervallbreite/Intervallzahl anpassen [⇨ Kap. 27.4.3]).

2.5.2 Kreuztabellen

In den meisten Fällen wird man auch den Zusammenhang von zwei und gegebenenfalls mehr Variablen untersuchen wollen. Das einfachste Verfahren dazu ist die Erstellung einer Kreuztabelle. Das Untermenü „Kreuztabellen..." bietet die dazu notwendigen Prozeduren. Darüber hinaus kann man auch Zusammenhangsmaße (Korrelationskoeffizienten) als statistische Kennzahlen errechnen lassen und die statistische Bedeutsamkeit (Signifikanz) eines Zusammenhanges überprüfen. Wir wollen als Beispiel den Zusammenhang zwischen Geschlecht (Variable GESCHL) und politischem Interesse (Variable POL) untersuchen.

▷ Wählen Sie dazu die Befehlsfolge „Analysieren", „Deskriptive Statistiken ▷", „Kreuztabellen...". Die Dialogbox „Kreuztabellen" öffnet sich (⇨ Abb. 2.13).

Hier kann man die Variablen für eine Kreuztabelle auswählen und gleichzeitig angeben, welche im Kopf und welche in der Vorspalte der Tabelle stehen soll. Wenn es der Umfang der Ausprägungen nicht anders verlangt, liegt es nahe, die unabhängige Variable in den Kopf der Tabelle zu nehmen. Das ist in unserem Falle das Geschlecht.

▷ Übertragen Sie die Variable GESCHL aus der Liste der Quellvariablen in das Auswahlfeld „Spalten:". GESCHL wird damit im Kopf der Tabelle stehen. Die

Ausprägungen „männlich“ und „weiblich“ werden die Spaltenüberschriften bilden.

▷ Markieren Sie dann POL, und übertragen Sie diese Variable in das Auswahlfeld „Zeilen:“.

Abb. 2.19. Dialogbox „Kreuztabellen: Zellen anzeigen“

Absolute Häufigkeiten sind im allgemeinen schwer zu interpretieren. Deshalb sollen sie in Prozentwerte umgerechnet werden. Bei einer Kreuztabelle ist zu entscheiden, in welcher Richtung die Prozentuierung erfolgen soll. Steht die unabhängige Variable im Kopf der Tabelle, ist eine spaltenweise Prozentuierung angemessen. Dadurch werden die verschiedenen Gruppen, die den Ausprägungen der unabhängigen Variablen entsprechen, vergleichbar.

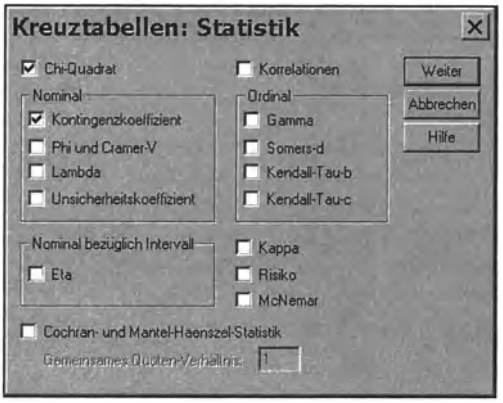


Abb. 2.20. Dialogbox „Kreuztabellen: Statistik“

▷ Klicken Sie auf die Schaltfläche „Zellen...“. Es öffnet sich eine Dialogbox (⇨ Abb. 2.19). Wählen Sie hier in der Gruppe „Prozentwerte“ die Option „Spaltenweise“. Bestätigen Sie die Auswahl mit „Weiter“.

▷ Klicken Sie dann auf die Schaltfläche „Statistik...“. Es öffnet sich eine Dialogbox (⇨ Abb. 2.20). Wählen Sie dort die Option „Chi-Quadrat“ und in der

Gruppe „Nominal" die Option „Kontingenzkoeffizient". Bestätigen Sie mit „Weiter". Starten Sie den Befehl mit „OK".

Tabelle 2.5. Kreuztabelle „Politisches Interesse nach Geschlecht"

Politisches Interesse * Geschlecht Kreuztabelle

			Geschlecht		
			männlich	weiblich	Gesamt
Politisches Interesse	Sehr stark	Anzahl	4	2	6
		% von Geschlecht	22,2%	14,3%	18,8%
	Stark	Anzahl	8		8
		% von Geschlecht	44,4%		25,0%
	Mittel	Anzahl	5	7	12
		% von Geschlecht	27,8%	50,0%	37,5%
	Wenig	Anzahl	1	5	6
		% von Geschlecht	5,6%	35,7%	18,8%
Gesamt		Anzahl	18	14	32
		% von Geschlecht	100,0%	100,0%	100,0%

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	11,344[a]	3	,010
Likelihood-Quotient	14,515	3	,002
Zusammenhang linear-mit-linear	6,269	1	,012
Anzahl der gültigen Fälle	32		

a. 6 Zellen (75,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 2,63.

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Kontingenzkoeffizient	,512	,010
Anzahl der gültigen Fälle		32	

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

Wir erhalten die in Tabelle 2.5 auszugsweise dargestellte Ausgabe. Zunächst sehen wir die eigentliche Kreuztabelle. Dort enthält jede Zelle eine Zeile, in der die Absolutzahlen („Anzahl") der jeweiligen Wertekombinationen angegeben sind. Darunter stehen die Spaltenprozente, ausgewiesen mit einer Kommastelle. Ein Vergleich der Prozentwerte für Männer und Frauen macht deutlich, dass Männer wesentlich häufiger angeben, an Politik sehr starkes oder starkes Interesse zu haben

als Frauen. Bei Männern betragen die Prozentwerte 22,2 und 44,4, bei Frauen dagegen lediglich 14,3 und 0.

In der zweiten Tabelle finden wir die Ergebnisse verschiedener Varianten des Chi-Quadrat-Tests. Er erlaubt es zu überprüfen, ob eine gefundene Differenz der Häufigkeiten als statistisch abgesichert angesehen werden kann (signifikant ist) oder nicht. Betrachten wir nur die erste, mit „Chi-Quadrat nach Pearson" beschriftete Reihe. Diese zeigt einen Chi-Quadrat-Wert von 11,344; 3 Freiheitsgrade („df") und eine Wahrscheinlichkeit, dass ein solches Ergebnis bei Geltung von H_0 (der Hypothese, dass kein Zusammenhang besteht) zustande kommt („Asymptotische Signifikanz (2-seitig)"), von 0,01. Üblicherweise erkennt man einen Unterschied erst als signifikant an, wenn dieser Wert 0,05 oder kleiner ist (signifikant) bzw. 0,01 oder kleiner (hoch signifikant). Also ist in unserem Falle der Unterschied tatsächlich hoch signifikant. Es ist statistisch abgesichert, dass Männer häufiger ein hohes bzw. sehr hohes Interesse an Politik haben.

Hinweis. Ein Problem ist die geringe Zahl der untersuchten Fälle. Der Chi-Quadrat-Test sollte eigentlich nicht durchgeführt werden, wenn sich für zu viele Zellen eine Besetzung von weniger als fünf erwarteten Fällen ergibt. Die Anmerkung zum Output gibt aber für sechs von acht Zellen ein Erwartungswert von < 5 an. In solchen Fällen bietet das neue Modul „Exakte Tests" eine genaue Testmöglichkeit (⇨ Kap. 29).

In einer weiteren Tabelle ist der Kontingenzkoeffizient von 0,512 angegeben. Er zeigt einen für sozialwissenschaftliche Untersuchungen durchaus beachtlichen Zusammenhang zwischen den beiden Variablen an. Da sie aus dem Chi-Quadrat-Test entwickelt ist, ergibt die Signifikanzprüfung für den Kontingenzkoeffizienten dasselbe Ergebnis wie der Chi-Quadrat-Test.

2.5.3 Mittelwertvergleiche

Wenn die Daten der abhängigen Variablen auf einer metrischen Skala gemessen wurden, die Daten der unabhängigen dagegen auf Nominalskalenniveau (oder bei höherem Messniveau zu Gruppen zusammengefasst wurden), kann die Option „Mittelwerte..." im Menü „Mittelwerte vergleichen ▷" eine ähnliche Funktion wie „Kreuztabellen" erfüllen. Allerdings werden hier für die Gruppen, die den Ausprägungen der unabhängigen Variablen entsprechen, die Mittelwerte der abhängigen Variablen verglichen. Das bietet sich z.B. an, wenn untersucht werden soll, ob Männer im Durchschnitt ein höheres Einkommen haben als Frauen.

- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Mittelwerte vergleichen ▷", „Mittelwerte...". Es öffnet sich die in Abbildung 2.21 dargestellte Dialogbox. Hier müssen Sie angeben, welche Variable die unabhängige und welche die abhängige sein soll.
- ▷ Übertragen Sie aus der Quellvariablenliste die abhängige Variable EINK in das Eingabefeld „Abhängige Variablen:". Übertragen Sie die unabhängige Variable GESCHL in das Eingabefeld „Unabhängige Variablen:".
- ▷ Klicken Sie die Schaltfläche „Optionen..." an. Es öffnet sich eine Dialogbox (⇨ Abb. 2.22). In der Liste „Zellenstatistik" sind die Optionen „Mittelwert", „Stan-

dardabweichung“ und „Anzahl der Fälle“ bereits ausgewählt. Weitere könnte man aus der Liste „Statistik“ übertragen. (Wir verzichten darauf.)

Abb. 2.21. Dialogbox „Mittelwerte“

▷ Mit „Weiter“ bestätigen wir die Eingabe. „OK“ startet den Befehl. Als Ausgabe erhalten wir die Tabelle 2.6.

Abb. 2.22. Dialogbox „Mittelwerte: Optionen“

Hier ist für alle gültigen Fälle („insgesamt“) sowie für die Vergleichsgruppen Männer und Frauen jeweils das arithmetische Mittel („Mittelwert“) für das Einkommen angegeben. Es beträgt bei Männern 2320,36 DM, bei Frauen 1370,00 DM. Wie man sieht, haben die Männer im Durchschnitt ein sehr viel höheres Einkommen. Außerdem sind die Standardabweichung für die Gesamtpopulation und die beiden Gruppen sowie die Fallzahlen („N“) enthalten. (Vorangestellt ist wieder eine Tabelle „Verarbeitete Fälle“, die hier nicht dargestellt wird.)

Tabelle 2.6. Ergebnis von „Mittelwerte" für Einkommen nach Geschlecht

Bericht

Monatliches Nettoeinkommen

Geschlecht	Mittelwert	N	Standardabweichung
männlich	2320,36	14	1093,72
weiblich	1370,00	5	1503,16
Insgesamt	2070,26	19	1245,35

2.6 Index bilden, Daten transformieren

Aus den vier Variablen RUHE, EINFLUSS, INFLATIO und MEINUNG soll ein zusammenfassender Index, der sogenannte Inglehart-Index, gebildet werden. (Dieser Index wurde von Ronald Inglehart [1971] entwickelt und spielt eine große Rolle in der sogenannten „Wertewandeldiskussion".) Bei allen vier Variablen ist festgehalten, ob der Befragte sie im Vergleich zu den anderen in der Wichtigkeit an die erste, zweite, dritte oder vierte Stelle setzt. Der Inglehart-Index soll die Befragten nach folgender Regel in vier Gruppen einteilen. Als „reine Postmaterialisten" (= 1) sollen diejenigen eingestuft werden, die EINFLUSS und MEINUNG in beliebiger Reihenfolge auf die beiden ersten Plätze setzten. Als „reiner Materialist" (= 4) dagegen soll eingestuft werden, wer bei der Einordnung der vier Aussagen nach Wichtigkeit RUHE und INFLATIO an die ersten beiden Stellen setzt, gleichgültig in welcher Reihenfolge. Dagegen sollen „tendenzielle Postmaterialisten" (= 2) diejenigen heißen, die entweder EINFLUSS oder MEINUNG an die erste und eine der beiden anderen Aspekte auf die zweite Stelle gesetzt haben. Schließlich seien „tendenzielle Materialisten" (= 3) solche, die von den beiden Aussagen RUHE und INFLATIO eine auf den ersten, von den beiden anderen eine auf den zweiten Platz setzen.

Die neue Variable INGL wird durch Transformation der vier alten Variablen gebildet. Dazu bedarf es einer relativ komplexen Befehlsfolge, da jeder Wert mit Hilfe eines Bedingungsausdruckes gebildet werden muss. Es ist in diesem Fall einfacher, nicht jeden Befehl einzeln auszuführen, sondern die ganze Befehlsfolge zunächst in ein Syntaxfenster zu übertragen und dann zusammen abzuarbeiten. Dazu gehen Sie wie folgt vor:

- ▷ Wählen Sie im Dateneditor im Menü „Transformieren" das Untermenü „Berechnen...". Es öffnet sich die in Abb. 2.23 dargestellte Dialogbox, in der auf verschiedene Weise Datentransformationen vorgenommen werden können.
- ▷ Tragen Sie als Namen der neuen Variablen INGL in das Eingabefeld „Zielvariable:" ein. Darauf 1 in das Eingabefeld „Numerischer Ausdruck:". Es ist der Wert, der vergeben werden soll, wenn die erste Bedingung erfüllt ist.
- ▷ Da Sie jetzt diesen Bedingungsausdruck bilden müssen, wählen Sie die Schaltfläche „Falls...". Es öffnet sich eine Dialogbox, in der die Bedingung definiert werden kann.

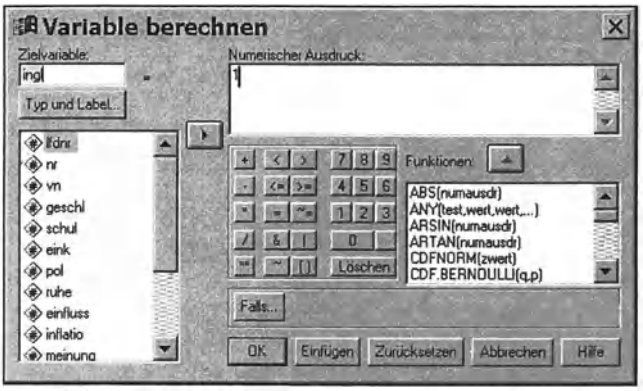


Abb. 2.23. Dialogbox „Variable berechnen"

▷ Wählen Sie „Fall einschließen, wenn Bedingung erfüllt ist:" .
Jetzt stellen Sie die erste Bedingung zusammen, so dass das in der Abb. 2.24 ersichtliche Ergebnis entsteht. Dazu gehen Sie wie folgt vor:

Abb. 2.24. Dialogbox „Variable berechnen: Falls Bedingung erfüllt ist"

▷ Wählen Sie zuerst in der Rechnertastatur die Doppelklammer „( )". Sie wird dadurch in das Definitionsfeld eingetragen. Wählen Sie dann in der Quellvariablenliste EINFLUSS aus und dann aus der Rechnertastatur nacheinander „=„ ; „1" und das „&" (= logisches „Und"). Danach übertragen Sie die Variable MEINUNG und aus der Rechnertastatur „=„ und „2".

▷ Setzen Sie den Cursor hinter die Klammer im Definitionsfeld, und wählen Sie zunächst | (das logische „Oder") aus. Dann nacheinander „()"; EINFLUSS; „=„; „2" ; „&"; MEINUNG; „=„ und „1". Mit „Weiter" bestätigen Sie die Eingabe. Es öffnet sich wieder die Dialogbox „Variable berechnen". Der Bedingungsausdruck ist jetzt auch hier neben der Schaltfläche „Falls..." eingetragen.

▷ Wählen Sie die Schaltfläche „Einfügen". Das Syntaxfenster öffnet sich, und der soeben gebildete Befehl ist in der SPSS-Befehlssyntax eingetragen. Zusätzlich der Befehl: „EXECUTE".

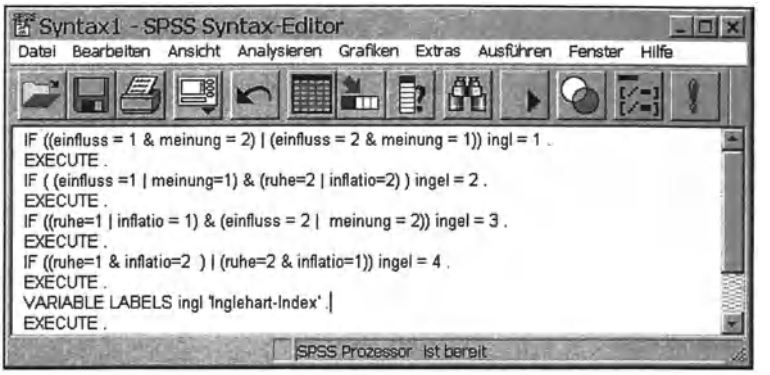


Abb. 2.25. Befehlssyntax für die Bildung des Inglehart-Index im Syntaxfenster

Wir führen diesen Befehl jedoch nicht aus, sondern bilden auf dieselbe Weise jetzt nach und nach die drei anderen Bedingungen und übertragen sie ebenfalls ins Syntaxfenster.

▷ Bevor Sie die letzte so gebildete Bedingung in das Syntaxfenster übertragen, wählen Sie in der Dialogbox „Variablen berechnen:" die Schaltfläche „Typ und Label...". Es öffnet sich eine Dialogbox. Tragen Sie dort in das Eingabefeld „Label" das Variablen-Label „Inglehart-Index" für die neue Variable ein. Bestätigen Sie diesen mit „Weiter" und übertragen Sie auch die letzte Definition in das Syntaxfenster. Das Ergebnis müsste mit Abbildung 2.25 übereinstimmen.

▷ Markieren Sie die gesamte Befehlsfolge im Syntaxfenster, und klicken Sie in der Symbolleiste auf ▶. Die gesamte Befehlsfolge wird jetzt abgearbeitet.

Tabelle 2.7. Häufigkeitstabelle für den Inglehart-Index

Inglehart-Index

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	reine Postmaterialisten	13	40,6	40,6	40,6
	tendenzielle Postmaterialisten	10	31,3	31,3	71,9
	tendenzielle Materialisten	7	21,9	21,9	93,8
	reine Materialisten	2	6,3	6,3	100,0
	Gesamt	32	100,0	100,0	

Wenn Sie in das Dateneditorfenster schalten, ist als letzte die neue Variable mit dem Namen INGL zu sehen, die Werte zwischen 1 und 4 enthält. Wenn Sie wollen, vergeben Sie auch noch die oben genannten Bezeichnungen als „Werte-La-

bels". Bilden Sie dann eine Häufigkeitstabelle für die neue Variable. Das Ergebnis zeigt Tabelle 2.7.

2.7 Gewichten

Erstellen Sie zunächst mit der Befehlsfolge „Analysieren", „Deskriptive Statistiken ▷", „Häufigkeiten" eine Häufigkeitstabelle für die Variable Geschlecht (GESCHL). Es ergibt sich die Tabelle 2.8.

Tabelle 2.8. Häufigkeitstabelle für die Variable Geschlecht

Geschlecht

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	männlich	18	56,3	56,3	56,3
	weiblich	14	43,8	43,8	100,0
	Gesamt	32	100,0	100,0	

Die Tabelle zeigt, dass in unserer kleinen Stichprobe 56,3 % der Fälle Männer sind und 43,8 % Frauen. Wir können dieses Ergebnis zur Überprüfung der Repräsentativität unserer Auswahl benutzen. Die „wahre" Verteilung können wir näherungsweise einer Sonderzählung des Mikrozensus 1989 entnehmen. Demnach waren 47,1 % der Zielbevölkerung des ALLBUS männlichen und 52,9 % weiblichen Geschlechts. Die Stichprobe enthält demnach zu viele Männer und zu wenige Frauen. Sie ist gegenüber der Grundgesamtheit verzerrt.

Man kann nun versuchen, diese Verzerrung durch Gewichtung zu beseitigen. Ein einfaches Verfahren besteht darin, einen Gewichtungsfaktor für jede Ausprägung der Variablen aus der Relation Soll zu Ist zu entwickeln.

$$G_i = \frac{SOLL}{IST} \tag{2.1}$$

Entsprechend errechnet sich für die Männer ein Gewichtungsfaktor:

$$G_M = \frac{47{,}1}{56{,}3} = 0{,}84$$

und für Frauen

$$G_W = \frac{52{,}9}{43{,}8} = 1{,}21$$

Wir wollen unsere Daten mit diesen Faktoren gewichten. Dazu muss zunächst eine Gewichtungsvariable GEWICHT gebildet werden, in die für Männer und Frauen jeweils das zugehörige Gewicht eingetragen wird. Dann werden die Daten mit diesen Gewichtungsfaktoren gewichtet.

Bilden Sie mit Hilfe der Befehlsfolge „Transformieren", „Berechnen...", „Variable berechnen" die Variable Gewicht. Benutzen Sie dazu das Syntaxfenster. Die Vorgehensweise ist dieselbe wie im vorigen Beispiel. Bei Dezimalzahlen muss ein Dezimalpunkt verwendet werden. Das Ergebnis im Syntaxfenster muss dem folgenden Bild entsprechen. (Sie können die Befehle im Syntaxfenster auch einfach eintippen.)

```
IF (geschl = 1) Gewicht = 0.84 .
EXECUTE .
IF (geschl = 2) Gewicht = 1.21 .
EXECUTE .
```

Zur Durchführung der Gewichtung :

- ▷ Wählen Sie die Befehlsfolge „Daten", „Fälle gewichten...". Es erscheint die Dialogbox „Fälle gewichten" (⇨ Abb. 2.26).
- ▷ Klicken Sie auf den Optionsschalter vor „Fälle gewichten mit der". Damit wird die Gewichtung für die folgenden Befehle eingeschaltet.
- ▷ Wählen sie aus der Variablenliste die Variable GEWICHT.
- ▷ Bestätigen Sie mit „OK". Die Gewichtung wird für nachfolgende Prozeduren durchgeführt.

Abb. 2.26. Dialogbox „Fälle gewichten"

Zur Überprüfung des Ergebnisses bilden Sie erneut eine Häufigkeitstabelle für die Variable Geschlecht (GESCHL). Diese zeigt nun für die Männer 47,2 % und die Frauen 52,8 % an, also beinahe die exakte Verteilung. Beachten Sie, dass die Gesamtzahl der Fälle nicht verändert ist.

Die Gewichtung, die zunächst nur auf den Abweichungen bei der Variablen Geschlecht beruht, wirkt sich selbstverständlich auch auf die anderen Variablen aus. Überprüfen Sie das, indem sie zwei Häufigkeitsverteilungen für die klassifizierten Einkommensdaten Variable EINK2 erstellen, zunächst mit den gewichteten Daten, dann ohne Gewichtung. Die Gewichtung schalten Sie aus, indem Sie mit der Befehlsfolge „Daten", „Fälle gewichten..." die Dialogbox „Fälle gewichten" öffnen und dort die Option „Fälle nicht gewichten" auswählen.

3 Definieren und Modifizieren einer Datendatei

SPSS kann Datendateien verarbeiten, die mit verschiedenen anderen Programmen erstellt wurden (⇨ Kap. 6). Nach dem Import erscheinen die Daten im Dateneditorfenster, Registerblatt „Datenansicht". Dieses ähnelt dem Arbeitsblatt eines Tabellenkalkulationsprogrammes. Hier können die importierten Daten auch weiter verarbeitet und geändert werden. Daneben enthält der Dateneditor das Registerblatt „Variablenansicht". Es ähnelt ebenfalls dem Arbeitsblatt eines Tabellenkalkulationsblattes, enthält aber die Variablendefinition.

Auf dem Registerblatt „Datenansicht" des Dateneditors von SPSS können aber auch die Daten eingetippt werden. Im folgenden wird die Definition einer Datenmatrix, die Dateneingabe und Bearbeitung im Editor besprochen. Die grundsätzliche Arbeitsweise des Dateneditors wurde schon ausführlich in Kapitel 2.3 erörtert. Das vorliegende Kapitel macht ergänzende Angaben.

Grundsätzlich werden die Daten auf dem Blatt „Datenansicht" in Form einer rechteckigen Matrix eingegeben. Jede Zeile entspricht einem Fall (z.B. einer Person), jede Spalte der Matrix einer Variablen. In den Zellen sind die Werte einzutragen.

Im folgenden wird davon ausgegangen, dass in der Ländereinstellung der Windows-Systemsteuerung das Komma als Dezimaltrennzeichen eingestellt ist .

3.1 Definieren von Variablen

Name, Format und Labels einer Variablen werden auf dem Blatt „Variablenansicht" des Dateneditors (⇨ Abb. 2.10) festgelegt. Man kann entweder alle oder einzelne Voreinstellungen akzeptieren oder Einstellungen ändern.

Vorgehensweise. Um eine Variable zu definieren:

- ▷ Gehen Sie gegebenenfalls auf das Blatt „Variablenansicht", indem Sie im Dateneditor das Registerblatt „Variablenansicht" anklicken oder im „Datenblatt" auf den Namen der zu definierenden Variablen klicken. Im letzteren Falle ist zugleich die Zeile der angewählten Variablen markiert.
- ▷ Tragen Sie im Feld „Variablenname:" den gewünschten Variablennamen ein.
- ▷ Um Variablentyp, fehlende Werte (Missings) oder Wertelabels zu definieren, müssen Sie jeweils eine Dialogbox öffnen. Zur Definition des Variablentyps aktivieren Sie z.B. die entsprechende Zelle der Spalte „Typ" durch Anklicken mit der linken Maustatste. Die Zelle wird hervorgehoben, auf der rechten Seite erscheint ein unterlegtes Kästchen mit drei Punkten. Wenn Sie das Kästchen

anklicken, erscheint die Dialogbox „Variablentyp definieren", in der Sie die weitere Definition vornehmen. Entsprechend ist das Vorgehen bei den anderen angegebenen Formatelementen.

- ▷ Die Definitionsspalten „Spaltenformat" (gibt die Zahl der Stellen bei der Variablendarstellung an), „Dezimalstellen", „Spalten" (gibt die im Datenblatt des Editors angezeigte Spaltenbreite an) werden die Angaben etwas anders bearbeitet. Zur Definition des Variablenbreite aktivieren Sie z.B. die entsprechende Zelle der Spalte „Variablenformat" durch Anklicken mit der linken Maustatste. Die Zelle wird hervorgehoben, auf der rechten Seite erscheinen zwei Pfeile. Durch Anklicken eines dieser Pfeile verringert oder erhöht man die angegebene Zahl. (Man kann auch nach Doppelklick auf die Zelle die Zahl einfach markieren und überschreiben.) Entsprechend ist das Vorgehen bei der Definition der anderen abgegebenen Formatelemente.
- ▷ In der Spalte Feld „Messniveau" wird bei allen Variablentypen außer „String" die Voreinstellung „Metrisch" eingestellt. Sollten ihre Daten dem nicht entsprechen, können Sie „Ordinal" oder „Nominal" wählen. Bei einer Stringvariablen dagegen ist „Nominal" eingestellt. Gegebenenfalls können Sie „Ordinal" wählen (zum Messniveau ⇨ Kap. 8.3.1). Dies geschieht nach Aktivieren der entsprechenden Zellen und Anklicken des auf deren rechten Seite erscheinenden Pfeils durch Markieren des gewünschten Messniveaus in der sich öffnende Auswahlliste.
- ▷ Neben der Spalte „Namen" enthält auch die Spalte „Variablenlabel" reine Eingabefelder. Hier tragen Sie den gewünschten Namen oder das gewünschte Label einfach ein.

Variablennamen. Es gelten folgende Regeln:

- ❒ Der Name darf maximal 8 Zeichen umfassen nicht mit einem Punkt enden.
- ❒ Der Name muss mit einem Buchstaben beginnen. Ansonsten gelten auch Ziffern, Punkte und die Symbole @, #, _ und $.
- ❒ Er darf keine Leerzeichen oder die Zeichen !, ?, ` und * enthalten.
- ❒ Ein Variablenname darf nur einmal auftreten.
- ❒ Groß und Kleinbuchstaben sind gleichwertig.
- ❒ Nicht verwendet werden können die Schlüsselwörter: ALL; NE; EQ; TO; LE; LT; BY; OR; GT; AND; NOT; GE; WITH.

Beachten Sie bitte: Alle Variablen erscheinen zur Auswahl für die statistische Analyse in der Quellvariablenliste. Bei Umbenennung erscheint der neue Name sofort in diesem Feld. In einem vorherigen Auswertungslauf ausgewählte Variablen werden jedoch nicht aktualisiert. Diese müssen erst aus der Liste ausgewählter Variablen entfernt werden.

Variablen- und Wertelabels (Etiketten). Ein Variablenlabel wird einfach in die entsprechende Zelle der Spalten „Variablenlabel" eingetragen. Es kann bis zu 256 Zeichen lang sein. Groß- und Kleinschreibung werden beachtet. Durch Klicken auf das unterlegte Quadrat auf der rechten Seite einer aktivierten Zelle der Spalte „Wertelabels" öffnet man die Dialogbox „Wertelabels definieren:". Dort werden Wertelabels festgelegt. Wertelabels können bis zu 60 Zeichen lang sein, werden aber in den meisten Prozeduren verkürzt ausgegeben. Auch hier werden Groß- und

Kleinschreibung beachtet. Ein Wertelabel wird festgelegt, indem zunächst der Wert in das Eingabefeld „Wert:" eingegeben wird. Anschließend schreiben Sie die zugehörige Wert-Etikette in das Eingabefeld „Wertelabel:". Klicken Sie dann auf die Schaltfläche „Hinzufügen". Die Etikette ist fixiert. Wiederholen Sie diese Schritte für alle Werte, denen eine Etikette zugeordnet werden soll. Bestätigen Sie die Eingabe mit „Weiter" und „OK". Eine bereits eingegebene Etikette ändern Sie, indem sie zunächst das Label in der Liste markieren. Geben Sie das neue Label und/oder den neuen Wert ein, und klicken Sie auf die Schaltfläche „Ändern". Das veränderte Label erscheint. Sie löschen ein Wertelabel, indem Sie das Label in der Liste markieren und die Schaltfläche „Entfernen" anklicken.

Fehlende Werte (Missing-Werte). Die Deklaration von fehlenden Werten ermöglicht es, diese Werte bei den verschiedenen Prozeduren gezielt von der Berechnung auszuschließen. Alle nicht ausgefüllten Zellen in einer Datenmatrix werden automatisch als systemdefinierte fehlende Werte (*System-Missing-Werte*) behandelt. In der Matrix werden sie durch ein Komma gekennzeichnet (wenn dieses in der Windows-Systemsteuerung als Dezimaltrennzeichen deklariert wurde). Der Benutzer kann aber auch selbst fehlende Werte festlegen (*nutzerdefinierte Missing-Werte*). Dies geschieht, indem Sie die entsprechende Zelle der Spalte „Fehlende Werte" auf das unterlegte Quadrat auf der rechten Seite der Zelle klicken. Es öffnet sich die Dialogbox „Fehlende Werte definieren:". Hier können entweder bis zu drei einzelne Werte oder ein Wertebereich oder ein Wertebereich mit zusätzlich einem einzelnen Wert als fehlende Werte deklariert werden. (Für lange Stringvariablen und Datumsvariablen können keine fehlenden Werte deklariert werden, für Stringvariablen nur einzelne fehlende Werte, aber kein Wertebereich.)

Variablentypen. In SPSS können acht verschiedene Datentypen verwendet werden. Die Einstellung erfolgt in der Dialogbox „Variablentyp definieren" (⇨ Abb. 3.1), die sich öffnet, wenn man das unterlegte Kästchen in einer aktivierten Zelle der Spalte „Typ" anklickt. Es handelt sich überwiegend um Varianten von numerischen Variablen. Die Unterschiede bestehen in der verschiedenartigen Darstellung der Zahlen. Das Grundformat „Numerisch" akzeptiert ausschließlich Ziffern, Plus-, Minus- und Dezimalzeichen, in anderen Formaten kommen Tausendertrenn- und/oder Währungszeichen hinzu. Oder sie verwenden die wissenschaftliche Notation. Stringvariablen dagegen arbeiten mit Zeichenketten, Datumsvariablen sind speziell für Datumsformate vorgesehen. Mit Ausnahme von String- und Datumsvariablen gilt, dass eine Zeichenbreite von acht Zeichen und zwei Dezimalstellen voreingestellt ist. Die Voreinstellung kann mit der Befehlsfolge „Bearbeiten", „Optionen..." im Register „Daten" geändert werden. Die gewünschten Werte werden in die entsprechenden „Eingabefelder" eingetragen. Maximal sind 40 Zeichen und 16 Dezimalstellen zulässig. Bei der Zahl der Zeichen sind Plus-, Minus-, Dezimalzeichen, Tausendertrennzeichen und Währungszeichen mitzurechnen. Die Einstellung der Breite und Dezimalstellen betrifft bei numerischen, Punkt- und Kommaformaten lediglich die Anzeige der Daten, intern werden die Nachkommastellen bis zur maximal zulässigen Zahl von 16 Stellen gespeichert und weiter verarbeitet. Lediglich in der Anzeige erscheinen sie als gerundeter Wert. Das Dezimalzeichen ist beim numerischen Format und bei der wissenschaftlichen Notation

ein Komma, wenn im Windows-Betriebssystem Deutschland bei der Ländereinstellung gewählt wurde. Bei anderen Ländereinstellungen kann es ein Punkt sein. Alle anderen Formate (Ausnahme Sekundenbruchteile bei den Datumsformaten) werden von der Ländereinstellung nicht berührt.

Zulässige Variablentypen sind (⇨ Abb. 3.1):

① *Numerisch.* Gültig sind Ziffern, vorangestelltes Plus- oder Minuszeichen und ein Dezimaltrennzeichen. *Beispiele:* +1660,50; 1000; -250,123. Dieser Variablentyp ist voreingestellt und ist auch für die meisten Zwecke am geeignetsten.

② *Komma, Punkt.* Komma und Punkt sind komplementär zueinander. Sie werden durch die Ländereinstellung der Windows-Systemsteuerung nicht verändert. Zusätzlich zu den im Format „Numerisch" zugelassenen Zeichen wird ein Tausendertrennzeichen verwendet. Komma entspricht der amerikanischen, Punkt der deutschen Schreibweise. Im Format Komma muss das Dezimalzeichen ein Punkt (!) sein, das Tausendertrennzeichen ein Komma. Umgekehrt muss im Format Punkt das Dezimaltrennzeichen ein Komma und das Tausendertrennzeichen ein Punkt sein. Die Tausendertrennzeichen werden automatisch eingefügt, sofern sie bei der Eingabe nicht eingetippt werden. Bei der Angabe der Breite müssen Vor- und Trennzeichen mit berechnet werden. *Beispiel:* -1,203.24 (Kommaformat) entspricht -1.203,24 (Punktformat).

③ *Wissenschaftliche Notation.* Diese wird gewöhnlich verwendet, wenn sehr große oder sehr kleine Zahlen zu verarbeiten sind. Eine Zahl wird dann als Dezimalzahl, multipliziert mit einer Zehnerpotenz dargestellt. *Beispiel:* 244.000 wird zerlegt in 2,4 mal 10^5 angezeigt als 2,4E+05. Dagegen wird 0,0005 zerlegt in 5,0 mal 10^{-4}, angezeigt als 5,0E-04.

④ *Datum.* Hier wird eine Liste von Formaten für Datums- und/oder Zeitangaben angeboten (⇨ Abb. 3.1).

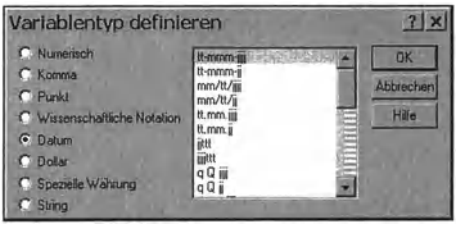


Abb. 3.1. Dialogbox „Variablentyp definieren"

Die Formate sind in allgemeiner Form in einer Auswahlbox angegeben. Dabei bedeutet:

t	Tag	h	Stunden
m	Monat	m	Minuten
j	Jahr	s	Sekunden (inklusive Bruchteile
Q	Quartal		von Sekunden)[1]

[1] Spezielle Formate geben auch die Wochen des Jahres an (ww). Andere sind für die Eingabe von Tages bzw. Monatsnamen gedacht.

Die Zahl der Buchstaben gibt an, mit wie vielen Stellen der jeweilige Teil angezeigt wird. Dreistellige Monatsangaben ergeben die Monatsabkürzung in Buchstaben (englische Abkürzungen !). Trennzeichen können Bindestrich, Punkt und Slash (/) sein. tt.mm.jjjj ist z.B. das Format der gängigen deutschen Datumsangabe. *Beispiel:* 12.12.1993. tt-mmm-jjjj ergäbe dagegen bei derselben Eingabe: 12. DEC. 1993. Drei Zeichen für den Tag ttt bedeuten, dass ganzjährige Tageszählung von 1 bis 365 benutzt wird. WK bedeutet, dass mit 53 Wochen des Jahres gearbeitet wird. Die 44. Woche 1993 wird entsprechend bei Format ww WK jjjj mit 44 WK 1993 eingeben. Für Quartale steht q Q; q Q jjjj erlaubt z.B. die Eingabe 1 Q 1993 für das erste Quartal 1993. Bei Formaten mit vierstelligen Jahreszahlen werden zweistellige Eingaben automatisch um 19 ergänzt. Bei Formaten mit wörtlicher Monatsbezeichnung werden Monatszahlen automatisch umgerechnet. Ebenso werden Eingaben in Buchstaben bei den Zahlenformaten automatisch in Zahlen umgewandelt. Außerdem können Monatsangaben voll ausgeschrieben oder abgekürzt eingegeben werden. Unabhängig von der Anzeige werden die Daten intern aber immer als Sekunden seit dem 14.10.1582 abgespeichert. Für die weitere Verarbeitung ist zu beachten, dass Quartals-, Monats- und Jahresdaten immer ab Mitternacht des ersten Tages des entsprechenden Zeitabschnitts interpretiert werden.

Einige Datumsvariablen sind für die Registrierung von Tageszeiten bzw. Tageszeiten zusätzlich zu Datumsangaben ausgelegt. Die Zeiten können unterschiedlich exakt, bis maximal auf eine Hundertstelsekunde genau, festgelegt werden.

Als Trennzeichen zwischen Stunden, Minuten und Sekunden wird der Doppelpunkt verwendet. hh:mm:ss,ss lässt die Eingabe von Zeiten auf die Hundertstelsekunde genau zu. *Beispiel:* 08:22:12,22. Die detaillierteste Information ergäbe eine Variable des Types tt-mmm-jjj hh:mm:ss,ss. *Beispiel:* 12-DEC-1993 18:33:12,23. Als Trennzeichen zwischen Stunden, Minuten und Sekunden kann auch ein Leerzeichen verwendet werden (der im Handbuch ebenfalls angegebene Punkt funktioniert bei der vorliegenden Version nicht). Angezeigt werden auf jeden Fall Doppelpunkte. Die Zeitangaben werden intern als Sekunden seit Beginn der jeweiligen Zeitperiode abgespeichert. (Näheres zum Beginn der Zeitperioden siehe Syntax Reference Guide.)

Bei den Datums- und Zeitformaten ist weiter zu beachten, dass einige Formate mehr Stellen zur Ausgabe als zur Eingabe benötigen. (Genaue Angaben enthält der Syntax Reference Guide.) Reichen die eingestellten Stellen zur vollständigen Ausgabe nicht aus, werden die Daten vollständig gespeichert, aber nur verkürzt angezeigt. Dabei wird gerundet.

⑤ *Dollar.* Entspricht der Option Komma mit einem ergänzend vorangestellten Dollarzeichen. Dollarzeichen und Tausendertrennzeichen werden, wenn nicht eingegeben, automatisch eingefügt. *Beispiel:* **$#,###.##** ergibt eine Dollarzahl mit neun Zeichen und zwei Dezimalstellen. Ein Dollarformat wird durch Anklicken im Auswahlfeld bestimmt. Die verschiedenen Dollarformate unterscheiden sich in erster Linie durch die Zahl der Stellen und die Zahl der Dezimalstellen (0 oder 2). Die Einstellungen von „Breite“ und „Dezimalstellen“ werden auto-

matisch in das entsprechende Anzeigefeld übernommen. Man kann sie aber dort auch unabhängig von den im Fenster angezeigten Formaten einstellen.

© *Spezielle Währung.* Hier können bis zu fünf selbst definierte Formate zur Verfügung. gestellt werden. Diese müssen allerdings zunächst an anderer Stelle, im Unter-Menü „Optionen" des Menüs „Bearbeiten", Register „Währung" definiert werden. Festgelegt werden damit das Dezimalzeichen, ein Prä- und Suffix und das Zeichen für einen Negativwert (als Prä- oder Suffix). Zur Definition wählen Sie im Menü „Bearbeiten" das Unter-Menü „Optionen..." und das Register „Währung". Die Registerkarte „Währung" (⇨ Abb. 3.2) erscheint. Hier definieren Sie die Formate. Diese werden unter den Bezeichnungen CCA, CCB, CCC, CCD und CCE abgelegt und stehen im folgenden für die Definition von Variablentypen zur Verfügung (⇨ Beispiel unten).

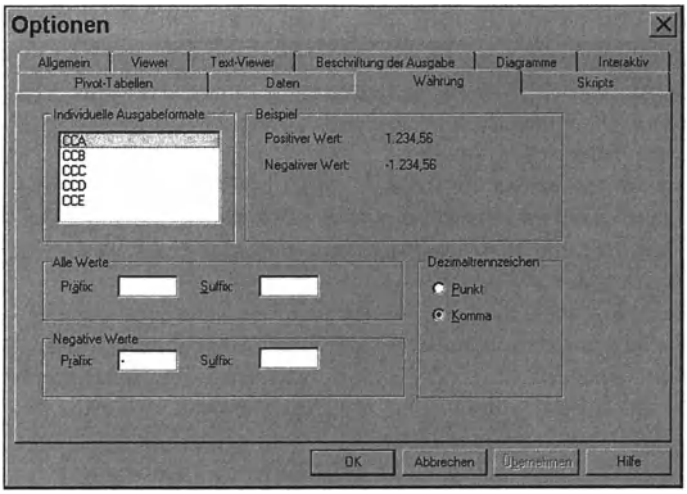


Abb. 3.2. Dialogbox „Optionen" mit Register „Währung"

In der Dialogbox „Variablentyp definieren" erscheint danach bei Auswahl der Option „Spezielle Währung" ein Auswahlfeld mit den Namen der Währungsformate (CCA usw.). Markiert man einen davon, wird im Feld „Beispiel" ein Beispiel für dieses Format angezeigt. Durch Markieren des gewünschten Namens und „Weiter" wird für die Variable dieses Format ausgewählt. Zusätzlich lässt sich die maximale Breite und die Zahl der Dezimalstellen einstellen.

Anmerkung. Steht das im Beispielfenster eingegebene Dezimalzeichen zur Ländereinstellung im Widerspruch, muss es gemäß der Ländereinstellung getippt werden getippt werden, erscheint aber auf dem Datenblatt gemäß der Währungsdefinition. Bei der Ausgabe der Ergebnisse von Statistikprozeduren wird überwiegend das Dezimalzeichen der länderpezifischen Einstellungen verwendet, bei speziellen Statistiken, wie Mittelwert im Menü Häufigkeiten aber auch schon einmal das im Währungsformat angegeben Dezimalzeichen.

⑦ *String* (Zeichenkette). Gültige Werte sind Buchstaben, Ziffern und Sonderzeichen. Die maximale Länge des Strings beträgt 255 Zeichen. Voreingestellt ist acht. Beträgt die maximale Zeichenlänge nicht mehr als acht Zeichen, handelt es sich um eine kurze String-Variable (wird in der Quellvariablenliste mit gekennzeichnet), ist sie größer als acht Zeichen, um eine lange (wird in der Quellvariablenliste mit gekennzeichnet). Stringvariablen werden rechts bis zur maximalen Länge mit Leerzeichen aufgefüllt. Bei der Interpretation des Strings kommt es auf die genaue Position des Zeichens an. *Beispiel:* 'Ja' ist nicht gleich 'Ja '. Lange Stringvariablen können bei den meisten Prozeduren gar nicht oder nur eingeschränkt gebraucht werden. Kurze Stringvariablen sind vielfältiger auswertbar. Jedoch sind auch sie in vielen Prozeduren nicht zu verwenden. Stringvariablen werden nur bei den Menüs in der Liste der Quellvariablen angezeigt, in denen sie verwendet werden können. Sie haben zwar den Vorteil, in den meisten Fällen direkt lesbar zu sein, man sollte sich aber in jedem Falle überlegen, ob man sie tatsächlich einsetzt. Ihre Werte können in der Regel ohne weiteres auch als numerische Werte kodiert werden. Dann kann man wesentlich mehr statistische Prozeduren mit ihnen ausführen. (Allerdings ist das Messniveau zu beachten.) Die Lesbarkeit kann durch Vergabe von „Wertelabels" erhalten bleiben. Man kann sich mit der Befehlsfolge „Ansicht", „Wertelabels" oder durch Anklicken von die Variablenwerte in Form der Werteetiketten im Dateneditorfenster anzeigen lassen. Die Wertelabels der jeweils in einer Liste markierten Variablen können bei Bedarf in einem durch Drücken der rechten Maustaste geöffneten Kontextmenü und Auswahl von „Info zu Variable" eingeblendet werden. In Datenbanken werden häufig Stringvariablen verwendet. Diese werden dann auch als solche importiert. In einem solchen Falle ist zu überlegen, ob die Stringvariable nicht in ein anderes Format umgewandelt werden soll.

Weiter erkennt SPSS:

- ❒ Implizites Dezimalformat
- ❒ Prozentformat
- ❒ Hexadezimales Format
- ❒ Spaltenbinäres Format.

Werden Daten mit einem solchen Format importiert, wird das Format der Liste verfügbarer Formate angehängt. Man kann dann entweder mit diesem Format weiter arbeiten oder die Daten in ein SPSS-Format umwandeln.
Mit der Definition des Variablentyps wird auch die Variablenbreite und bei numerischen Formaten die Zahl der Dezimalstellen festgelegt. Diese Definition wirkt sich automatisch auf die Breite der angezeigten Matrixspalte aus.

Spaltenbreite und -Ausrichtung. Die angezeigte Spaltenbreite der Datenmatrix kann geändert werden, wenn man in der „Variablenansicht" die entsprechende Zelle der Spalte „Spalten" aktiviert. Durch Anklicken der Pfeile, die am rechten Rand der Zelle erscheinen, vergrößert oder verkleinert man die Spaltenbreite. Dies berührt die Variablenbreite nicht. Ist der definierte Wert länger als die Spaltenbreite, wird er abgeschnitten angezeigt. Auch kann die Ausrichtung der Anzeige

auf linksbündig (Voreinstellung für Stringvariablen), rechtsbündig (Voreinstellung für alle anderen Variablen) oder zentriert gesetzt werden. Dies geschieht durch Auswahl aus einer Drop-Down-Liste, die sich beim Anklicken des Pfeiles öffnet, der bei Aktivieren einer Zelle in der Spalte „Ausrichtung" auf der rechten Seite der Zelle erscheint.

Ein Übungsbeispiel. In der Übungsdatei in Kap. 2 wurde bewusst nur ein Datentyp, nämlich „Numerisch" verwendet. Dies dürfte für die meisten Zwecke hinreichen und, zusammen mit der Deklaration von Variablen- und Werteetiketten, der häufigste Weg zur Definition einer Datenmatrix sein. Die Veranschaulichung der verschiedenen Variablentypen soll jetzt anhand einer anderen Datei erfolgen. Es handelt sich um den Auszug einer Datei, die sich bei einer Untersuchung über Überschuldung von Verbrauchern bei der Verbraucherzentrale Hamburg ergab. Die Datei (Dateiname VZ.SAV) soll die in Tabelle 3.1 dargestellten Variablen enthalten: In dieser Datei sind alle angebotenen Formate, mit Ausnahme der typisch amerikanischen Formate Komma und Dollar. Die Kreditbeträge sollen in Verbindung mit der Währungseinheit eingegeben und angezeigt werden. Es soll sich um DM-Beträge handeln. Definiert werden hier nur Variablenname und der Variablentyp. Zinsbeträge sollen mit % als Zusatz angezeigt werden. Dazu müssen zwei Formate unter dem Generalformat „Spezielle Währung" definiert werden.

Tabelle 3.1. Variablen des Datensatzes VZ.SAV

Variable	Variablennamen	Variablentyp	Breite/Dezimalstellen
Fallnummer	NR	Numerisch	8/0
Name des Schuldners	NAME	String	15
Datum: Erster Kontakt mit der Beratungsstelle	KONTAKT	Datum	-
Datum: Beginn der Überschuldung	BEG_UEB	Datum	-
Zeitraum zwischen Überschuldung und Kontakt mit der Beratungsstelle	ZEIT_BER	Wissenschaftliche Notation	11/2
Zinsen Kredit 2	ZINS2	Andere Währung	8/2
Monatseinkommen	EINK	Punkt	12/2
Summe Kredit 1	KREDIT1	Andere Währung	14/2
Zinsen Kredit 1	ZINS1	Andere Währung	8/2
Summe Kredit 2	KREDIT2	Andere Währung	14/2

Zur Definition dieser Datendatei gehen Sie wie folgt vor:

▷ Eröffnen Sie mit „Datei", „Neu ▷" und „Daten" ein neues Dateneditorfenster und wechseln gegebenenfalls durch Anklicken der Registerkarte in die „Variablenansicht".

Danach definieren Sie zuerst das gewünschte „DM"-Format.

▷ Öffnen Sie dazu mit „Bearbeiten", „Optionen..." und „Währung" das Registerblatt „Währung" (⇨ Abb. 3.2).
▷ Markieren Sie die erste Bezeichnung CCA.
▷ Tragen Sie in der Gruppe „Alle Werte" in das Feld „Suffix" DM ein.
▷ Tragen Sie in der Gruppe „Negative Werte" in das Feld „Präfix" ein Minuszeichen ein.
▷ Klicken Sie in der Gruppe „Dezimaltrennzeichen" auf den Optionsschalter „Komma".
▷ Wählen Sie „Übernehmen" (ändert das Format, ohne das Registerblatt zu verlassen).
▷ Definieren Sie auf gleiche Weise das „Prozent"-Format. Im Unterschied zum „DM"-Format markieren Sie als Bezeichnung CCB, tragen in der Gruppe „Alle Werte" in das Feld „Suffix" % ein und markieren den Optionsschalter „Komma".
▷ Bestätigen Sie am Schluss alle Definitionen mit „OK".

Die definierten Formate sind jetzt unter ihren Bezeichnungen bei der Auswahl des Variablentyps abrufbar.

Jetzt können die einzelnen Variablen definiert werden. Zur Definition der Variablen NR gehen Sie wie folgt vor:

▷ Aktivieren Sie in der Datenansicht in der ersten Zeile die Zelle der Spalte „Namen" und tragen Sie dort NR ein. ,
▷ Aktivieren Sie die Zelle der Spalte „Typ" und klicken Sie auf die Schaltfläche. Die Dialogbox „Variablentyp definieren" erscheint (⇨ Abb. 3.1).
▷ Wählen Sie den Optionsschalter „Numerisch".
▷ Ändern Sie die Werte des Feldes „Breite" auf 4 und des Feldes „Dezimalstellen" auf 0.
▷ Bestätigen Sie mit „OK".

Für die Definition der anderen Variablen verfahren Sie ebenso. Im folgenden wird lediglich der Eintrag in der Dialogbox „Variablentyp definieren" besprochen. Einträge in anderen Dialogboxen werden nur dann dargestellt, wenn diese zum ersten Mal auftreten.

Die Variable NAME soll eine lange Stringvariable mit 15 Zeichen Maximallänge sein:

▷ Wählen Sie in der Dialogbox „Variablentyp definieren" die Option „Datum".
▷ Markieren Sie im dann erscheinenden Auswahlfeld die Option tt.mm.jjjj, die dem in Deutschland üblichen Datumsformat entspricht.

Die weiteren Variablendefinitionen bis zur Variablen KREDIT1 sollten Sie selbst vornehmen können. KREDIT1 soll eine Währungsvariable mit dem zu Beginn definierten DM-Format sein.

- ▷ Wählen Sie in der Dialogbox „Variablentyp definieren" die Option „Spezielle Währung". Es öffnet sich ein Auswahlfeld.
- ▷ Markieren Sie dort die Bezeichnung „CCA" (unter der unser oben definiertes DM-Format gespeichert ist). In der Informationsgruppe „Beispiel" werden zwei Beispiele für die Darstellung in diesem Format angezeigt. Ändern Sie den Wert für die Breite im Feld „Breite" auf 14.
- ▷ Bestätigen Sie die Eingabe mit „OK".

Die weiteren Variablen sollten Sie jetzt selbst definieren können. In Abb. 3.3 sehen Sie eine Datenmatrix mit den Variablen des Übungsbeispiels und den Daten der vier ersten Fälle. Sie können zur Übung diese Daten eingeben. Die Werte der Variablen ZEIT_BER lassen Sie am besten zunächst offen und berechnen sie später in der Dialogbox „Variable berechnen" (⇨ Abb. 5.1) durch Bildung der Differenz zwischen KONTAKT und BEG_UEB. Testen Sie dabei auch die unten geschilderten Möglichkeiten zur Auswahl von Eingabebereichen und zum Editieren der Daten.

	nr	name	kontakt	beg_ueb	zeit_ber	eink	kredit1	zins1	kredit2	zins2
1	1	Frederi	17.10.89	01.10.1986	9,61E+07	1200,0	4.000,00DM	11,2%	2.500,00DM	10,3%
2	2	Birgid	08.01.89	01.11.1982	1,95E+08	1798,0	2.600,00DM	10,3%	2.000,00DM	11,5%
3	3	Ronald	01.02.88	01.01.1988	2,68E+06	2050,0	15.000,00DM	12,4%	9.700,00DM	12,9%
4	4	Gertru	08.06.89	01.11.1980	2,71E+08	2000,0	100.000,00DM	11,4%	163.000,00DM	10,6%

Abb. 3.3. Dateneditorfenster mit den vier ersten Fällen von VZ.SAV

3.2 Variablendefinitionen kopieren und übernehmen

3.2.1 Variablendefinitionen kopieren

Haben einige Variablen dasselbe oder ähnliche Formate, kann man sich die Definition durch Kopieren erleichtern. Das Verfahren wurde bereits in Kap. 2.3.3 erläutert. In unserem Beispiel sollen KREDIT2 und KREDIT3 gleich definiert werden. Man erstellt zunächst die Definition einer Variablen (hier Kredit 2). Dann markiert man in der Variablenansicht die Zeile mit den Definitionen dieser Variablen, indem man auf die Zeilennummer am linken Rand klickt. Man wählt die Befehlsfolge „Bearbeiten", „Kopieren". Darauf markiert man die Zeile für die Variablendefinition der neuen Variablen (hier KREDI3). Es können auch mehrere nebeneinaderliegende Variablen gleichzeitig markiert werden. Dann wählt man die Befehlsfolge „Bearbeiten" und „Einfügen". Die Definition ist übernommen, mit Ausnahme des Variablennamens. Dieser wird, falls nicht schon vorher eingetragen, von SPSS automatisch generiert. (Zum Kopieren und Einfügen kann auch das Kontextmenü, das sich beim Klicken auf die rechte Maustaste öffnet, verwendet werden.)

Unterscheidet sich die Definition der neuen Variablen in einigen Elementen von der der Ausgangsvariablen, kann man dies jetzt nachträglich anpassen. Oder aber

man kopiert von vorne herein nur die Definitionselemente, die übernommen werden sollen. In diesem Falle muss die jeweilige Zelle der Ausgangsvariable markiert und kopiert und in die entsprechende Zelle der Zielvariable(n) eingefügt werden.

3.2.2 Variablendefinition aus einer bestehenden Datei übernehmen

Möchten Sie eine neue Datendatei erstellen, die Variablen enthält, die schon in einer bestehenden Datei vorhanden sind, dann können Sie die Definition vereinfachen. Sie können die Definition aus der alten Datei übernehmen. Dazu müssen allerdings die Namen der Variablen in der neuen Datei identisch mit denen in der alten Datei sein. Wählen Sie dann „Datei", „Datenlexikon zuweisen...". Es öffnet sich die Dialogbox „Datenlexikon zuweisen". Stellen Sie dort Laufwerk und Verzeichnis ein, in dem sich die alte Datei befindet. Wählen Sie den zutreffenden Dateityp aus der Auswahlliste, und wählen Sie aus der Dateiliste den Namen der alten Datei aus oder tragen Sie diesen in das Eingabefeld ein. Bestätigen Sie mit „Öffnen". Die Variablen sind nach dem Lexikon der alten Datei definiert. (Die neue Datei kann daneben auch Variablen enthalten, die nicht in der alten Datei vorhanden sind und eigenständig definiert werden müssen.)

3.3 Eingeben von Daten

Eingabe und Korrektur. Die Daten werden wie in Kap. 2.3.1 geschildert in die Zellen der Datenmatrix (auf dem Blatt „Datenansicht") eingegeben. Dazu wird zunächst die Eingabezelle (aktive Zelle) markiert. Dies geschieht durch Anklicken der Zelle mit der Maus oder durch Bewegung des Cursors mit der Richtungstaste auf eine Zelle. Innerhalb eines durch Variablendefinition und eingefügte Fälle bezeichneten Bereichs werden Zeilennummer und Variablennamen der aktiven Zelle zusätzlich in der oberen linken Ecke der Zelleneditorzeile angezeigt, einer Zeile unterhalb der Menü- bzw. Symbolleiste. Darauf wird der Wert eingegeben. Er erscheint zunächst im Zelleneditor. Durch Drücken der <Enter>-Taste (oder Anwählen einer anderen Zelle) wird der Wert bestätigt und in die Zelle eingetragen. Bei Bestätigung mit der <Enter>-Taste rückt gleichzeitig der Cursor eine Zelle nach unten. Bestätigung mit der Taste <Tab> verschiebt den Cursor eine Zelle nach rechts (nur, wenn schon Variablen definiert sind, sonst eine Zeile nach unten). Eingabe und Verschiebung des Cursors kann auch mit den Pfeiltasten bewirkt werden. Bei einer Eingabe in einer neuen Zeile entsteht automatisch ein neuer Fall. Alle Zellen dieser Zeile werden zunächst automatisch als „System-Missing-Wert" behandelt, bis ein Wert eingegeben worden ist.

Ein bereits eingegebener Wert kann ersetzt oder geändert werden. Dazu wird die betreffende Zelle markiert. Der Wert erscheint dann im Zelleneditor. Geben Sie entweder den neuen Wert ein oder ändern Sie den vorhandenen Wert auf die übliche Weise. Mit Bestätigung des Werte auf eine der angegeben Weisen wird der neue bzw. der veränderte Wert in die Zelle eingetragen.

Eingabe in ausgewählten Bereichen. Sind Variablen bereits definiert, durchläuft bei Verwendung der <Tabulator>-Taste zur Bestätigung der Eingabe der Cursor die Zeilen von links nach rechts und springt nach Eingabe des Wertes für die letzte Variable automatisch auf den Beginn der nächsten Zeile.

Einschränken der Datenwerte. Der Editor bietet insofern eine gewisse Kontrolle bei der Dateneingabe, als er weitgehend nur Daten im Rahmen des festgelegten Formats akzeptiert. Werden nicht erlaubte Zeichen eingegeben, trägt der Editor diese nicht ein. Bei Stringvariablen kann die Zeichenlänge nicht überschritten werden. Wird bei der Eingabe numerischer Variablen bei ganzzahligen Werten die definierte Variablenbreite überschritten, so werden diese mit wissenschaftlicher Notation angezeigt. Zahlen mit Nahkommastellen werden gerundet angezeigt. Es werden aber immer bis zu 16 Kommastellen intern verarbeitet. Durch Veränderung der Variablenbreite kann eine exakte Anzeige erreicht werden. Weitere Einschränkungen des Datenbereichs (wie sie zur Begrenzung von Eingabefehlern bei Verwendung von Data-Entry oder Datenbankprogrammen vorgenommen werden können), sind nicht möglich.

3.4 Editieren der Datenmatrix

Die Datenmatrix kann editiert werden, indem man:

- die Datenwerte ändert,
- Datenwerte ausschneidet, kopiert und einfügt,
- Fälle hinzufügt oder löscht,
- Variablen hinzufügt oder löscht,
- die Reihenfolge der Variablen ändert,
- Variablendefinitionen ändert.

(Für einen großen Teil dieser Funktionen stehen auch „Kontextmenüs" zur Verfügung. Diese öffnen sich, wenn man nach Markieren des gewünschten Bereichs die rechte Maustaste drückt. Probieren Sie es aus.)

Die Änderung der Datenwerte wurde bereits erläutert. Ebenso die Änderung der Variablendefinition (⇨ Kap. 2.3.3). Sind schon Werte eingegeben und wird anschließend die Definition der Variablen geändert, können Probleme auftauchen, wenn die bereits eingegebenen Werte dem neuen Format nicht entsprechen. SPSS konvertiert soweit möglich die Daten in das neue Format. Ist das nicht möglich, werden sie durch System-Missing-Werte ersetzt. Führt die Konvertierung zum Verlust von Wertelabels oder nutzerdefinierter fehlender Werte, dann gibt SPSS eine Warnung aus und fragt nach, ob die Änderung abgebrochen oder fortgesetzt werden soll.

Einfügen und Löschen neuer Fälle und Variablen, Verschieben von Variablen. Jede Eingabe eines Wertes in eine neue Zeile erzeugt einen neuen Fall. Ein Fall kann zwischen bestehende Fälle eingefügt werden. Dazu markieren sie eine beliebige Zelle in der Zeile unterhalb des einzufügenden Falles und wählen „Daten", „Fall einfügen" oder klicken auf das Symbol . Alternativ können Sie „Fall

einfügen" aus einem Kontextmenü wählen, das erscheint, wenn Sie mit der rechten Taste auf die Fallnummer des Falles klicken, vor dem Sie den Wert einfügen möchten. Wählen Sie dort die Option „Fall einfügen".

Jedes Einfügen eines Wertes in eine neue Spalte erzeugt automatisch eine Variable mit einem voreingestellten Variablennamen und dem voreingestellten Format. Schließt sich die neue Variable nicht unmittelbar an die bisher als Variablen definierten Spalten an, werden auch alle dazwischen liegenden Spalten zu Variablen mit vordefiniertem Namen und Format. Vorläufig werden System-Missing-Werte eingesetzt. Zum Einfügen einer neuen Variablen markieren Sie eine beliebige Zelle in der Spalte rechts neben der einzufügenden Variablen und wählen „Daten", „Variable einfügen", oder klicken Sie auf das Symbol  . „Variable einfügen" können Sie auch aus einem Kontextmenü auswählen, das sich öffnet, wenn Sie mit der rechten Maustaste den Namen derjenigen Variablen anklicken, vor der die neue Variable eingefügt werden soll.

Eine Variable verschieben Sie durch Ausschneiden und Einfügen. Erzeugen Sie zunächst an der Einfügestelle eine neue Variable. Markieren Sie dann die zu verschiebende Variable, indem Sie den Variablennamen im Kopf der Spalte anklicken. Wählen Sie „Bearbeiten", „Ausschneiden". Markieren Sie die neu eingefügte Variable, indem Sie den Namen anklicken. Wählen Sie „Bearbeiten", „Einfügen". (Sie können auch die entsprechenden Kontextmenüs verwenden.)

Fälle löschen Sie, indem Sie zunächst den Fall markieren. Klicken Sie dazu auf die Fallnummer am linken Rand. Wählen Sie dann „Bearbeiten", „Löschen".

Analog löscht man eine Variable durch Markieren der entsprechenden Spalte und Auswahl von „Bearbeiten", „Löschen". Mit dem Löschen von Variablen werden die Quellvariablenlisten für die verschiedenen Prozeduren unmittelbar korrigiert. (Beides geht auch über entsprechende Kontextmenüs.) Aus einer vorher erzeugten Liste ausgewählter Variablen werden sie jedoch erst durch „Zurücksetzen" oder Markieren und Anklicken von  entfernt.

Ausschneiden, Kopieren und Einfügen von Werten. Sind bei der Dateneingabe Fehler passiert, sollen bestimmte Variablen dupliziert werden oder kommen dieselben Werte häufig vor, so kann die Eingabe der Werte durch die Möglichkeit, Werte auszuschneiden oder zu kopieren und gegebenenfalls wieder einzufügen, erleichtert werden. Man markiert dazu die Werte, die ausgeschnitten, kopiert oder verschoben werden sollen. Sollen sie lediglich ausgeschnitten oder verschoben werden, wählen Sie „Bearbeiten", „Ausschneiden". Die Daten verschwinden dann. Sollen sie verschoben werden, markiert man daraufhin die Einfügestelle und wählt „Bearbeiten", „Einfügen". Sollen Werte kopiert werden, markiert man die Zellen und wählt „Bearbeiten", „Kopieren". Setzen Sie dann den Cursor auf die Einfügestelle, und wählen Sie „Bearbeiten", „Einfügen". (Alle Funktionen können auch über Kontextmenüs ausgewählt werden.)

Beim Verschieben und Kopieren muss der Zielbereich nicht dieselbe Zahl an Zellen umfassen, wie der ausgeschnittene bzw. kopierte Bereich. Das kann man sich zunutze machen und auf einfache Weise Werte vervielfältigen. So wird der Wert einer ausgeschnittenen/kopierten Zelle in sämtliche Zellen des markierten Zielbereiches eingefügt. Ebenso können die Werte mehrerer nebeneinander liegender Zellen einer Zeile in mehrere Zellen hinein kopiert werden. Dasselbe gilt um-

gekehrt für Spalten. Wird dagegen ein ganzer Bereich (mehrere Zeilen und Spalten) kopiert/verschoben, werden die Daten abgeschnitten, wenn der markierte Zielbereich in einer Richtung oder beiden Richtungen kleiner ist, und es werden System-Missing-Werte eingesetzt, wenn er in einer oder beiden Richtungen größer ist als der ausgeschnittene/kopierte Bereich. Da die Daten in der Zwischenablage (Clipboard) verbleiben bis ein neuer Ausschneide-/Kopiervorgang erfolgt, kann das Einfügen auch an unterschiedlichen Stellen wiederholt werden. Wird dabei der bereits definierte Datenbereich überschritten, fügt SPSS automatisch neue Werte und/oder neue Variablen ein und füllt die noch nicht bearbeiteten Zellen mit System-Missing-Werten.

Schließlich können die Daten über das Clipboard auch in das Syntax-, das Ausgabefenster (dort allerdings nur in eine Textzeile) oder in andere Anwendungsprogramme übertragen werden.

Finden von Variablen, Fällen und Datenwerten. Zum Editieren kann es nötig sein, gezielt auf bestimmte Fälle, Variablen und/oder Datenwerte zu zugreifen. So kann es etwa sein, dass für einen bestimmten Fall ein noch fehlender Wert nachzutragen oder ein Wert zu ändern ist. Häufig wird es auch vorkommen, dass man in einer Auszählung einen nicht gültigen Wert für eine Variable entdeckt hat. Dann wird man in der Matrix diese Variable suchen (was z.B. bei großen Datenmatrizen oder, wenn die Sortierreihenfolge unübersichtlich ist, schwer sein kann).

Um einen speziellen Fall nach der automatisch vergebenen Fallnummer (Zeilennummer) zu finden, wählen Sie „Daten", „Gehe zu Fall...", oder klicken Sie auf  und geben Sie in der sich dann öffnenden Dialogbox „Gehe zu Fall" die Fallnummer ein. Bestätigen Sie mit „OK". Der Cursor springt auf die Zeile mit der gewählten Fallnummer. (Beachten Sie: Diese Nummer ist nicht unbedingt identisch mit der vom Forscher selbst vergebenen Fallnummer. Wird diese zum Suchen benutzt, verfahren Sie wie bei der Suche von Datenwerten in Variablen.)

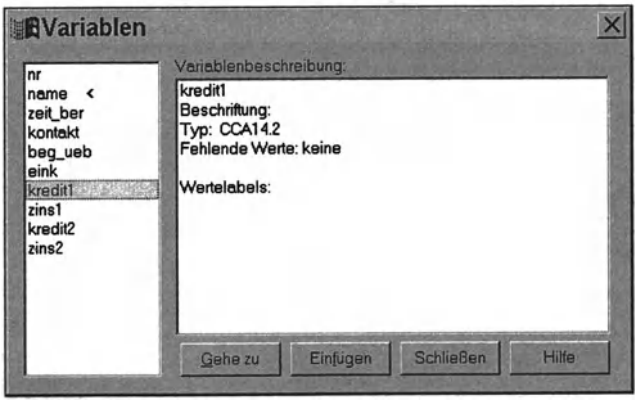


Abb. 3.6. Dialogbox „Variablen"

Variablen können Sie auf folgende Weise anspringen: Wählen Sie „Extras", „Variablen...", oder klicken Sie auf  . Es öffnet sich die Dialogbox „Variablen" (⇨ Abb. 3.6). Markieren Sie dort in der Quellvariablenliste die gewünschte Variable, und klicken Sie auf die Schaltfläche „Gehe zu".

Die Dialogbox schließt sich, und der Cursor befindet sich in der Spalte der gewählten Variablen in der Datenansicht des Daten-Editors.

Einen Datenwert für eine Variable können Sie ausgehend von einer beliebigen Zelle in der Spalte dieser Variablen suchen. Markieren Sie eine Zelle. Wählen Sie „Bearbeiten", „Suchen", oder klicken Sie auf  . Es öffnet sich die Dialogbox „Daten in Variablen suchen". Tragen sie dort in das Eingabefeld „Suchen nach:" den gesuchten Wert ein, und klicken Sie dann auf die Schaltfläche „Weitersuchen". Der Cursor springt auf die erste Zelle, die diesen Wert enthält. Kommt der Wert mehrmals vor, muss die Suche wiederholt werden. Bei Stringvariablen kann weiter festgelegt werden, ob Groß- und Kleinschreibung bei der Suche berücksichtigt werden soll (Voreinstellung: nicht).

Auswirkung offener Transformationen. Um Rechenzeit zu sparen, kann im Menü „Bearbeiten", „Optionen", Registerblatt „Daten" festgelegt werden, dass bestimmte Datentransformationen (Umkodieren, Berechnen) und Dateitransformationen (neue Variablen, neue Fälle) erst dann durchgeführt werden, wenn ein Befehl einen Datendurchlauf erfordert (Option „Werte vor Verwendung berechnen"). Bis dahin handelt es sich um sogenannte offene Transformationen. So lange solche Transformationen geöffnet sind, können Variablen weder eingefügt, noch gelöscht, noch neu geordnet werden. Ebenso kann weder ein Variablenname noch der Variablentyp geändert werden. Werden Werte geändert, können sie bei der späteren Transformation überschrieben werden. In einem solchen Falle erscheint eine Sicherheitsabfrage, mit der entschieden werden kann, ob die offenen Transformationen durchgeführt werden sollen oder nicht.

3.5 Einstellungen für den Dateneditor

In den Menüs „Ansicht" kann man einige Einstellungen des Dateneditors ändern. So kann man:

- in den Zellen die Wertelabels (nur in der Datenansicht) anstelle der Werte anzeigen lassen,
- die Gitterlinien in der Anzeige und/oder für den Druck ausschalten,
- die Schriftart der Anzeige und/oder des Drucks ändern.

Anzeigen von Wertelabels. Man kann z.B. für eine Variable „Geschlecht" den Variablentyp „String" definieren und die selbsterklärenden Werte „männlich" und „weiblich" vergeben. Es spricht aber vieles dafür, statt dessen lieber eine numerische Variable mit den Werten 1 und 2 zu verwenden. Um dennoch lesbare Outputs zu erhalten, ordnet man dann den Werten die Wertelabels 1 = männlich, 2 = weiblich zu. Vielfach erleichtert es die Eingabe und die Kontrolle, wenn auch in der Tabelle des Editors anstelle der Werte die Etiketten angezeigt werden.

Lassen Sie sich zur Übung einmal die Wertelabels der Datei ALLBUS.SAV. anzeigen. Laden Sie zunächst die Datendatei und gehen Sie dann wie folgt vor:

- ▷ Wählen Sie das Menü „Ansicht".
- ▷ Klicken Sie auf die Option „Wertelabels". Diese wird jetzt mit einem Häkchen gekennzeichnet und das Menü verschwindet. Durch Anklicken des Symbols kann ebenfalls zwischen diesen beiden Anzeigearten umgeschaltet werden.

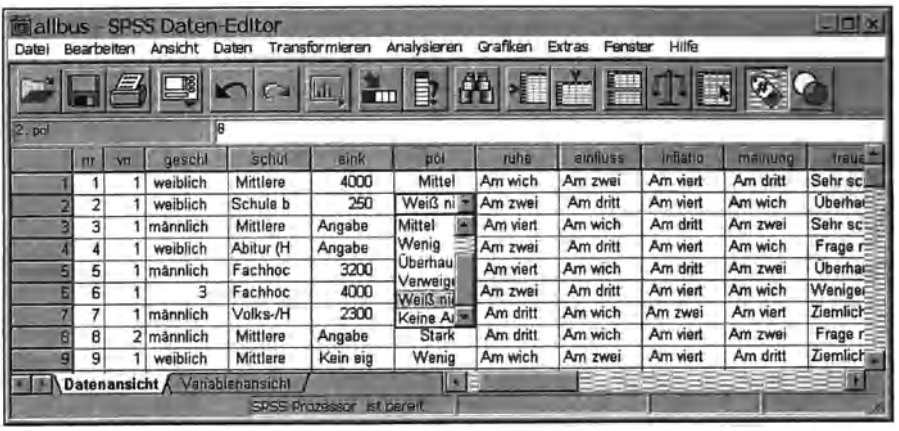


Abb. 3.7. Datenmatrix mit Anzeige der Wertelabels und des Auswahlfensters

Die Datendatei zeigt jetzt die Wertelabels an (Abb. 3.7). Wie man sieht, allerdings nur mit der Zahl der Stellen, die der Spaltendefinition entspricht. Sollen längere Werteetiketten vollständig angezeigt werden, muss man die Spaltenbreite anpassen. Gibt man nun die Werte (nicht die Labels !) in der üblichen Weise ein, so werden diese in der Anzeige sofort als Labels angezeigt. Zusätzlich kann man sich bei dieser Anzeigeart alle Wertelabels einer ausgewählten Variablen in einer Drop-Down-Liste anzeigen lassen. Man hat dadurch eine Art Kodeplan Online verfügbar. Dazu setzt man in der Zeile eines bereits existierenden Falles den Cursor auf die Zelle der interessierenden Variablen. Es erscheint am rechten Rand der Zelle ein Pfeil. Beim Anklicken des Pfeils (alternativ: <Shift> + <F2>) öffnet sich eine Drop-Downs Liste mit den Wertelabels.

Die Wertelabels werden dann (anders als in früheren Versionen) leider nur bis zur durch die Spaltenbreite vorgegebenen Stelle angezeigt. Maximal sind sechs Labels gleichzeitig im Fenster zu sehen. In der üblichen Weise kann man in dem Fenster scrollen und so die weiteren Wertelabels sichtbar machen. Soll ein Wert aus dieser Liste in die Zelle übertragen werden:

- ▷ Klicken Sie auf das ausgewählte Label.

Ohne Übernahme eines Wertes verlassen Sie das Auswahlfenster durch Anklicken irgendeines Feldes in der Tabelle.

Gitterlinien ausschalten. Die Gitterlinie der Editortabelle schalten Sie durch Anklicken der Option „Gitter" im Menü „Ansicht" aus. Das Häkchen neben der Option verschwindet.

Schriftarten ändern. Schriftarten für Anzeige auf dem Bildschirm und für den Druck können Sie mit der Befehlsfolge „Ansicht", „Schriftarten..." in der Dialogbox „Schriftart" ändern (alternativ über das Kontextmenü „Schriftart für Gitter"). Einstellen lässt sich „Schriftart", „Auszeichnung" (Schriftschnitt), „Größe" und gegebenenfalls unter „Skript" ein spezielles Sprachskript (eine auf eine Landessprache abgestellte Variante) für die gewählte Schriftart.

Die Optionen „Werte-Labels anzeigen" und „Gitter" sind Ein-Ausschalter. Durch erneutes Anklicken wird die Einstellung jeweils wieder umgeschaltet.

3.6 Drucken, Speichern, Öffnen, Schließen einer Datendatei

Drucken. Den Inhalt des Dateneditors können Sie ausdrucken. Das ist möglich, wenn der Dateneditor das aktive Fenster ist (⇨ Kap. 28.1).

Speichern. Eine Datendatei kann gespeichert werden, wenn das Dateneditorfenster aktiv ist. Soll die Datei unter dem alten Namen gespeichert werden, wählen Sie:

▷ „Datei", „Speichern", oder klicken Sie auf  .

Die Datei wird dann unter ihrem alten Namen gespeichert (für eine neu geöffnete Datei – der voreingestellte Name ist „Unbenannt" – wird automatisch die Dialogbox „Daten speichern unter" geöffnet, in der zuerst ein Name zu vergeben ist).

Soll die Datei unter einem neuen Namen oder einem neuen Format gespeichert werden, wählen Sie:

▷ „Datei", „Speichern unter...". Die Dialogbox „Daten speichern unter" öffnet sich (⇨ Abb. 2.7).
▷ Setzen Sie in das Eingabefeld „Dateiname" den gewünschten Dateinamen ein, und wählen Sie gegebenenfalls im Auswahlfeld „Speichern" das gewünschte Verzeichnis aus. Bestätigen Sie mit „Speichern".

Es ist jetzt auch möglich, beim Speichern nur einen Teil der Variablen auszuwählen. Möchten Sie dies, öffnen Sie vor dem Abspeichern durch Anklicken der Schaltfläche „Variablen" in der Dialogbox „Datenspeichern unter" die Unterdialogbox „Daten speichern als: Variablen". Dort finden Sie eine Auswahlliste aller Variablen. Ganz links sind in der Spalte „Beibehalten" alle zum Speichern ausgewählten Variablen durch ein Kreuz gekennzeichnet. Wenn man dieses Kreuz durch Anklicken löscht, wird die entsprechende Variable nicht gespeichert. Durch erneutes Anklicken kann man das Auswahlkreuz wieder erstellen. Je nachdem, wie viele Variablen man zum speichern auswählt, kann es günstiger sein, zuerst alle als ausgewählt zu markieren und die auszuschließenden Variablen anzuklicken oder umgekehrt erst alle auszuschließen und die ausgewählten anzuklicken. Durch Anklicken der Schaltfläche „Alle verwerfen" schließt man zunächst alle aus, umgekehrt schließt man durch Anklicken der Schaltfläche „Alle beibehalten" zunächst

alle ein. Außerdem kann man die Reihenfolge der Variablen in der Liste ändern. Klickt man auf die Bezeichnung der Spalte „Name", werden sie alphabetisch nach dem Variablennamen sortiert, klickt man auf die Spaltenüberschrift „Beschriftung" alphabetisch nach dem Variablenlabel, klickt man schließlich auf „Reihenfolge", werden die Variablen in umgekehrter Reihenfolge sortiert. Dies ist allerdings nur eine Hilfe für die Selektion der Variablen, auf die gespeicherte Datenmatrix selbst wirkt sich dies nicht aus.

Abb. 3.7. Dialogbox zur Auswahl von Variablen beim Speichern

Soll ein anderes als das SPSS-Windows Dateiformat zum Abspeichern benutzt werden, öffnen Sie durch Anklicken des Pfeiles neben dem Eingabefeld „Dateityp" eine Auswahlliste. Wählen Sie eines der angebotenen Formate durch Anklicken des Namens aus. Wird in das Format eines Tabellenkalkulationsprogramms übertragen, kann festgelegt werden, ob die Variablennamen mit übernommen werden sollen. Wird dieses gewünscht, markieren Sie das Kontrollkästchen „Variablennamen im Arbeitsblatt speichern".

Öffnen und Schließen von Dateien. SPSS kann immer nur eine geöffnete Datendatei bearbeiten. Deshalb wird auch beim Öffnen einer anderen Datei mit der Option „Öffnen" oder einer neuen Datei mit der Option „Neu" die bisher geöffnete Datei geschlossen. Wählt man dagegen „Beenden", wird zugleich SPSS verlassen. Wurden in der Datei Änderungen vorgenommen, erscheint immer die Sicherheitsabfrage danach, ob die Änderungen gespeichert werden sollen oder nicht. (Die Abfrage betrifft alle geöffneten und veränderten Fenster, also neben Dateneditorfenster auch Syntax- und Ausgabefenster.) Man kann dies getrennt für die verschiedenen Dateien bestätigen oder das Programm ohne Speichern verlassen. Geöffnet wird eine neue Datei mit der Option „Neu", „Daten" (sie erhält automatisch die Bezeichnung „Unbenannt"). Eine bestehende Datei öffnet man mit „Datei", „Öffnen". Es öffnet sich dann eine Dialogbox, in der man Laufwerk, Verzeichnis und die gewünschte Datei durch Anklicken in Auswahllisten auswählt. Man kann aber auch den Dateinamen (gegebenenfalls inklusive Pfad) direkt in das Feld „Dateinamen:" eintragen. Bestätigen Sie mit „Öffnen". Außerdem werden die zuletzt verwendeten Dateien im Menü „Datei" in der vorletzten Gruppe angezeigt (Option „Zuletzt verwendete Daten"). Sie können diese durch Anklicken direkt öffnen.

4 Arbeiten im Ausgabe- und Syntaxfenster

Einige SPSS-Fenster sind Textfenster, so das „Syntaxfenster", der „Skript-Editor" und der „Text-Viever". Darin enthaltene Texte können mit einigen Editierfunktionen bearbeitet und als Textdateien gespeichert werden. Die dort erzeugten Texte kann man in Textverarbeitungsprogramme übernehmen. Umgekehrt können auch die in SPSS selbst oder in einem anderen Programm geschriebenen Textdateien im ASCII-Format eingelesen werden. Einige Editierungsfunktionen – wie Kopieren, Ausschneiden, Einfügen, Suchen und Ersetzen – stehen in allen drei Fenstern zur Verfügung. Im Text-Viewer kann weiter die Schrift formatiert und der Text bearbeitet werden. Außerdem erleichtern Symbole das Blättern in der Ausgabe. Der Skript-Editor bietet umfangreiche Hilfen für die Überprüfung des Skripts.

Das eigentliche Ausgabefenster, der „Viewer" ist dagegen grafisch orientiert. Hier werden (wenn die Ausgabe nicht durch Änderung der Voreinstellung oder die Befehle „Datei", „Neu" und „Textausgabe" ein den Textviewer geleitet wird) automatisch alle statistischen Ergebnisse und einige Meldungen der SPSS-Sitzung angezeigt. Sie können dort bearbeitet und gespeichert werden. Das Arbeiten in diesem Fenster und im Syntaxfenster wird in diesem Kapitel dargestellt. Es ist davon auszugehen, dass der Textviewer selten benutzt wird. Wir beschränken uns darauf zu zeigen, wie er aufgerufen werden kann. Editieren eines Skripts ist nicht Gegenstand dieses Buches.

4.1 Arbeiten mit dem Viewer

Alle Ergebnisse statistischer Prozeduren, Diagramme und einige Meldungen der SPSS-Sitzung werden im „SPSS-Viewer" angezeigt (wir bezeichnen ihn auch als Ausgabefenster). Dieser besteht aus zwei Ausschnitten. Der linke Ausschnitt wird als *Gliederungsansicht* bezeichnet. Diese enthält eine Gliederung der im anderen Ausschnitt, dem *Inhaltsfenster*, enthaltenen Ausgaben. Die Gliederungsansicht dient dazu, schnell innerhalb der Ausgabe zu navigieren, Teile der Ausgabe ein- und auszublenden oder zu verschieben. Alles dies ist, umständlicher, auch im Inhaltsfenster möglich. Darüber hinaus kann man dort Tabellen pivotieren sowie Tabellen und Texte weiter bearbeiten. Zur Bearbeitung der Diagramme dient dagegen der Diagramm-Editor (⇨ Kap. 27.1). Die Ausgabe kann als Datei gespeichert und später wieder geladen sowie in andere Programme übertragen werden. Umgekehrt können aus anderen Programmen Texte und Objekte übernommen werden. Weil es sich beim SPSS-Viewer um ein grafisch orientiertes Fenster handelt, erfolgt der Austausch mit anderen Programmen in der Regel in

Form von Objekten. Für spezielle Zwecke ist auch ein Austausch bestimmter Inhalte im anderen Formaten möglich.

4.1.1 Öffnen von Dateien in einem oder mehreren Ausgabefenstern

Öffnen und Blättern. Mit der ersten Ausgabe einer SPSS-Sitzung wird (falls nicht durch Optionen anders festgelegt) automatisch ein Ausgabefenster mit dem Namen „Ausgabe1" geöffnet. In dieses werden die statistischen Ergebnisse geleitet, solange nicht weitere Fenster geöffnet und zum Hauptfenster bestimmt werden. Weitere Ausgabefenster können Sie öffnen mit: „Datei", „Neu" und „Ausgabe". Die weiteren Fenster heißen dann „Ausgabe2" usw.. In das jeweils gewünschte Fenster schaltet man mit „Fenster" und durch Anklicken des Namens des interessierenden Fensters in der sich öffnenden Liste oder durch Anklicken des Registerkarte dieses Fenster in der Task-Leiste. Die Ergebnisse werden jeweils in das „Hauptfenster" (dezidierte Fenster) geleitet. Das ist, so lange nicht anders festgelegt, immer das zuletzt geöffnete Fenster. Man ändert das Hauptausgabefenster, indem man in das gewünschte Fenster schaltet und in der Symbolleiste das Zeichen  anklickt. Dass ein Fenster als Hauptfenster gewählt wurde, erkennt man daran, dass dort das Ausrufezeichen nicht fett oder farbig dargestellt ist.

Weiter ist es möglich, bereits existierende Ausgabedateien in das Ausgabefenster zu laden. Das ist auf verschiedene Weise möglich.

Dazu gehen Sie wie folgt vor:

Wählen Sie die Befehlsfolge „Datei", „Öffnen" und „Ausgabe...". Wählen Sie dann in der sich öffnenden Dialogbox auf die übliche Weise Laufwerk, Verzeichnis und Datei aus. Sie laden diese durch Anklicken von „Öffnen". Die Datei erscheint dann auf jeden Fall in einem neuen Ausgabefenster.

Sollten Sie diese Datei erst vor kurzem verwendet haben, befindet sich deren Namen u.U. noch in der Liste der zuletzt verwendeten Dateien, die sie als Option im Menü „Datei" finden. Dann können Sie die Datei auch durch Klick auf ihren Namen in dieser Liste öffnen.

Symbolleiste. Die Symbolleiste des Ausgabefensters enthält einige zusätzliche Schaltflächen.

 *Seitenansicht.* Zeigt in einem Fenster die Ausgabe in der Ansicht von Druckseiten. In diesem Fenster kann man die Ansicht vergrößern und verkleinern, zwei Seiten nebeneinander betrachten sowie seitenweise blättern und drucken. Außerdem kann man in einem Dialogfenster „Seite einrichten", d.h. Größe, Format und Seitenränder bestimmen.

 *Exportieren.* Öffnet ein Dialogfenster, mit dem der Export einer Ausgabedatei gesteuert werden kann. Es ist möglich, Tabellen und Diagramme zusammen oder einzeln in verschiedenen Formaten in Dateien zu exportieren. Dabei können entweder alle Objekten, alle sichtbaren Objekten und nur ausgewählte Objekte exportiert werden.

 *Letzte Ausgabe auswählen.* Springt von einer beliebigen Stelle des Ausgabefensters aus den Beginn der zuletzt erstellten Ausgabe an.

Speichern. Sie können den Inhalt des Fensters speichern. Dazu muss das Ausgabefenster aktiv sein. Wählen Sie dazu „Datei" und „Speichern unter", oder klicken Sie auf . Es öffnet sich die Dialogbox „Speichern unter". Wählen Sie auf die übliche Weise das gewünschte Verzeichnis aus, und tragen Sie den Dateinamen im Feld „Namen" ein (gegebenenfalls können Sie eine existierende Datei aus der Liste auswählen). Bestätigen Sie mit „Speichern". Der Inhalt des Ausgabefensters wird als „Viewer-Datei" (Extension „spo") gespeichert.

4.1.2 Arbeiten mit der Gliederungsansicht

Der linke Ausschnitt des Viewers wird als *Gliederungsansicht* bezeichnet. Diese bietet eine knappe Inhaltsangabe der im rechten Ausschnitt, dem *Inhaltsfenster*, enthaltenen Ausgabe. Die Gliederungsansicht dient der schnellen Orientierung in der Ausgabe. Man kann in ihr Ausgabestellen anwählen, die Ausgabe in verschiedene Ebenen gliedern, Ausgabeteile umstellen, sie aus- bzw. einblenden, löschen oder Textfelder einfügen. Einige dieser Aktivitäten sind auch im Inhaltsfenster möglich, aber schwieriger zu bewerkstelligen. Die Aktionen werden zudem durch die spezielle Symbolleiste „Viewer-Gliederung" unterstützt, bzw. können auch über die Menüs „Bearbeiten", "Ansicht" und „Einfügen" bzw. mit dem lokalen Menü ausgeführt werden. (Der Weg über die Menüs wird hier nicht besprochen.)

Das Arbeiten mit der Gliederungsübersicht üben Sie am besten anhand einer umfangreichen Ausgabe. Erstellen Sie z.B. eine Grundauszählung für sämtliche Variablen von ALLBUS.SAV. Einen Teil des Ergebnisses sehen Sie in Abb. 4.1.

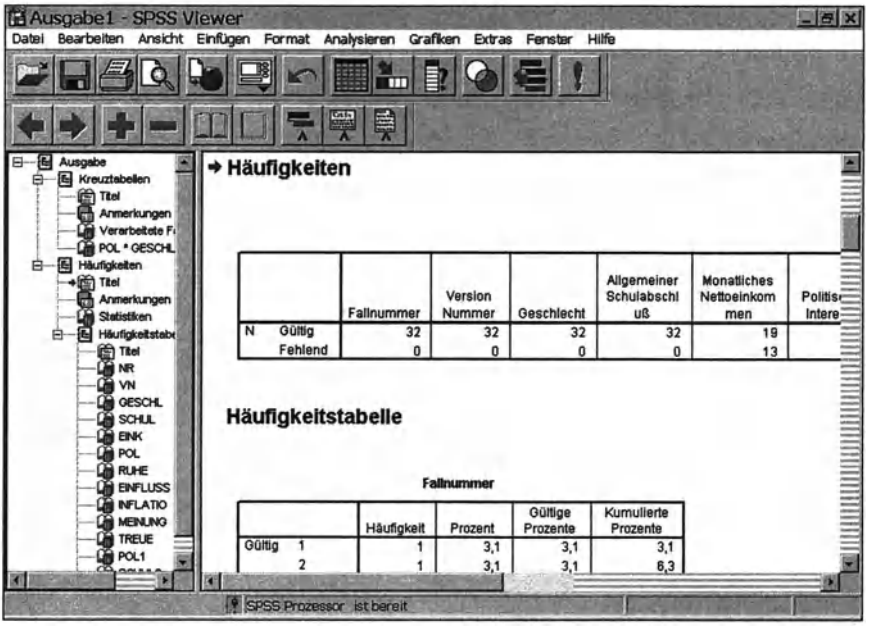


Abb. 4.1. Viewer-Fenster

Das linke Fenster enthält die Gliederungsansicht. Dessen Breite können Sie ändern, indem Sie mit dem Cursor auf dessen rechten Rahmen zeigen, bis sich die Form des Cursors zu einem Doppelpfeil ändert. Ziehen Sie dann den Cursor mit gedrückter linker Maustaste bis an die gewünschte Stelle. Mit Hilfe der Bildlaufleiste bewegen Sie sich im Gliederungsfenster. Wenn Sie auf ein Element in der Gliederungsansicht klicken, sehen Sie im Inhaltsfenster die dazugehörige Tabelle bzw. das entsprechende Diagramm. Sie können ein Objekt ausblenden, ohne es zu löschen, indem Sie auf das Buchsymbol vor dem Namen dieses Objektes doppelklicken. Aus dem offenen wird gleichzeitig ein geschlossenes Buch. Man kann auch die Ergebnisse ganzer Prozeduren ausblenden. Dafür muss man auf das Symbol für diese Prozedur (eine Gliederungsebene höher) doppelklicken. Umgekehrt kann durch Klicken auf das entsprechende Symbol auch das Objekt wieder eingeblendet werden.

Verschiebung der Position eines Objektes (einer Prozedur) ist ebenfalls möglich. Klicken Sie dazu auf das Symbol dieses Objektes (der Prozedur) und ziehen Sie den Cursor bis zur gewünschten Einfügestelle.

Die Symbolleiste „Viewer-Gliederung" unterstützt ebenfalls das Ein- und Ausblenden von Objekten der Ausgabe. Daneben kann man die verschiedenen Objekte der Ausgabe in der Gliederung um Gliederungsstufen herab- und hinaufstufen. Daneben kann man Fenster zum Eingeben zusätzlicher Texte und Überschriften öffnen.

Heraufstufen/Herabstufen. In einer hierarchischen Struktur des Outputnavigators wird ein markierter Gliederungspunkt hinauf- bzw. herabgestuft.

Erweitern/Reduzieren. Ermöglicht es, einzelne Gliederungspunkte des Outputs auszublenden oder einzublenden.

Einblenden/Ausblenden. Ermöglicht es, einzelne Objekte des Outputs ein- oder auszublenden.

Überschrift einfügen/Titel einfügen/Text einfügen. Öffnen Textfelder, in die Überschriften, Titel oder Texte zur Ergänzung der Ausgabe eingetragen werden können.

4.1.3 Aufrufen von Informationen und Formatieren von Pivot-Tabellen

Im Ausgabefenster finden Sie die Tabellen, Diagramme, aber auch Überschriften, Erläuterungen usw.. Bei den Tabellen handelt es sich um sogenannte Pivot-Tabellen, die sich in besonderer Weise bearbeiten lassen.

Erläuterungen zu Pivot-Tabellen. Zu den Tabellen können Sie sich weitere Erläuterungen geben lassen. Zunächst können Sie Erläuterungen zu einigen Begriffen der Tabelle abrufen. Dazu wählen Sie die Tabelle durch Doppelklicken aus. Sie erscheint dann in einem gerasterten Rahmen. (Einfaches Anklicken wählt die Tabelle ebenfalls aus. Sie wird dann durch einfachen Rahmen gekennzeichnet. Dies ist z.B. für das Kopieren oder Löschen der ganzen Tabelle Voraussetzung.) Setzen Sie den Cursor auf das Element, zu dem Sie eine Erläuterung wünschen, drücken

Sie die rechte Maustaste, und wählen Sie im sich öffnenden lokalen Menü (falls aktiv) die Option „Direkthilfe“ (⇨ Abb. 4.2). Es öffnet sich ein Pop-Up-Fenster mit einer Erläuterung zu diesem Element. Eine zweite Möglichkeit besteht darin, im Menü „Hilfe“ die Option „Ergebnis-Assistent“ aufzurufen. Dadurch gelangen Sie in eine kurze Hilfesequenz – ähnlich dem Lernprogramm –, in dem die wichtigsten Elemente der Haupttabellen der entsprechenden Prozedur erläutert werden.

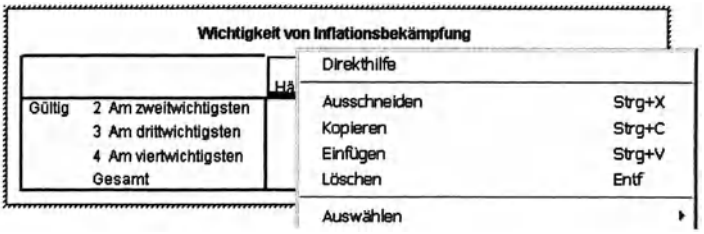
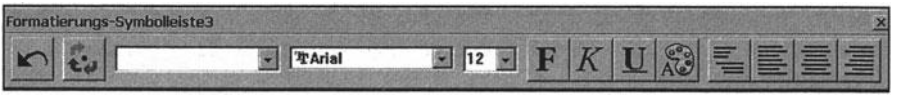


Abb. 4.2. Ausschnitt aus dem Lokales Menü zu einer Pivot-Tabelle im Viewer

Ausblenden von Zeilen und Spalten. Sie können, ohne sie zu löschen, einzelne Zeilen und/oder Spalten aus der Tabelle ausblenden. Dazu Doppelklicken Sie zunächst auf die Tabelle, um sie zu aktivieren. Drücken Sie <Strg>+<Alt>, und klicken Sie dann auf die Beschriftung der Zeile oder Spalte. Die ganz Zeile oder Spalte ist dann markiert. Drücken Sie auf die rechte Maustaste, und wählen Sie aus dem sich öffnenden Kontextmenü die Option „Kategorie ausblenden“. Sie können die Zeile oder Spalte wieder anzeigen lassen, indem Sie im Menü „Ansicht“ die Option „Alles einblenden“ wählen.

Formatieren der Tabellen. Um das Schriftformat zu ändern, klickt man in der schon ausgewählten Tabelle auf das Element, dessen Schriftformat verändert werden soll. Man kann dann auf eine „Formatierungs-Symbolleiste“ zur Änderung von Schriftart, Größe, Auszeichnung, Farbe und Absatzausrichtung zurückgreifen. Um die Symbolleiste zu öffnen, drückt man die rechte Maustaste und wählt in dem sich öffnenden lokalen Menü die Option „Symbolleiste“.



Weitere Formatierungsmöglichkeiten sind über die Menüs „Format“, „Schriftart“ verfügbar. Insbesondere sei aber auf das Menü „Format“, „Zelleneigenschaften“ hingewiesen. Dort können u.a. Formate für das Anzeigen der Werte in den Zellen gewählt werden. Dies wird sicher häufig gebraucht, u.a. um die Zahl der angezeigten Nachkommastellen zu bestimmen. Auch die Ausrichtung innerhalb der Zelle lässt sich bestimmen.

Andere Textobjekte der Ausgabe (Überschriften, Erläuterungen etc.) können ebenfalls nach Doppelklick auf diese Elemente formatiert werden. Verändert werden können Schriftattribute und Ausrichtung des Absatzes.

Weiter können im Menü Format u.a. Spaltenüberschriften gedreht, Fußnoten formatiert und Umbrüche festegelegt werden.

Ändern von Text. Doppelklicken Sie auf den Text, den Sie ändern möchten. Danach erscheint er markiert. Ist das nicht der Fall, markieren Sie ihn noch durch Ziehen des Cursors mit gedrücktem linkem Mauszeiger über den Text. Sie können dann den Text löschen und neuen Text eingeben. Ändern Sie ein Element, das in der Tabelle mehrmals vorkommt, z.B. einen Wert, wird er automatisch an allen Stellen durch den neuen Namen ersetzt. Beachten Sie, dass Veränderung eines numerischen Ergebnisses in der Tabelle nicht zur Neuberechnung anderer, dieses Ergebnis beinhaltender, Werte führt (etwa der Gesamtsumme).

Ändern der Spaltenbreite. Die Standardspaltenbreite können Sie ändern, wenn Sie im Menü „Format", mit der Option „Breite der Datenzelle" die Dialogbox „Breite der Datenzelle einstellen" öffnen. Die Breite jeder einzelnen Spalte lässt sich verändern, indem man den Cursor auf den Spaltenrand führt bis sich ein Doppelpfeil bildet und dann den Rand mit gedrückter linker Maustaste verschiebt.

Grundeinstellungen der Ausgabe können im Register „Viewer", „Pivot-Tabellen", „Diagrammen" des Menüs „Bearbeiten", „Optionen" geändert werden (⇨ Kap. 28.5). Um ungleichmäßigen Darstellung von Daten innerhalb einer Tabelle zu vermeiden sei hier empfohlen, im selben Menü Register „Allgemein" die Optionsschaltfläche „Keine wissenschaftliche Notation für kleine Zahlen in Tabellen" zu markieren.

4.1.4 Pivotieren vonTabellen

Tabellen Pivotieren heißt, ihren Aufbau in Spalten, Zeilen und Schichten zu verändern. Das Pivotieren üben Sie am besten mit einer dreidimensionalen Kreuztabelle. Erstellen Sie z.B. aus ALLBUS90.SAV eine dreidimensionale Kreuztabelle: Abhängige Variable INGL, unabhängige GESCHL, Testvariable SCHUL2 (die letztere muss in das Feld „Schicht 1 von 1" übertragen werden ⇨ Kap 10.1, Abb. 10.1). In der Dialogbox „Kreuztabelle: Zellen anzeigen" wählen Sie neben „Beobachtete" Häufigkeiten „Spaltenweise" Prozentwerte. Ergebnis ist eine Tabelle, die vorerst etwas anders aussieht als in Abb. 4.3, weil über den Prozentwerten jeweils die Absolutwerte in den Zellen des Tabellenkörpers zu sehen sind.

Diese Tabelle kann man auf verschiedene Weise pivotieren. Möglich ist dies über das Menü „Pivot" und seine Optionen. Dies wird zusätzlich verfügbar, wenn Sie eine Pivot-Tabelle durch Doppelklick aktivieren. Anschaulicher gestaltet sich das Pivotieren bei Verwendung der „Pivot-Leisten", was hier dargestellt wird. Nachdem Sie eine Tabelle ausgewählt haben, öffnen Sie die „Pivot-Leisten" entweder über das Menü „Pivot", Option „Pivot-Leisten" oder über dieselbe Option des lokalen Menüs. Pivotleisten sind immer wie in Abb. 4.3 aufgebaut. Rechts befindet sich eine Randleiste „Spalte", unten eine „Zeile" und links eine mit der Bezeichnung „Schicht". Auf diesen Leisten wird durch Kästchen angezeigt, wie die gerade ausgewählte Tabelle formal aufgebaut ist. Die Kästchen repräsentieren in der Regel eine Variable, in Ausnahmefällen weitere Beschriftungen. Wenn Sie

z.B. den Cursor auf das Kästchen in der Leiste „Spalte“ setzen, öffnet sich ein Drop-Down Fenster, in dem der Variablennnamen erscheint. Das ist hier GESCHL und zeigt uns an, dass die Variable GESCHL in der Tabelle als Spaltenvariable erscheint. Das linke Kästchen in der Leiste „Zeile“ repräsentiert die Variable SCHUL2, gibt also an, dass SCHUL2 die erste Zeilenvariable dieser Tabelle ist, das nächste Kästchen steht für INGL. Dies ist die nächste Zeilenvariable. In der ursprünglichen Tabelle steht daneben noch ein Kästchen „Statistik“, weil als drittes in den Zeilen „Anzahl“ und „Prozent“ unterschieden sind. Die Tabelle in Abb. 4.3 dagegen ist schon pivotiert. Dieses Kästchen wurde nämlich in die Leiste „Schichten“ verschoben. Um ein Kästchen zu verschieben, klickt man mit der linken Maustaste darauf und zieht es mit gedrückter Taste an die gewünschte Stelle. In dem Moment, in dem man auf die Taste drückt, sieht man übrigens die Beschriftung der entsprechenden Zeilen bzw. Spalten der Tabelle zur besseren Orientierung schraffiert unterlegt. Wenn ein Kästchen in der Zeile „Schicht“ steht, wird es zugleich um Pfeile auf der linken und rechten Seite erweitert.

INGL * GESCHL * SCHUL2 Kreuztabelle

Statistik	Anzahl			
			GESCHL	
SCHUL2			MAENNLICH	WEIBLICH	Gesamt
Hauptschule	INGL	POSTMATERIALISTEN	10	6	16
		PM-MISCHTYP	20	16	36
		M-MISCHTYP	26	30	56
		MATERIALISTEN	12	23	35
	Gesamt		68	75	143
Mittelschule	INGL	POSTMATERIALISTEN			
		PM-MISCHTYP			
		M-MISCHTYP			
		MATERIALISTEN			
	Gesamt				
Fachh/Abi	INGL	POSTMATERIALISTEN			
		PM-MISCHTYP			
		M-MISCHTYP			
		MATERIALISTEN			
	Gesamt		34	35	69

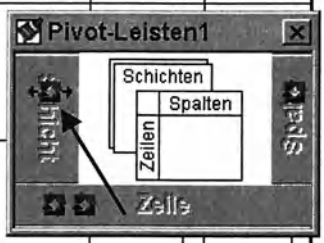


Abb. 4.3. Dreidimensionale geschichtete Tabelle mit „Pivot-Leisten“

Werden Schichten gebildet, so heißt das, dass für jede Ausprägung der Schichtungsvariablen eine eigene Tabelle für die Kombinationen der anderen Variablen gebildet wird. In unserem Beispiel wurde keine eigentliche Untersuchungsvariable, sondern „Statistik“ zum Schichten verwendet. Diese Variable hat die Ausprägungen „Anzahl“ und „Prozent von Geschlecht“. So wurde eine Tabelle mit den „Anzahl“-Werten und eine mit den „Prozentwerten“ für den Zusammenhang Geschlecht, Schulbildung und Materialismus gebildet. Selbstverständlich kann man auch anders schichten. So etwa SCHUL2 zur Schichtungsvariablen machen. Dann erhält man eine eigene Tabelle für jede Schulbildungsgruppe. (Dies könnte durchaus mit der Schichtungsvariablen „Statistik“ kombiniert werden, wodurch sich 6

eigene Tabellen ergäben.) Wurden Schichten gebildet, erscheint derer Name/die Namen der Schichtungsvariablen im Kopf der Tabelle. An der Seite dieses Feldes befindet sich ein Auswahlpfeil. Klicken Sie auf diesen, dann öffnet sich eine Auswahlliste mit den Werten der Schichtungsvariablen. Durch Klicken auf den Namen eines dieser Werte können Sie die Tabelle der zu diesem gehörenden Schicht öffnen. Zwischen den Schichten kann man auch in den „Pivot-Leisten" wechseln. Klicken Sie dazu auf den linken oder rechten Pfeil an dem entsprechenden Kästchen für die Schichtungsvariable auf der Randleiste „Schicht".

Schichtenbildung ist eine Möglichkeit des Pivotierens. Häufiger werden aber Spalten zu Zeilen umdefiniert werden und/oder Zeilen zu Spalten. Dies geschieht ebenfalls durch Ziehen des Variablensymbols von einer Leiste in die andere. So könnte man in unserem Beispiel etwa Geschlecht zur Zeilen und Schulbildung zur Spaltenvariablen machen. (Die Prozentuierungsrichtung wird sachlich zutreffend angepasst.) Die Reihenfolge innerhalb einer Leiste kann ebenfalls entsprechend geändert werden. So könnte man in unserem Beispiel etwa die Reihenfolge der Zeilenvariablen INGL und SCHUL2 ändern. Probieren Sie am besten alle Pivotierungsmöglichkeiten aus. Die wichtigsten Varianten wie „Zeilen und Spalten vertauschen", „Schichten in Zeilen bzw. Spalten verschieben" können auch über das Menü „Pivot" gewählt werden. Vor allem kann man dort auch „Pivots auf Standartwerte" zurücksetzen und damit die Ausgangstabelle wieder erzeugen.

4.1.5 Ändern von Tabellenformaten

Bei der äußeren Gestaltung der Tabellen sind Sie weitgehend auf die von SPSS gelieferten Tabellenformate angewiesen. Jedoch bietet das Programm neben dem voreingestellten Format zahlreiche weitere zur Auswahl. Um eine Tabelle in einem dieser Formate zu formatieren, gehen Sie wie folgt vor. Wählen Sie die Tabelle durch Doppelklicken zum Pivotieren aus. Wählen Sie „Format", „Tabellenvorlagen". Es öffnet sich die Dialogbox „Tabellenvorlagen". Im Auswahlfeld „Dateien für Vorlagen" finden Sie eine Liste der verfügbaren Vorlagen (evtl. müssen Sie über das Schaltfeld „Durchsuchen" erst die Dialogbox „Öffnen" anwählen und dort das Verzeichnis einstellen, in dem sich die Vorlagen befinden. Das Verzeichnis heißt per Voreinstellung „Look", die Dateien haben die Extension „tlo"). Wenn Sie den Namen einer der Vorlagen markieren, sehen Sie im Fenster „Vorschau" eine Darstellung der äußeren Gestalt einer Tabelle mit dieser Vorlage. Markieren Sie den Namen der gewünschten Vorlage und bestätigen Sie mit „OK".

In begrenztem Rahmen kann man auch eigene Tabellenvorlagen erstellen. Dazu markieren Sie wiederum in der Dialogbox „Tabellenvorlagen" den Namen einer Vorlage, die Ihren Wünschen am nächsten kommt. Durch Anklicken der Schaltfläche „Tabellenvorlage bearbeiten" öffnen Sie die Dialogbox „Tabelleneigenschaften für". Dort können Sie in verschiedenen Registern Veränderungen vornehmen. So kann im Register „Allgemein" etwa die Spaltenbreite verändert werden. Weiter sind einstellbar: das „Zellenformat" (Schrift, Ausrichtung, Rahmen und Farbe), Eigenschaften von „Fußnoten" und „Rahmen" (Strichart, Stärke und Farbe) sowie bestimmte Druckoptionen. Sie bestätigen die Veränderungen mit „OK" und speichern die neue Vorlage entweder mit „Vorlage speichern" unter dem alten Namen

oder mit „Speichern unter“ durch Eingabe von Verzeichnis und Namen in der gleichnamigen Dialogbox als neues Tabellenformat.

Letztlich ist es möglich, ein anderes als das voreingestellte Tabellenformat zum Standardtabellenformat zu bestimmen. Dazu wählen Sie „Bearbeiten“, „Optionen“ und das Register „Pivot-Tabellen“. Dort markieren Sie im Auswahlfenster „Tabellenvorlagen“ den Namen des gewünschten Formates (evtl. müssen Sie über das Schaltfeld „Durchsuchen“ erst die Dialogbox „Öffnen“ anwählen und dort das Verzeichnis einstellen, in dem sich die Vorlagen befinden). Bestätigen Sie das ausgewählte Tabellenformat mit „OK“. Es wird jetzt automatisch auf jede neu erstellt Tabelle angewendet (⇨ Kap. 28.5).

4.1.6 Arbeiten mit dem Textviewer

Anstelle des grafisch orientierten Viewers kann auch ein Textviewer für die Ausgabe benutzt werden. Dort werden alle Ausgaben – mit Ausnahme der Diagramme, die als nicht weiter bearbeitbare Grafiken erscheinen – im ASCII-Format ausgegeben. Ein Vorteil des Textviewers besteht darin, dass solche Ausgabedateien weniger Speicherplatz benötigen. Zum anderen lassen sich die Ergebnisse in einige andere Programme nur im ASCII-Format übertragen, oder es geht zumindest leichter. Soll der Text-Viewer schon beim Start als reguläres Ausgabefenster festgelegt werden, kann man dies im Menü „Bearbeiten“, „Optionen“, Register „Allgemein“ einstellen (⇨ Kap. 28.5). Während der Sitzung kann man ein Text-Viewer Fenster ebenfalls öffnen. Wählen Sie dazu „Datei“, „Neu“ und in der sich öffnenden Liste „Textausgabe“. Die Ausgabeergebnisse werden von nun an in das Text-Viewer Fenster ausgegeben, es sei denn, ein anderes Fenster wird als dezidiertes Fenster deklariert. Viewer- und Text-Viewer-Fenster können nebeneinander geöffnet sein.

4.2 Arbeiten im Syntaxfenster

4.2.1 Erstellen und Ausführen von Befehlen

Ein Syntaxfenster öffnet sich automatisch mit der Befehlssyntax dieses Befehls, wenn man in einer Dialogbox die Schaltfläche „Einfügen“ anklickt. Eine bereits bestehende Syntaxdatei kann man in den Syntaxeditor über die Befehlsfolge „Datei“, „Öffnen“ , „Syntax“ auf die übliche Weise laden. Auch das „Speichern“ unterscheidet sich nicht vom Vorgehen beim Speichern der Inhalte anderer Fenster. Für das Festlegen des Hauptfensters gelten zunächst dieselben Regeln, die auch für das Ausgabefenster zutreffen. Es sei daher auf die Ausführungen in Abschnitt 4.1.1 verwiesen. Der Unterschied liegt lediglich darin, dass beim Öffnen als Dateityp „Syntax...“ zu wählen ist, gegebenenfalls ebenso beim Speichern. Die jeweiligen Dialogboxen heißen „Datei öffnen“ bzw. „Speichern unter“, die voreingestellte Extension SPS. Ansonsten ist genauso, wie unter Abschnitt 4.1.1 dargestellt, zu verfahren. SPSS-Befehle können im Syntaxfenster selbst geschrieben oder aus einer in einem anderen Programm erstellten Textdatei importiert werden. Sie können auch mit der Option „Einfügen“ aus der Dialogbox übertragen werden. Auch aus dem Hilfesystem zur Befehlssyntax können die Befehle durch Kopieren in die

Zwischenablage (markieren und mit „Optionen", „Kopieren" in die Zwischenablage übernehmen) und „Einfügen" übertragen werden. Schreibt man die Befehle im Syntaxfenster selbst, ist es hilfreich, Variablennamen aus der Variablenliste zu übernehmen. Wählen Sie dazu:

- „Extras", „Variablen...". Es öffnet sich die Dialogbox „Variablen".
- Markieren Sie den oder die Variablennamen in der Quellvariablenliste dieser Dialogbox, und übertragen Sie ihn/sie durch Anklicken von „Einfügen".

Editiert wird auf die gleiche Weise wie in einem einfachen Schreibprogramm. Texte können eingefügt oder überschrieben werden. Gelöscht wird mit den Löschtasten. Texte können über das Menü „Bearbeiten" ausgeschnitten, kopiert und eingefügt werden. Mit „Bearbeiten" und „Suchen" oder durch Anklicken des Fernglassymbols öffnet man die Dialogbox „Suchen und ersetzen", mit der man im Register „Suchen" eine Suche nach gewünschten Zeichenketten im Syntaxtext durchführen kann. Im Register „Ersetzen" kann man gleichzeitig die Suchbegriffe durch andere Begriffe ersetzen. Die Schrift im Syntaxfenster kann in der Dialogbox „Schriftart" geändert werden. Sie öffnet sich bei der Befehlsfolge „Ansicht", „Schriftarten".

Befehle werden über das Menü „Ausführen" gestartet. Wählt man die Option „Alles", werden sämtliche im Syntax-Editor befindlichen Befehle gestartet. Will man nur einen Teil davon abschicken, muss man anders verfahren. Befindet sich der Cursor in einer Befehlszeile und wählt man die Option „Aktueller Befehl", wird nur der zu dieser Zeile gehörige Befehl ausgeführt. „Bis Ende" führt alle Befehle ab dem Befehl, in dessen Zeile sich der Cursor befindet, aus. Schließlich kann man auch Befehle durch Ziehen des Cursors markieren und mit „Auswahl" abschicken. Es werden nur die markierten Befehle ausgeführt.

Symbolleiste. Die Symbolleiste enthält speziell für das Syntaxfenster zwei weitere Befehle:

Aktuellen Befehl ausführen. Führt die im Syntaxfenster markierten Befehle aus. Ist kein Befehl markiert, wird der Befehl ausgeführt, in dem sich der Cursor befindet.

Hilfe zur Syntax. Führt zu einer kontextsensitiven Hilfe für die Syntaxbefehle. Durch Anklicken des Symbols öffnet sich ein Fenster, das ein Syntaxdiagramm für die Befehlszeile enthält, in der der Cursor sich gerade befindet (⇨ Abb. 4.4). Überschrieben ist es mit der englischen Bezeichnung des Befehls. Ist in dem Bereich, in der sich der Cursor befindet, kein Befehl enthalten, wird eine Gesamtliste aller Befehle angezeigt. Markieren Sie die Zeile „command syntax" zu einem dieser Befehle, und klicken Sie auf „Anzeigen". Das Syntaxdiagramm dieses Befehls erscheint.

4.2.2 Charakteristika der Befehlssyntax

In der Regel wird in diesem Buch davon ausgegangen, dass SPSS für Windows mit Hilfe des Dialogsystems und der für sie charakteristischen Fenstertechnik bedient wird. Es kann jedoch sinnvoll sein, auch unter dieser Oberfläche mit Befehlsdateien zu arbeiten, die in der üblichen SPSS-Syntax programmiert sind und im SPSS-Syntaxfenster ablaufen können. Das gilt, wenn Befehle genutzt werden sollen, die nur bei Gebrauch der Befehls-Syntax zur Verfügung stehen. Auch wenn Befehlssequenzen häufig wiederholt oder wenn umfangreiche Routinen bearbeitet werden, empfiehlt sich die Nutzung von Stapeldateien. Die Befehle können überwiegend in den Dialogboxen erzeugt und in das Syntax-Editorr übertragen werden. Routinierte Programmierer werden diese aber häufig auch selbst schreiben. Unerlässlich ist dies bei Verwendung von nur in der Syntax verfügbaren Befehlen.

Hier ist nicht der Platz, die gesamte Befehlssyntax zu beschreiben. Ausführlich findet man sie im „SPSS Base System Syntax Reference Guide". Dieser wird auf der Installations CD-ROM mitgeliefert und kann, wenn installiert, im Hilfemenü mit der Option „Syntax Guide" und „Base" aufgerufen werden. Der Syntax Guide wird dann mit dem mitgelieferten Programm „Acrobat Reader" lesbar. Dieses enthält zum Suchen – ähnlich dem Viewer – ein Gliederungsfenster neben dem eigentlichen Inhaltsfenster. Der Reference Guide enthält neben den Befehlsdiagrammen ausführliche Erläuterungen.

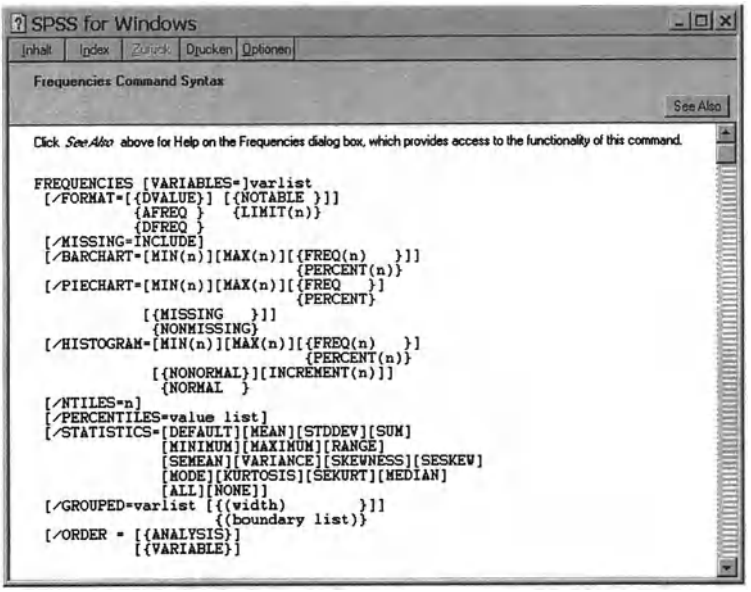


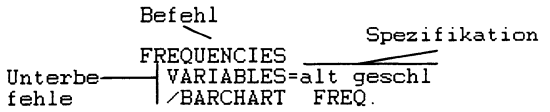
Abb. 4.4. Hilfefenster für die Befehlssyntax mit Syntax für den Befehl „Frequencies"

Im Hilfesystem sind jedoch die verfügbaren Befehle auch in Form von Befehlsdiagrammen dargestellt. Häufig reicht es aus, diese lesen zu können. Daher sollen hier kurz die Konventionen dieser Diagramme erläutert werden.

Eine Befehlsdatei besteht aus einem oder mehreren Befehlen. Jeder Befehl beginnt in einer neuen Zeile. Er wird durch einen Punkt abgeschlossen. Die Syntaxdiagramme geben jeweils die Syntax eines Befehles wieder. Dabei wird der Befehl in allen möglichen Varianten angegeben. Aus diesen wird man beim Programmieren lediglich eine Auswahl treffen. Der Befehl ist lauffähig, wenn er die Mindestangaben enthält.

Der gesamte Befehl kann aus mehreren Teilen zusammengesetzt sein. Obligatorisch ist das eigentliche *Befehlswort*. Zusätzlich können *Unterbefehle* erforderlich sein. Diese werden in der Regel durch / abgetrennt. Weiter kann ein Befehl *Spezifikationen* erfordern. Insbesondere müssen die Variablen angegeben werden, auf die sich der Befehl bezieht. Andere Angaben wie Bereichsgrenzen u.ä. werden bisweilen ebenfalls benötigt. Für Befehle, Unterbefehle und einige Spezifikationen sind *Schlüsselwörter* reserviert, die in der angegebenen Form verwendet werden müssen. Allerdings reicht für das Befehlswort, die Unterbefehle und sonstigen Schlüsselworte fast immer eine auf drei Zeichen abgekürzte Angabe aus. Das gilt nur dann nicht, wenn dadurch keine eindeutige Unterscheidung zustande kommt, so nicht bei zusammengesetzten Befehlen (z.B. FILE LABEL) und den INFO-Spezifikationen.

Beispiel. Ein Befehl, der für die Variablen „ALT" und „GESCHL" eine Häufigkeitsauszählung ausführt und ein Balkendiagramm auf Basis der Prozentwerte erstellt:



Schlüsselwörter sind: FREQUENCIES; VARIABLES; BARCHART und FREQ.

Beispiel für einen Minimalbefehl: FRE alt.

Das Beispiel zeigt einen lauffähigen Befehl. Der Befehl FREQUENCIES wird durch das abgekürzte Schlüsselwort „FRE" aufgerufen. „alt" ist ein Variablennamen. Der Befehl wird durch einen Punkt abgeschlossen.

Zu beachten ist: Variablennamen müssen immer ausgeschrieben sein. Eine Befehlszeile darf maximal 80 Zeichen umfassen. Als Dezimalzeichen muss immer der Punkt verwendet werden. In Apostrophe oder Anführungszeichen gesetzte Texte dürfen sich nur innerhalb einer Zeile befinden. Kommandos, Unterkommandos, Schlüsselwörter und Variablennamen können in großen oder kleinen Buchstaben geschrieben werden. Sie werden automatisch in Großbuchstaben transformiert. Dagegen wird bei allen anderen Spezifikationen die Schreibweise beachtet.

Das Syntaxdiagramm ist nach folgenden Konventionen aufgebaut:

- Alle Schlüsselwörter sind in Großbuchstaben geschrieben. (z.B. FREQUENCIES; BARCHART; MIN; MAX usw.).

- Angaben in Kleinschrift bedeuten, dass hier Spezifikationen durch den Nutzer erwartet werden. (*Beispiel:* varlist bedeutet, dass eine Liste der Variablen eingegeben werden muss, für die der Befehl ausgeführt werden soll.)
- In eckige Klammern gesetzte Angaben können wahlweise gemacht werden, müssen aber nicht. (*Beispiel:* Der Unterbefehl „VARIABLES=„ muss nicht angegeben werden. Man kann auch die Variablenliste ohne ihn eingeben.)
- Kann zwischen mehreren Alternativen gewählt werden, werden die Alternativen in geschweiften Klammern untereinander angegeben. (*Beispiel:* Im Unterkommando FORMAT – das nicht unbedingt benutzt werden muss – kann man zwischen den Alternativen DVALUE, AFREQ und DFREQ wählen.)
- Werden Angaben verwendet, die in der Syntax in runden Klammern, Apostrophen oder Anführungszeichen angegeben werden, so sind diese Zeichen auf jeden Fall mit anzugeben. *Beispiel:* MIN(10) beim Unterbefehl BARCHART besagt, dass ein Wert unterhalb der Grenze zehn nicht ausgedruckt werden soll.
- Fett gedruckte Angaben zeigen, dass diese die Voreinstellung sind. *Anmerkung.* Fettdruck ist im Hilfediagramm nicht zu erkennen, wohl aber im „Syntax Guide". (*Beispiel:* **FREQ** beim Unterbefehl Barchart zeigt, dass die Balken des Diagramms per Voreinstellung die Absolutwerte und nicht die Prozentwerte repräsentieren.)

  Man kann zwei Arten von Voreinstellung unterscheiden. Im einen Fall handelt es sich um die Voreinstellung, die eingehalten wird, wenn der Unterbefehl gänzlich ausgelassen wird. Gekennzeichnet wird dies mit **. (*Beispiel:* **TABLE**** im Unterkommando MISSING bei CROSSTABS bedeutet, dass auch dann, wenn der Unterbefehl MISSINGS gar nicht genannt wird, per Voreinstellung die fehlenden Werte aus der Tabelle ausgeschlossen werden.) Im anderen Falle wird die Voreinstellung dann benutzt, wenn der Unterbefehl ohne weitere Spezifikation Verwendung findet. (*Beispiel:* im Unterbefehl BARCHART von FREQUENCIES wird verwendet, wenn nichts anderes angegeben, d.h. die Balkenhöhe des Diagramm entspricht den absoluten Häufigkeiten. Sollte sie den Prozentwerten entsprechen, müsste PERCENT ausdrücklich angegeben werden.)
- *var* bedeutet, ein Variablennamen muss eingegeben werden, *varlist*, eine Liste von Variablennamen. Häufig ist beides alternativ möglich.

Beim Arbeiten im Produktionsmodus (⇨ Kap. 28.6) benutzt man häufig den INCLUDE-Befehl. Für Befehlsdateien, die den INCLUDE-Befehl benutzen, gilt abweichend: Jeder Befehl muss in der ersten Spalte einer neuen Zeile beginnen. Fortsetzungszeilen müssen mindestens um ein Leerzeichen eingerückt werden.

Beispiel:

```
DATA LIST FILE 'Daten.dat' FIXED / v1 1 v2 to v6 2-11 v7 12 v8 to v9 13-16
  v10 17.
FREQUENCIES VARIABLES=v1.
```

Benutzen von Protokoll- und Ausgabedateien für das Programmieren mit der Befehlssyntax. Wenn Sie bei den Optionen von „Bearbeiten" im Register „Allgemein" „Befehlssyntax in Journaldatei aufzeichnen" gewählt haben (⇨ Kap. 28.5), wird in der Protokolldatei die Befehlssyntax aller in ihrer Sitzung abgearbeiteten

Befehle protokolliert. Für das Erstellen einer Syntaxdatei können Sie dann die Protokolldatei (Standardname SPSS.JNL) benutzen. Sie befindet sich im Verzeichnis, das Sie für die temporären Dateien bestimmt haben. Laden Sie dazu die Protokolldatei in das Syntaxfenster. (Sie wird im Auswahlfenster der Dialogbox „Datei öffnen" mit angezeigt, wenn sie als „Dateityp" „Alle Dateien" wählen.) Bearbeiten Sie diese, bis nur die gewünschte Befehlsfolge übrig bleibt und starten Sie den Lauf. Dasselbe ist möglich bei Benutzung der Ausgabedatei. Dazu muss allerdings die Ausgabe auch die Befehlssyntax umfassen. Das ist möglich, wenn im Menü „Bearbeiten", „Optionen" im Register „Viewer" das Auswahlkästchen „Befehle im Log anzeigen" markiert haben (⇨ Kap. 28.5, Abb. 28.10).

Auch hier müssen Sie die Datei so bearbeiten, dass nur die Befehlssyntax verbleibt und diese in ein Syntaxfenster übertragen. Sie können dazu z.B. die einzelnen Befehlsteile aus der Ausgabedatei herauskopieren. Im Syntaxfenster starten Sie den Lauf.

5 Transformieren von Daten

SPSS bietet eine Reihe von Möglichkeiten, Daten zu transformieren. Damit kann man in erster Linie Berechnungen durchführen. Aus den Werten verschiedener Variablen können neue Ergebnisvariablen berechnet werden. Das wird man z.B. verwenden, wenn ein Überschuss oder Verlust aus der Differenz zwischen Einnahmen und Ausgaben zu ermitteln ist. Oder man berechnet die monatlich für einen Kredit zu zahlende Rate aus Kredithöhe und Zins. Die Berechnung kann sich auch auf die Zuweisung eines festen Wertes beschränken. Weiter kann man Datentransformationen benötigen, wenn die Daten nicht den Bedingungen der statistischen Analyse entsprechen, z.B. keine linearen oder orthogonal Beziehungen zwischen den Variablen bestehen, oder wenn unvergleichbare Maßstäbe bei der Messung verschiedener Variablen verwendet wurden. Verschiedene Transformationsmöglichkeiten, wie z-Transformation, Logarithmieren u.ä. können hier Abhilfe schaffen (solche Funktionen stellen auch verschiedene Statistikprozeduren zur Verfügung). Es ist auch möglich, solche Berechnungen jeweils für ausgewählte Fälle, die eine bestimmte Bedingung erfüllen, durchzuführen. Das benötigt man beispielsweise, um eine Gewichtungsvariable zu konstruieren (⇨ Kap. 2.7). Von großer Bedeutung ist schließlich die Möglichkeit, Daten umzukodieren. Man kann dabei anstelle der alten Werte neue Werte setzen. Dies nutzt man insbesondere zur Zusammenfassung mehrerer Werte oder großer Wertebereiche zu Werteklassen.

5.1 Berechnen neuer Variablen

Nehmen wir an, in der Datei VZ.SAV, die in Kap. 3.1 zur Illustration der Datendefinition benutzt wurde, soll aus den Angaben über die Kreditbeträge und die Zinshöhen die monatliche Zinsbelastung berechnet und in einer neuen Variablen MON_ZINS gespeichert werden. Die neue Variable soll zudem ein Variablen-Label „monatliche Zinszahlung" erhalten. Alle Schuldner müssen in dieser Datei zwei Kredite bedienen, deren Höhe in den Variablen KREDIT1 und KREDIT2 und deren jährliche Zinshöhe in Prozent in den Variablen ZINS1 und ZINS2 gespeichert ist. Die monatliche Zinsbelastung in DM ergibt sich demnach als:

MON_ZINS = ((KREDIT1 * ZINS1) / 100 + (KREDIT2 * ZINS2) / 100)) / 12

Für eine Berechnung wählen Sie die Befehlsfolge:

▷ „Transformieren", „Berechnen...". Es öffnet sich die Dialogbox „Variable berechnen" (⇨ Abb. 5.1).

- ▷ Geben Sie in das Eingabefeld „Zielvariable:“ den Namen der Variablen ein, die das Ergebnis der Berechnung erhalten soll. Es kann eine neue Variable oder eine bereits existierende sein. Im letzteren Falle wird immer eine Warnmeldung ausgegeben: „Wollen Sie eine existierende Variable ändern?“ und die Transformation wird erst nach Bestätigung mit „OK“ ausgeführt.
- ▷ Stellen Sie im Eingabefeld „Numerischer Ausdruck:“ die Berechnungsformel zusammen. Es kann sich dabei um einen einfachen Wert, aber auch um sehr komplexe Formeln unter Einbezug von Variablenwerten, arithmetischen, statistischen und logischen Funktionen und Verwendung verschiedener Arten von Operatoren handeln.

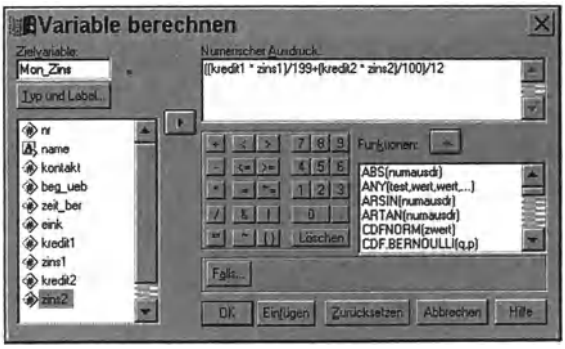


Abb. 5.1. Dialogbox „Variable berechnen“ mit Ausdruck für die Variable ´MON_ZINS´

In unserem Beispiel benutzen wir dazu lediglich die sogenannte Rechnertastatur, das sind die grau unterlegten Knöpfe in der Mitte der Dialogbox, und die Variablenliste. Wir klicken zunächst auf die Doppelklammer in der Rechnertastatur und wiederholen das, so dass zwei Klammerpaare ineinander geschachtelt stehen. Wir setzen den Cursor in die innere Klammer. Dann markieren wir die Variable KREDIT1 in der Variablenliste und übertragen sie durch Anklicken von ▶ (oder Doppelklick auf den Variablennamen) in die Klammer. Durch Anklicken von * übertragen wir den Multiplikationsoperator. Dann übertragen wir auf die angegebene Weise die Variable ZINS1. Durch Anklicken von / übernehmen wir den Divisionsoperator und geben dann den Wert 100 ein. Der erste Klammerausdruck der Formel ist gebildet. Neben die innere Klammer setzen wir das Pluszeichen. Um den zweiten Klammerausdruck zusammenzusetzen, fügen wir zunächst eine Doppelklammer neben dem Pluszeichen ein und übertragen dann in der beschriebenen Weise die Variablennamen KREDIT2, ZINS2, die Operatoren und die Zahl 100. Zum Abschluss fügen wir hinter die äußere Klammer das Divisionszeichen und die 12 an.

Wenn Sie den voreingestellten Variablentyp ändern und/oder Variablen-Label vergeben wollen, gehen Sie wie folgt vor:

- ▷ Klicken Sie auf die Schaltfläche „Typ und Label“. Die Dialogbox „Variablen berechnen: Typ und Label“ öffnet sich (⇨ Abb. 5.2).

Man kann zwischen numerischen und Stringvariablen wählen. Als Variablenlabel kann eine in das Eingabefeld „Label" einzugebende Zeichenkette oder aber der im Feld „Numerischer Ausdruck:" enthaltene Ausdruck dienen.

▷ Bestätigen Sie mit „Weiter" und „OK". Die neuen Werte werden berechnet und in die Variable eingetragen.

Abb. 5.2. Dialogbox „Variablen berechnen: Typ und Label" mit Variablenlabel

Hinweis. In einer Funktion müssen Dezimalzahlen immer mit Punkt als Dezimaltrennzeichen eingegeben werden. Stringwerte müssen in Hochkommas oder Anführungszeichen gesetzt werden.

Operatoren. Die Option „Berechnen" bietet drei Arten von Operatoren. Sie sind auf der „Rechnertastatur" in der Dialogbox enthalten und können von ihr übertragen, aber auch normal über die PC-Tastatur eingegeben werden.

- *Arithmetische Operatoren.* Sie ermöglichen die üblichen Rechenarten: Addition (+), Subtraktion (-), Multiplikation (*), Division (/) und Potenzieren (**). Die Abarbeitung folgt den üblichen Regeln, zunächst Potenzieren, dann Punktrechnung, schließlich Strichrechnung. Aber Funktionen werden vorab berechnet. Die Reihenfolge kann durch Klammern, die ebenfalls auf der Tastatur vorhanden sind, verändert werden.
- *Relationale Operatoren* (Vergleichsoperatoren). Mit ihrer Hilfe werden zwei Werte verglichen. Sie werden insbesondere im Zusammenhang mit bedingten Transformationen gebraucht. Relationale Operatoren sind: < (kleiner), > (größer), <= (kleiner/gleich), >= (größer/gleich), = (gleich) und ~= (ungleich).
- *Logische Operatoren.* Mit ihnen verbindet man zwei relationale Ausdrücke oder kehrt den Wahrheitswert eines Bedingungsausdrucks um. Auch sie werden vornehmlich im Zusammenhang mit bedingten Ausdrücken gebraucht. Logische Operatoren sind:

  
  „Logisches Und". Beide Ausdrücke müssen wahr sein.

  
  „Logisches Oder" (im Sinne von entweder oder). Einer der beiden Ausdrücke muss wahr sein.

  
  „Logisches Nicht". Kehrt den Wahrheitswert des Ausdrucks um.

Funktionen. Die Option „Berechnen" stellt eine umfangreiche Reihe von Funktionen zur Verwendung in numerischen Ausdrücken zur Verfügung. Sie sind alle im

Auswahlfeld „Funktionen“ enthalten. Um zur gesuchten Funktion zu gelangen, muss man u.U. in diesem Auswahlfeld scrollen (beim Eintippen eines Buchstabens springt der Cursor auf die erste Funktion mit diesem Buchstaben als Anfangsbuchstaben). Die Funktion überträgt man in das Feld „Numerischer Ausdruck“, indem man sie zuerst markiert und dann ▲ anklickt (oder durch Doppelklicken auf die Funktionsbezeichnung). Gegebenenfalls müssen noch Werte, Variablen etc. in die Funktion eingesetzt werden. Die Funktionen sind im folgenden dargestellt. Die kursiv gedruckte Angabe zeigt das Format der Ausgabevariablen an.

© *Arithmetische Funktionen.*

ABS(numausdr). *Numerisch.* Ergibt den Absolutbetrag eines numerischen Ausdrucks. *Beispiel:* ABS(-5-8) ergibt 13.
RND(numausdr). *Numerisch.* Rundet zur nächstgelegenen ganzen Zahl. *Beispiel*: 3.8 ergibt 4.
TRUNC(numausdr). *Numerisch.* Schneidet Dezimalstellen ab. *Beispiel:* 3,8 ergibt 3.
MOD(numausdr,modulus). *Numerisch.* Der Rest einer Division eines Arguments durch ein zweites (Modulus). *Beispiel:* MOD(930,100) ergibt 30. Die Argumente werden durch Komma getrennt.
SQRT(numausdr). *Numerisch.* Quadratwurzel des Ausdrucks.
EXP(numausdr). *Numerisch.* Exponentialfunktion, gleich $e^{(numausdr.)}$.
LG10(numausdr). *Numerisch.* Logarithmus zur Basis 10.
LN(Numausdr). *Numerisch.* Natürlicher Logarithmus, Basis e.
ARSIN(numausdr). *Numerisch.* Arkussinus. Ergebnisse in Bogenmaß.
ARTAN(numausdr). *Numerisch.* Arkustangens. Ergebnisse in Bogenmaß.
SIN(radiant). *Numerisch.* Sinus. Argumente müssen in Bogenmaß eingegeben werden.
COS(radiant). *Numerisch.* Kosinus. Argumente müssen in Bogenmaß eingegeben werden.

Bei den numerischen Ausdrücken kann es sich um einzelne Zahlen, aber auch komplexe Ausdrücke handeln. Gewöhnlich werden auch Variablen in ihnen enthalten sein.

Beispiel: Die Werte der Variablen EINK sollen logarithmiert und in der neuen Variablen LOGEINK gespeichert werden:

Abb. 5.3. Rechnen mit einer arithmetischen Funktion

- ▷ Tragen Sie in das Eingabefeld „Zielvariable“ den neuen Variablennamen LOGEINK ein.
- ▷ Markieren Sie im Feld Funktionen die Funktion LG10(numausdr) und übertragen Sie sie in das Feld „Numerischer Ausdruck“. Es erscheint LG10(?).

▷ Markieren Sie das Fragezeichen in der Funktion, markieren Sie in der Quellvariablenliste EINK und übertragen Sie die Variable in den Ausdruck.
▷ Bestätigen Sie mit „OK".

Eine Funktion verlangt immer das Einsetzen von *Argumenten*. Per Voreinstellung enthält sie bei Übertragung so viele Fragezeichen wie die Mindestzahl der Argumente beträgt. Argumente trägt man ein, indem man das Fragezeichen markiert und das Argument danach eingibt. Wird (bei statistischen und logischen Funktionen) mehr als die Mindestzahl an Argumenten verwendet, fügt man die zusätzlichen Argumente durch Komma getrennt in die Argumentliste ein.

① *Statistische Funktionen.*

SUM(numausdr,numausdr,...). *Numerisch.* Summe der Werte über die Argumente. *Beispiel:* SUM(kredit1,kredit2) ergibt die gesamte Kreditsumme für die beiden Kredite.
MEAN(numausdr,numausdr,...). *Numerisch.* Arithmetisches Mittel über die Argumente.
SD(numausdr,numausdr,...). *Numerisch.* Standardabweichung über die Argumente.
VARIANCE(numausdr,numausdr,...). *Numerisch.* Varianz über die Argumente.
CFVAR(numausdr,numausdr,...). *Numerisch.* Variationskoeffizient über die Argumente.
MIN(wert,wert,...). *Numerisch oder String.* Kleinster Wert über alle Argumente. (Bei Stringvariablen der in alphabetischer Reihenfolge erste Wert.)
MAX(wert,wert,...). *Numerisch oder String.* Größter Wert über alle Argumente. (Bei Stringvariablen der in alphabetischer Reihenfolge letzte Wert.)

Alle Ausdrücke haben mindestens zwei durch Komma getrennte Argumente. Gewöhnlich ergeben sich diese Argumente aus Variablenwerten verschiedener Variablen. Im Unterschied zur Berechnung statistischer Maßzahlen zur Beschreibung eindimensionaler Häufigkeitsverteilungen, geht es hier um die Zusammenfassung mehrerer Argumente/Variablen jeweils eines Falles, sei es in Form einer Summe, eines arithmetischen Mittels, einer Standardabweichung usw.. Deshalb sollte man genau prüfen, inwiefern dies nützlich ist. In unserem Beispiel aus der Datei VZ.SAV kann das für die Summenbildung bejaht werden. Der Gesamtbetrag mehrerer Kredite ist eine wichtige Information. Bei allen anderen Maßzahlen wäre das fraglich. Was können wir aus dem arithmetischen Mittel zweier Kredite oder aus Streuungsmaßen, die sich auf nur zwei Kredite beziehen, entnehmen? Eher nutzbringende Verarbeitungsmöglichkeiten wären schon für die Ergebnisse der Funktion MIN und MAX denkbar. Haben wir dagegen zahlreiche Messungen einer latenten Variablen, etwa eine Testbatterie bei psychologischen Tests, vorliegen, kann durchaus der Durchschnittswert ein sinnvoller zusammenfassender Wert sein. Varianz, Standardabweichung und Variationskoeffizient sind vielleicht brauchbare Maße für die Homogenität bzw. Heterogenität der verschiedenen Messungen.
Bei den Funktionen SUM bis MAX können Sie auch eine *Mindestzahl gültiger Argumente* angeben. Wird diese Zahl unterschritten, setzt SPSS in der Ergebnisvariablen einen System-Missing-Wert ein. Dazu wird zwischen den Funktionsnamen

und der ersten öffnenden Klammer ein Punkt und die gewünschte Mindestzahl gesetzt. *Beispiel:* Sum.2(kredit1,kredit2,kredit3) berechnet nur dann eine Summe, wenn mindestens für zwei Kredite ein gültiger Wert vorliegt. Ansonsten wird ein System-Missing-Wert eingesetzt.

② *Logische Funktionen.*

RANGE(test,min,max[,min,max...]). *Logisch.* Dient dazu zu prüfen, ob ein Wert innerhalb eines oder mehrerer Bereiche liegt. „Test" steht gewöhnlich für einen Variablennamen, „min" und „max" für die Grenzen des Bereiches. *Beispiel:* Es soll geprüft werden, ob jemand in die Gruppen der „Armen" fällt. Dies sei der Fall bei einem Einkommen von 0 bis 2000 DM. Entsprechend gälte die logische Funktion: Range(EINK,0,2000). Wahr ergibt den Wert 1, nicht wahr eine 0.
ANY(test,wert,wert,...). *Logisch.* Kann die Werte wahr (= 1) und nicht wahr (= 0) annehmen. Ist wahr, wenn der Wert des ersten Arguments (gewöhnlich eines Tests oder einer Variablen) mit irgendeinem der folgenden Argumente übereinstimmt. Das erste Argument (Test) ist gewöhnlich ein Variablennamen. *Beispiel:* Es sollen aus der Datei einer Wahlumfrage auf Basis der Angaben in Variable PART_91 alle die Fälle ausgewählt werden, die irgendeine konservative Partei wählen wollen. Diese seien CDU = 2, REP = 5, DVU = 8. Entsprechend gälte: ANY (part_91,2,5,8).

③ *Funktionen für Zufallszahlen.*

Es kann sein, dass man für irgendeinem Zweck Fällen Zufallszahlen zuweisen möchte. Etwa könnte man in Kombination mit einem Bedingungsausdruck fehlende Werte durch Zufallszahlen ersetzen (Imputation). Es gibt zwei Funktionen, die das ermöglichen:

NORMAL(stdabw). *Numerisch.* Jedem Fall wird eine Pseudo-Zufallszahl aus einer Normalverteilung mit dem Mittelwert 0 und einer nutzerdefinierten Standardabweichung zugewiesen. *Beispiel:* Normal(100).
UNIFORM(max). *Numerisch.* Jedem Fall wird eine Pseudo-Zufallszahl aus einer Gleichverteilung mit dem Minimum Null und einem nutzerdefinierten Maximum zugewiesen. *Beispiel:* UNIFORM(200).

Drei weitere Arten von Funktionen stehen zur Verfügung, die im folgenden nur knapp dargestellt werden können. Genaue Auskunft geben in englischer Sprache der Syntax Reference Guide und das Hilfesystem (Index-Stichwort „Funktionen"), die sämtliche Funktionsdefinitionen enthalten.

④ *Funktionen für fehlende Werte.*

Man kann damit festlegen, dass die fehlenden Werte ignoriert werden sollen oder auch gerade nutzerdefinierte Missing-Werte oder System-Missing-Werte heraussuchen. Schließlich kann man über eine Argumentliste (Variablenliste) die Zahl der fehlenden Werte oder der gültigen Werte auszählen.

VALUE(variable). *Numerisch oder String.* Überträgt die Werte einer Variablen in eine neue und löscht dabei die Definition fehlender Werte. Wurde z.B. 5 = Son-

stige in der alten Variable als fehlender Wert behandelt, ist das in der neuen Variablen ein gültiger Wert.
MISSING(variable). *Logisch.* Setzt die fehlenden Werte der Argumentvariablen 1, alle anderen 0. *Beispiel:* War 0 als fehlender Wert deklariert, erhalten Fälle mit 0 eine 1, alle anderen eine 0.
SYSMIS(numvar). *Logisch.* Setzt für die System-Missing-Werte der Argumentvariablen eine 1, für alle anderen eine 0. Geht nur, wenn die Argumentvariable numerisch ist.
NMISS(variable,...). *Numerisch.* Zählt aus, wievielmal ein fehlender Werte in den Argumentvariablen auftritt. Minimal kann eine Variable als Argument benutzt werden, sinnvoll ist der Einsatz allerdings nur bei mehreren Argumenten. Sind z.B. drei Variablen als Argumente eingesetzt, so können 0, 1, 2, oder 3 mal Missing-Werte auftreten.
NVALID(variable,...). *Numerisch.* Zählt umgekehrt aus, wieviel Argumentvariablen einen gültigen Wert haben.

⑤ *Datums- und Zeitaggregationsfunktionen.*

Dienen dazu, in unterschiedlichen Variablen gespeicherte Datumsangaben in einer Datumsvariablen zusammenzufassen.

DATE.DMY(tag,monat,jahr). *Numerisch im SPSS-Datumsformat.* Wenn Tag, Monat und Jahr in drei verschiedenen Variablen als Integerzahlen gespeichert sind, kann man sie damit in eine neue Variable mit Datumsformat überführen. Die Jahreszahl muss größer als 1528 und vierstellig oder zweistellig angegeben sein. Bei zweistelliger Angabe wird 19 ergänzt. *Beispiel:* Vom Geburtsdatum sind der Tag in der Variablen GBTAG, der Monat in der Variablen GBMONAT und das Jahr in der Variablen GBJAHR gespeichert. In einer Variablen fasst man sie zusammen mit dem Befehl DATE.MDY(GBTAG,GBMONAT,GBJAHR).
DATE.MDY(monat,tag,jahr). *Numerisch im SPSS-Datumsformat.* Wie vorher, jedoch mit anderer Reihenfolge der Eingabe von Monat, Tag und Jahr. Die neue Variable muss ebenfalls in das gewünschte Datumsformat umdefiniert werden.
DATE.MOYR(monat,jahr). *Numerisch im SPSS-Datumsformat.* Dasselbe, allerdings ohne Tagesangabe.
DATE.QYR(quartal,jahr). *Numerisch im SPSS-Datumsformat.* Gleiche Voraussetzungen wie bei den vorherigen Formaten. Eingegeben werden jedoch Quartal und Jahr.
DATE.WKYR(wochenum,jahr). *Numerisch im SPSS-Datumsformat.* Ebenso, jedoch Eingabe einer Wochennummer zwischen 1 und 52 und einer Jahreszahl.
DATE.YRDAY(jahr,tagnum). *Numerisch im SPSS-Datumsformat.* Ebenso, jedoch Eingabe einer Jahreszahl und einer Tagesnummer zwischen 1 und 366.

Die nächsten Funktionen dienen dazu, auf verschiedene Variablen verteilte Zeitangaben zusammenzufassen.

TIME.HMS(std,min,sek). *Numerisch im SPSS-Zeitintervall-Format.* Wenn von einer Zeitangabe Stunden, Minuten und Sekundenangaben in verschiedenen Variablen als Integerzahlen gespeichert sind, können sie in einer Zeitvariablen zusammengefasst werden. Die Variable mit den Sekundenangaben kann auch Sekunden-

bruchteile als Nachkommastellen enthalten. *Beispiel:* Die Variable STUNDE enthält die Stunden-, die Variable MINUTE die Minuten- und SEKUNDE die Sekundenangabe. Die Zusammenfassung erfolgt mit TIME.HMS(STUNDE,MINUTE, SEKUNDE). Bei Ausgabe in eine neue Variable muss diese in ein passendes Zeitformat umdefiniert werden.
TIME.HMS(std,min). Ebenso, jedoch werden nur Stunden und Minuten eingegeben.
TIME.HMS(std). Ebenso, jedoch werden nur Stunden angegeben.
TIME.DAYS(Tage). *Numerisch im SPSS-Zeitintervall-Format.* Eine Tagesangabe wird in ein Zeitintervall umgerechnet. Die neue Variable muss in ein passendes Zeitformat umdefiniert werden.

Die Umwandlung einer einzigen Angabe wie einer Stunden- oder Tagesangabe kann aus verschiedenen Gründen sinnvoll sein. So kann das neue Format genauere Angaben, wie die Angabe von Sekundenbruchteilen zulassen. Außerdem werden Zeitintervalle angegeben. Tagesangaben werden etwa in Stunden seit dem Monatsbeginn umgewandelt. Die Datums- und Zeitvariablen können gut zur Differenzenbildung benutzt werden, da sie intern immer von einem festen Referenzzeitpunkt aus gerechnet werden. Generell ist dies der wichtigste Vorteil der Speicherung in einer Datumsvariablen anstelle von getrennten Variablen für die Einzelangaben.

© *Datums- und Zeitkonvertierungsfunktionen.*

YRMODA(jahr,monat,tag). *Numerisch.* Aus den als Integerwerte in drei Variablen gespeicherten Datumsangaben, Jahr, Monat und Tag berechnet man die Zahl der Tage seit den 15. Oktober 1582. (Dieser Tag wird in SPSS allgemein als Referenztag verwendet.) *Beispiel:* 18.4.1945 ist 132404 Tage vom Referenzzeitpunkt entfernt.

Die folgenden Konvertierungsfunktionen sind speziell für Zeitformate vorgesehen, funktionieren aber auch mit Datumsformaten. Der Referenzzeitpunkt variiert entsprechend dem benutzten Format.

CTIME.DAYS(zeit[1]). *Numerisch.* Dazu muss eine Variable in einem SPPS-Datums- oder Zeitformat vorliegen. Dann wird die Zahl der Tage, einschließlich Bruchteilen von Tagen seit dem Referenzzeitpunkt ausgegeben. Der Referenzzeitpunkt ist je nach Art der Zeitvariablen unterschiedlich. Bei Datumsangaben ist das der 15. Oktober 1582, bei reinen Zeitangaben dagegen 0 Uhr Mitternacht usw. (⇨ Syntax Reference Guide „Date and Time in SPSS").
CTIME.HOURS(zeit). *Numerisch.* Ebenso, jedoch wird der Abstand zum Referenzzeitpunkt in Stunden (einschließlich Bruchteilen von Stunden) ausgegeben.
CTIME.MINUTES(zeit). *Numerisch.* Ebenso, jedoch wird der Abstand zum Referenzzeitpunkt in Minuten (einschließlich Minutenbruchteilen) angegeben.
CTIME.SECONDS(zeit). *Numerisch.* Ebenso, jedoch wird der Zeitabstand zum Referenzzeitpunkt in Sekunden (einschließlich Sekundenbruchteilen) angegeben.

[1] Das Argument „zeit" verlangt immer die Eingabe von Werten im SPSS-Zeitformat.

⑦ *Datums- und Zeit-Extraktionsfunktionen.*

Diese Funktionen dienen dazu, aus einer im SPSS-Datums- bzw. Zeitformat vorliegenden Variablen eine Teilinformation zu extrahieren, z.B. aus einer Variablen, die Datum, Stunden und Sekunden enthält, ausschließlich das Datum. Allen so gewonnenen Variablen muss noch durch Umdefinieren ein geeignetes Variablenformat zugewiesen werden.

XDATE.DATE(datum[2]). *Numerisch im SPSS-Datumsformat.* Aus einer SPSS-Datumsvariable werden alleine die Datumsinformationen extrahiert. *Beispiel:* 12.7.1945 22:30 wird 12.7.1945.
XDATE.HOUR(datum). *Numerisch.* Aus einer SPSS-Datumsvariablen werden alleine die Stundenangaben extrahiert. *Beispiel*: 12.7.1945 22:30 wird 22.
XDATE.JDAY(datum). *Numerisch.* Aus einer SPSS-Datumsvariablen wird ermittelt, um den wievielten Tag des Jahres es sich handelt (ergibt eine Zahl zwischen 1 und 366). *Beispiel:* Der 12.7.1945 ist der 193. Tag des Jahres.
XDATE.MDAY(datum). *Numerisch.* Es wird ermittelt, um den wievielten Tag eines Monats es sich handelt (ergibt eine ganze Zahl zwischen 1 und 31).
XDATE.MINUTE(datum). *Numerisch.* Extrahiert die Minutenangaben aus einer SPSS-Datumsvariablen (ergibt eine ganze Zahl zwischen 0 und 59).
XDATE.MONTH(datum). *Numerisch.* Extrahiert die Monatszahl aus einer SPSS-Datumsvariable (gibt eine ganze Zahl zwischen 1 und 12).
XDATE.QUARTER(datum). *Numerisch.* Bestimmt aus einer SPSS-Datumsvariablen, um welches Quartal im Jahr es sich handelt und gibt den Wert (eine ganze Zahl zwischen 1 und 4) aus. Die Ausgangsvariable muss selbst in ihrem Format keine Quartalsangabe enthalten.
XDATE.SECOND(datum). *Numerisch.* Extrahiert die Sekunden aus einer SPSS-Datumsvariablen (eine Zahl zwischen 0 und 59).
XDATE.TDAY(zeit). *Numerisch.* Rechnet eine Zeitangabe in ganze Tage um (ergibt eine ganze Zahl). Bei Anwendung auf Datumsangaben ergibt sich die Zahl der Tage seit 15. Okt. 1582.
XDATE.TIME(datum). *Numerisch.* Extrahiert die Tageszeit aus einer SPSS-Datumsvariablen und gibt sie als Sekunden seit Mitternacht aus. Eine so kreierte Variable muss erst in ein adäquates Datumsformat umdefiniert werden.
XDATE.WEEK(datum). *Numerisch.* Ermittelt aus einer SPSS-Datumsvariablen, um die wievielte Woche des Jahres es sich handelt (gibt eine ganze Zahl zwischen 1 und 53 aus).
XDATE.WKDAY(datum). *Numerisch.* Extrahiert aus einer SPSS-Datumsvariablen den Wochentag (gibt eine ganze Zahl zwischen 1 und 7 aus).
XDATE.YEAR(datum). *Numerisch.* Extrahiert die Jahreszahl aus einer SPSS-Datumsvariablen (gibt sie als vierstellige ganze Zahl aus).

⑧ *Cross-Case Funktionen.* (Sind nur für Zeitreihen sinnvoll.)

LAG(variable). *Numerisch oder String.* Diese Funktion verschiebt die Werte für die Fälle einer Variablen. Der erste Fall bekommt einen System-Missing-Wert, je-

[2] Das Argument „datum" verlangt immer die Eingabe von Werten im SPSS-Datumsformat.

der weitere Fall jeweils den Wert seines Vorgängers. Diese Funktion ist wichtig für die Analyse von Zeitreihen, z.B. zur Berechnung von Wachstumsraten. Kann auch auf Stringvariablen angewendet werden.
LAG(variable,n). *Numerisch oder String.* Hat dieselbe Funktion. Die Anzahl der Fälle (n) bestimmt aber, wie weit die Werte verschoben werden. *Beispiel:* LAG(NR,2) bewirkt, dass die ersten beiden Fälle einen System-Missing-Wert erhalten, der dritte bekommt den Wert des ersten, der vierte des zweiten usw.

⑨ *Wahrscheinlichkeits- und Verteilungsfunktionen.*

Im Prinzip lassen sich Wahrscheinlichkeitsverteilungen durch zwei auseinander ableitbaren Typen von Funktionen beschreiben[3]:

- *Wahrscheinlichkeitsfunktion, Wahrscheinlichkeitsdichte.* Diese Funktion gibt bei diskreten Verteilungen an, wie wahrscheinlich bei gegebener Verteilungsform mit gegebenen Parametern das Auftreten eines bestimmten diskreten Ergebnisses q (gebräuchlicher ist die Symbolisierung als x) ist. Bei kontinuierlichen Vereilungen lässt sich die Wahrscheinlichkeit p für das Auftreten eines konkreten Wertes nicht sinnvoll bestimmen. An dessen Stelle tritt die Wahrscheinlichkeitsdichte, das heißt der Grenzwert der Wahrscheinlichkeit eines Intervalls an dieser Stelle x mit Intervallbreite nahe Null.
- *Verteilungsfunktion.* Diese Funktion gibt die kumulierte Wahrscheinlichkeit dafür an, dass ein Ergebnis < einem bestimmten Wert q eintritt.

In beiden Fällen lässt sich die Betrachtung auch umkehren und für eine gegebene Wahrscheinlichkeit p der dazugehörige Wert q ermitteln.

Mit wenigen Ausnahmen (z.B. Bernoulli) bestimmt die Funktion die Grundform einer Schar von Verteilungen, deren genaue Form durch die variablen Parameter bestimmt wird. So haben z.B. alle Normalverteilungen die charakteristische Glockenform. Die Parameter μ und Sdtv. bestimmen aber, bei welchem Wert das Zentrum der Verteilung liegt und wie breit sie verläuft Abb. 5.4.a stellt die Wahrscheinlichkeitsfunktion einer Normalverteilung mit $\mu = 2000$ und Stdv. = 500 dar, Abb. 5.4.b. deren Verteilungsfunktion..

SPSS bietet im Grunde um *vier Funktionen* an, die allerdings mit bis zu 20 Verteilungen kombiniert werden können. Die Auswahlliste enthält jede dieser Kombinationen gesondert. Daher nehmen in ihr die Verteilungsfunktionen einen sehr breiten Raum ein. An dieser Stelle kann nicht jede Kombination, sondern nur das Aufbauprinzip erklärt werden. Sie können dies Beispiele anhand der Daten von ALLBUS90.SAV nachvollziehen.

Die ersten beiden beziehen sich auf die Wahrscheinlichkeits- bzw. Dichtefunktion.

---

[3] Die Verteilungsfunktionen stellen Beziehungen zwischen konkreten Ergebnissen q und deren Wahrscheinlichkeit p beim Vorliegen einer bestimmten Verteilungsform mit gegebenen Spezifikationsparametern her. In SPSS werden die konkreten Ergebnisse z.T. als q (in den Auswahllisten), z.T. als x (in der Kontexthilfe) bezeichnet. Auch die Bezeichnung der Parameter variiert. Wir bezeichnen im folgenden das Ergebnis mit q.

- *RV-Funktionen.* Sie erzeugen für jeden Fall einen Zufallswert aus der angegebenen Verteilung. B.: RV.NORMAL(2096,1134) weist den einzelnen Fällen Zufallszahlen aus einer Normalverteilung mit dem Mittelwert 2096 und der Standardabweichung 1134 zu. Die Wahrscheinlichkeit, eines Wertes zugewiesen zu werden, hängt von der Wahrscheinlichkeitsdichte an der entsprechenden Stelle der Verteilung ab.
- *PDF-Funktion.* Gibt die Wahrscheinlichkeit bzw. die Wahrscheinlichkeitsdichte für einen bestimmten Wert q aus. *B:* PDF.NORMAL(eink,2096,1134). Gibt bei jedem Fall für seinen konkreten Wert q in der Variablen EINK die Wahrscheinlichkeitsdichte aus, wenn man davon ausgeht, dass die Daten normalverteilt sind mit dem Mittelwert 2096 und der Standardabweichung 1134. Das wäre z.B. bei einem Fall, der ein Einkommen von 2096 DM aufweist 0,000352.

Die beiden anderen beziehen sich auf die Verteilungsfunktion.

- *CDF-Funktion.* Gibt die kumulierte Wahrscheinlichkeit dafür an, dass ein Ergebnis < einem bestimmten Wert q eintritt.
  B.: CDF.NORMAL(eink,2096,1134). Gibt bei jedem Fall für seinen konkreten Wert q in der Variablen EINK die Wahrscheinlichkeitsdichte aus, wenn man davon ausgeht, dass die Daten normalverteilt sind mit dem Mittelwert 2096 und der Standardabweichung 1134. Das wäre z.B. für einen Fall, der das Einkommen 2096 hat, 0,5.
- *IDF-Funktion.* Gibt umgekehrt für eine Wahrscheinlichkeit p den Wert q aus, unterhalb dessen die kumulierte Wahrscheinlichkeit bei gegebener Verteilung p beträgt. *B.:* IDF.NORMAL(einkcdf,2096,1134). Gibt bei jedem Fall für seine konkret angegebene kumulierte Wahrscheinlichkeit p den Wert x, unter dem bei der konkreten Verteilung die Wahrscheinlichkeiten auf p kumulieren würden. Im Beispiel stammen die Wahrscheinlichkeiten aus einer Variablen EINKCDF, die mit der CDF-Funktion gebildet wurde. Die Rückrechnung führt wieder zu den Ausgangswerten, die in der Variablen EINK stehen. Ein p-Wert von 0,5 führt z.B. zu einem Wert q = 2096.

Darüber hinaus gibt es für einige Funktionen (BETA, CHISQ, F und T, Varianten für nicht zentrale Verteilungen.

- *NPDF-Funktion.* Gibt wie eine PDF-Funktion die Wahrscheinlichkeit bzw. die Wahrscheinlichkeitsdichte für einen bestimmten Wert q aus. Zu den Parametern dieser Verteilungen aus den PDF-Funktionen kommt jeweils ein Parameter nz für die Nichtzentralität hinzu. Die Nichtzentralität bezieht sich auf die Stelle q. Ein nz von 0 ergibt eine zentrale Verteilung. Mit steigendem nz verschiebt sich die Mitte der Verteilung nach rechts. *Beispiel:* Einem Chi-Quadrat Wert von 3,84 entspricht bei df = 1 in einer zentralen Verteilung (PDF.CHISQ(3.84,1)) eine Wahrscheinlichkeitsdichte von 0,298. In einer nicht zentralen Verteilung mit nz = 1 (NPDF.CHISQ(3.84,1,1)) einer Wahrscheinlichkeitsdichte von 0,665.
- *NCDF-Funktion.* Gibt wie eine CDF-Funktion die kumulierte Wahrscheinlichkeit dafür an, dass ein Ergebnis < einem bestimmten Wert q eintritt. Zu den Pa-

rametern dieser Verteilungen aus den CDF-Funktionen kommt jeweils ein Parameter nz für die Dezentralität hinzu. Dieser muss jeweils größer gleich 0 und kleiner sein als q. *Beispiel:* Bei einer zentralen Chi-Quadrat Verteilung mit df = 1 (CDF.CHISQ(3.84,1)) beträgt die kumulierte Wahrscheinlichkeit aller Wert > 3,84 0,95. Bei einer dezentralen mit nz = 1 (NCDF.CHISQ(3.84,1,1)) dagegen 0,83.

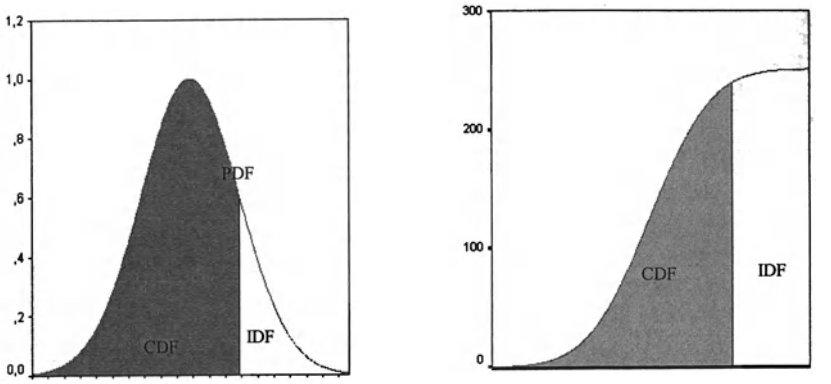


Abb. 5.4.a und b. Wahrscheinlichkeitsdichte- und Verteilungsfunktion einer Normalverteilung

In Abbildung 5.4.a. entspricht die Höhe der Säule über einem minimalen Bereich bei q der Wahrscheinlichkeit p dieses Wertes (genauer Wahrscheinlichkeitsdichte eines Minimalintervalls um diesen Wert) (Ergebnis von PDF). Die (grau eingefärbte) Fläche zwischen dem Minimalwert - ∞ und q gibt die kumulierte Wahrscheinlichkeit aller Werte ≤ q (CDF), umgekehrt die (weiße) Fläche von dort bis + ∞ die kumulierte Wahrscheinlichkeit aller Werte >q (ICDF). In der Verteilungsfunktion entspricht die linke (grau eingefärbte) Fläche zwischen dem Minimalwert - ∞ und q der kumulierten Wahrscheinlichkeit aller Werte ≤ q (CDF), die (weiße) Fläche von dort bis + ∞ die kumulierte Wahrscheinlichkeit aller Werte >q (ICDF).

Verfügbare Verteilungen. Die verfügbaren Verteilungen sind in Tabelle 5.1 zusammengestellt. In ihr sind die Parameter (Kennwerte) für alle drei Verteilungsarten angegeben. Sie müssen daraus die für die jeweilige Funktion geltenden zusammenstellen (⇨ Anmerkungen zur Tabelle). Es sind auch die Bereichsgrenzen für die jeweiligen Spezifikationsparameter angegeben. *Beispiel:* Aus der Tabelle entnehmen Sie für die Chi-Quadrat-Verteilung die Spezifikationen q, p und df; dabei ist df = Zahl der Freiheitsgrad der für die Chi-Quadrat-Funktion charakteristische einzige Parameter; q dagegen steht für das jeweilige konkrete Ergebnis q und p für dessen Wahrscheinlichkeit bzw. die kumulierte Wahrscheinlichkeit aller Werte kleiner q (diese Spezifikationen kommen bei allen Verteilungen in der entsprechenden Form vor). Für die RV-Funktion ist ausschließlich der charakteristischer Parameter der Chi-Quadrat-Verteilung die Zahl der Freiheitsgrade df relevant, denn es sollen Zufallszahlen aus der so charakterisierten Verteilung generiert

werden. Die charakteristischen Verteilungsparameter sind selbstverständlich bei allen Funktionstypen relevant. Die anderen Spezifikationsparameter aber wechseln. Für eine PDF-Funktion ist neben df auch q relevant, denn für bestimmte Werte q soll die Wahrscheinlichkeitsdichte p ermittelt werden. Dasselbe gilt für die CDF-Funktion. Hier soll die kumulierte Wahrscheinlichkeit p für alle Werte kleiner q ermittelt werden. Die IDF-Funktion dagegen verlangt die Angabe von p, der kumulierten Wahrscheinlichkeit, für die dann der Wert q ermittelt werden soll unter dem die Werte liegen, deren Wahrscheinlichkeit zusammen p ergibt.

Beachten Sie weiter: Alle Funktionen geben *numerische Werte* aus. RV.Verteilungen führen zur Zuweisung von Pseudozufallszahlen. Wie bei der Verwendung des Zufallszahlengenerators generell, sollte man, wenn man die neu gebildete Verteilung reproduzieren möchte, zunächst immer unter „Transformieren", „Startwert für Zufallszahlen..." den Startwert setzten, von der die Auswahl ausgehen soll.

Die Funktion CDFNORM(zWert), die unter anderen Funktionen aufgeführt ist, erfüllt für die Standardnormalverteilung dieselbe Aufgabe wie die CDF.Normal, setzt aber voraus, dass z-transformierte Daten als q-Werte eingesetzt werden

Tabelle 5.1. Verfügbare Verteilungen der Verteilungsfunktionen

Verteilung	Spezifikationen, Spezifikationsbereiche
BERNOULLI[2)]	(q, p). $0 \geq p \geq 1$. p ist die Wahrscheinlichkeit für einen Erfolg.
BETA[1)3)]	(q, p, form1, form2). $0 \geq q \geq 1$, $0 \geq p \geq 1$, form > 0.
BINOM[2)]	(q, n, p). n ist Zahl der Versuche, p die Wahrscheinlichkeit, bei einem Versuch Erfolg zu haben. q muss eine positive Integerzahl sein. $0 \leq p \leq 1$.
CAUCHY[1)]	(q, p, lage, Skala). $0 < p < 1$, skala > 0.
CHISQ (Chi-Quadrat)[1) 3)]	(q, p, df). $q \geq 0$, $0 \leq p < 1$, $df \geq 0$.
EXP (Exponential)[1)]	(q, p, form). $q \geq 0$, $0 \leq p \geq 1$, form > 0.
F[1) 3)]	(q, p, df1, df2). $q \geq 0$, $0 \leq p < 1$, df1 und df2 > 0
GAMMA[1)]	(q, p, form, skala). $q \geq 0$, $0 \leq p < 1$, form und skala > 0.
GEOM (Geometrische)[2)]	(q, p). $0 < p < 1$. q ist die Zahl der Versuche (inklusive dem letzten m), die benötigt werden, bevor ein Erfolg erzielt wird, p die Wahrscheinlichkeit für einen Erfolg in einem einzigen Versuch.
HALFNRM (Halbnormal)[1)]	(q, p, schwelle, skala), $0 < p < 1$, schwelle > 0.

HYPER (Hypergeometrische)[2]	(q, gesamt, stichpr, treffer). gesamt = Zahl der Objekte in der Grundgesamteit, stichprobe = Größe einer Zufallsstichprobe, gezogen ohne Zurücklegen, treffer = Zahl der Objekte mit der festgelegten Eigenschaft in der Grundgesamtheit (alle drei müssen positive ganze Zahlen sein), q = die Zahl der Objekte mit dieser Eigenschaft in der Stichprobe.
IGAUSS (inverse Normalverteilung)[1]	(q, p, mittel, skala) $0 < p < 1$, mittlel > 0, skala > 0.
LAPLACE (Doppelex-ponentialverteilung)[1]	(q, p, mittel, skala). $0 < p < 1$, skala > 0.
LOGISTIC (Logistische)[1]	(q, p, mittel, skala). $0 < p < 1$, skala > 0.
LNORMAL (Lognormal)[1]	(q, p, a, b). $q \geq 0$, $0 \leq p < 1$, a, b > 0.
NEGBIN (negative Binomial)[2]	(q, p, schwelle). $0 < p \leq 1$. a ist die Erfolgswahrscheinlichkeit bei einem Versuch. Schwelle ist die Zahl der Erfolge, muss eine ganze Zahl sein. q ist die Zahl der Versuche (inklusive dem letzten), bevor die durch Schwelle angegebene Zahl von Erfolgen beobachtet werden. (Beim Schwellenwert 1 identisch mit der geometrischen Verteilung.)
NORMAL[1]	(q, p, mittel, stdAbw). $0 < p < 1$, stdAbw > 0.
PARETO[1]	(q, p, schwelle, form). $q \geq$ a, $0 \leq p < 1$, schwelle, form > 0.
POISSON[2]	(q, mittel). Mittel > 0. mittel ist die Zahl der Ereignisse eines bestimmten Typs, die im Durchschnitt in einer festgelegten Zeitperiode eintreten. q muss eine positive Integerzahl sein.
SMOD (studentisiertes Maximalmodul)[1][4]	(q, p, größe, df). a und $df \geq 1$.
SRANGE (studentisierte Spannweite)[1][4]	(q, p, größe, df). größe und $df \geq 1$. q = studentisierte Spannweite, größe = Zahl der Fälle (verglichenen Stichproben).
T[1][3]	(q, p, df). $0 < p < 1$. df > 0.
UNIFORM[1]	(q, p, min, max). $0 \leq q \leq 1$, $0 \leq p \leq 1$.
WEIBULL[1]	(q, p, a, b). $q \geq 0$, $0 \leq p < 1$, a und b > 0.

1) Kontinuierliche Funktionen. Bei Verwendung von PDF, CDF, NPDF und NCDF entfällt der Parameter p, bei IDF der Parameter q, bei Verwendung von RV entfallen jeweils p und q (q = quantity, Wert, für den die Wahrscheinlichkeit gesucht wird; p = probability, Wahrscheinlichkeit, für die der Wert gesucht wird).

2) Diskrete Funktionen. Es existieren nur CDF und RV-Funktionen (a ist jeweils ein Wahrscheinlichkeitsparameter; q entfällt bei Verwendung von RV).

3) Auch als nicht-zentrale Verteilung (NCDF) verfügbar. Für diese gelten dieselben Spezifikationsparameter wie für CDF-Funktionen, jedoch ergänzt durch den Nichtzentralitätsparameter nc.

4) Nur CDF und IDF-Funktion.

Dabei bedeutet gewöhnlich: Lage = Mitte der Verteilung (arithmetisches Mittel), Skala = Streuung (Standardabweichung) oder einen Skalenparameter λ, Schwelle = Wert, ab dem die Verteilung beginnt.

⑩ *Andere Verteilungsfunktionen.*

CDFNORM(zWert). *Numerisch.* Gibt die Wahrscheinlichkeit dafür an, dass eine normalverteilte Zufallsvariable mit einem arithmetischen Mittel von 0 und einer Standardabweichung von 1 (= Standardnormalverteilung) unter dem vom Benutzer definierten z-Wert liegt. Der z-Wert kann zwischen 0 und etwa 3 variieren. (Dezimalstellen müssen mit Punkt eingegeben werden.) Diese Funktion muss auf bereits z-transformierte Variable angewendet werden. *Beispiel:* Die Variable ZEINK enthält z-transformierte Einkommenswerte. CDFNORM(zeink) gibt dann für jeden Fall die Wahrscheinlichkeit aus, mit der ein Fall einen geringeren Wert als der Fall selbst erreicht oder anders ausgedrückt, welcher Anteil der Fälle ein geringeres Einkommen besitzt. (*Voraussetzung:* Die Normalverteilungsannahme ist zutreffend.) *Beispiel:* In der Datei ALLBUS90 hat der 3. Fall ein Einkommen von 1450 DM, daraus – sowie aus dem Mittelwert und der Streuung der Verteilung – errechnet sich ein z-Wert –0,57046 und dieser ergibt wiederum die Wahrscheinlichkeit von 0,28 dafür, dass ein anderer Fall ein geringeres Einkommen besitzt.

CDF.BVNOR(q1, q2, Korr). *Numerisch.* Diese Funktion gibt die kumulative Wahrscheinlichkeit zurück, mit welcher ein Wert aus der bivariaten Standardnormalvorteilung mit dem angegebenem Parameter r = Korrelation kleiner als q1 und q2 ist.

PROBIT(p). *Numerisch.* Kehrt die Berechnung um und ermittelt den z-Wert, unter dem mit einer nutzerdefinierten Wahrscheinlichkeit (p) ein Wert der Standardnormalverteilung liegt. Die eingesetzte Wahrscheinlichkeit muss zwischen 0 und 1 liegen. Verfügt man über eine Variable, in der für Fälle jeweils angegeben ist, mit welcher Wahrscheinlichkeit ihr Wert über dem anderer Fälle liegt, so kann man mit dieser Funktion diese Wahrscheinlichkeiten in z-Werte umrechnen, sicher eine seltene Anwendungsmöglichkeit. (Dezimalstellen müssen mit Punkt eingegeben werden.) Der ausgegebene z-Wert liegt zwischen 0 und ungefähr 3.

⑩ *String-Funktionen.*

ANY(test,wert,wert,...). *Logisch.* Ergibt 1, wenn der Wert des ersten Arguments mit mindestens einem Wert der restlichen Argumente übereinstimmt. Das erste Argument ist gewöhnlich eine Variable, das kann aber auch für die anderen zutreffen. *Beispiel*: Eine String-Variable BANK1 enthält die Namen der Banken, bei denen ein erster Kredit aufgenommen ist, eine zweite Variable BANK2 enthält die Namen der Banken, bei denen ein zweiter Kredit aufgenommen wurde, BANK3 die der Banken eines dritten Kredits. Dann ergibt ANY(bank1,bank2,bank3) den Wert 1 für die Fälle, die mehr als einen Kredit bei derselben Bank aufgenommen haben. 0 ergibt sich, wenn alle Kredite bei unterschiedlichen Banken aufgenommen wurden. Die Funktion erfordert mindestens zwei Argumente.

CONCAT(strausdr,strausdr,...). *String.* Verbindet die String-Werte aller Argumente zu einem resultierenden String-Wert. *Beispiel:* In einer String-Variablen LAND stehen die Namen von Ländern, in einer anderen Stringvariablen ENTW

wird deren Entwicklungsstand (z.B. „unterentwickelt") festgehalten. Die Funktion CONCAT(Land,Entw) fasst beides in einem neuen String zusammen (etwa wird aus „Bangla Desh" und „unterentwickelt" „Bangla Desh unterentwickelt"). Diese Funktion benötigt mindestens zwei Argumente, die String-Werte sein müssen.

INDEX(heuhaufen,nadel). *Numerisch.* Diese Funktion hilft dabei, einen bestimmten Ausdruck (Nadel) in einer Stringvariablen (Heuhaufen) zu finden. *Beispiel:* in der Stringvariablen LAND sei Deutschland in unterschiedlicher Weise gespeichert, z.B. als „Bundesrepublik Deutschland" und „Deutschland". Die Funktion INDEX(„Land",Deutschland) gibt in einer neuen numerischen Variablen für jeden Fall einen numerischen Wert aus, bei dem der String „Deutschland" in der Variablen LAND an irgendeiner Stelle vorkommt. Der numerische Wert beträgt z.B. 16, wenn der gesamte String „Bundesrepublik Deutschland" lautet. Die 16 besagt, dass das gesuchte Wort Deutschland an der 16. Stelle im gefundenen String beginnt. Alle Fälle, in denen der String nicht vorkommt, erhalten den Wert 0 zugewiesen. Die Ergebnisvariable muss numerisch sein. Bei allen Heuhaufen- Nadel-Argumenten muss der Stringausdruck „Nadel" in Anführungszeichen eingegeben werden.

INDEX(heuhaufen,nadel,teiler). *Numerisch.* Wie die vorhergehende Funktion. Das Argument „Teiler" kann wahlweise verwendet werden. Mit ihm kann der String „nadel" in einzelne zu suchende Teilstrings unterteilt werden. Der ganzzahlige Wert von „teiler" muss den String „nadel" so teilen, dass kein Rest verbleibt.

LENGTH(strausdr). *Numerisch.* Ergibt die Länge des Stringausdrucks (strAusdr). Das Ergebnis ist die definierte, nicht die tatsächliche Länge. Für eine Stringvariable LAND mit 30 Zeichen Länge gibt die Funktion ohne Bezug auf den Stringwert eines Falles immer das Ergebnis 30 aus. Wenn Sie die tatsächliche Länge erhalten möchten, geben Sie den Ausdruck in folgender Form an: LENGTH(RTRIM(strausdr)). *Beispiel:* LENGTH(RTRIM(Land)) gibt bei dem genannten Beispiel für den Stringwert „Deutschland" das Ergebnis 11, für „England" das Ergebnis 7 aus.

LOWER(strausdr). *String.* Wandelt alle Großbuchstaben in „strausdr" in Kleinbuchstaben um. Alle anderen Zeichen werden nicht verändert.

LPAD(strausdr,länge). *String.* Ergibt einen String, in dem der Ausdruck, der in „strausdr" enthalten ist (kann eine Variable sein) von links mit Leerzeichen aufgefüllt wird, bis die im Argument „länge" angegebene Gesamtlänge erreicht ist. „Länge" muss eine positive ganze Zahl zwischen 1 und 255 sein. *Beispiel:* LPAD(land,50) macht aus der bisher 30stelligen Stringvariablen LAND eine 50stellige und speichert die Stringwerte nicht mehr linksbündig, sondern (allerdings links beginnend in der gleichen Spalte, orientiert am längsten Wert) am rechten Rand der Variablen.

LPAD(strausdr,länge,zeichen). *String.* Ergibt einen String, in dem der in „strausdr" enthaltene String (kann auch eine Variable sein) von links mit dem im Argument „zeichen" angegebenen Zeichen aufgefüllt wird, bis die im Argument „länge" angegebene Gesamtlänge erreicht ist. Länge muss eine positive ganze Zahl zwischen 1 und 255 sein. „Zeichen" kann ein einzelnes Zeichen, eingeschlossen in Anführungszeichen, sein oder das Ergebnis einer String-Funktion, die ein einzelnes

Zeichen liefert. *Beispiel:* LPAD(land,50,"x"). Erbringt dasselbe Ergebnis wie oben. Aufgefüllt werden aber nicht Leerzeichen, sondern x-Zeichen.

LTRIM(strausdr). *String.* Entfernt vom im Argument „strausdr" enthaltenen String alle führenden Leerzeichen. Beispiel: ' Deutschland' ergibt 'Deutschland'.

LTRIM(strausdr,zeichen). *String.* Entfernt vom im Argument „strausdr" enthaltenen String die im Argument „zeichen" angeführten führenden Zeichen. Zeichen kann ein einzelnes Zeichen, eingeschlossen in Anführungszeichen, sein oder das Ergebnis einer String-Funktion, die ein einzelnes Zeichen liefert. *Beispiel:* LTRIM(Land2,"x"). Aus 'xxxDeutschland' wird 'Deutschland'.

Beachten Sie: Immer wenn in Stringvariablen ausgegeben wird, muss, bevor der Befehl abgeschickt wird, in der nach Anklicken der Schaltfläche „Typ und Label..." geöffneten Dialogbox „Variable berechnen: Typ und Label" in der Gruppe „Typ" die Option „String" gewählt und eine Stringbreite eingetragen werden.

NUMBER(strausdr,format). *Numerisch.* Diese Funktion benutzt man, um in einer Stringvariablen gespeicherte Zahlenangaben (!) in einen numerischen Ausdruck zu verwandeln. Das kann insbesondere bei Übernahme von Zahlen aus Fremdprogrammen interessant sein. Numerische Variablen sind in SPSS besser zu verarbeiten. Das Argument „format" gibt das Einleseformat des numerischen Ausdrucks an. Die Formatangabe erfolgt entsprechend den dazu vorgesehenen Syntaxbefehlen. Festes Format der Breite 8 z.B. wird mit f8 eingegeben. Das bedeutet Folgendes: Wenn der Stringausdruck acht Zeichen lang ist, entspricht NUMBER(Strausdr, f8) dem numerischen Einleseformat. Wenn der Stringausdruck nicht mit dem angegebenen Format eingelesen werden kann oder wenn er nicht interpretierbare Zeichen enthält, ist das Ergebnis ein System-Missing-Wert. Wenn die angegebene Länge n des numerischen Formats kleiner als die Länge des Stringausdrucks ist, werden nur die ersten n Zeichen zur Umwandlung herangezogen. (*Beachten Sie:* Sie müssen der Ergebnisvariablen vorher ein numerisches Format geben. Es können so auch Zahlen mit Kommastellen eingelesen werden. Das Einleseformat gibt die Kommastellen nicht an, wenn diese explizit sind. Das Ergebnisformat bildet sie automatisch. Bei impliziten Kommastellen dagegen müssten Sie im Einleseformat angegeben werden. Evtl. muss die Ergebnisvariable noch in geeigneter Form umdefiniert werden.)*Beispiel.* Eine Stringvariable EINKSTR, Breite 15, enthält Einkommensdaten. Sie sollen in eine numerische Variable überführt werden. Ein Stringwert '3120,55 ' wird dann durch NUMBER(einkstr, f15) in einen numerischen Wert 3120,55 überführt, der wie alle numerischen Werte auch rechtsbündig gespeichert ist.

RINDEX(heuhaufen,nadel). *Numerisch.* Diese Funktion hilft wie die Funktion Index dabei, einen bestimmten Ausdruck (nadel) in einer Stringvariablen (heuhaufen) zu finden. Sie ergibt einen ganzzahligen Wert für das letzte Auftreten des Strings „nadel" im String „heuhaufen". Das ganzzahlige Ergebnis gibt die Stelle des ersten Zeichens von „nadel" in „heuhaufen" an. Ergibt 0, wenn „nadel" nicht in „heuhaufen" vorkommt. Die Funktion RINDEX(„Land",Deutschland) gibt z.B. in einer neuen numerischen Variablen für jeden Fall einen numerischen Wert aus, bei dem der String „Deutschland" in der Variablen LAND vorkommt. Kommt Deutschland mehrmals vor, dann wird die Stelle des ersten Buchstabens vom letzten Vorkommen zum numerischen Wert. „Lieferland Deutschland Empfangsland

Deutschland" z.B. ergibt den Wert 37, weil das letzte Auftreten von Deutschland" an Position 37 beginnt.

RINDEX(heuhaufen,nadel,teiler). *Numerisch.* Wie die vorhergehende Funktion. Das optionale dritte Argument „teiler" wird von SPSS verwendet, um den String „nadel" in einzeln zu suchende Teilstrings zu unterteilen. Der ganzzahlige Wert von „teiler" muss den String „nadel" so teilen, dass kein Rest verbleibt.

RPAD(strausdr,länge). *String.* Ergibt einen String, in dem der im Ausdruck „strausdr" enthaltene String (kann eine Variable sein) von rechts mit Leerzeichen aufgefüllt wird, bis die im Argument „länge" angegebene Gesamtlänge erreicht ist. Länge muss eine positive ganze Zahl zwischen 1 und 255 sein. Zeichen kann ein einzelnes Zeichen (eingeschlossen in Anführungszeichen) sein oder das Ergebnis einer String-Funktion, die ein einzelnes Zeichen liefert.

RPAD(strausdr,länge,zeichen). *String.* Identisch mit der vorherigen Funktion, jedoch wird der String von rechts mit dem im Argument „zeichen" angegebenen Zeichen aufgefüllt.

RTRIM(strausdr). *String.* Entfernt vom im Argument „strausdr" angegebenen String (kann auch eine Variable sein) alle nachstehenden Leerzeichen.

RTRIM(strausdr,zeichen). *String.* Entfernt vom im Argument „strausdr" angegebenen String (kann auch eine Variable sein) jedes Vorkommen des im Argument definierten Zeichens als nachstehendes Zeichen.

STRING(numausdr,format). *String.* Wandelt einen numerischen Ausdruck in einen String um. Das Argument „format" muss ein gültiges numerisches Darstellungsformat sein. *Beispiel:* STRING(eink,F10.2) ergibt für einen in der numerischen Variablen EINK gespeicherten Wert 1250,55 den Stringausdruck '1250,55'.

SUBSTR(strausdr,pos). *String.* Liefert einen Teil des Ausdrucks im Argument „strausdr" ab der Stelle „pos" bis zum Ende von „strausdr". *Beispiel:* SUBSTR(land,8) ergibt für den String „Deutschland" in der Variablen LAND den neuen Wert „land".

SUBSTR(strausdr,Pos,länge). *String.* Liefert den Teil des Ausdrucks „strausdr", der an der Stelle „pos" beginnt und die im Argument „länge" angegebene Länge hat. *Beispiel:* SUBSTR(land,8,3) ergäbe anstelle von „Deutschland" „lan".

UPCAS(strausdr). *String.* Wandelt alle Kleinbuchstaben des im Argument **„strausdr"** enthaltenen Strings in Großbuchstaben um. Alle anderen Zeichen werden nicht verändert.

5.2 Verwenden von Bedingungsausdrücken

Es kommt häufig vor, dass man in Abhängigkeit von bestimmten Bedingungen Fällen unterschiedliche Werte zuweisen muss. Das ist z.B. der Fall, wenn man eine Gewichtungsvariable konstruiert. In Kap. 2.7 wurde z.B. für unsere Datei mit Hilfe der Gewichtungsvariablen GEWICHT Männern das Gewicht 0,84, Frauen dagegen das Gewicht 1,21 zugewiesen. Der zugewiesene Wert kann auch das Ergebnis einer z.T. umfangreichen Berechnung sein. Auch die Zuweisung von Werten in Abhängigkeit von logischen Bedingungen ist durch Verwendung von Bedingungsausdrücken möglich.

Beispiel. Die Daten aus der Schuldnerberatung sollen daraufhin ausgewertet werden, welche Zahlungspläne den Gläubigern angeboten werden können. Die Informationen dazu finden sich in einer Datei KLIENTEN.SAV. Es soll der monatlich zur Zinszahlung und Tilgung eingesetzte Betrag ermittelt und in der Variablen MON_ZAHL (monatlicher Zahlungsbetrag) gespeichert werden. Zur Ermittlung von MON_ZAHL stehen die Angaben in EINK (monatliches Einkommen), SOZBED (Sozialhilfebedarf = Pfändungsfreibetrag) und MON_FORD (monatliche Forderung) zur Verfügung.

Der monatliche Zahlungsbetrag ist je nach gegebenen Bedingungen unterschiedlich zu ermitteln. Unterschieden werden drei Fallgruppen:

- ❒ Das Einkommen ist geringer oder gleich dem Pfändungsfreibetrag (EINK – SOZBED ≤ 0). Dann steht überhaupt kein Geld für einen Zahlungsplan zur Verfügung. Die Zielvariable MON_ZAHL bekommt den Wert 0.
- ❒ Das Einkommen ist größer als der Pfändungsfreibetrag (EINK – SOZBED > 0). Es steht also ein Betrag zur Zahlung zur Verfügung. Allerdings sind hier zwei Fälle zu unterscheiden:
  - ● Der verfügbare Betrag ist kleiner oder gleich den monatlichen Forderungen (EINK – SOZBED ≤ MON_FORD). Dann muss der gesamte Betrag (EINK – SOZBED) für die Zahlung eingesetzt werden.
  - ● Der verfügbare Betrag ist größer als die monatlichen Forderungen (EINK – SOZBED > MON_FORD). Dann wird nur der zur Begleichung der Forderungen erforderliche Betrag (MON_FORD) eingesetzt.

Für die genannten drei Fallgruppen wird der monatliche Zahlungsbetrag auf Basis eines entsprechenden Bedingungsausdrucks getrennt ermittelt und in die Variable MON_ZAHL eingelesen. Um den Wert für die erste Fallgruppe zu berechnen, gehen Sie wie folgt vor:

- ▷ Wählen Sie „Transformieren", „Berechnen...". Es öffnet sich die Dialogbox „Variable berechnen" (⇨ Abb. 5.1).
- ▷ Geben Sie in das Eingabefeld „Zielvariable" den neuen Variablennamen (MON_ZAHL) ein. Und tragen Sie in das Eingabefeld „Numerischer Ausdruck:" den Wert 0 ein. Demnach wird der neuen Variablen der Wert 0 gegeben, wenn die jetzt zu formulierende Bedingung gilt.
- ▷ Um die Bedingung zu formulieren, klicken Sie auf die Schaltfläche „Falls...". Es erscheint die in Abb. 5.5 dargestellte Dialogbox „Variable berechnen: Falls Bedingung erfüllt ist".
- ▷ Klicken Sie auf den Optionsschalter „Fall einschließen, wenn Bedingung erfüllt ist".
- ▷ Formulieren Sie im Eingabefeld die Bedingung. Dazu können Sie die Variablen, die im Rechnerbereich angegebenen logischen und arithmetischen Zeichen, Klammern und die Funktionen in der bereits oben kennengelernten Form verwenden. Das Eingabefeld der Dialogbox muss dann wie in Abb. 5.5. aussehen.
- ▷ Bestätigen Sie mit „Weiter" (der Bedingungsausdruck wird zur Information jetzt auch in der Box „Variable berechnen" angezeigt) und „OK". Der Wert 0 wird in der Datenmatrix bei den zutreffenden Fällen eingesetzt. Die anderen Fälle erhalten einen System-Missing-Wert zugewiesen.

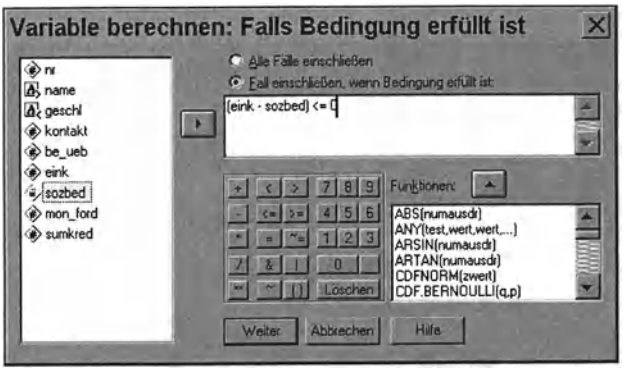


Abb. 5.5. Dialogbox „Variable berechnen: Falls Bedingung erfüllt ist"

Dieselbe Prozedur muss jetzt für die beiden anderen logischen Bedingungen wiederholt werden.

Für die zweite Fallgruppe:

- ▷ Setzen Sie in der Dialogbox „Variable berechnen" denselben Variablennamen, aber anstelle von 0 in „Numerischer Ausdruck:" (EINK – SOZBED) ein. Dieser berechnet den später für die zutreffenden Fälle einzutragenden Betrag.
- ▷ Öffnen Sie mit „Falls..." die Dialogbox „Variable berechnen: Falls Bedingung erfüllt ist". Wählen Sie die Option „Fall einschließen, wenn Bedingung erfüllt ist".
- ▷ Geben Sie die komplexere Bedingung ein: ((EINK – SOZBED) > 0) & ((EINK – SOZBED) < MON_FORD). Bestätigen Sie mit „Weiter" und „OK". Da neue Werte in eine schon bestehende Variable eingetragen werden sollen, erscheint eine Warnmeldung.
- ▷ Bestätigen Sie mit „OK", dass Sie eine bestehende Variable verändern wollen. Die neuen Werte werden errechnet und bei den zutreffenden Fällen eingetragen.

Für die dritte Fallgruppe wiederholt sich der gesamte Prozess. Allerdings muss als „Numerischer Ausdruck:" MON_FORD eingesetzt und als Bedingung: ((EINK – SOZBED)>0) & ((EINK – SOZBED) >= MON_FORD).

Ergänzungen zur Berechnung einer neuen Variablen. Alle neu berechneten Variablen sind per Voreinstellung numerisch. Soll eine Stringvariable berechnet werden, geschieht das über die Dialogbox „Variable berechnen: Typ und Label" (⇨ Abb. 5.2). Diese öffnen Sie in der Dialogbox „Variable berechnen" durch Anklicken der Schaltfläche „Typ und Label" (⇨ Abb. 5.1).

Hier kann der Typ in „String" geändert werden. Die voreingestellte Stringlänge acht ist gegebenenfalls abzuändern. (Im jetzt unbenannten Eingabefeld „String" der Dialogbox „Variable berechnen" werden die Stringwerte festegelegt.) Für jede Art von Variablen kann zudem ein Variablenlabel vergeben werden. Dieses wird im Feld „Label" eingegeben. Man kann aber auch den zur Berechnung der Variablen verwendeten Ausdruck durch Auswahl der Option „Ausdruck als Label verwenden" zur Etikettierung nutzen. Sollen andere Variablentypen verwendet werden,

muss man den Typ der Variablen nachträglich durch Umdefinition im Dateneditor generieren.

Bei der Bildung von Ausdrücken ist weiter zu beachten: Das Dezimalzeichen in Ausdrücken muss immer ein Punkt sein; Werte von Stringvariablen müssen in doppelte oder einfache Anführungszeichen gesetzt werden, enthält der String selbst Anführungszeichen, benutzt man einfache (*Beispiel:* 'BR „Deutschland"'). Argumentlisten sind in Klammern einzuschließen; Argumente müssen durch Kommata getrennt werden. Argumente in Bedingungsausdrücken müssen vollständig sein, d.h. insbesondere, dass bei mehrfacher Verwendung einer Variablen der Variablennamen wiederholt werden muss (*Beispiel:* EINK>0 & EINK<1000; nicht: EINK>0 & <1000).

Startwert für Zufallszahlen. Innerhalb bestimmter Funktionen (*Beispiel:* Normal, Uniform) werden Zufallszahlen verwendet. Diese führen zu wechselnden Ergebnissen. Will man das vermeiden, setzt man mit der Befehlsfolge „Transformieren", „Startwert für Zufallszahlen" einen festen Startwert ein (⇨ Kap. 7.4.2).

5.3 Umkodieren von Werten

Wohl am häufigsten werden Daten durch Umkodieren modifiziert. Man benutzt diese Möglichkeit zur Zusammenfassung von Kategorien bzw. Bildung von Klassen bei metrischen Daten. Bisweilen wird man dadurch auch eine ungeeignete Reihenfolge der Kategorien ändern. Beim Umkodieren kann man entweder die Werte einer bestehenden Variablen verändern oder eine neue Variable mit den veränderten Werten erzeugen. In den meisten Fällen ist es ratsam, eine neue Variable zu erzeugen, um die Ausgangsdaten nicht zu verlieren. Um eine Umkodierung vorzunehmen, gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Transformieren", „Umkodieren ▷". Es öffnet sich eine Auswahlliste.
- ▷ Wählen Sie durch Anklicken einer der Optionen „In dieselben Variablen..." oder „In andere Variablen...", ob sie die veränderten Daten in dieselbe oder eine neue Variable eingeben möchten. Je nach ihrer Wahl öffnet sich entweder die Dialogbox „Umkodieren in dieselben Variablen" oder „Umkodieren in andere Variablen" (⇨ Abb. 2.17).

Die beiden Dialogboxen unterscheiden sich dadurch, dass in der ersten lediglich der Name der umzukodierenden Variablen zu wählen ist, in der zweiten muss natürlich zusätzlich ein Name für die neue Variable in der Gruppe „Ausgabevariable" eingesetzt und mit „Ändern" bestätigt werden. Das Feld „Eingabevar. → Ausgabevar.:" ändert die Überschrift in „Numerische Var. → Ausgabevar.:" und zeigt die so festgelegte Übergabe an. (Beim Umkodieren von String-Variablen heißt das Feld dagegen durchgängig „String-Variable → Ausgabevar.:".) Zusätzlich kann im Feld „Label" ein Variablen-Label für die neue Variable vergeben werden. Die weiteren Schritte sind unabhängig davon, ob die Umkodierung in dieselbe oder in eine neue Variable erfolgt.

- ▷ Sie können die Umkodierung auf einen Teil der Fälle beschränken (z.B. schließen Sie bei der Umkodierung von Einkommensdaten diejenigen Fälle aus, die kein Einkommen haben). Dies ist durch Verwendung eines Bedingungsausdrucks möglich. Durch Anklicken der Schaltfläche „Falls...“ öffnet sich die Dialogbox „Umkodieren in (dieselbe) eine andere Variable: Falls Bedingung erfüllt ist“. Diese Dialogboxen haben mit Ausnahme der Überschrift dasselbe Aussehen wie die Dialogbox in Abb. 5.5. Der Bedingungsausdruck wird auf die oben geschilderte Weise gebildet und ausgeführt. Bedenken Sie: Wenn Sie die Umkodierung auf diese Weise auf einen Teil der Fälle beschränken, werden allen anderen Fällen System-Missing-Werte zugewiesen.
- ▷ Für das eigentliche Umkodieren klicken Sie in der Dialogbox „Umkodieren in dieselbe bzw. in eine andere Variable“ auf die Schaltfläche „Alte und neue Werte...“. Die Dialogbox „Umkodieren in dieselbe/andere Variable: Alte und neue Werte“ erscheint. (⇨ Abb. 5.6. Gegenüber dieser Abbildung fehlen bei „Umkodierung in dieselbe Variable“ die Option „Alte Werte kopieren“ und die beiden Kontrollkästchen für die Umwandlung des Variablentyps.)

In dieser Dialogbox werden die Umkodierungsvorschriften definiert. Da eine Variable mehrere Werte umfasst, sind es in der Regel auch mehrere Umkodierungsvorschriften, die nacheinander definiert werden. Die Dialogbox besteht aus zwei Teilen. Im linken Teil wird jeweils der alte Wert bzw. der alte Wertebereich angegeben, im rechten der neue Wert definiert (es kann sich hier nur um einen Einzelwert, keinen Wertebereich handeln). Durch Anklicken von „Hinzufügen“ wird die Definition abgeschlossen und das Ergebnis der jeweiligen Umkodierungsvorschrift im Feld „Alt → Neu:“ angezeigt. (Die Anzeige erfolgt unter Benutzung der englischen Begriffe aus der Syntaxsprache.) Dies wird so lange wiederholt, bis alle alten Werte umdefiniert sind. Beachten Sie: Wenn Sie in eine neue Variable umkodieren, werden alle nicht umdefinierten Werte in System-Missing-Werte umgewandelt. Sie müssen also auch solche Werte umkodieren, für die sie die alten Werte beibehalten wollen. (Das gilt jedoch nicht für Umkodierung in dieselbe Variable.)

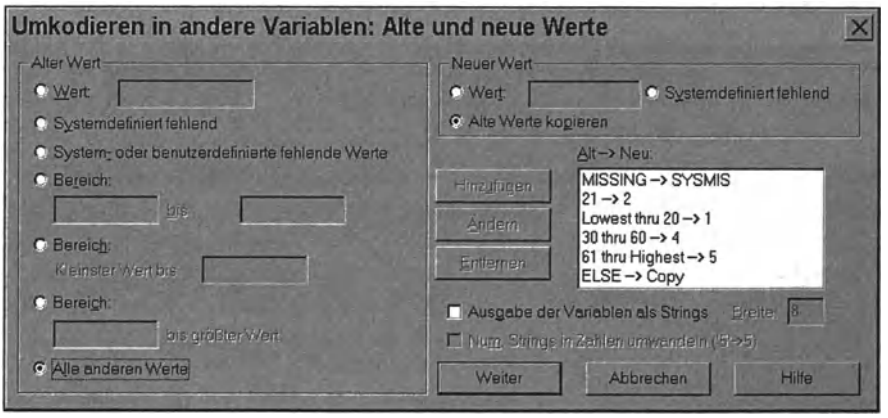


Abb. 5.6. Dialogbox „Umkodieren in andere Variablen: Alte und neue Werte“

Angeben der Ausgangswerte (alte Werte).

- *Wert.* Benutzt man für die Umkodierung einzelner Werte (z.B. 12 soll zu 2 werden).
- *Bereich.* Benutzt man, wenn mehrere nebeneinander liegende Werte einen gemeinsamen neuen Wert erhalten sollen. (*Beispiel:* 30 bis 60 soll 4 ergeben.) Für offene Randklassen kann man die zwei Varianten benutzen, die jeweils vom kleinsten bis zu einem nutzerdefinierten oberen bzw. vom größten bis zu einem nutzerdefinierten unteren Wert reichen (*Beispiel:* Kleinster Wert bis 20 soll 1, 60 bis größter Wert soll 5 werden).
- *Alle anderen Werte.* Vereinfacht die Zuordnung nicht zusammenhängender Werte zu einem neuen Wert. (*Beispiel:* Man hat alle Werte bis auf die Werte zwischen 22 und 29 umkodiert. Diese sollen unter dem neuen Wert 3 zusammengefasst werden.)
- Außerdem gibt es zwei Optionen, mit denen man systemdefinierte fehlende Werte bzw. alle fehlende Werte zusammen als umzudefinierende Werte deklarieren kann.

Festlegen der neuen Werte. Auch für die Festlegung der neuen Werte stehen drei Möglichkeiten zur Verfügung:

- *Wert.* Man klickt auf den Optionsschalter und gibt den neuen Wert im Eingabefeld ein (Wertbereiche sind nicht möglich). Diese Möglichkeit wird für die überwiegende Zahl der Umkodierungen benutzt.
- *Systemdefiniert fehlend.* Werte kann man nur durch Auswahl dieser Option in systemdefinierte fehlende Werte umwandeln.
- *Alte Werte kopieren.* Bei Anklicken dieses Optionsschalters kopiert man die alten Werte für den in „Alter Wert" ausgewählten Bereich.

In Abb. 5.6 finden sich Anweisungen zur Umkodierung einer Altersvariablen. Die wichtigsten Varianten sind darin enthalten. Zum Zwecke der Demonstration wurden auch unzweckmäßige Kodierungen vorgenommen. Die erste Anweisung macht aus allen fehlenden Werten (system- und nutzerdefinierten) System-Missing-Werte, die zweite verschlüsselt den einzelnen Wert 21 als neuen Wert 2. Die dritte Anweisung überführt den Wertebereich vom kleinsten (Lowest) Wert bis 20 in 1, die nächste den Bereich 30 bis 60 in 4, die nächste von 60 bis zum größten Wert (Highest) in 5. Schließlich werden alle noch nicht umkodierten Werte (ELSE) kopiert, d.h. mit ihrem alten Wert übernommen (es handelt sich im Beispiel um die Werte 22 bis 29).

Umwandeln des Variablentyps (nur bei Umkodierung in eine neue Variable). Schließlich bietet die Dialogbox auch noch die Möglichkeit, durch Anklicken der entsprechenden Kontrollkästchen bei der Umkodierung eine Umwandlung von numerischen in Stringvariablen (Stringlänge muss festgelegt werden) oder von (numerischen) Stringvariablen in numerische vorzunehmen. Eine numerische Stringvariable enthält Zahlen im Stringformat. Im letzten Falle reicht es, einen einzigen Wert umzukodieren und das zutreffende Kästchen anzukreuzen. Dann werden alle Werte in numerische umgewandelt. (Beides ist auch im Menü „Berechnen" mit Hilfe von String-Funktionen möglich.)

5.4 Zählen des Auftretens bestimmter Werte

Unter Umständen kann es von Interesse sein, eine neue Variable zu bilden, in der das Auftreten desselben Wertes oder derselben Werte über mehrere Variablen ausgezählt ist. *Beispiel.* In der Datei ALLBUS90.SAV sind vier Variablen gespeichert, die erfassen, wie Befragte bestimmte Arten „kriminellen" Verhaltens beurteilen, nämlich Steuerbetrug (STEUERA), Schwarzfahren (SCHWARZ), Kaufhausdiebstahl (KAUFHAUS), Alkohol am Steuer (ALKOHOL). Alle vier Variablen sind mit den Werten 1 = „sehr schlimm", 2 = „ziemlich schlimm", 3 = „weniger schlimm" und 4 = „überhaupt nicht schlimm" verschlüsselt.[Wenn Sie das Beispiel nachvollziehen möchten, downloaden Sie die Datei von den zum Buch gehörigen Internetseiten (⇨ Anhang B) und öffnen Sie diese im Daten-Editor.]. Durch Zusammenfassung der Angaben soll eine neue Variable moralischer Rigorismus (MORAL) gewonnen werden. Es wird jemand als moralisch umso rigoroser angesehen, je mehr Fragen er mit „sehr schlimm" (=1) beantwortet hat. Die neuen Werte können von 4 = „sehr rigoros" bis 0 = „gar keine moralischen Ansprüche" schwanken. Um eine solche Zählvariable zu bilden, gehen Sie wie folgt vor:

▷ Wählen Sie „Transformieren" und „Zählen...". Die Dialogbox „Häufigkeiten von Werten in Fällen zählen" (⇨ Abb. 5.7) öffnet sich.

Abb. 5.7. Dialogbox „Häufigkeiten von Werten in Fällen zählen"

▷ Geben Sie den Namen der neuen Variablen (hier: MORAL) im Eingabefeld „Zielvariable:" ein.
▷ Geben Sie, wenn gewünscht, ein Label für die Zielvariable im Feld „Label" ein.
▷ Übertragen Sie die Variablen, über die das Auftreten des Wertes ausgezählt werden soll, aus der Quellvariablenliste in das Eingabefeld „Variablen:". Es ändert damit seinen Namen je nach Variablenart in „Numerische Variablen:" oder „String-Variablen:".
▷ Klicken Sie auf „Werte definieren". Die Dialogbox „Werte in Fällen zählen: Welche Werte?" erscheint (⇨ Abb. 5.8).
▷ Legen Sie hier fest, für welchen Wert/welche Werte die Häufigkeit des Vorkommens ausgezählt werden soll (hier: 1).

Bei der Festlegung des zu zählenden Wertes bzw. der zu zählenden Werte wird ähnlich verfahren wie beim Umkodieren. Man kann in linken Teil der Dialogbox

einzelne Werte oder Wertebereiche eingeben. Dazu ist der Optionsschalter anzuklicken und der zu zählende Wert (oder Wertebereich) in ein Eingabefeld bzw. zwei Eingabefelder einzutragen. Auch „Systemdefiniert" und „System- oder benutzerdefinierte fehlende"-Werte können per Optionsschalter gewählt werden.

▷ Durch Klicken auf „Hinzufügen" werden die zu zählenden Werte jeweils in die Liste „Zu zählende Werte:" übertragen. Es kann also eine längere Liste definiert werden.

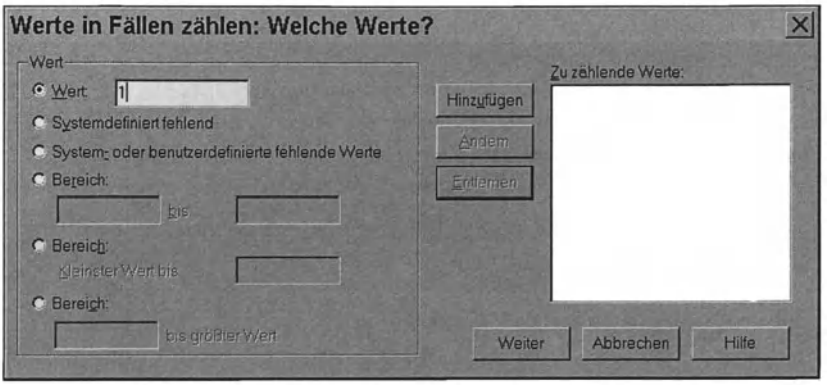


Abb. 5.8. Dialogbox „Werte in Fällen zählen: Welche Werte?"

Es wird immer nur festgestellt, ob irgendeiner der genannten Werte (logisches Oder) in der Variablen auftritt. Ist das der Fall, wird der Zähler um 1 nach oben gesetzt. In unserem Beispiel wird für jeden Fall ausgezählt, wie häufig in den vier Variablen eine 1 auftritt. Das kann keinmal bis viermal sein. Ergebniswerte sind 0 bis 4. Hätten wir als zu zählende Werte 1 und 2 eingesetzt, würde für jeden Befragten ausgezählt werden, wie häufig in den vier Variablen eine 1 oder eine 2 auftritt. Ergebniswerte können nach wie vor 0 bis 4 sein. Allerdings wird z.B. 4 häufiger auftreten, weil ja alle Fälle, die vier mal 1 oder 2 angegeben haben, diesen Wert erhalten.

Beschränken auf eine Teilmenge der Fälle. Gegebenenfalls können Sie das Auszählen auf einen Teil der Fälle beschränken. Dazu klicken Sie auf die Schaltfläche „Falls...". Es erscheint Dialogbox „Zählen: Falls Bedingung erfüllt ist". Definieren Sie dort auf die bekannte Art einen Bedingungsausdruck.

5.5 Transformieren in Rangwerte

Manchmal kann es von Interesse sein, für die Analyse anstelle der ursprünglichen Messwerte die Rangplätze der Fälle zu verwenden. Das heißt, man setzt anstelle des ursprünglichen Messwertes für einen Fall den Rang, den diese Untersuchungseinheit in einer nach den Messwerten geordneten Reihe der Fälle einnimmt. Will man z.B. auf Ordinalskalenniveau gemessene Variablen miteinander korrelieren,

ist das unerlässlich. Dasselbe gilt, wenn Signifikanztests für solche Daten durchgeführt werden. SPSS führt allerdings bei Verwendung entsprechender Statistiken die Rangtransformation automatisch durch, so dass nicht unbedingt die Rangtransformationsoption zur Anwendung kommen muss. Es kann aber auch sein, dass die Informationen, die man aus Rangplätzen gewinnt, aufschlussreicher als die originären Messdaten sind. So interessiert z.B. am Ergebnis einer Leistungsmessung weniger der Wert, den eine Person auf der entsprechenden Skala erlangt, als die relative Position, die diese Person innerhalb einer Population einnimmt. Das Untermenü „Rangfolge bilden..." bietet eine Reihe unterschiedlicher Möglichkeiten, Messwerte in absolute oder relative Rangplätze umzuwandeln oder auch in Perzentilgruppen einzuordnen.

Zur Illustration seien die Noten von neun Schülern einer Klasse in einem Fach herangezogen (Datei: NOTEN.SAV). Sie sehen die Noten in der ersten Spalte von Abb. 5.11. In der zweiten Spalte sehen Sie das Geschlecht der Schüler(innen). Um die Noten in Rangplätze umzuwandeln, gehen Sie wie folgt vor:

▷ Wählen Sie die Befehlsfolge „Transformieren" und "Rangfolge bilden...". Die Dialogbox „Rangfolge bilden" (⇨ Abb. 5.9) erscheint.

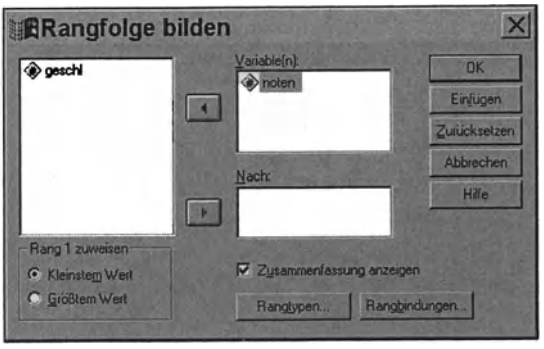


Abb. 5.9. Dialogbox „Rangfolge bilden"

▷ Übertragen Sie den Namen der Variablen, für deren Werte die Transformation vorgenommen werden soll, in das Feld „Variable(n):"

▷ Sie können jetzt in der Gruppe „Rang 1 zuweisen" bestimmen, in welcher Richtung die Fälle geordnet werden sollen.

- ❒ *Kleinstem Wert.* Der Fall mit dem kleinsten Wert erhält den Rang 1 (im Beispiel der Fall mit der Note 1).
- ❒ *Größtem Wert.* Der Fall mit dem größten Wert erhält den Rang 1 (im Beispiel der Fall mit der Note 4).

▷ Anklicken des Kontrollkästchens „Zusammenfassung anzeigen" sorgt dafür, dass im Ausgabefenster eine Meldung darüber erfolgt, welche Variable in welche neue Variable nach welcher Funktion transformiert wurde.

Beispiel:

From variable	New variable	Label
NOTEN	RNOTEN	RANK of NOTEN

Per Voreinstellung werden die Werte in absolute Rangplätze transformiert. Bei *Bindungen*, d.h. Fällen mit gleichem Wert, wird jedem Fall der mittlere Rangplatz all dieser Fälle zugeordnet. Die transformierten Werte werden automatisch in einer neuen Variablen gespeichert. Dieser wird automatisch ein neuer Variablennamen zugeordnet. Er besteht aus dem alten Namen und einem bzw. mehreren vorangestellten Buchstaben, die das zur Transformation verwendete Verfahren symbolisieren oder, wenn der Name bereits vergeben ist, diesem/diesen Buchstaben mit einer nachgestellten Ziffernfolge (beginnend mit 001). Außerdem wird ein Label vergeben. Sollen andere als der voreingestellte Rangtyp benutzt werden oder sollen Bindungen anders behandelt werden, muss durch Klicken auf „Rangtypen..." bzw. „Rangbindungen..." eine Auswahl erfolgen.

Rangtypen. Um einen Rangtyp auszuwählen, gehen Sie wie folgt vor:

▷ Klicken Sie auf die Schaltfläche „Rangtypen...". Die Dialogbox „Rangfolge bilden: Typen" erscheint. Sie können durch Anklicken der Auswahlkästchen die gewünschten Rangtypen auswählen. (In Abb. 5.10. wurden alle ausgewählt. Das Ergebnis ist in Abb. 5.11 zu sehen.)

Abb. 5.10. Dialogbox „Rangfolge bilden: Typen" nach Anklicken von „Mehr>>"

Die verschiedenen Typen werden anhand des ersten Falles (Schülers) in Abb. 5.11 erläutert. Typen sind:

- *Rang* (Voreinstellung). Absoluter Rangwert. (RNOTEN. Fall 1 hat den Rang 3).
- *Savage-Wert.* Rangplätze, die auf einer Exponentialverteilung basieren. (SNOTEN. Die Rangplätze werden in Exponentialscores umgewandelt. Im Beispiel laufen diese von −0,8889 für den Rangplatz 1 bis 1,829 für den Rangplatz 9. Fall 1 bekommt den Score −0,6210).
- *Relative Rangfolge.* Der Rangplatz wird durch die Zahl der gültigen Fälle dividiert (RFR001. Fall 1 = Rang 3 dividiert durch 9 = 0,3333).

- *Relative Rangfolge in %.* Dasselbe, multipliziert mit 100. (PNOTEN. Fall 1 = 3/9*100 = 33,33). In beiden Fällen geht es um relative Rangplätze. Jeweils wird angegeben, welche relative Position ein Fall in der Population einnimmt. 33,33% besagt z.B., dass ein Drittel der Population einen geringeren Rangplatz hat.
- *Summe der Fallgewichtungen.* Interessiert nur dann, wenn die Ränge für Untergruppen vergeben werden. Die Untergruppen werden durch Eingabe einer Gruppierungsvariablen in das Eingabefeld „Nach:" gebildet. Dann ermittelt jede Art der Rangbildung den Rang eines Falles immer nur als Rang innerhalb seiner Untergruppe. Die Auswahl der Option „Summe der Fallgewichtungen" sorgt dafür, dass die Zahl der Fälle in der jeweiligen Untergruppe (die Fallgewichte) ausgegeben werden (N0001). (In unserem Beispiel existiert nur eine Gruppe von 9 Fällen, deshalb hat jeder Fall als Summe der Fallgewichte 9.)
- *N-Perzentile.* Der Benutzer legt durch Eintrag in das Eingabefeld fest, in wie viele Perzentilgruppen er die Population eingeteilt haben will (Voreinstellung 4). Jeder Fall bekommt den Wert der Perzentilgruppe zugewiesen, der er zugehört. (NNOTEN. Im Beispiel wurden vier Perzentilgruppen gewählt. Fall 1 fällt mit der Note 2 ins zweite Viertel, also die Perzentilgruppe 2.)

1:rnoten

	noten	geschl	rnoten	snoten	rfr001	pnoten	nnoten	nti001
1	2,00	2,00	3,000	-,6210	,3333	33,33	9	2
2	1,00	1,00	1,000	-,8889	,1111	11,11	9	1
3	2,50	1,00	4,500	-,3544	,5000	50,00	9	2
4	3,00	2,00	6,500	,1623	,7222	72,22	9	3
5	2,50	2,00	4,500	-,3544	,5000	50,00	9	2
6	3,00	1,00	6,500	,1623	,7222	72,22	9	3
7	4,00	1,00	9,000	1,8290	1,0000	100,00	9	4

Abb. 5.11. Ausgangsdaten und transformierte Werte der Datei NOTEN.SAV.

Rangtransformationen unter Annahme einer Normalverteilung. Durch Anklicken der Schaltfläche „Mehr>>„ in der Dialogbox „Rangfolge bilden: Typen" werden in einem zusätzlichen Bereich am unteren Rande der Dialogbox zwei weitere Rangtypen verfügbar. Es geht dabei um kumulierte Anteile unter der Voraussetzung, dass man eine Normalverteilung der Daten unterstellen kann:

- *Anteilsschätzungen.*
- *Normalrangwerte.* Angabe der Anteilswerte in Form von z-Scores.

Für die Schätzung beider Arten von Werten können vier verschiedene Berechnungsarten verwendet werden: „Blom", „Tukey", „Rankit" und „Van der Waerden". Alle vier Verfahren schätzen den kumulativen Anteil für einen Rangwert als Anteil der Fläche unter der Normalverteilungskurve bis zu diesem Rang. Dabei

werden etwas unterschiedliche Formeln verwendet, die zu leicht differierenden Ergebnissen führen. Formeln und Beispielsberechnungen der folgenden Übersicht beziehen sich auf Anteilsschätzungen.

- *Blom.* $(r - 3/8)/(n + 1/4)$. Dabei ist r der Rangplatz, n die Anzahl der Beobachtungen. *Beispiel:* Fall 1 hat, wie oben gesehen, den Rangplatz 3. Die Zahl der Fälle (n) beträgt 9. Also beträgt die Anteilsschätzung für den ersten Fall $(3 - 3/8)/(9 + 1/4) = 2{,}625/9{,}25 = 0{,}2838$.
- *Tukey.* $(r - 1/3)/(n + 1/3)$. Im Beispiel $(3 - 1/3)/(9 + 1/3) = 0{,}2857$.
- *Rankit.* $(r - 1/2)/n$. Im Beispiel $(3 - 0{,}5)/9 = 0{,}2778$.
- *Van der Waerden.* $r/(n + 1)$. Im Beispiel: $3/(9 + 1) = 0{,}30$.

Normalrangwerte. Die Berechnung der Normalrangwerte basiert auf den Anteilsschätzungen. In einer Tabelle der Standardnormalverteilung kann abgelesen werden, bei welchem z-Wert der geschätzte kumulierte Anteil der Fläche unter der Kurve erreicht wird. Illustriert sei das für den ersten Fall für das Verfahren nach Blom. Für den ersten Fall betrug der kumulierte Anteil 0,2838. Aus einer hinreichend genauen Tabelle der Standardnormalverteilung kann man ablesen, dass diesem Anteil ein z-Wert von 0,5716 entspricht, der, da er links der Kurvenmitte liegt, negativ sein muss.

Rangbindungen (Ties). Haben mehrere Fälle den gleichen Wert, kann ihnen auf unterschiedliche Weise ein Rang zugewiesen werden. Dies kann beeinflusst werden in der Dialogbox „Rangfolge bilden: Rangbindungen", die sich beim Anklicken der Schaltfläche „Rangbindungen" öffnet.

- ❒ *Mittelwert* (Voreinstellung). In unserem Beispiel haben die Fälle 3 und 5 dieselbe Note 2,5. In einer vom untersten Wert her geordneten Rangreihe würden sie die Plätze 4 und 5 einnehmen. Stattdessen bekommen sie beide den Rang 4,5.
- ❒ *Minimaler Rang.* Alle Werte erhalten den niedrigsten Rangplatz. In unserem Beispiel bekämen beide den Rangplatz 4.
- ❒ *Maximaler Rang.* Alle Fälle bekommen den höchsten Rangplatz. Im Beispiel bekämen beide den Rang 5.

Bei diesen Verfahren bekommen die nächsten Fälle jeweils den Rang, den sie bekommen würden, wenn jeder der gebundenen Werte einen einzelnen Rangplatz einnehmen würde. Im Beispiel sind demnach – unabhängig von den für die gebundenen Fälle vergebenen Werten – die Rangplätze 4 und 5 besetzt. Der nächste Fall kann erst den Rangplatz 6 bekommen.

- ❒ *Rangfolge fortlaufend vergeben.* Alle gebundenen Fälle erhalten den gleichen Rang (wie bei Minimum). Der nächste Fall bekommt aber die nächsthöhere ganze Zahl. Im Beispiel erhalten die Fälle 3 und 5 den Platz 4, der nächste Fall den Platz 5.

In der Regel ist das Mittelwertverfahren angemessen. In der Praxis gibt es aber auch andere Fälle. So werden im Sport gewöhnlich Plätze nach dem Minimumverfahren vergeben. Bei der Preisverleihung in der Kunst kommt es dagegen häufig vor, dass man nach dem Maximumverfahren vorgeht (keiner bekommt den ersten, aber drei den dritten Preis). Auch das letzte Verfahren mag mitunter angemessen

sein. Nehmen wir z.B. an, in einer Klasse erhalten zehn Schüler die Note 2, der nächste eine 2,5. Nach allen anderen Verfahren würde er trotz des augenscheinlich geringen Unterschieds immer weit hinter den anderen rangieren (am krassesten bei Anwendung des Minimumverfahrens), bei Vergabe fortlaufender Ränge dagegen läge er nur einen Rang hinter allen anderen.

Die Art der Behandlung von Bindungen beeinflusst auch die Ergebnisse der verschiedenen Rangbildungsverfahren. So erreicht man mit der Option „Maximaler Rang" in Verbindung mit relativen Rängen (Prozenträngen) eine empirische kumulative Verteilung.

Rangplätze für Untergruppen. Wahlweise ist es auch möglich, jeweils für Untergruppen Rangplätze zu ermitteln. In unserem Beispiel könnte man etwa Ränge getrennt für Männer und Frauen ermitteln. Dazu wird in der Dialogbox „Rangfolge bilden" die Variable, aus der sich die Untergruppen ergeben, in das Eingabefeld „Nach:" übertragen. Ansonsten bleibt die Prozedur dieselbe.

Ergänzung bei Benutzen der Syntaxsprache. Benutzt man die Syntaxsprache, kann man anstelle der automatisch gebildeten Variablennamen einen eigenen Variablennamen definieren. Dazu verwenden Sie das Unterkommando INTO und geben den Variablennamen ein.

5.6 Automatisches Umkodieren

Einige SPSS-Prozeduren können keine langen Stringvariablen und/oder nicht fortlaufend kodierte Variablen verarbeiten. Deshalb existiert eine Möglichkeit, numerische oder Stringvariablen in fortlaufende ganze Zahlen umzukodieren.

Beispiel. Wir haben eine Datei mit einer Zufriedenheitsvariablen (ZUFRIED). Die Werte sind z.T. als ganze Zahlen, z.T. als Dezimalzahlen angegeben und dadurch nicht fortlaufend kodiert. Eine weitere Variable ist eine Stringvariable mit den Namen der Befragten (NAME). Beide sollen in Variablen mit fortlaufenden ganzen Zahlen umgewandelt werden. Dazu wählen Sie:

▷ „Transformieren" und „Automatisch umkodieren...". Die Dialogbox „Automatisch umkodieren" erscheint (⇨ Abb. 5.12).

Abb. 5.12. Dialogbox „Automatisch umkodieren"

- ▷ Übertragen Sie die Variablennamen der umzukodierenden Variablen in das Feld „Variable → Neuer Name".
- ▷ Markieren Sie eine der ausgewählten Variablen. Setzen Sie den Cursor in das Eingabefeld „Neuer Name". Geben Sie einen neuen Namen ein. Klicken Sie auf die Schaltfläche „Neuer Name". Der neue Name erscheint im Auswahlfeld hinter dem alten. Wiederholen Sie das gegebenenfalls mit weiteren Variablen.
- ▷ Wählen Sie durch Anklicken der Optionsschalter „Kleinstem Wert" oder „Größtem Wert" in der Gruppe „Umkodierung beginnen bei", ob dem kleinsten oder größten Wert der Wert 1 zugewiesen und entsprechend die anderen Werte in fallender oder steigender Folge kodiert werden.
- ▷ Bestätigen Sie mit „OK".

Es werden die neuen Variablen gebildet. Die Sortierung geschieht bei Stringvariablen in alphabetischer Folge. Großbuchstaben gehen vor Kleinbuchstaben. *Beispiel.* „Albert" kommt vor „albert" und beide vor „alle". Die Wertelabels der alten Variablen werden übernommen. Sind keine vorhanden, werden die alten Werte als Wertelabels eingesetzt. *Beispiel:* In der Variablen ZUFRIED wird der alte Wert 1,5 zu 2, als Wertelabel wird 1,5 eingesetzt. In der Variablen NAME wird aus „Alfred" 1. Im Ausgabefenster erscheint ein Protokoll, das die alten und neuen Namen und die alte und neue Kodierung der Variablen angibt.

5.7 Transformieren von Zeitreihendaten

Das Basismodul von SPSS enthält auch spezielle Routinen zur Bearbeitung von Zeitreihen. Sie befinden sich einerseits im Menü „Transformieren", andererseits im Menü „Daten".

Generieren von Datumsvariablen. Das Menü „Daten" enthält die Option „Datum definieren...", die es erlaubt, Datumsvariablen zu generieren. Mit dieser Option kann man einer Zeitreihe Datumsvariablen hinzufügen, die die Termine der Erhebungszeitpunkte enthalten. Diese Variablen werden erst erzeugt, nachdem die Daten der Zeitreihe bereits vorliegen. Die so generierten Daten können als Labels für Tabellen und Grafiken benutzt werden. Vor allem sind sie aber mit den Zeitreihendaten so verknüpft, dass das Programm ihnen die Periodizität der Daten entnehmen kann. Bei Benutzung der später zu besprechenden Transformation „Saisonale Differenz" sind sie unentbehrlich. Alle anderen Zeittransformationsfunktionen benötigen nicht zwingend die vorherige Generierung von Datumsvariablen.

Datenmatrizen mit Zeitreihen haben gegenüber den ansonsten benutzten Matrizen die Besonderheit, dass die Zeilen (Fälle) den verschiedenen Erhebungszeitpunkten entsprechen, für die jeweils für jede Variable in den Spalten eine Messung vorliegt. Die Messungen sollten in (möglichst) gleichmäßigen Abständen erfolgen. Für jeden Messzeitpunkt muss eine Messung vorliegen, und sei es ein fehlender Wert. Ist das nicht der Fall, werden die angebotenen Transformationen weitgehend sinnlos und die Generierung von Datumsvariablen führt zu falschen Ergebnissen.

Beispiel: In einer Datei ALQ.SAV sind in der Spalte ALQ_E die Arbeitslosenquoten für die alten Bundesländer der Jahre 1989 bis 1993 zu den jeweiligen

Quartalsenden gespeichert. Eine Variable TERMIN enthält im Datumsformat jeweils das Stichdatum. Man darf eine solche normale Datumsvariable nicht mit einer mit der Option „Datum definieren..." erzeugten Datumsvariablen verwechseln. TERMIN ist keine für die Zeitreihe generierte Datumsvariable. Es sollen jetzt Datumsvariablen generiert werden, die die Jahres- und Quartalsangaben enthalten. Vorausgesetzt ist, dass eine lückenlose Reihe mit gleichen Abständen vorliegt.

- ▷ Dazu wählen Sie die Befehlsfolge „Daten", „Datum definieren...". Die Dialogbox „Datum definieren" öffnet sich (⇨ Abb. 5.13).
- ▷ Im Auswahlfeld „Fälle entsprechen:" müssen Sie jetzt markieren, was für Zeitperioden die Zeilen enthalten. In unserem Beispiel sind es Quartale verschiedener Jahre. Die Daten sind also zuerst nach Jahren und innerhalb der Jahre nach Quartalen geordnet. Zu markieren ist daher „Jahre, Quartale".

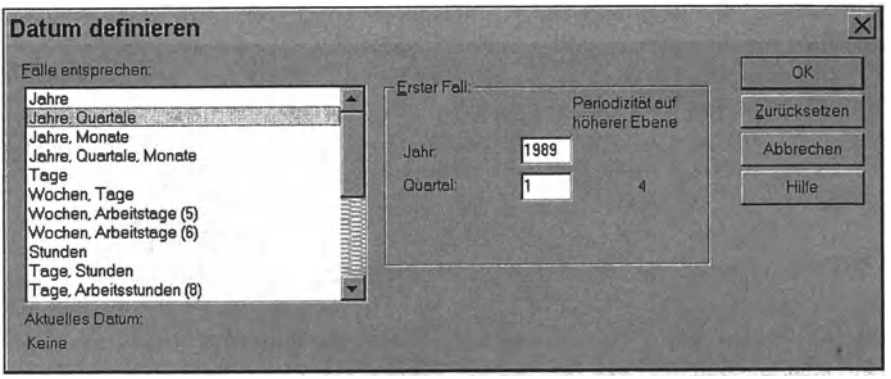


Abb. 5.13. Dialogbox „Datum definieren" mit Eintragungen

- ▷ Im Eingabefeld „Erster Fall:" muss nun angegeben werden, welches Datum genau für den ersten Fall zutrifft. Je nach der Art der ausgewählten Zeitperiode gestaltet sich das Feld „Erster Fall:" anders. Die Datumsangaben können Jahre, Quartale, Wochen, Tage, Stunden, Minuten und Sekunden enthalten. Für die im Format jeweils enthaltenen Einheiten werden Eingabefelder angezeigt, in die der Wert für den ersten Fall einzutragen ist. Gleichzeitig ist die Eingabe auf Werte innerhalb sinnvoller Grenzen (bei Quartalen z.B. ganze Zahlen von 1 bis 4) beschränkt, deren höchster Wert neben dem Eingabefeld angegeben ist. In unserem Beispiel enthält die Periodizität nur Jahre und Quartale, entsprechend erscheint je ein Eingabefeld für das Jahr („Jahr:") und das Quartal („Quartal:"). Unsere erste Eingabe ist die Arbeitslosenquote für das 1. Quartal 1989. Entsprechend tragen wir bei „Jahr:" 1989 und bei „Quartal:" 1 ein.
- ▷ Bestätigen Sie die Eingabe. Die neuen Variablen werden generiert. SPSS weist den Fällen, ausgehend von dem ersten, Datumsangaben zu. Das Programm setzt dabei gleichmäßige Abstände voraus.

Es erscheint das Ausgabefenster mit einer Meldung über die vollzogene Variablengenerierung.

The following new variables are being created:

Name	Label
YEAR_	YEAR, not periodic
QUARTER_	QUARTER, period 4
DATE_	DATE. FORMAT: „QQ YYYY“

Mehrere Variablen werden gleichzeitig generiert, für jedes Element der Datumsangabe eine eigene, im Beispiel sowohl eine für die Jahresangabe (YEAR) als auch eine für die Quartalsangabe (QUARTER). (Letztere wird für Periodisierungen verwendet.) Außerdem entsteht eine Variable, die alle Angaben zusammenfasst (DATE). Im Dateneditorfenster sind die neuen Variablen nun hinzugefügt (⇨ Abb. 5.14).

	termin	alq_e	year_	quarter_	date_
1	31.03.1989	8,4	1989	1	Q1 1989
2	30.06.1989	7,4	1989	2	Q2 1989
3	31.10.1989	7,3	1989	3	Q3 1989
4	31.12.1989	8,0	1989	4	Q4 1989

Abb. 5.14. Ergebnis der Generierung von Datumsvariablen

Transformieren von Zeitreihenvariablen. Im Menü „Transformieren“ stellt SPSS eine Reihe von Datentransformationsverfahren für Zeitreihen zur Verfügung. Damit kann dreierlei bewirkt werden:

- ❒ Aus den Ausgangsdaten werden die Differenzen zwischen den Werten verschiedener Zeitpunkte ermittelt.
- ❒ Die Werte der Zeitreihe werden verschoben.
- ❒ Die Zeitreihe wird geglättet.

Zur Glättung einige Bemerkungen. Die einzelnen Werte einer Zeitreihe können typischerweise als Kombination der Wirkung verschiedener Komponenten gedacht werden. In der Regel betrachtet man sie als Ergebnis der Verknüpfung einer Trendkomponente mit zyklischen Komponenten (etwa Konjunktur- oder Saisonschwankungen) und einer Zufallskomponenten. Die Analyse von Zeitreihen läuft weitgehend auf den Versuch hinaus, die Komponenten durch formale Datenmanipulationen voneinander zu trennen. Um eine Zeitreihe in eine neue zu transformieren, gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Transformieren“ und „Zeitreihen erstellen...“. Es öffnet sich die Dialogbox „Zeitreihen erstellen“ (⇨ Abb. 5.15).
- ▷ Übertragen Sie aus der Quellvariablenliste die Variable, die transformiert werden soll, in das Eingabefeld „Neue Variable(n):“. Automatisch wird in diesem Feld eine Transformationsgleichung generiert. Diese enthält auf der linken Seite den neuen Variablennamen. Standardmäßig wird dieser aus dem alten Namen und einer zusätzlichen laufenden Nummer gebildet (*Beispiel:* ALQ_E wird

ALQ_E_1). Auf der rechten Seite steht die verwendete Transformationsfunktion, gefolgt von den Argumenten (eines davon ist der alte Variablennamen). Der Funktionsname ist jeweils eine Abkürzung der amerikanischen Bezeichnung (⇨ unten).

Abb. 5.15. Dialogbox „Zeitreihen erstellen" mit Transformationsgleichungen

Der voreingestellte Namen und die voreingestellte Funktion müssen nicht übernommen werden. Bei Bedarf ändern Sie den Namen der neuen Variablen und/oder die verwendete Funktion:

- ▷ Markieren Sie dazu die zu ändernde Gleichung.
- ▷ Tragen Sie in das Eingabefeld den gewünschten neuen Namen ein.
- ▷ Klicken Sie auf den Pfeil an der Seite des Auswahlfeldes „Funktion". Eine Auswahlliste erscheint.
- ▷ Markieren Sie die gewünschte Funktion (⇨ verfügbare Funktionen siehe unten).
- ▷ Ändern Sie gegebenenfalls die Werte in „Ordnung:" und „Spanne:".
- ▷ Übertragen Sie die veränderten Angaben durch Anklicken von „Ändern" in die Liste „NeueVariable(n)".

Sie können auf diese Weise mehrere Transformationen nacheinander definieren. Diese können sich auch alle auf dieselbe Ausgangsvariable beziehen. Starten Sie die Transformation mit „OK". Es erscheint eine Meldung im Ausgabefenster, die u.a. den neuen Namen, die verwendete Transformationsfunktion und die Zahl der verbleibenden gültigen Fälle mitteilt (⇨ Tabelle 5.2).

Die neuen Variablen erscheinen im Dateneditorfenster. Abb. 5.16 zeigt die ersten sieben Fälle.

Die verfügbaren Funktionen werden nun erläutert. Zur Illustration werden sämtliche Funktionen (mit Ausnahme von „Lag" und „Lead") auf die Variable ALQ_E angewandt. Die Erläuterung bezieht sich jeweils auf den ersten gültigen Wert in der durch die Transformation neu gebildeten Variablen. Verfügbare Funktionen (in Klammern die Abkürzung) sind:

- ❒ *Differenz* (DIFF). Bildet die Differenz zwischen den Werten zweier aufeinanderfolgender Zeitpunkte (*Beispiel:* ALQ_E_1). In „Ordnung:" kann die Ord-

nung der Differenzen eingestellt werden. Voreingestellt ist die erste Ordnung. Zweite Ordnung bedeutet z.B., dass die Differenz der Differenzen der ersten Ordnung gebildet wird. (*Beispiel:* ALQ_E_2. Die Differenz erster Ordnung war für das zweite Quartal 89 -1,0, für das dritte -0,1. Die Differenz zweiter Ordnung beträgt: -0,1 - (-1,0) = +0,9.) Am Beginn einer Zeitreihe lassen sich keine Differenzen bilden. Zu Beginn einer neuen Reihe werden daher so viele Fälle als System-Missings ausgewiesen, wie durch den Ordnungswert festgelegt wurde.

Tabelle 5.2. Meldung des Ergebnisses einer Transformation von Zeitreihenvariablen

Result Variable	Missing Values Replaced	First Non-Miss	Last Non-Miss	Valid Cases	Creating Function
ALQ_E_1	2	19	18		DIFF(ALQ_E,1)
ALQ_E_2	3	19	17		DIFF(ALQ_E,2)
ALQ_E_3	5	19	15		SDIFF(ALQ_E,1,4)
ALQ_E_4	2	18	17		MA(ALQ_E,3,3)
ALQ_E_5	4	19	16		PMA(ALQ_E,3)
ALQ_E_6	2	18	17		RMED(ALQ_E,3,3)
ALQ_E_7	1	19	19		CSUM(ALQ_E)
ALQ_E_8	1	19	19		T4253H(ALQ_E)

	alq_e_1	alq_e_2	alq_e_3	alq_e_4	alq_e_5	alq_e_6	alq_e_7	alq_e_8
1	.	.	.	.	.	.	8,4	8,19
2	-1,0	.	.	7,70	.	7,40	15,8	7,89
3	-,1	,9	.	7,57	.	7,40	23,1	7,69
4	,7	,8	.	7,67	7,70	7,70	31,1	7,56
5	-,3	-1,0	-,7	7,53	7,57	7,70	38,8	7,39
6	-,8	-,5	-,5	7,07	7,67	6,90	45,7	7,11
7	-,3	,5	-,7	6,77	7,53	6,80	52,3	6,79

Abb. 5.16. Ergebnisse von Zeitreihen-Transformationen

❒ *Saisonale Differenz* (SDIFF). Es werden jeweils die Differenzen zwischen denselben Phasen zweier verschiedener Perioden berechnet. In unserem Beispiel sind solche Phasen die Quartale verschiedener Jahre. Üblicherweise wird man die Differenzen der Werte zweier aufeinanderfolgender Perioden berechnen (Ordnung: 1). Mit Ordnung kann man aber auch größere Abstände bestimmen. Ordnung: 2 würde z.B. die Differenz zwischen den Phasenwerten eines Jahres und den Werten derselben Phasen zwei Jahre voraus ermitteln. (*Beispiel:* ALQ_E_3. Die Differenz zwischen dem Wert des ersten Quartals 1990 und dem des ersten Quartals 1989 beträgt 7,7 - 8,8 = -0,7. Die Arbeitslosenquote ist zwischen dem ersten Quartal 1989 und dem ersten Quartal 1990 gesunken.) Die

Zahl in „Ordnung" bestimmt wiederum, wie viele Werte am Beginn der Zeitreihe als System-Missing ausgewiesen werden: Ordnungszahl (= Zahl der Perioden) multipliziert mit der Zahl der Phasen. Diese Transformation verlangt, dass vorher eine Datumsvariable kreiert wurde, aus der die Peridozität hervorgeht. Ist das nicht der Fall, wird die Ausführung mit einer Fehlermeldung abgebrochen.

- ❒ *Zentrierter gleitender Durchschnitt* (gleitende Mittelwerte) (MA). Die Zeitreihe wird geglättet, indem anstelle der Ausgangswerte Mittelwerte aus einer Reihe benachbarter Zeitpunkte berechnet werden. Im Eingabefeld „Spanne:" wird festgelegt, wieviel benachbarte Werte zusammengefasst werden (Mittelungsperiode = m). Wird eine ungerade Mittelungsperiode verwendet, berechnet man das arithmetische Mittel der m benachbarten Werte und setzt den Mittelwert anstelle des in der Mitte der Mittelungsperiode liegenden Wertes (*Beispiel:* ALQ_E_4. Spanne war 3. Der Wert für das 2. Quartal 1989 ergibt sich aus der Rechnung: (8,4 + 7,4 + 7,3) : 3 = 7,7.) Legt die Spanne (Mittelungsperiode) allerdings eine gerade Zahl von Fällen zur Mittelung fest, dann existiert kein Fall in der Mitte. Man benutzt daher dennoch eine ungerade Zahl von Fällen (m+1) zur Mittelung, behandelt aber die beiden Randfälle als halbe Fälle, d.h. ihre Werte gehen nur zur Hälfte in die Mittelung ein. (*Beispiel:* Bei einer Spanne 4 ergäbe sich für das 3. Quartal 1989 folgende Berechnung: (7,4/2 + 7,3 + 8 + 7,7 + 6,9/2) : 4 = 7,69.) Die Zahl der System-Missings in der neuen Variablen ist bei ungerader Größe der Spanne (n – 1) : 2 bei gerader Spanne n : 2 an jedem Ende der Zeitreihe.
- ❒ *Zurückgreifender gleitender Durchschnitt* (PMA). Es werden auf die beschriebene Weise gleitende Mittelwerte gebildet, und gleichzeitig werden die errechneten Mittelwerte um die für die Mittelwertbildung benutzte Spanne nach hinten verschoben. (*Beispiel:* ALQ_E_5. Es wurde die Spanne 3 verwendet. Der Wert 7,7 für den Zeitpunkt 4. Quartal 1989 ergibt sich aus der Mittelung der Werte der drei vorangegangenen Perioden: (8,4 + 7,4 + 7,3) : 3.) Entsprechend dem Wert der Spanne treten am Anfang und am Ende der neue Zeitreihe System-Missings auf.
- ❒ *Gleitende Mediane* (RMED). Die Originalwerte werden durch den Medianwert einer durch die Spanne definierten Zahl von Werten um den zu ersetzenden Fall herum (inklusive dieses Falles) ersetzt. Setzt die Spanne eine ungerade Zahl von Fällen fest, ist der Medianwert der Wert des mittleren Falles. (*Beispiel:* ALQ_E_6. Die Spanne war 3. Der Wert für das zweite Quartal ist der mittlere Wert der geordneten Werte 8,4; 7,4 und 7,3, also 7,4. Das ist hier zufällig der Wert des zu ersetzenden Quartils selbst.) Wird eine gerade Zahl von Fällen als Spanne festgesetzt, gibt es keinen mittleren Fall. Dann wird ebenfalls eine ungerade Zahl von Fällen (m+1) benutzt. Von diesen wird zunächst aus den ersten m Fällen ein Medianwert ermittelt. Es ist das arithmetische Mittel der beiden mittleren Werte der geordneten Reihe dieser m Fälle. Dann bildet man für die letzten m Fälle auf die gleiche Weise den Medianwert. Aus den beiden so gebildeten Medianwerten wird wiederum das arithmetische Mittel als endgültiger zentrierter Medianwert berechnet. *Beispiel:* Bei Benutzung der Spanne 4 errechnet man als ersten gleitenden Medianwert den Wert für das 3. Quartal 1989.

Dazu werden die Werte vom ersten Quartal 1998 bis zum 1. Quartal 1990 (einschließlich) benutzt. Man bildet zuerst den Median für die ersten vier Werte dieser Reihe. Geordnet lauten diese 8,4; 8,0; 7,4; 7,3. Der Medianwert daraus beträgt (8,0 + 7,4) : 2 = 7,7. Die geordnete Reihe der zweiten vier Werte ist 8,0; 7,7; 7,4; 7,3. Deren Medianwert beträgt (7,7 + 7,4) : 2 = 7,55. Der zentrierte Medianwert für das 3. Quartil ist somit (7,7 + 7,55) : 2 = 7,63.

- *Kumulierte Summe* (CSUM). Kumulierte Summe der Zeitreihenwerte bis zu einem Zeitpunkt, inklusive des Wertes dieses Zeitpunkts. (*Beispiel:* ALQ_E_7. Für das 3. Quartal ergibt sich der Wert aus der Summe der Werte des ersten, zweiten und dritten Quartals: 8,4 + 7,4 + 7,3 = 23,1. Im Beispiel ist das keine sinnvolle Anwendung. Sinnvolle Anwendungen lassen sich denken bei Variablen, deren Werte sich faktisch in der Zeit kumulieren, etwa gelagerte Abfälle u.ä.)
- *Lag.* Die Werte werden um die in Ordnung angegebene Zahl der Zeitpunkte in der Zeitreihe nach hinten verschoben. (*Beispiel:* Ordnung ist 2. Der Wert des 1. Quartals 1989 wird zum Wert des 3. Quartals.) Die am Beginn der Reihe entstehen Zahl Missing-Werte entspricht dem in „Ordnung" angegebenen Wert.
- *Lead.* Die Werte werden um die in „Ordnung" eingegebene Zahl der Zeitpunkte in der Zeitreihe nach vorne verschoben. (*Beispiel:* Ordnung ist 2. Der Wert des 3. Quartals 1989 wird zum Wert des 1. Quartals usw.) Die am Ende Zeitreihe entstehende Zahl der Missing-Werte entspricht dem Wert in „Ordnung".
- *Glätten* (Glättungsfunktion). (T4253H). Die neuen Werte werden durch eine zusammengesetzte Prozedur gewonnen. Zunächst werden Medianwerte mit der Spanne 4 gebildet, die wiederum durch gleitende Medianwerte der Spanne 2 zentriert werden. Die sich daraus ergebende Zeitreihe wird wiederum geglättet durch Bildung von gleitenden Medianwerten der Spanne 5, darauf der Spanne 3 und schließlich gleitender gewogener arithmetischer Mittel. Aus der Differenz zwischen Originalwerten und geglätteten Werten errechnet man Residuen (Reste), die wiederum in einem zweiten Durchgang selbst demselben Glättungsprozess unterworfen werden. Die endgültigen Werte gewinnt man, indem man zu den gleitenden Werte des ersten Durchgangs die geglätteten Residuen des zweiten addiert. Das Schlüsselwort dieser Prozedur heißt „T4253H", wobei die Ziffern die festgelegte Spannweite der einzelnen Glättungsschritte repräsentieren. (*Beispiel:* ALQ_E_8 enthält die Ergebnisse dieser Glättungsprozedur.)

Zusätzliche Möglichkeiten bei Verwenden der Befehlssyntax. Störend wirkt sich bei fast allen genannten Transformationen aus, dass an einem oder beiden Enden der neuen Datenreihe fehlende Werte entstehen, je nach Ordnung bzw. Rang mehr oder weniger. Die Syntaxsprache lässt es daher bei den Funktionen *„Zentrierter gleitender Durchschnitt"* (Centered moving averages, MA) und *„Gleitende Mediane"* (Running medians, RMDE) zu, eine zweite Spanne (minumum span) anzugeben. Diese muss einen Wert annehmen, der zwischen 1 und dem Wert der ersten Spanne liegt. Durch diese Definition werden in den Randbereichen, wenn keine ausreichende Zahl von Fällen mehr für die Mittelwertbildung gemäß der ersten Spanne existieren, Mittelwerte aus Fällen einer verkleinerten Spanne bis minimal der niedrigsten in der zweiten Spanne angegebenen Fallzahl gebildet. Auf diese Weise werden in den Randbereichen zusätzliche Werte gewonnen. Außer-

dem stehen zwei weitere Funktionen „Fast Fourier transform" (FFT) und „Inverse fast Fourier transform" (IFFT) zur Verfügung. Erstere produziert zwei neue Serien, die eine mit dem Sinus-Teil, die andere dem Cosinus-Teil einer Ausgangsverteilung. Die zweite Funktion bildet umgekehrt aus zwei Ausgangsreihen, deren eine den Sinus-, die andere den Cosinus-Anteil enthält, eine neue Zeitreihe.

Ersetzen von fehlenden Werten in Zeitreihen. Fehlen innerhalb einer Zeitreihe Werte, so wirkt sich das auf die Berechnung neuer Zeitreihen störend aus. Bei Differenzenbildung ergibt jede Berechnung einen fehlenden Wert, wenn einer der Ausgangswerte fehlt. Bei der Berechnung von gleitenden Durchschnitten bzw. Medianwerten gibt jede Berechnung, bei der irgendein Wert innerhalb der angegebenen Spanne fehlt, einen fehlenden Wert in der neuen Reihe. In diesen Fällen vermehrt sich die Zahl der fehlenden Werte in der neuen Zeitreihe. Bei Verwendung der Lag- und Lead-Funktion ergeben fehlende Werte wieder fehlende Werte. Die Zahl bleibt gleich. Die „Glättungsfunktion" lässt keine eingebetteten fehlenden Werte zu. Ist diese Bedingung verletzt, werden lauter System-Missings erzeugt. Bei der Bildung einer kumulierten Summe wird lediglich zum Zeitpunkt des fehlenden Wertes ein System-Missing eingesetzt. In der Folge summiert das Programm weiter.

Sind eingebettete fehlende Werte vorhanden, so müssen diese zur Anwendung der „Glättungsfunktion" ersetzt werden. Aber auch bei der Berechnung gleitender Mittelwerte kann das notwendig sein, um eine zu große Zahl von fehlenden Werten zu vermeiden. Eine „Imputation" (Ersetzung) fehlender Werte kommt jedoch nur in Frage, wenn die Gewähr gegeben ist, dass die Ersatzwerte nicht zu stark von den wirklichen (fehlenden) Werten abweichen. Fehlt in einer Zeitreihe nur gelegentlich ein Wert, so kann man das bei Auswahl eines geeigneten Verfahrens zumeist bejahen. SPSS bietet verschiedene Möglichkeiten, fehlende Werte in Zeitreihen zu ersetzen.

Beispiel. In unserer Zeitreihe fehle der Wert für das 3. Quartal 1989. Er soll ersetzt werden. Um einen Wert zu ersetzen, gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Transformieren", „Fehlende Werte ersetzen...". Die Dialogbox „Fehlende Werte ersetzen" (⇨ Abb. 5.17) erscheint. Die weitere Eingabe erfolgt analog zum Verfahren bei der Transformation von Zeitreihen. Nur werden hier nicht alle Werte der Zeitreihe, sondern nur die fehlenden Werte ersetzt.
- ▷ Übertragen Sie die Variable, bei der ein fehlender Wert ersetzt werden soll. Im Feld „Neue Variable(n)" erscheint automatisch eine Gleichung mit einem neuen Variablennamen auf der linken und der zuletzt verwendeten Funktion und dem alten Variablennamen als eines der Argumente auf der rechten Seite.

Wollen Sie am Namen oder der Funktion etwas ändern (⇨ unten verfügbare Funktionen), gehen Sie analog zu obigen Ausführungen vor. Bei den Funktionen „Mittel der Nachbarpunkte" und „Median der Nachbarpunkte" ist gegebenenfalls noch eine Spanne durch Anklicken der Optionsschalter „Anzahl" und Eingabe einer Zahl oder durch Anklicken der Optionsschalter „Alle" vorzugeben. Neue Variablen und Funktionen sind mit „Ändern" zu bestätigen. Sie können auch wieder mehrere Transformationen für verschiedene Variablen nacheinander definieren

und/oder mit unterschiedlichem Verfahren zum Ersetzen der fehlenden Werte für dieselbe Variable arbeiten. Die Ausführung starten Sie mit „OK".

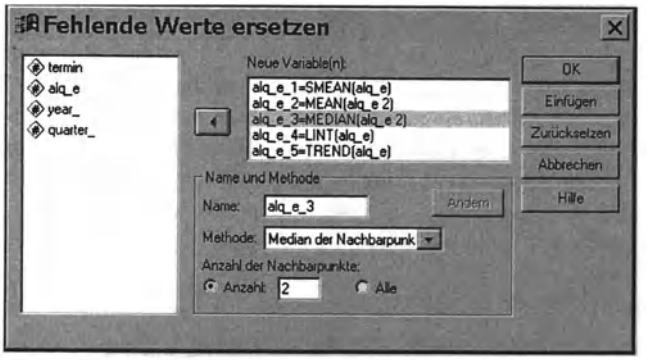


Abb. 5.17. Dialogbox „Fehlende Werte ersetzen"

Im Ausgabefenster erscheint eine Meldung über die Ausführung des Befehls. Sie enthält u.a. wiederum den neuen Namen, das Verfahren, sowie die Zahl der gültigen Werte (⇨ Tabelle 5.3).

Tabelle 5.3. Meldung beim Ersetzen fehlender Werte in einer Zeitreihe

Result Variable	Missing Values Replaced	First Non-Miss	Last Non-Miss	Valid Cases	Creating Function
ALQ_E_1	0	1	19	19	SMEAN(ALQ_E)
ALQ_E_2	0	1	19	19	MEAN(ALQ_E,2)
ALQ_E_3	0	1	19	19	MEDIAN(ALQ_E,2)
ALQ_E_4	0	1	19	19	LINT(ALQ_E)
ALQ_E_5	0	1	19	19	TREND(ALQ_E)

	date_	alq_e_1	alq_e_2	alq_e_3	alq_e_4	alq_e_5
1	Q1 1989	8,40	8,40	8,40	8,40	8,40
2	Q2 1989	7,40	7,40	7,40	7,40	7,40
3	Q3 1989	7,08	7,88	7,85	7,70	7,15
4	Q4 1989	8,00	8,00	8,00	8,00	8,00
5	Q1 1990	7,70	7,70	7,70	7,70	7,70
6	Q2 1990	6,90	6,90	6,90	6,90	6,90

Abb. 5.18. Ergebnis des Ersetzens eines fehlenden Wertes mit verschiedenen Verfahren

Im Dateneditorfenster erscheinen die neuen Variablen mit ersetzten fehlenden Werten (⇨ Abb. 5.18).

Die verfügbaren Verfahren werden nun am Beispiel erläutert. Ersetzt wird jeweils der fehlende Wert für das 3. Quartal 1989.

- *Zeitreihen-Mittelwert* (SMMEAN). Ersetzt den fehlenden Wert durch das arithmetische Mittel der ganzen Serie (siehe ALQ_E_1).
- *Mittel der Nachbarpunkte* (MEAN). Arithmetisches Mittel der dem fehlenden Wert benachbarten Zeitpunkte. Durch Eingabe einer Zahl in das Feld „Anzahl" bestimmt man, wie viele Nachbarpunkte jeweils auf beiden Seiten herangezogen werden sollen (2 bedeutet demnach vier Nachbarpunkte insgesamt). Die Auswahl von „Alle" ergäbe dasselbe Ergebnis wie „Zeitreihen-Mittelwerte" (siehe ALQ_E_2). Die Spanne darf nicht größer angesetzt werden als gültige Werte um den fehlenden zur Verfügung stehen. Sonst wird der fehlende Wert nicht ersetzt.
- *Median der Nachbarpunkte* (MEDIAN). Median der dem fehlenden Wert benachbarten Zeitpunkte. Wiederum kann die Spanne über „Anzahl" oder „Alle" festgelegt werden. „Anzahl" legt die Zahl der Fälle auf jeder Seite des Medianwertes fest. (*Beispiel:* ALQ_E_3. „Anzahl" war 2. Nach der Größe geordnet ergeben die vier Werte die Reihe: 8,4; 8,0; 7,7; 7,4. Der Medianwert ist das arithmetische Mittel der beiden mittleren Werte 8,0 und 7,7, also 7,85.) Die Spanne darf nicht größer angesetzt werden als gültige Werte um den fehlenden zur Verfügung stehen. Sonst wird der fehlende Wert nicht ersetzt.
- *Lineare Interpolation.* (LINT). Ausgehend von dem ersten gültigen Wert vor und nach dem/den fehlenden Werten wird interpoliert. Fehlt nur ein Wert, ist das identisch mit dem arithmetischen Mittel zwischen diesen beiden Werten. (*Beispiel:* ALQ_E_4. Die Differenz zwischen 7,4 und 8,0 = 0,6. Die Hälfte davon = 0,3 wird bei der Interpolation der 7,4 zugeschlagen = 7,7, um den Wert für das 3. Quartal 1989 zu ermitteln.) Liegen mehrere fehlende Werte nebeneinander, muss die Differenz zwischen den Nachbarwerten in entsprechend viele gleich große Anteile zerlegt werden.
- *Linearer Trend an dem Punkt* (TREND). Dazu wird zunächst eine Zeitvariable mit den Werten 1 bis n für die Zeitpunkte gebildet. Danach wird eine Regressionsgerade für die Voraussagevariable auf dieser Zeitvariablen gebildet. Aus der so gewonnen Regressionsgleichung wird der Voraussagewert für den fehlenden Wert errechnet und an dessen Stelle eingesetzt. (In unserem Beispiel ergibt die Regressionsanalyse die Regressionsgleichung $y = 1{,}177 - 0{,}009x$. Den Zeitpunkt 3 für x eingesetzt, ergibt 7,15, den Wert in ALQ_E_5.)

5.8 Offene Transformationen

Per Voreinstellung werden Transformationen sofort ausgeführt. Um bei einer Vielzahl von Transformationen und großen Datenmengen Rechenzeit zu sparen, kann man diese Einstellung im Menü „Optionen", Register „Daten" ändern, so dass Transformationen erst dann durchgeführt werden, wenn ein Datendurchlauf erforderlich ist (⇨ Kap 28.5). Im letzteren Falle kann man die Transformationen jederzeit mit der Befehlsfolge „Transformieren" und „Offene Transformationen ausfüh-

ren“ ausführen lassen. Ansonsten werden Sie automatisch beim Aufruf einer Statistikprozedur vorgenommen.

5.9 Variable kategorisieren

Mit der Prozedur "Variablen kategorisieren" kann eine Variable mit stetigen numerische Daten in eine kategoriale Variable mit einer diskreten Anzahl von Kategorien umgewandelt werden. Dies kann für bestimmte statistische Analysen notwendig sein. z.B. verlangt eine Varianzanalyse, bei der Alter die unabhängige, Einkommen die abhängige Variable sein soll, dass die unabhängige Variable Einkommen in eine beschränkte Zahl vergleichbarer Gruppen kategorisiert ist. Dies wäre auch durch Prozeduren wie „Unkodieren“ erreichbar. Die Prozedur „Variable kategorisieren“, löst die Aufgabe aber besonders elegant, wenn das Umkodieren zu einer festgelegten Zahl etwas gleich stark besetzter Gruppen führen soll.

Nach Festelegung der Zahl der Kategorien werden die Fälle in eine entsprechende Zahl von Perzentilen gruppiert. Alle Fälle eines Perzentils erhalten denselben Wert. Eine Einteilung in beispielsweise 4 Gruppen würde Fällen unter dem 25. Perzentil den Wert 1, Fällen zwischen dem 25. und dem 50. Perzentil den Wert 2, zwischen dem 50. und dem 75. Perzentil den Wert 3 und Fällen über dem 75. Perzentil den Wert 4 zuweisen. Jede Gruppe umfasst im Prinzip die gleiche Anzahl von Fällen. Da aber alle Fälle mit gleichem Wert derselben Gruppe zugeordnet werden und die Grenzen der Perzentile nicht immer genau dazu passen, kann es zu etwas unterschiedlichen Gruppengrößen kommen.

- ▷ Laden Sie z.B. ALLBUS90.SAV. Wählen Sie „Transformieren“, „Variablen kategorisieren“. Die Dialogbox „Variable kategorisieren“ erscheint. Um die Daten der Variable ALT in einer neuen Variablen im vier Kategorien einzuteilen:
- ▷ Übertragen Sie ALT aus der „Quellvariablenliste“ in die Liste „Kategorien erstellen für:“. Tragen Sie in das Feld „Anzahl der Kategorien“ die Zahl 4 ein. Bestätigen Sie mit „OK“. Im Datenblatt des Dateneditors erkennen Sie, dass eine neue Variable „NALT“ erstellt wurde, in der nur die Kategorien 1 bis 4 auftreten. Dabei steht 1 für das jüngste Viertel, 2 für das zweitjüngste Viertel der Befragten usw..

Abb. 5.19. Dialogbox „Variable kategorisieren“

6 Daten mit anderen Programmen austauschen

Datendateien können mit SPSS für Windows selbst erstellt, im SPSS Windows-Format gespeichert und wieder geladen werden. Man kann aber auch in anderen Programmen erstellte Datendateien in den Dateneditor von SPSS für Windows laden und verarbeiten. Die Datei wird dann innerhalb der Arbeitsdatei in das SPSS-Windows-Format umgewandelt. Bei Bedarf kann die neue Datei auch in diesem Format gespeichert werden. Umgekehrt kann SPSS für Windows Datendateien für die Weiterverarbeitung in anderen Programmen in deren Formate umwandeln und speichern. Das Einlesen und Ausgeben von Fremdformaten erfordert die Auswahl weniger Menüpunkte und ist weitgehend unproblematisch. Jedoch müssen insbesondere beim Einlesen von Daten mit Fremdformaten einige Dinge berücksichtigt werden, damit keine fehlerhaften Dateien entstehen. Übernommen werden können:

① Über die Befehlsfolge „Datei öffnen", „Daten":

- ❒ *SPSS-Dateien* aus anderen Betriebssystemen.
- ❒ Dateien des Statistikprogramms *SYSTAT*.
- ❒ Dateien des Statistikprogramms *SAS* (der verschiedensten Plattformen).
- ❒ Dateien aus *Tabellenkalkulationsprogrammen* (unmittelbar übernommen werden können Daten aus Lotus 1-2-3 [Versionen 2.0, 3.0 und 1A], Excel und aus Dateien, die das SYLK-Format benutzen wie Multiplan).
- ❒ Dateien des Datenbankprogramms dBase (Versionen II, IIIPlus, III und IV).
- ❒ *Textdateien-Dateien* und SPSS *Datendateien* als ASCII-Dateien.

② Über die Befehlsfolge „Datei", „Datenbank öffnen":

- ❒ Dateien aus *Datenbankprogrammen* (und Excel Version 5) können über die ODBC-Schnittstelle übernommen werden, wenn man über den entsprechenden Treiber für dieses Programm verfügt. (Viele Treiber werden auf der SPSS-CD mitgeliefert, andere bietet z.B. das Microsoft Data Access Pack.)

③ Über die Befehlsfolge „Datei", „Textdaten einlesen":

- ❒ *ASCII-Dateien*. (Dabei können verschiedene Trennzeichen für Variablen benutzt werden. Sind bestimmte Bedingungen eingehalten, kann man auch andere ASCII-Dateien verwenden.) Diese Befehlsfolge führt zu identischem Ergebnis wie mit der Auswahl des Typs „Text" in der Option „Datei", „Daten".

Da es sich bei den aufgeführten Tabellenkalkulations- und Datenbankprogrammen um Standardprogramme handelt, sind fast alle gängigen Programme in der Lage,

Daten in deren Formate zu exportieren. Daher ist die Übernahme von Daten aus anderen externen Programmen über den Umweg des Exports in Formate der aufgeführten Programme oder das ASCII-Format möglich. Das Programm selbst muss dazu nicht installiert sein. Es genügt, wenn die Datendatei in einem entsprechenden Format vorliegt.

6.1 Übernehmen von Daten aus Fremddateien

Außer bei der Benutzung Datenbank-Schnittstelle oder Übernahme von ASCII-Daten über die Option „Textdaten einlesen", gehen Sie zum Laden von Daten aus einer Datei in einem der zulässigen Formate wie folgt vor:

▷ Wählen Sie die Befehlsfolge „Datei", „Öffnen", „Daten". Es öffnet sich die Dialogbox „Datei öffnen" (⇨ Abb. 6.1).

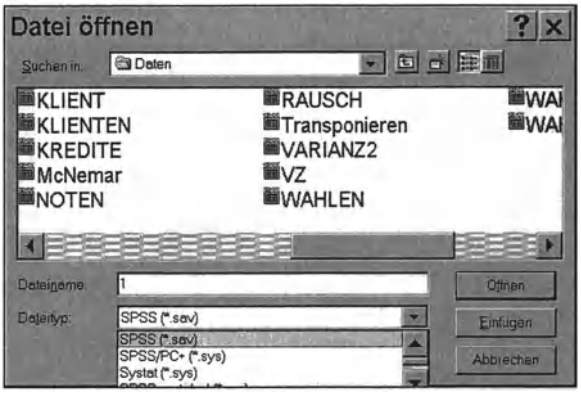


Abb. 6.1. Dialogbox „Datei öffnen" mit geöffneter Dateitypliste

▷ Wählen Sie im Feld „Suchen in:" zunächst das Laufwerk, in dem die gewünschte Datei steht.
▷ Wählen Sie dort weiter über die Auswahlliste das Verzeichnis, in dem die gewünschte Datei steht. Standardmäßig zeigt SPSS dann jeweils die Dateien mit der Extension SAV (SPSS-Windows-Dateien) an. (Wenn der richtige Dateityp ausgewählt ist, können Sie auch die Datei einschließlich Laufwerk und Verzeichnis in das Eingabefeld „Dateiname:" eintragen.)
▷ Öffnen Sie durch Anklicken des Pfeils am Auswahlfeld „Dateityp" die Liste der verfügbaren Dateiformate, und klicken Sie das gewünschte Format an. Im Dateiauswahlfeld erscheinen die Dateien mit der zu diesem Format zugehörigen Standardextension.

Standardextensionen sind: *SYS* (SPSS/PC+ und Systat), *POR* (SPSS PORTABLE), *XLS* (Excel), *W** (Lotus 1-2-3), *SLK* (SYLK für Multiplan und optional Excel-Dateien), *DBF* (dBASE), *TXT* (ASCII-Dateien), *DAT* (ASCII-Dateien mit Tabulator

als Trennzeichen) sowie *SAV* (SPSS für Windows und für UNIX). Dateien mit beliebiger Extension werden bei Auswahl von „Alle Dateien (*.*)" angezeigt. Sie können sich aber auch Dateien mit einer beliebigen anderen Extension anzeigen lassen. Tragen Sie dazu in das Eingabefeld „Dateiname:" „*.Extension" ein, und bestätigen Sie mit „Öffnen". *Beachten Sie:* Eine Datei muss das ausgewählte Format besitzen, aber nicht unbedingt die Standardextension im Namen haben. SPSS erkennt das Format auch nicht an der Extension.

- ▷ Wählen Sie die gewünschte Datei aus der Liste, oder tragen Sie den Dateinamen in das Eingabefeld „Dateiname:" ein und bestätigen Sie mit „Öffnen".
- ▷ Je nach Dateiart öffnet sich evtl. eine zusätzliche Dialogbox mit den Optionen „Variablennamen einlesen" und/oder „Bereich". Stellen Sie diese Optionen entsprechend ein.

6.1.1 Übernehmen von Daten mit SPSS Portable-Format

SPSS-Dateien, die mit der MacIntosh-, der Unix- oder einer Großrechnerversion erstellt wurden, können nicht unmittelbar eingelesen werden. Man muss sie zunächst in das SPSS Portable-Format exportieren. SPSS für Windows ist danach in der Lage, eine solche Datei zu importieren.

Beispiel. Die Daten des ALLBUS können von SPSS-Nutzern vom Zentralarchiv für empirische Sozialforschung in Köln als SPSS-Exportdatei erworben werden. Für den ALLBUS des Jahres 1990 hat diese den Namen S1800.EXP. (Beachten Sie, dass der Name nicht die Standardextension POR hat. Andere SPSS-Versionen benutzen im übrigen als Standardextension für portable Dateien EXP.) Sie stehe im Verzeichnis C:\ALLBUS\ALLBUS90. Um diese Datei zu importieren, wäre wie folgt vorzugehen:

- ▷ Wählen der Befehlsfolge „Datei", „Öffnen", „Daten".
- ▷ Auswählen von Laufwerk und Verzeichnis (hier C:\ALLBUS\ALLBUS90).
- ▷ Auswahl des Dateityps „SPSS portable".
- ▷ Eingabe des Dateinamens „S1800.EXP" in das Eingabefeld „Dateiname:" oder: Auswahl des Dateityps „Alle Dateien (*.*)" und Auswahl von „S1800.EXP"aus der Dateiliste. Bestätigen mit „Öffnen".

Hinweis. Wird eine SPSS/PC+-Datei importiert, die in Stringvariablen in Windows-Programmen nicht verfügbare Sonderzeichen benutzt, müssen diese umgewandelt werden. Dies geschieht automatisch beim Import, funktioniert aber dann nicht immer fehlerfrei, wenn der Zeichensatz der bei Erstellung der Datei vorhandene DOS-Version nicht identisch ist mit der bei der Installation von SPSS für Windows benutzten.

6.1.2 Übernehmen von Daten aus einem Tabellenkalkulationsprogramm

Beispiel. Die Daten einer Schuldenberatungsstelle über überschuldete Verbraucher sind in einer Excel-Datei VZ.XLS gespeichert. Zeilen enthalten die Fälle, Spalten die Variablen. In den Zeilen 1 und 2 stehen Überschriften (⇨ Abb. 6.2). Die Daten sollen in SPSS weiterverarbeitet werden. Übernommen werden sollen die ersten zehn Fälle (Zeile 3 bis 12). Die Überschriften in Zeile 2 werden als Variablennamen benutzt.

Um diese Datei zu importieren, gehen Sie wie folgt vor:

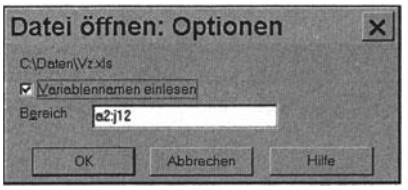
- ▷ Wählen Sie die Befehlsfolge „Datei", „Öffnen", „Daten". Die Dialogbox „Datei öffnen" erscheint (⇨ Abb. 2.5).
- ▷ Wählen Sie das gewünschte Laufwerk und Verzeichnis (hier: C:\DATEN).
- ▷ Wählen Sie den Datentyp „Excel", und wählen Sie die Datei aus der Liste aus oder geben Sie den Dateinamen ein (hier: VZ.XLS).

	A	B	C	D	E	F	G	H	I	J
1		Erstkontakt						Beg.	Übersch.	
2	Nr.	Tag	Monat	Jahr	Vorname	Geschl.	Eink.	Jahr	Monat	Ges. Schuld
3	1	17	10	89	Frederic	2	1200	10	86	6500
4	2	9	1	89	Birgid	3	1798	11	82	4600
5	3	1	2	88	Ronald	1	2050	1	88	24700
6	4	8	6	89	Gertrud	3	2000	11	80	163000
7	5	17	7	89	Carola	1	9999	0	0	999999
8	6	1	9	88	Alfred	1	1950	7	82	33200
9	6	6	11	87	Manfred	2	1800	7	86	32000
10	7	21	7	89	Jürgen	1	1750	12	81	14500
11	8	5	11	88	Hildegard	3	1050	2	83	9086
12	9	28	1	88	Tom	2	1400	10	87	44740

Anmerkung. Vorname bezieht sich hier auf Schuldner, Geschlecht auf Ratsuchende, Geschlecht = 3 bedeutet, dass ein Paar gemeinsam die Beratungsstelle aufsuchte.

Abb. 6.2. Excel-Datei VZ.XLS

- ▷ Klicken Sie auf „Öffnen". Es öffnet sich die Dialogbox „Datei öffnen: Optionen". (Die Auswahl von „Einfügen" hat die gleiche Wirkung.)
- ▷ Klicken sie auf das Kontrollkästchen „Variablennamen einlesen".
- ▷ Geben Sie den Zellenbereich der Excel-Datei (hier: a2 [linke obere Ecke] bis j12 [rechte untere Ecke]) ein und bestätigen Sie mit „OK".



Die Daten erscheinen im SPSS-Dateneditor als Datei unter dem Namen UNBENANNT. Die Variablennamen entsprechen den Spaltenüberschriften. Da Jahr und Monat doppelt auftreten, werden die Variablennamen beim zweiten Auftreten durch die SPSS-Standardvariablennamen V8 und V9 ersetzt.

Die Option „Variablennamen lesen" steht nur für Excel-, Sylk-, Lotus-, und Tab-delimited (d.h., den Tabulator als Trennzeichen nutzende ASCII-)Dateien zur Verfügung. Die erste Zeile der Datei (oder des vom Benutzer definierten Zellenbereichs) wird als Variablennamen interpretiert. Namen von mehr als acht Zeichen Länge werden abgeschnitten, nicht eindeutige Namen modifiziert. Mit dieser Option kann man sich die Definition von Variablennamen ersparen. Zugleich verhin-

dert sie, dass die Datenformate nach dem Wert in der ersten Zeile definiert werden. Verwendung findet dann der Wert in der zweiten Zeile.
Die Option „Bereich" steht für Lotus-, Excel- und Sylk-Dateien zur Verfügung, nicht aber für ASCII-Dateien. Dateien von Excel 5 oder Nachfolgeversionen können mehrere Arbeitsblätter enthalten. In der Standardeinstellung liest der Daten-Editor das erste Arbeitsblatt. Wenn Sie ein anderes Arbeitsblatt einlesen möchten, wählen Sie es aus der Drop-Down-Liste aus.

Um eine fehlerhafte Datenübernahme zu verhindern, müssen die Regeln beachtet werden, nach denen SPSS Daten aus Tabellenkalkulationsblättern übernimmt. Generell liest SPSS Daten aus Tabellenkalkulationsprogrammen wie folgt:

Aus der Tabelle wird ein rechteckiger Bereich, der durch die Bereichsgrenzen festgesetzt ist, als SPSS-Datenmatrix gelesen. Die Koordinatenangaben für den Zellenbereich variieren nach Ausgangsformaten. *Beispiel:* Lotus (A1..J10), Excel (A1:J10) und Sylk (R1C1:R10C10). Zeilen werden Fälle, Spalten Variablen (sollte dies der Datenstruktur nicht entsprechen, muss die Matrix später gedreht werden ⇨ Kap. 7.1.2). Enthält eine Zelle innerhalb der Bereichsgrenzen keinen gültigen Wert, wird ein System-Missing gesetzt. Verzichtet man auf die Angabe von Bereichsgrenzen, ermittelt SPSS diese selbständig. Dies sollte man jedoch nur bei Tabellen ohne Beschriftung verwenden. Die Übernahme von Spalten unterscheidet sich danach, ob Spaltenüberschriften als Variablennamen gelesen werden oder nicht. Werden Spaltenüberschriften als Variablennamen verwendet, nimmt SPSS nur solche Spalten auf, die mit einer Überschrift versehen sind. Die letzte Spalte ist die letzte, die eine Überschrift enthält. Werden keine Überschriften verwendet, vergibt SPSS selbständig Variablennamen. Je nach Herkunftsformat sind sie identisch mit dem Spaltenbuchstaben oder mit der Spaltennummer mit einem vorangestellten C. Die letzte übernommene Spalte ist dann diejenige, die als letzte mindestens eine ausgefüllte Zelle enthält. Die Zahl der übernommenen Fälle ergibt sich aus der letzten Zeile, die mindestens eine ausgefüllte Zelle innerhalb der Spaltenbegrenzung enthält. Der Datentyp und die Breite der Variablen ergeben sich in beiden Fällen aus der Spaltenbreite und dem Datentyp der ersten Zelle der Spalte, falls Variablennamen gelesen werden, der zweiten Zelle. Werte mit anderem Datentyp werden in System-Missings umgewandelt. Leerzeichen sind bei numerischen Variablen System-Missings, bei Stringvariablen dagegen ein gültiger Wert.

Fehler können vor allem aus folgenden Quellen stammen:

- Der Datentyp wechselt innerhalb der Spalte. Das führt zu unerwünschten Missing-Werten.
- Leerzeilen, die aus optischen Gründen im Kalkulationsblatt enthalten sind, werden als Missing-Werte interpretiert.
- Sind nicht alle Spalten mit Überschriften versehen, werden Variablen evtl. unerwünschterweise nicht mit übernommen.
- Bei Import aus DOS-Programmen werden in String-Variablen enthaltene Sonderzeichen nicht mit übernommen.

Passen Sie vor der Übernahme die Kalkulationsblattdaten den Regeln entsprechend an, damit keine Fehler auftreten.

6.1.3 Übernehmen von Daten aus einem Datenbankprogramm

6.1.3.1 Übernehmen aus dBASE-Dateien

SPSS für Windows verfügt über eine Option zum Lesen von Daten aus dem Datenbankklassiker.

DBASE-Dateien werden ähnlich wie Tabellenkalkulationsdateien übernommen. Die Option befindet sich daher auch in demselben Untermenü. Zur Übernahme von dBase-Daten gehen Sie wie folgt vor:

- ▷ Wählen Sie „Datei", „Öffnen", „Daten".
- ▷ Wählen Sie das gewünschte Verzeichnis.
- ▷ Wählen Sie den Dateityp „dBASE".
- ▷ Wählen Sie in der Dateiliste die gewünschte Datei aus, oder geben Sie in das Feld „Name:" den gewünschten Namen ein. Und bestätigen Sie mit „Öffnen".

Die Daten werden gelesen und automatisch übernommen. Dabei ist folgendes zu beachten: Feldnamen werden automatisch in SPSS-Variablennamen übersetzt. Sie sollten daher der SPSS-Konvention über Variablennamen entsprechen. Feldnamen von mehr als acht Zeichen Länge schneidet das Programm ab. Achtung: Entsteht dadurch ein mit einem früheren Feld identischer Name, so wird das Feld ausgelassen. Doppelpunkte im Feldnamen werden zu Unterstreichungen. In dBASE zum Löschen markierte, aber nicht gelöschte Fälle werden übernommen. Es wird jedoch eine Stringvariable D_R erstellt, in der diese Fälle durch einen Stern gekennzeichnet sind. Umlaute können nicht erkannt werden. Deshalb kann es sinnvoll sein, vor dem Import erst entsprechende Änderungen vorzunehmen. *Hinweis:* dBASE-Daten können auch über die Option „Datenbank öffnen" gelesen werden. Dann ist es möglich, Variablen und Fälle zu selektieren.

6.1.3.2 Übernehmen über die Option „Datenbank öffnen"

Jede Datenbank, bei der ODBC-Treiber (Open Database Connectivity) verwendet werden, kann direkt von SPSS eingelesen werden, wenn ein entsprechender Treiber installiert ist (solche liefert z.B. SPSS selbst auf der Installations-CD oder z.B. Microsoft). Bei lokaler Analyse muss der jeweilige Treiber auf dem lokalen PC installiert sein (bei verteilter in der Netzwerkversion, auf die wir hier nicht eingehen, auf dem Remote-Server). Zum Laden der Datenbankdateien steht das Menü „Datenbank einlesen" zur Verfügung. (Es ist auch zur Übernahme von Daten aus der Excel Version 5 geeignet.) Das Öffnen der Datenbankdateien wird von einem Datenbank-Assistenten unterstützt und verläuft in 5 (beim Laden einer Tabelle) oder 6 Schritten (beim Laden mehrerer Tabellen).

Beispiel. Eine Microsoft Access Datenbank-Datei mit Namen VZ.MDB befindet sich im Verzeichnis C:\DATEN. Sie enthält dieselben Daten wie die bisher verwendete Schuldnerdatei. Die Access-Eingabemaske mit den Daten des Falles 1 sehen Sie in Abb. 6.3. Diese Daten sollen in SPSS für Windows importiert werden. Die zwei Variablen TAG für den Tag des Erstkontaktes und GESCHL für Geschlecht des Ratsuchenden sollen nicht interessieren und werden daher nicht übernommen. Ausgeschlossen werden sollen auch Fälle ohne eigenes Einkommen (in solchen Fällen wurde in der Variablen EINK den Wert 9999 eingetragen).

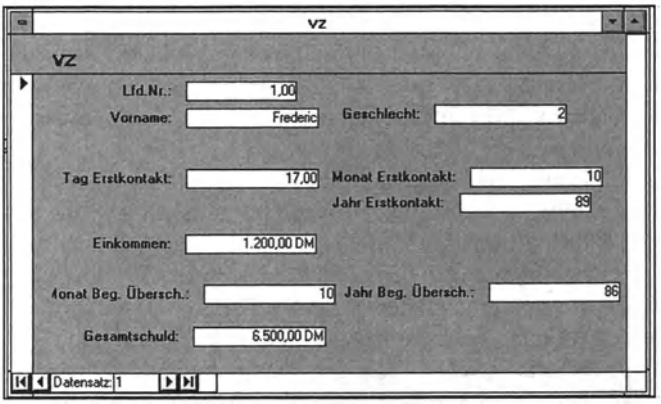


Abb. 6.3. Beispiel einer ACCESS-Eingabemaske

Um diese Daten in SPSS einzulesen, gehen Sie wie folgt vor:

- ▷ Wählen Sie „Datei", „Datenbank öffnen". Es öffnet sich eine Auswahlliste mit den Optionen „Neue Abfrage", „Abfrage bearbeiten", „Abfrage ausführen". Mit den letzten beiden Optionen werden früher durchgeführte und gespeicherte Abfragen bearbeitet und wiederholt.
- ▷ Wählen Sie die gewünschte Option (im Beispiel „Neue Abfrage"). Es öffnet sich die Dialogbox „Datenbankassistent" (⇨ Abb. 6.4). Dort sind die verfügbaren Quellen, d.h. Datenbanken samt zugehörigem Treiber, aufgeführt. (Sollte für die von Ihnen benötigte Datenbank noch kein Treiber installiert sein, müssen Sie dies zunächst nachvollziehen, indem Sie z.B. das Microsoft Data Access Pac von der entsprechenden CD aus starten.)

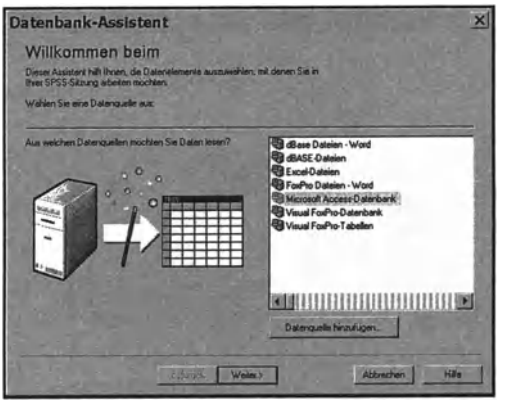


Abb. 6.4. Dialogbox „Datenbank-Assistent"

- ▷ Markieren Sie dort in der Liste die benötigte Datenquelle (im Beispiel „Microsoft Access-Datenbank") und klicken Sie auf die Schaltfläche „Weiter". Wenn

Sie keine bestimmt Datei mit der Quelle verbunden haben, öffnet sich die Dialogbox „Anmeldung des ODBC-Treibers". (Diese sieht je nachdem, welches Datenbankprogramm Sie verwenden, z.T. unterschiedlich aus.) Hier müssen Sie zumindest die Datei eingeben, die geöffnet werden soll. Sie können entweder Pfad und Dateiname eintragen oder durch Anklicken der Schaltfläche „Durchsuchen" die Dialogbox „Datei öffnen" nutzen.

- ▷ Wählen Sie dort auf die übliche Weise im Auswahlfeld „Suchen in" das gewünschte Laufwerk und Verzeichnis aus, und übertragen Sie aus der Auswahlliste den Namen der gewünschten Datei Eingabefeld „Dateiname". (Wenn bei der Datenbank ein Paßwort erforderlich ist oder das Netzwerk weitere Angaben erfordert, werden diese in weiteren Feldern oder Dialogboxen abgefragt.)
- ▷ Bestätigen Sie mit „Öffnen" und „OK". Es erscheint die Dialogbox „Daten auswählen" (⇨ Abb. 6.5). (Wenn man eine bestimmte Datenbank als Quelle definiert hat, erscheint diese Dialogbox sofort.) In ihr kann man sowohl die gewünschte Tabelle als auch die gewünschten Felder innerhalb dieser Tabelle auswählen.

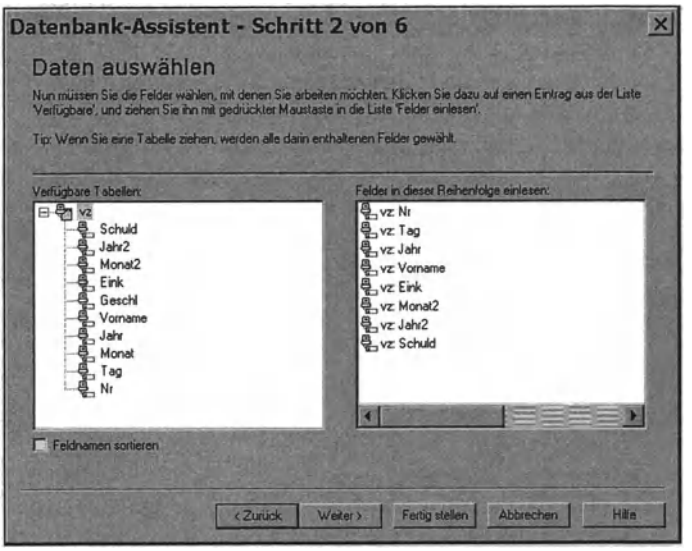


Abb. 6.5. Dialogbox „Daten auswählen"

- ▷ Zur Auswahl der Tabelle markieren Sie in der Auswahlliste „verfügbare Tabellen" die gewünschte Tabelle.
- ▷ Felder können auf unterschiedliche Weise ausgewählt werden. Doppelklicken auf den Namen der Tabelle überträgt unmittelbar sämtliche Felder dieser Tabelle in die Liste „Felder in dieser Reihenfolge einlesen". Aus dieser Liste kann man, durch Anklicken und Ziehen in die Liste „Verfügbare Tabellen" oder durch Doppelklick auf ihren Namen, Felder entfernen. Beim zweiten Verfahren klickt man auf das +-Zeichen vor der ausgewählten Tabelle. Dann werden

sämtliche Felder dieser Tabelle in der Liste „Verfügbare Tabellen angezeigt" (ist das Kontrollkästchen „Feldnamen sortieren" angewählt, in alphabetischer Folge, sonst in der Reihenfolge der Eingabe). Man kann diese durch Anklicken und Ziehen oder durch Doppelklick auf den Namen in beliebiger Reihenfolge in die Liste „Felder in dieser Reihenfolge einlesen" übertragen.

Sollen spezielle Fälle ausgewählt werden:

- Klicken Sie auf die Schaltfläche „Weiter." Die Dialogbox „Beschränkung der gelesenen Fälle" (⇨ Abb. 6.6) öffnet sich. Formulieren Sie darin die Auswahlbedingung. Dazu stellen Sie die Bedingung(en) in den seitlichen Feldern „Kriterien" zusammen. In unserem Beispiel sollen alle Fälle mit einem Einkommen unter dem Wert 9999 ausgewählt werden. Wir übertragen deshalb zunächst den Variablennamen EINK in das Feld „Ausdruck 1". Das geschieht durch Markieren des Feldes. Es erscheint dann ein Pfeil an der Seite des Feldes. Klicken Sie auf diesen Pfeil und wählen Sie den Variablennamen in der sich dann öffnenden Auswahlliste. Daraufhin geben Sie „<" in das Feld „Relation" ein. Dies geschieht auf gleiche Weise. Dann schreiben wir „9999" in das Feld „Ausdruck 2".
- Durch Anklicken von „Fertig stellen" laden wir die Datei. (Hätten wir keine Fälle ausgewählt, hätte auch schon im Dialogfenster „Daten auswählen" durch Anklicken von „Fertigstellen" die Datei geladen werden können. Umgekehrt könnten durch Klicken von „Weiter" zwei weitere Schritte eingeleitet werden.)

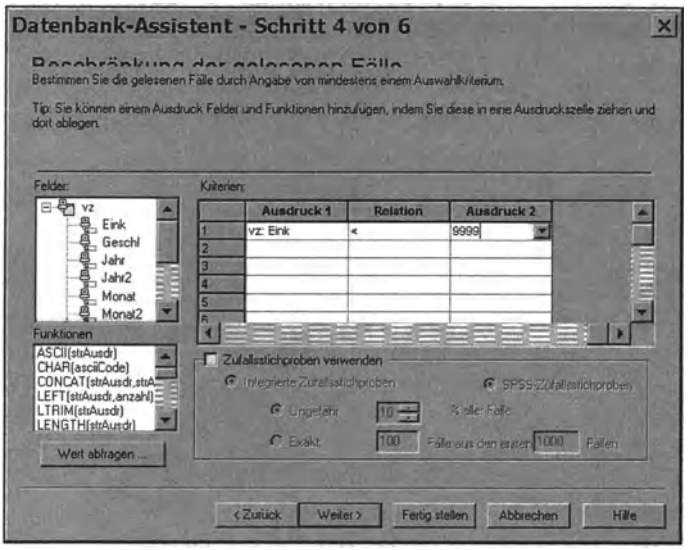


Abb. 6.6. Dialogbox „Beschränkung der gelesenen Fälle"

SPSS übernimmt die ersten 8 Zeichen der Bezeichnung eines Datenbankfeldes als Variablennamen, wenn sie mit den SPSS-Konventionen für Variablennamen entsprechen, ansonsten erstellt SPSS automatisch einen gültigen Namen. Die Be-

zeichnung eines Datenbankfeldes wird in jedem Falle als Variablenlabel übernommen.

Zur Bildung von Bedingungsfunktionen stehen weitere Möglichkeiten zur Verfügung:

- ❒ Zur Bildung der Bedingungen steht eine Liste von *Funktionen* in einem Auswahlfeld „Funktionen" zur Verfügung. Es handelt sich um arithmetische, logische und Stringfunktionen sowie Zeit- und Datumsfunktionen.
- ❒ Die Bedingung kann in den Feldern *„Abfragen"* enthalten sein. D.h., der Nutzer wird während der Ausführung des Datenbankzugriffs nach Werten gefragt. Dadurch kann die Abfrage variabel gehalten werden. In unserem Beispiel könnte man es etwas offen halten, wie groß das Einkommen sein soll, unter dem die Fälle in die Analyse einbezogen werden. Man würde dann im Ausdruck 2 statt des Werte 9999 eine Abfrage eintragen.

Dazu verfahren Sie wie folgt:

- ▷ Markieren Sie „Ausdruck 2". Klicken Sie auf die Schaltfläche „Wert abfragen...". Es öffnet sich die Dialogbox „Wert abfragen" (⇨ Abb. 6.7).
- ▷ Geben Sie in das Feld „Aufforderungstext" einen geeigneten Text ein (Voreinstellung „Enter value:„).
- ▷ Geben Sie in das Feld „Standardwert" einen Wert ein, der am häufigsten verwendet wird und deshalb als Option zuerst angezeigt werden soll.
- ▷ Geben Sie gegebenenfalls durch Anklicken von „Auswahl aus Liste durch Benutzer" und Eingabe weiterer Werte eine Liste von Werten ein, aus denen der Benutzer auswählen kann (der Standardwert muss in ihr enthalten sein).
- ▷ Stellen Sie bei „Datentyp" den richtigen Datentyp „String" (Zeichenkette) oder „Number" (numerisch) ein. Bestätigen Sie mit „OK". Sie werden in Zukunft beim Ausführen einer Abfrage aufgefordert, einen entsprechenden Wert einzugeben.

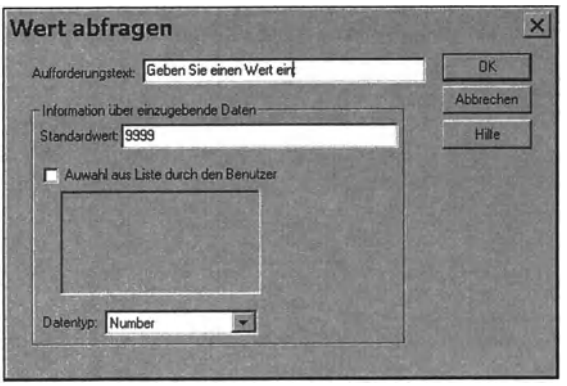


Abb. 6.7. Dialogbox „Wert abfragen"

❒ Wenn gewünscht, kann aus den Daten auch nur eine *Zufallsstichprobe* gezogen werden. Dazu markieren Sie das Auswahlkästchen „Zufallsstichprobe". Falls die Datenbank selbst über eine Option zum Ziehen von Zufallsstichproben verfügt, wird die Optionsschaltfläche „Integrierte Stichproben" aktiv. In diesem Fall können Sie zwischen einer im Datenbankprogramm selbst gezogenen Zufallsstichprobe und einer „SPSS-Stichproben" wählen. Ansonsten ist die Optionsschaltfläche für die „SPSS-Stichproben" automatisch markiert.

- *Ungefähr.* Markieren dieser Option und Eingabe einer Prozentzahl zwischen 1 und 100 führt zu einer Zufallstichprobe der angegebenen Größenordnung.
- *Exakt.* Durch Auswahl dieser Option und Angabe eines genauen Zahlenwertes bewirkt man die Ziehung einer Stichprobe in der exakt angegebenen Größe. Die Ziehung geschieht aus den ersten x Fällen. In einem zweiten Kästchen muss ein Wert x größer als die Zahl der auszuwählenden Fälle eingetragen werden.

Die zwei möglichen weiteren Schritte im Datenbank-Assistent bewirken folgendes:

❒ Zunächst kann ein Fenster „Werte definieren" geöffnet werden. In diesem die können Variablennamen geändert werden. Außerdem ist es möglich Stringvariablen (hier als alphabetische Variablen bezeichnet) in numerische umzuwandeln und dabei die ursprünglichen Werte als Labels zu verwenden. Dazu muss bei der entsprechenden Variablen ein Kontrollkästchen „Wertelabels" aktiviert werden.

❒ In einem weiteren Schritt kann das Ergebnis des Auswahlprozesses als Syntax in eine Dialogbox „Ergebnisse" übertragen werden. Dort kann dann entweder die Datei geladen oder die Syntax zur weiteren Bearbeitung in ein Syntaxfenster übertragen werden. Oder aber die Abfrage wird gespeichert. (Die Datei hat die Extension „spq".) Sie kann dann jederzeit mit der Befehlsfolge „Datei", „Datenbank öffnen", „Abfrage ausführen" aufgerufen oder mit „Abfrage bearbeiten" in ein Syntaxfenster geladen, dort bearbeitet und ausgeführt werden. Schließlich bewirkt die Auswahl des Kontrollkästchens „Daten zwischenspeichern", dass eine temporäre Datei auf der Festplatte eingerichtet wird, in der sich die Daten während der Sitzung befinden. Dadurch kann bei großen Dateien ein Beschleunigung der Bearbeitungsgeschwindigkeit erreicht werden.

Übernahme von Daten aus mehreren Tabellen. Enthält eine Datenbank mehrere Tabellen, die gemeinsame Primärschlüssel besitzen, können diese Tabellen verknüpft und kombiniert ausgewertet werden.

Beispiel. Eine Access Datenbank „Schulden" im Verzeichnis „c:\Daten" enthält 3 Tabellen. In der ersten (KUNDEN) sind die Adressen der Schuldner samt Personennummer (PERSNR) als Primärschlüssel enthalten. Die zweite (BANKEN) enthält die Angaben zu den Banken mit Banknummer (BANKNR) als Primärschlüssel. Eine dritte Tabelle (KREDITE) enthält Kreditdaten und die Personennummer des jeweiligen Kreditnehmers, die Banknummer der jeweiligen Bank sowie als Primärschlüssel eine Kreditnummer. Eine Person kann mehrere Kredite bei mehreren Banken haben. Man kann daraus *eine* SPSS-Datendatei bilden, in der alle Daten enthalten sind. Dabei wird aus jedem Kredit ein Fall. Den Kreditdaten werden die dazugehörigen Personen und Bankdaten zugeordnet.

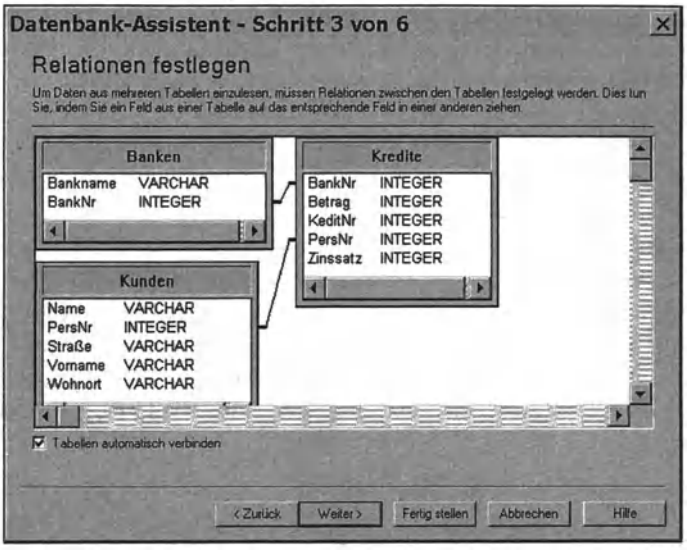


Abb. 6.8. „Datenbank-Assistent: Relationen Festlegen" bei mehreren Tabellen

Um diese verbunden als SPSS-Datendatei zu laden, gehen Sie wie oben beschrieben vor. Wählen Sie im ersten Schritt einfach „Microsoft Access-Datenbank" als Quelle. Im Fenster „Anmeldung des ODBC-Treibers" müssen Sie „C:\DATEN\SCHULDEN.MDB" eintragen bzw. über die Dialogbox „Datei öffnen" auswählen. Im zweiten Schritt stehen dann in der Dialogbox „Daten auswählen" alle drei Tabellen im Feld „verfügbare Tabellen". Aus allen dreien übertragen sie alle Felder (zumindest aber einige, insbesondere die Schlüsselfelder) in das Fenster „Felder in dieser Reihenfolge einlesen". Klickt man jetzt auf die Schaltfläche „Weiter", erscheint die Dialogbox „Relationen festlegen" (⇨ Abb. 6.8). Hier werden in drei Kästen die ausgewählten Felder der drei Tabellen angezeigt. Über Primärschlüssel verbundene Felder sind durch eine Linie verbunden. So führt in die Datei Kredite eine Verbindung aus „Banken" über „BankNr" und aus „Kunden" über „PersNr". Diese Verbindungen sind automatisch erstellt worden. Man kann diese Verbindung aufheben, indem man die Linie markiert und auf die Taste „Entfernen" drückt. (Automatische Verbindungen werden auch aufgehoben, wenn man die Markierung des Auswahlkästchens „Tabelle automatisch verbinden" aufhebt.) Durch Ziehen von einem Feld der einen Tabelle zu einem der anderen kann man eine neue Verbindung definieren, sofern die Felder vom selben Typ sind. Bei mehr als zwei Tabellen sind nur *„innere Verbindungen"* zulässig. Bei solchen Verbindungen werden nur solche Zeilen (Datensätze) der Tabellen übernommen, bei denen die Werte der verbundenen Zellen der verbundenen Tabellen übereinstimmen. *„Äußere (linke oder rechte) Verbindungen"* dagegen benutzen alle Datensätze der einen (linken oder rechten) Tabelle, aber nur die Datensätze der anderen Tabelle, bei denen die Werte der verbundenen Zelle übereinstimmen. In diesem Falle müssen Sie bei jeder Verbindung die Art der Verbindung zusätzlich definieren. (Dies geschieht in einem weiteren Dialogfeld „Eigenschaften der Beziehung".

Diese öffnet man durch Doppelklicken auf die jeweilige Verbindungslinie.) Durch Anklicken von „Fertig stellen" erzeugen Sie eine SPSS-Datendatei.

Hinweis. Excel 5 Dateien lassen sich auch über die ODBC-Schnittstelle einlesen. Dazu muss aber vorher für den Zellenbereich, in dem sich die Daten befinden, ein Name definiert sein.

6.1.4 Übernehmen von Daten aus ASCII-Dateien

Viele Datenbank-, Tabellenkalkulations- und Textverarbeitungsprogramme bieten auch Möglichkeiten, die Daten im ASCII-Format auszugeben. Dies ist eine Möglichkeit, auf einem Umweg auch Daten aus Programmen mit nicht kompatiblem Format in SPSS zu importieren. Man sollte davon aber nur Gebrauch machen, wenn die oben beschriebenen Möglichkeiten nicht bestehen. In der Textdatei selbst können die Daten in verschiedenem Format vorliegen:

- Durch *Trennzeichen* strukturierte Datei. In diesem Fall zeigen Trennzeichen (z.B. Tabulator, Kommata, Leerzeichen) an, wo eine Variable endet und damit eine neue beginnt. Zusätzlich beginnt jeder neue Fall in einer neuen Datenzeile. (Durch Trennzeichen strukturierte Dateien, bei denen ein Fall mehr als eine Zeile einnimmt, müssen wie Dateien im freien Format behandelt werden.)
- Datei mit *festem Format.* Hier stehen die Werte einer bestimmten Variablen bei allen Fällen immer an derselben Stelle einer Zeile.
- Datei mit *freiem Format.* Bei diesem Format werden die Variablen ebenfalls durch Trennzeichen gekennzeichnet. Allerdings können die Fälle unmittelbar aneinander anschließend gespeichert werden. Damit das Programm erkennen kann, wo ein neuer Fall beginnt, muss ihm mitgeteilt werden, wieviele Variablen ein Fall enthält. Es zählt dann die Variablen mit und erkennt nach Beendigung der letzten Variablen des ersten Falles die nächste Variable als erste des zweiten Falles usw.

In allen drei Fällen werden die Daten in 6 Schritten unter Anleitung des „Assistenten für Textimport" durchgeführt. Je nach Datenformat unterscheiden sich die Eingaben bei bestimmten Schritten. Der gesamte Ablauf wird im folgenden für eine durch Tabulatorzeichen als Trennzeichen strukturierte Textdatei dargestellt. Für die andren Varianten folgt dann eine Erörterung der differierenden Schritt.

ASCII-Dateien mit Trennzeichen. Abb. 6.9 zeigt die Daten der Schuldenberatung als ASCII-Datei mit Tabulator als Trennzeichen (tab-delimited). Diese kann über die Befehlsfolge „Datei", „Textdaten einlesen" in der oben beschriebenen Weise geöffnet werden. (Die Befehlsfolge „Datei", „Öffnen", „Daten" hat denselben Effekt, wenn Sie in dem sich öffnenden Fenster „Datei öffnen" je nach Extension der gewünschten Datei den Dateityp „Text" oder „Daten" wählen.) Sie wählen in der Dialogbox „Datei öffnen" in der üblichen Weise Verzeichnis und Namen (in unserem Beispiel heißt sie VZ.TXT) der zu öffnenden Datei und klicken auf „Öffnen". Die Dialogbox „Assistent für Textimport – Schritt 1 von 6" erscheint (⇨ Abb. 6.10).

```
NR→TAG→MONAT→JAHR→VORNAME→GESCHL→EINK→MONAT1→JAHR1→GES.SCHU¶
1→17→10→89→Frederic→2→1200→10→86→6500¶
2→9→1→89→Birgid→3→1798→11→82→4600¶
3→1→2→88→Ronald→1→2050→1→88→24700¶
4→8→6→89→Gertrud→3→2000→11→80→163000¶
5→17→7→89→Carola→1→9999→0→0→999999¶
6→1→9→88→Alfred→1→1950→7→82→33200¶
6→6→11→87→Manfred→2→1800→7→86→32000¶
7→21→7→89→Jürgen→1→1750→12→81→14500¶
8→5→11→88→Hildegard→3→1050→2→83→9086¶
9→28→1→88→Tom→2→1400→10→87→44740¶
```

Abb. 6.9. Tab-delimited ASCII-Datei VZ.TXT

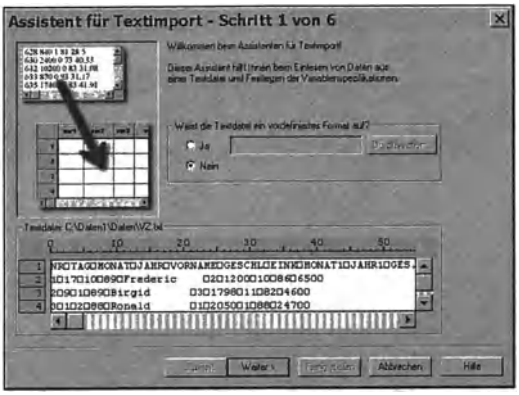


Abb. 6.10. Dialogbox „Assistent für Textimport – Schritt 1 von 6"

Diese enthält wie alle folgenden Dialogboxen ein Feld, in dem der Beginn der Datendatei bei derzeitigen Bearbeitungsstand zu erkennen ist. Ansonsten im Feld „Weist Textdatei ein vordefiniertes Format auf?" die Optionsschalter „Ja" und „Nein". Beim erstmaligen Einlesen einer Textdatei ist hier „Nein" zutreffend. (Um nicht jedes Mal beim Einlesen einer Textdatei das Format erneut bestimmen zu müssen, kann man am Ende eines Einlesevorganges das definierte Format speichern und bei späteren Einlesevorgängen verwenden. Ist dies geschehen, wäre hier „Ja" zu wählen.) Nach Anklicken der Schaltfläche „Weiter" erscheint die Dialogbox für den 2ten Schritt (⇨ Abb. 6.11).

Abb. 6.11. Dialogbox „Assistent für Textimport – Schritt 2 von 6"

In dieser wird im Feld „Wie sind die Variablen angeordnet?" mitgeteilt, ob es sich um durch Trennzeichen strukturierte Daten handelt bzw. Daten in freiem Format - in beiden Fällen ist die Optionsschaltfläche „Mit Trennzeichen" zu wählen – oder um Daten im festem Format – dann wäre „Festes Format" zu wählen. (Im Beispiel ist „Mit Trennzeichen" zutreffend.)

Außerdem ist im Bereich „Enthält die erste Zeile der Datei die Variablennamen?" anzugeben, ob dies der Fall ist oder nicht. (In unserem Beispiel ist dies der Fall, denn in der ersten Zeile stehen die Namen „NR. „TAG", „MONAT" etc.. Deshalb wird die Option „Ja" ausgewählt. Dadurch werden die Eintragungen der ersten Zeile zu Variablennamen [evtl. gekürzt und angepasst].) Nach Anklicken der Schaltfläche „Weiter" erscheint die Dialogbox für den 3ten Schritt. (⇨ Abb. 6.12).

Abb. 6.12. Dialogbox „Assistent für Textimport – Schritt 3 von 6"

In einem Auswahlkästchen ist zunächst anzugeben, in welcher Zeile der Textdatei der erste Fall beginnt. In unserem Beispiel ist die 2te Zeile, da sich in der ersten die

Datennamen stehen. Als nächstes wird zwischen den Optionsschaltern „Jede Zeile stellt einen Fall dar" und „Folgende Anzahl von Variablen stellt einen Fall dar" gewählt. Der erste Schalter trifft in unserem Beispiel zu. Er gilt für durch Trennzeichen strukturierte Daten zu. Der zweite Schalter dagegen ist bei „freiem Format" gültig. Weiter kann ausgewählt werden, ob alle Fälle oder nur ein bestimmter Teil eingelesen werden (letzteres wird man bei sehr großen Dateien für Testläufe nutzen). Soll nur ein Teil eingelesen werden, kann man entweder die ersten x Fälle (wobei x eine ganze Zahl kleiner n) wählen oder eine Zufallsauswahl der Fälle treffen lassen, die ungefähr einem einzugebenden Prozentsatz entspricht. Mit „Weiter" gelangt man in die Dialogbox zu Schritt 4 (⇨ Abb. 6.13).

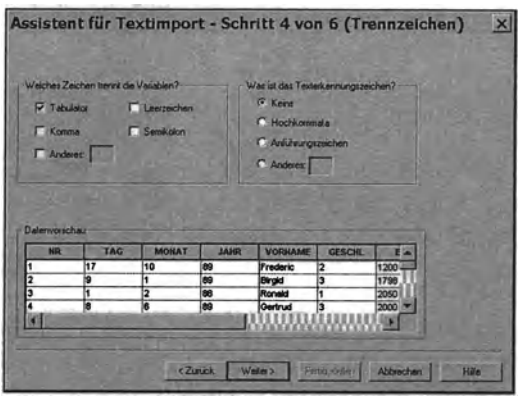


Abb. 6.13. Dialogbox „Assistent für Textimport – Schritt 4 von 6"

Hier gibt man an, welches „Trennzeichen" verwendet wird und gegebenenfalls, welches „Texterkennungszeichen" Verwendung findet. Texterkennungszeichen benötigt man, wenn das Trennungszeichen auch in Variablenwerten auftritt. Z.B. „-„ sei Trennungszeichen, kann aber auch in einer String-Variablen bei den Werten auftreten, etwas dem Namen „Meier-Müller". Dann würde das Programm die Daten falsch einlesen, wenn nicht durch ein Texterkennungszeichen (z.B. Hochkommata) gekennzeichnet ist, dass „-„ im Namen Meier-Müller kein Variablentrennzeichen ist, sondern Bestandteil des Wertes. (Entsprechend müssen die Textdateien vor dem Einlesen evtl. überarbeitet werden.)

In der Dialogbox zu Schritt 5 (⇨ Abb. 6.14).sind die Daten schon gemäß der bisherigen Angaben formatiert. Man kann hier noch die Variablendefinition bearbeiten. Dazu markiert man die jeweils umzudefinierende Variable. In den Eingabe- und Auswahlfeldern erscheint die derzeitige Definition. Im Feld „Variablennamen" kann man den Namen ändern. Im Feld „Datenformat" kann der Typ geändert werden. SPSS erkennt automatisch numerische und Stingvariablen, weshalb sich häufig eine Umdefinition erübrigt.

Verfügbare Formate. Die Daten müssen in der ASCII-Datei einem der folgenden Formate entsprechen. Sie werden dann in ein entsprechendes SPSS-Format übernommen. Mit Anklicken eines Formats werden im Informationsfeld der Gruppe „Datentyp" zugleich Beispiele für dessen Interpretation angegeben.

- *Numerisch.* (*Beispiel:* 123=123 oder 1,23=1,23.) Es ist eine Zahl, evtl. mit vorangestelltem Plus- oder Minus- und mit Dezimaltrennzeichen. Das Dezimaltrennzeichen ist das im Windows-Betriebssystem festgelegte länderspezifische (bei Einstellung auf Deutschland das Komma). Dezimaltrennzeichen müssen in der ASCII-Datei explizit angegeben sein. Die Zahlen werden so gelesen, wie sie dort angegeben sind. Im Dateneditor wird das Ergebnis jedoch u.U. ohne Kommastellen angezeigt, wenn die Feldbreite zur Anzeige nicht ausreicht. Die Datendefinition muss im Editor dann für deren Anzeige geändert werden.
- *Dollar (DOLLAR).* (*Beispiel:* 123=$123 und 1,23=$123, dagegen 1.23=$1.) Numerische Variable mit Dollarzeichen. Beachten Sie, dass hier die Angaben in amerikanischer Schreibweise erwartet werden (Komma ist Tausendertrennzeichen, Punkt Dezimaltrennzeichen). Bei deutscher Schreibweise werden die Daten verfälscht. Die Daten werden im Dateneditor ohne Dezimalstellen angezeigt. Durch Umdefinition des Variablenformats kann dies jedoch geändert werden.
- *Komma.* Gültige Werte sind Zahlen mit Dezimaltrennzeichen Punkt und Tausendertrennzeichen Komma.
- *Punkt:* Gültige Werte sind Zahlen mit Dezimaltrennzeichen Komma und Tausendertrennzeichen Punkt.

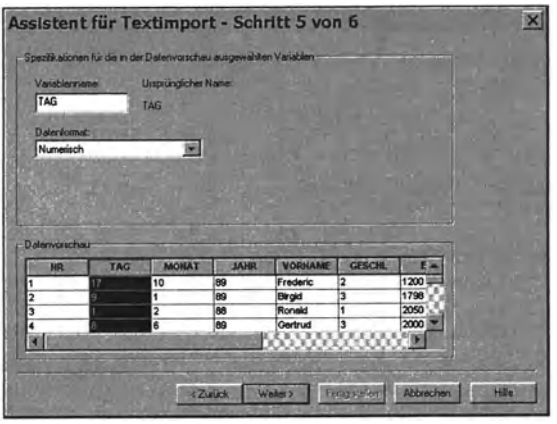


Abb. 6.14. Dialogbox „Assistent für Textimport – Schritt 5 von 6"

- *String (A).* Beliebige Zeichenketten werden gelesen, bis zu einer Zeichenbreite von acht Zeichen als Kurzstring, sonst als Langstring.
- Datum/Uhrzeit. Es handelt sich um verschiedene Varianten von Formaten für Datums- und Zeitvariablen zur Darstellung von Datum und Zeit und für Transformationen mit Datums- und Zeitfunktionen. Diese Formate sollten mit Vorsicht verwendet werden. Intern werden sie als sehr große Zahlen gespeichert, die für die statistischen Zwecke erst umgewandelt werden müssen. Bei manchen Funktionen, wie der Zeichnung von Histogrammen und Scatterplots, geschieht dies nicht und führt zu uninterpretierbaren Ergebnissen. In solchen Fällen müs-

sen die Datums- und Zeitangaben zuerst mit der Befehlsfolge „Transformationen", „Berechnen" und durch Verwendung einer der Funktionen des Typs XDATE.xxx umgerechnet werden. Auf die Darstellung dieser Formate wird hier verzichtet.

Darüber hinaus sind weitere Formate wie Komma-Formate (Komma als Tausendertrennzeichen), Punktformate, Prozentformate, wissenschaftliche Notation in der Befehlssyntax verfügbar (⇨ SPSS Base System Syntax Reference Guide).

Beim Markieren von „String" erscheint ein weiteres Auswahlkästchen zum Bestimmen der „Zeichenzahl". Markiert man „Datum/Zeit" erscheint ein Auswahlkästchen zur genaueren Festlegung des Datums- bzw. Zeitformats.

❒ Im Auswahlfeld „Datenformat" besteht aber auch die Möglichkeit, durch Auswahl von „nicht importieren" Variablen vom Einleseprozess auszuschließen und damit eine Selektion vorzunehmen.

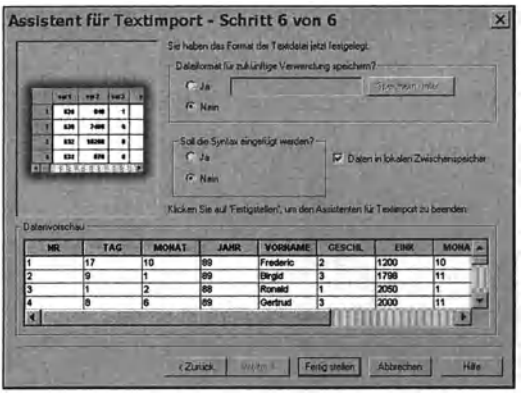


Abb. 6.15. Dialogbox „Assistent für Textimport – Schritt 6 von 6"

Die Datendefinition ist damit abgeschlossen. Im 6ten Schritt werden die Daten eingelesen. Dies geschieht durch Anklicken der Schaltfläche „Fertigstellen". Zuvor können noch einige interessante Optionen gewählt werden.((⇨ Abb. 6.15). Setzt man den Optionsschalter im Feld „Datei für zukünftige Verwendung speichern?" auf „Ja" und klickt daraufhin auf die Optionsschaltfläche „Speichern unter" öffnet sich ein Fenster, in dem man in üblicher Weise eine Datei mit den gerade getroffenen Formatierungsangaben speichern kann (Extension: tpf). Beim Zukünften Einlesen der Textdatei kann man es sich dann ersparen, den gesamten Prozess erneut zu durchlaufen, indem man im Schritt 1 den Optionsschalter „Ja" im Feld „Weist die Textdatei ein vordefiniertes Format auf?" einstellt und die zutreffende Datei im Auswahlfeld markiert. Weiter kann mit Hilfe eines Optionsschalters die Syntax des ganzen Definitions- und Einlesevorgangs in ein Syntaxfenster geleitet werden. Daraus ergibt sich eine weitere Möglichkeit, wiederholtes Einlesen der Textdatei zu vereinfachen. Schließlich steht noch das Kontrollkästchen „Daten in lokalen Zwischenspeicher" zur Verfügung. Wählt man es aus, wird eine Kopie des Daten-

satzes auf einem temporären Speicherplatz auf der Festplatte erstellt. Bei sehr großen Dateien kann dies die Bearbeitung beschleunigen.

Hinweis. Generell werden Werte mit in dem ausgewählten Format nicht zugelassenen Zeichen in System-Missing-Werte umgewandelt. Wenn z.B. in einer als numerisch definierten Variablen ein Stringwert auftaucht, wird dieser automatisch in einen System-Missing-Wert umgewandelt.

ASCII-Dateien in festem Format. Festes Format heißt: Die Werte für eine bestimmte Variable sind jeweils an derselben Stelle eines Datensatzes eingetragen, d.h. sie befinden sich in einem festgelegten Spaltenbereich. Falls die Daten für einen Fall sich über mehrere Zeilen erstrecken, müssen sich die Angaben für eine Variable auch in derselben Zeile (bezogen auf den Fall) befinden. Es können leere Zellen auftreten.

```
1  17  10  89  Frederic     2  1200  10  86    6500¶
2   9   1  89  Birgid       3  1798  11  82    4600¶
3   1   2  88  Ronald       1  2050   1  88   24700¶
4   8   6  89  Gertrud      3  2000  11  80  163000¶
5  17   7  89  Carola       1  9999   0   0  999999¶
6   1   9  88  Alfred       1  1950   7  82   33200¶
6   6  11  87  Manfred      2  1800   7  86   32000¶
7  21   7  89  Jürgen       1  1750  12  81   14500¶
8   5  11  88  Hildegard    3  1050   2  83    9086¶
9  28   1  88  Tom          2  1400  10  87   44740¶
```

Abb. 6.16. Schuldnerdatei in festem ASCII-Format VZ1.TXT

Beispiel. Die Schuldnerdatei würde als ASCII-Datei in festem Format in etwa aussehen wie in Abb. 6.16. Die Daten eines Falles stehen in einer Zeile. Die Variablen, zunächst formal mit den Namen V1 bis V10 bezeichnet, stehen in folgenden Spaltenbereichen: V1 1-2, V2 4-5, V3 8-9, V4 12-13, V5 17-28, V6 31, V7 34-37, V8 41-42, V9 46-47 und V10 49-55. Die Daten sollen nun importiert werden und dabei dieselben Namen erhalten, wie wir sie aus den bisherigen Beispielen kennen. Die Namensvariable soll als String, die Einkommensvariablen als numerische mit zwei Kommastellen und die restlichen als numerische, ohne Kommastellen definiert werden.

Der Import dieser Datei vollzieht sich in den 6 oben angegebenen Schritten mit gewissen Unterschieden bei Schritt 2, 3 und 4.

- Bei Schritt 2 wählen sie den Optionsschalter „Feste Breite".
- Dadurch ergibt sich in der Dialogbox zu Schritt 3 eine Änderung. Anstelle der Gruppe „Wie sind die Fälle dargestellt?", steht jetzt ein Auswahlkästchen „Wie viele Zeilen stellen einen Fall dar?". Hier muss angegeben werden, über wie viele Zeilen sich die Angaben zu einem Fall erstrecken. In unserem Beispiel ist dies nur eine Zeile.
- In der Dialogbox des vierten Schrittes sind die Daten in der Datenvorschau anders dargestellt. Die Grenzen der Variablen sind durch senkrechte Linien eingezeichnet. Falls diese nicht mit den tatsächlichen Grenzen übereinstimmen, kann man eine Anpassung vornehmen. Die Linien können durch Ziehen verschoben

werden. Zieht man eine Linie aus der Datenvorschau heraus, wird sie gelöscht. Durch Anklicken eines Punktes innerhalb des Vorschaufensters, kann man eine neue Trennlinie einfügen.

Die Veränderung von Namen und Datentyp erfolgt in Schritt 5 wie oben angegeben. Nur wurden in diesem Beispiel keine Variablennamen aus der Textdatei übernommen, sondern SPSS-Variablennamen automatisch generiert. Definieren Sie Namen und Typ wie bei Datei VZ.TXT.

- ▷ Markieren Sie dazu in der Datenvorschau V1 und ändern Sie den Namen im Feld „Variablenname" in NR.
- ▷ Tragen Sie auf gleiche Weise den gewünschten Variablennamen für alle weiteren Variablen ein.

Hinweis. Bei Vergabe der Variablennamen gelten die in Kap. 3.1 dargestellten Regeln.

ASCII-Dateien in freiem Format. Bei variablem Format sind die Variablen bei den verschiedenen Fällen in derselben Reihenfolge, nicht aber unbedingt in derselben Spalte gespeichert. Das Programm erkennt den Beginn einer neuen Variablen an einem Trennzeichen. Mehrere Fälle können in derselben Reihe abgespeichert werden. SPSS interpretiert nach Abarbeiten einer Variablenliste einen neuen Wert als ersten Wert des neuen Falles. Alle Variablen müssen definiert werden. Für jede Variable muss sich bei jedem Fall ein Eintrag finden, der nicht dem Trennwert entspricht. Sonst wäre das Programm nicht in der Lage, die Variablen richtig abzuzählen.

Unsere Beispielsdaten könnten etwa wie in Abb. 6.17 aussehen. Wie Sie am besten an den Namen sehen, sind die Fälle einfach aneinander anschließend abgespeichert. Die Zahl der Leerstellen zwischen den Variablen kann, wie in unserem Beispiel, durchaus variieren. Auch Tabulator oder andere Zeichen sind als Trennzeichen zulässig.

```
1 17 10 89 Frederic 2 1200 10 86 6500 2  9  1 89  Birgid  3 1798  11  82  4600 3  1  2
88  Ronald  1 2050  1  88 24700 4  8  6 89 Gertrud  3 2000  11  80 163000 5 17  7
89  Carola  1 9999  0   0 999999 6  1  9 88  Alfred  1 1950   7  82 33200 6  6 11
87  Manfred  2 1800  7  86 32000 7 21  7 89  Jürgen  1 1750  12  81 14500 8  5
11 88  Hildegard 3 1050  2  83  9086 9 28  1 88  Tom  2 1400  10  87 44740¶
```

Abb. 6.17. Daten der Schuldnerdatei VZ2.DAT in freiem Format

Das Einlesen der Daten folgt vollkommen den Schritten beim Einlesen einer durch Trennzeichen strukturierten Datei. Lediglich in Schritt 3 ergibt sich eine Änderung. Im Auswahlkästchen „Folgende Anzahl von Variablen stellt einen Fall dar" muss nun angegeben werden, wie viele Variablen ein Fall umfasst. In unserem Beispiel sind es 10 Variablen. Diese Zahl wird eingegeben. Im Schritt 4 ist wiederum das verwendete Trennzeichen anzugeben. Im Beispiel ist es das Leerzeichen. Die folgenden Schritte entsprechen exakt den für durch Trennzeichen strukturierten Dateien beschriebenen.

6.2 Daten in externe Formate ausgeben

Um im Format eines anderen Programms als SPSS für Windows zu speichern:

▷ Wählen Sie „Datei“ und „Speichern unter...“. Es öffnet sich die Dialogbox „Daten speichern unter“ (⇨ Abb. 6.18).

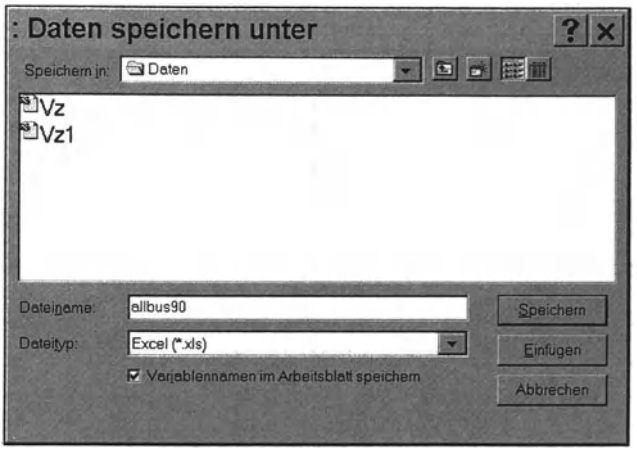


Abb. 6.18. Dialogbox „Daten speichern unter“

▷ Öffnen Sie durch Klicken auf den Pfeil neben dem Feld „Dateityp:“ die Liste der verfügbaren Dateiformate, und wählen Sie das gewünschte Format aus, in dem die Datei neu abgespeichert werden soll. Im Eingabefeld „Dateiname:“ geben Sie den gewünschten Namen ein. SPSS vergibt automatisch die Standardextension dieses Formats.

▷ Wählen Sie in der üblichen Weise das Verzeichnis, in das die neue Datei geschrieben werden soll.

▷ Markieren Sie gegebenenfalls das Auswahlkästchen „Variablennamen im Arbeitsblatt speichern“. (Dies bewirkt bei den Formaten Lotus, Excel, Sylk und Tab-delimited, dass die Variablennamen in die erste Zeile der Tabelle geschrieben werden.)

▷ Bestätigen Sie mit „Speichern“.

Für die Ausgabe stehen folgende Formate zur Verfügung:

❒ *SPSS-Formate.* Neben dem SPSS für Windows-Format und dem speziellen Format der Version 7.0 das Format SPSS/PC+ der DOS-Version und das Exportformat SPSS Portable für den Austausch mit SPSS-Versionen für andere Betriebssysteme.

❒ *ASCII-Formate.* ASCII-Format mit „Tab“ als Trennzeichen (Tabulator-getrennt), ASCII-Datei mit festem Format.

❒ *Tabellenkalkulationsformate.* Excel, Lotus 1-2-3 (WKS, WK1, WK3 für die Versionen 1.0 bis 3.0) und SYLK für spezielle Excel- und Multiplan-Dateien.

❒ *Datenbankformate.* dBASE für die Versionen II bis IV.

Beachten Sie bitte einige Einschränkungen für den Datenaustausch (⇨ Tab. 6.1):

Tabelle 6.1. Einschränkungen für den Datenaustausch

Datei-Format	Standardextension	maximale Variablenzahl	maximale Fallzahl
SPSS/PC+	sys	500	
SPSS Portable	por		16384
Excel	xls	256	
Lotus, Version 3.0	wk3	256	
Lotus, Version 2.0	wk1	256	9192
Lotus, Version 1.A	wks	256	2048
Sylk	slk	256	4095
dBASE, Version IV	dbf	255	1 Milliarde*)
dBASE, Version III	dbf	128	1 Milliarde*)
dBASE, Version II	dbf	32	65535
Tab-delimited	dat		
ASCII festes Format	dat		

*) Abhängig vom Speicherplatz

Weitere Hinweise. Beim Austausch von Daten zwischen verschiedenen SPSS-Plattformen sind verschiedene Restriktionen zu beachten. 1. Die DOS-Versionen sind nur in der Lage, bis zu 500 Variablen zu verarbeiten. 2. Die Zahl der nutzerdefinierten fehlenden Werte variiert. SPSS/PC+ kann z.B. nur einen nutzerdefinierten fehlenden Wert verarbeiten. Ist bei Übergabe einer SPSS für Windows Datei an eine SPSS/PC+-Datei mehr als ein fehlender Wert vom Nutzer definiert, werden die später definierten Werte automatisch in den ersten nutzerdefinierten einzelnen fehlenden Wert umkodiert, bei Austausch über eine portable-Datei dagegen in den untersten Wert eines Wertebereichs. 3. Umlaute in Variablen- und Wertelabels können beim Austausch mit SPSS/PC+ Version 4.0, bei Verwendung von portable-Dateien und MacIntosh-Dateien nicht korrekt übertragen werden. 4. Bei Übertragung auf MacIntosh oder UNIX-Workstations muss bei Verwendung von portable-Dateien die Recordebegrenzung von CR/LF auf CR geändert werden. (⇨ Bernhard Krüger, Heiner Ritter, Cornelia Züll).

Bei allen Dateien, die nicht in einem der SPSS-Formate gespeichert sind, gehen SPSS-spezifische Informationen wie Werte-Labels und Missing-Werte verloren. Bei Tab-delimited ASCII-Dateien werden die Werte durch Tab-Zeichen getrennt. ASCII-Dateien in festem Format speichern die Variablen in durch die Variablenbreite vorgegebenen festen Abständen. An die maximal zulässige Variablenzahl passen Sie die Daten an, indem Sie entweder im Dateneditor die überzähligen Variablen löschen oder die Befehlssyntax benutzen. Verwenden Sie im letztgenannten Fall den Befehl SAVE TRANSLATE – mit dem Unterbefehl /DROP.

7 Transformieren von Dateien

7.1 Daten sortieren, transponieren und umstrukturieren

7.1.1 Daten sortieren

Für verschiedene Zwecke ist es nützlich oder unerlässlich, die Daten in einer bestimmten Sortierung vorliegen zu haben. Datenbereinigungen lassen sich z.B. besser in einer nach der Fallnummer sortierten Datei durchführen. Für das Auflisten von Fällen wird man ebenfalls nach Fallnummer sortieren. Manche Prozeduren verlangen sogar nach bestimmten Kriterien geordnete Dateien. So muss für die Zusammenfassung von Dateien unter Verwendung von Schlüsselvariablen die Datenmatrix nach der Schlüsselvariablen sortiert sein. Erstellt man zusammenfassende Berichte mit Break-Variablen (Gruppierungsvariablen), muss die Datei nach den Kategorien der Break-Variablen geordnet vorliegen. Ebenso erfordert die Aufteilung von Dateien eine Sortierung nach den Gruppierungsvariablen. Die genannten Prozeduren stellen zwar selbst eine Sortieroption zur Verfügung, unabhängig davon kann man aber auch im Menü „Daten" das Untermenü „Fälle sortieren..." für Sortiervorgänge auswählen. Es öffnet sich dann die Dialogbox „Fälle sortieren" (⇨ Abb. 7.10). Darin sind zunächst aus der Quellvariablenliste die Sortiervariablen auszuwählen (Stringvariablen sind in der Liste mit 🔤 bzw. 🔤 gekennzeichnet). Werden mehrere Sortiervariablen verwendet, wird die Sortierung in der Reihenfolge der Eintragung in das Feld „Sortieren nach:" vorgenommen. Die Sortierung einer Datei nach Geschlecht (männlich = 1; weiblich = 2) und dann nach Alter in aufsteigender Ordnung bewirkt z.B., dass zuerst die Datei nach Männern und Frauen sortiert wird, danach innerhalb der Kategorien Männer und Frauen jeweils nach aufsteigendem Alter. Als Sortierreihenfolge kann „Aufsteigend" (vom kleinsten Wert zum größten bzw. bei Stringvariablen vom ersten Buchstaben des Alphabets zum letzten) oder „Absteigend" gewählt werden.

7.1.2 Transponieren von Fällen und Variablen

Die Prozedur „Transponieren" wird benötigt, wenn eine Datenmatrix in ihrem Aufbau nicht den SPSS-Bedingungen entspricht. Transponieren heißt, Zeilen in Spalten und Spalten in Zeilen umzuwandeln, also die Datenmatrix zu drehen. Besonders nach der Übernahme von Daten aus anderen Programmen ist es häufig erforderlich, die Datenmatrix für die Weiterverarbeitung in SPSS zu transponieren.

Nehmen wir als Beispiel die Datenmatrix in Abb. 7.1. Sie enthält die Fälle spaltenweise, und zwar so, dass die Werte des Falles 1 in der Spalte VAR00002, die

des Falles 2 in der Spalte VAR00003 stehen usw.. Es sind für die Fälle die Variablen „Fallnummer“ (NR), „Geschlecht“ (GESCHL) und „Konfession“ (KONF) erfasst (Datei TRANSPONIEREN.SAV). Die Matrix soll gedreht werden.

	var00001	var00002	var00003	var00004
1	nr	1,00	2,00	3,00
2	geschl	1,00	1,00	1,00
3	konf	2,00	2,00	1,00

Abb. 7.1. Datenmatrix (Spalten: Fälle, Zeilen: Variablen)

Dazu gehen Sie wie folgt vor:

▷ Wählen Sie die Befehlsfolge „Daten“, „Transponieren...“. Es öffnet sich die Dialogbox „Transponieren“ (⇨ Abb. 7.2).

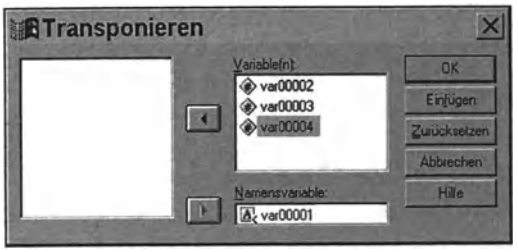


Abb. 7.2. Dialogbox „Transponieren“

▷ Übertragen Sie aus der Liste der Quellvariablen alle Variablen, die zu einem Fall (einer Zeile) der neuen Matrix werden sollen, in das Auswahlfeld „Variable(n):“.

Falls in einer der Ausgangsvariablen die Namen der zukünftigen Variablen als Werte stehen, kann man diese Namen übernehmen. Am günstigsten ist es, wenn die Namen in einer Stringvariablen vorliegen. Wird dagegen eine numerische Variable zur Namensbildung herangezogen, bildet SPSS Variablennamen, die sich aus dem Buchstaben V und dem Variablenwert zusammensetzen. Sind die Namen, die so entstehen, nicht eindeutig, weil z.B. der Wert 2 doppelt vorkommt, vergibt SPSS eindeutige Namen, indem es an den Wert eine fortlaufende Zahl anhängt. Im angegebenen Beispiel würde die erste Variable den Namen V2, die zweite den Namen V21 erhalten. Werte mit mehr als 8 Stellen werden abgeschnitten. (Enthält die Variable Nachkommastellen, werden auch die im Namen nach einem Unterstrich berücksichtigt. *B.:* V2_10.) Wird keine Variable zur Definition der Variablennamen verwendet, vergibt SPSS per Voreinstellung automatisch die Variablennamen V001, V002 usw.. Die Fälle bekommen automatisch als Fallnummern (case_lbl) die Nummer ihrer Ursprungs-

spalte zugewiesen. Wollen Sie die Variablennamen aus einer Ausgangsvariablen übernehmen:

- ▷ Markieren Sie die Variable in der Quellvariablenliste (hier: V00001) und übertragen Sie sie in das Eingabefeld „Namensvariable:".
- ▷ Bestätigen Sie mit „OK". Das Ergebnis der Transponierung der in Abb. 7.1 dargestellten Matrix mit der Einstellung nach Abb. 7.2 sehen Sie in Abb. 7.3.

	case_lbl	nr	geschl	konf
1	VAR00002	1,00	1,00	2,00
2	VAR00003	2,00	1,00	2,00
3	VAR00004	3,00	1,00	1,00

Abb. 7.3. Transponierte Datenmatrix

Behandlung fehlender Werte. Beim Transponieren werden alle nutzerdefinierten fehlenden Werte in System-Missings umgewandelt. Will man das verhindern, sollte man vor dem Transponieren die Datendefinition so ändern, dass keine nutzerdefinierten fehlenden Werte auftreten.

7.1.3 Daten umstrukturieren

Das Menü „Umstrukturieren" dient ebenfalls der Datentransformation, verfügt aber über mehr Wahlmöglichkeiten und wird durch einen „Assistent(en) für die Datenumstrukturierung" unterstützt.

In dieses Menü gelangt man mit der Befehlsfolge „Daten", „Umstrukturieren". Es öffnet sich eine erste Dialogbox des „Assistenten".

In dieser Dialogbox werden drei Varianten der Datenumstrukturierung geboten:

- ❒ Umstrukturieren ausgewählter Variablen in Fälle.
- ❒ Umstrukturieren ausgewählter Fälle in Variablen.
- ❒ Transponieren sämtlicher Daten.

Die letzte Option ist der einfachste Fall und erbringt dieselbe Leistung wie das Menü „Transponieren" (⇨ Kap. 7.1.2) und öffnet dieselbe Dialogbox (⇨ Abb. 7.2). Auf sie wird daher hier nicht mehr eingegangen.

Die beiden anderen Optionen dienen der Umstrukturierung von komplexeren Daten, die nicht der grundsätzlichen Form einer SPSS-Datenmatrix mit Variablen in den Spalten und den Fällen in Zeilen entsprechen und auch nicht durch Tauschen von Spalten und Reihen in diesen Form gebracht werden können und sollen. Solche Datenstrukturen findet man häufig bei experimentell erhoben Daten, insbesondere bei Messwiederholung.

Die zu besprechenden Optionen dienen dazu, zwei spezielle Datenstrukturen zu erzeugen, wobei diese insofern miteinander korrespondieren, als jeweils die eine Option von der Datenstruktur ausgeht, die die andere erzeugt.

Zunächst seien daher die Datenstrukturen vorgestellt. Als Beispiel werden fiktive Daten einer Untersuchung mit Mehrfachmessung verwendet. Bei 10 Probanden

seien Blutdruck und Hämatokrit-Wert zu drei verschiedenen Zeitpunkten (die vielleicht einem Belastungsfaktor entsprechen) gemessen. Die Daten könnten nun in zwei verschiedenen Varianten organisiert sein. Wir können Blutdruck und Hamatokrit-Wert als Variable, Zeit als Faktor mit drei Faktorstufen bezeichnen.

❒ *Fallgruppen.* Die Daten sind in Fallgruppen geordnet (⇨ Abb. 7.4). D.h., die Messungen für einen Fall sind nicht in einer sondern mehreren Zeilen enthalten, wobei in jeder Zeile die Messungen der Variablen für eine Faktorstufe enthalten sind. Im Beispiel enthalten die Zeilen 1-3 (= die erste Fallgruppe) die Werte des Falles 1, die Zeile 1 diejenigen zum Zeitpunkt 1, die Zeile 2 die zum Zeitpunkt 2 etc..

In Fallgruppen müssen Daten z.B. geordnet sein für einen t-Test bei unabhängigen Stichproben, nichtparametrische Tests, die Erstellung eines OLAP-Würfels und einfache Varianzanalysen (nicht Messwiederholungen). Diese Prozeduren benötigen immer eine unabhängige und eine abhängige Variable. Es muss also eine gesonderte Spalte für die unabhängige Variable (den Faktor) vorhanden sein und die Daten der abhängigen (der Gruppe der abhängigen Variablen) müssen ebenfalls in einer einzigen Spalte stehen. Dafür nimmt man in Kauf, dass pro Fall mehrere Zeilen benötigt werden und die Daten einfacher Variablen vervielfältigt auftreten.

	patient	zeit	bldr	häma	geschl
1	1	1	90,90	36,98	w
2	1	2	97,49	31,81	w
3	1	3	92,64	30,85	w
4	2	1	109,63	47,29	m
5	2	2	108,02	44,29	m
6	2	3	94,22	37,33	m

Abb. 7.4. Daten der beiden erste Fälle der Datei BLUTDR1.SAV (als Fallgruppen)

❒ *Variablengruppen* (*Spaltengruppen*). Sind die Daten in Variablengruppen geordnet (⇨ Abb. 7.5), dann enthält eine Zeile die Messung eines Falles, aber jeweils mehrere Spalten (Variablengruppe oder Spaltengruppen enthalten die Daten im Grund einer Variablen (im Beispiel drei Blutdruck- bzw. Hämatokrit-Wert für die verschiedene Faktorstufen [hier Zeitpunkte]). Die Zeile wird deshalb auch oft als Gruppenvariable bezeichnet. SPSS kann auch solche Datenstrukturen verarbeiten, jedoch hängt es von den Prozeduren ab, welche der beiden Formen erwartet werden.

	patient	bltdr1	bltdr2	bltdr3	häm	häm1	häm2	häm3	geschl
1	1	90,90	97,49	92,64	37,00	36,98	31,81	30,85	w
2	2	109,63	108,02	94,22	42,00	47,29	44,29	37,33	m
3	3	117,24	141,52	151,68	47,00	40,08	35,21	44,18	w
4	4	126,40	127,68	120,01	52,00	50,60	47,63	58,01	m

Abb. 7.5. Daten der vier erste Fälle der Datei BLUTDR2.SAV (als Variablengruppen)

In *Spalten-/Variablengruppen* müssen die Daten bei allen Analysen von Messwiederholungen organisiert sein. Beispiele sind t-Test für gepaarte Stichproben, nonparametrische Tests mit verbundene Stichproben, Varianzanalyse mit Messwiederholung. Bei diesen Prozeduren werden immer die Werte zweier Variablen, die in verschiedenen Spalten der Matrix stehen, gepaart oder die Werte mehrerer Variablen verbunden.

Umstrukturieren ausgewählter Fälle in Variablen.

Beispiel. Daten der Blutuntersuchung liegen in der Fallgruppenstruktur vor. Um eine t-Test für abhängige Stichproben durchführen zu können, benötigen wir sie in der Struktur „Spaltengruppen". Um dies zu erreichen, gehen Sie, nachdem die Datei BLUTDR1.SAV geladen ist, wie folgt vor:

- ▷ Wählen Sie „Daten", „Umstrukturieren" und im der sich öffnenden ersten Dialogbox des „Assistenten für die Datenumstrukturierung" „Umstrukturieren ausgewählter Fälle in Variable". Bestätigen Sie mit „Weiter". Die zweite Dialogbox des Assistenten erscheint.
- ▷ Jetzt müssen aus der Gruppe „Variablen in der aktuellen Datei" Variablen in die Felder „Bezeichnervariable(n)" und „Indexvariable(n)" übertragen werden. Es handelt sich dabei gerade nicht um die Variablen, aus denen eine Gruppe neuer Variablen erstellt werden soll.
  - Eine *Bezeichnervariable* ist eine Variable, die angibt, welche Zeilen zusammen einen Fall ausmachen. In unserem Beispiel sind pro Fall drei Zeilen vorhanden, welche zusammengehören, erkennt man an den Werten der Variablen PATIENT. Die ersten drei Zeilen enthalten alle die Ziffer 1, d.h. sie gehören zum Fall /Patienten 1 etc.. Aus diesen drei Zeilen wird nach dem Umstrukturieren eine einzige. „Patient" ist also im Beispiel die Bezeichnervariable. (Die Datei muss nach der Bezeichnervariable sortiert sein. Ist dies nicht der Fall holen Sie das nach.)
  - Eine *Indexvariable* ist eine Variable, aus der zu erkennen ist, welche Faktorstufe jeweils eine Zeile angibt. Aus jeder dieser Faktorstufen wird beim Umstrukturieren eine eigene Spalte. Im Beispiel ist die Variable „Zeit" die Indexvariable. Die Faktorstufen sind nämlich die Zeitpunkte 1, 2 und 3. Für jeden dieser Zeitpunkte wird beim Umstrukturieren automatisch eine neue Variable sowohl für BLTDR als auch für HÄM erstellt.
- ▷ Übertragen Sie also PATIENT in das Feld „Bezeichnervariable" und ZEIT in das Feld „Indexvariable". Bestätigen Sie mit „Weiter".
- ▷ In der folgenden Dialogbox bestimmen Sie, ob die neu entstehende Datei nach Bezeichner- und Indexvariable sortiert werden soll oder nicht.
- ▷ Darauf folgt eine Dialogbox „Optionen". Hier kann zunächst die Reihenfolge der neu gebildeten Variablen bestimmt werden. Die Option „Nach Originalvariable gruppieren" führt dazu, dass alle neuen Variablen, die derselben Originalvariablen entspringen, in nebeneinander liegende Spalten gruppiert werden. Der Index bestimmt die Reihenfolge innerhalb der Gruppe (im Beispiel erst BLTDR1, BLTDR2, BLTDR3, dann HÄM1, HÄM2, HÄM3). Wird nach Index gruppiert, folgt auf BLTDR1, HÄM1, BLTDR2 etc.. Außerdem kann man

eine Variable abfordern, die zählt, aus wie viel Fällen der Originaldatei ein Fall der neuen Datei entsteht (im Beispiel sind dies 3).

- ▷ Eine letzte Dialogbox ermöglicht es, die Umstrukturierung entweder fertig zu stellen oder die Syntax zur weiteren Bearbeitung oder späteren Nutzung in ein Syntaxfenster zu übertragen. (Werden nicht alle Schritte benötigt, kann die Umstrukturierung auch schon in einem früheren Fenster fertiggestellt werden.)

Umstrukturieren ausgewählter Variablen in Fälle.

Beispiel. Daten der Blutuntersuchung liegen in der Variablengruppenstruktur vor. Um einen OLAP-Würfel zu erstellen, benötigen wir sie in der Struktur „Fallgruppen". Um dies zu erreichen, gehen Sie, nachdem die Datei BLUTDRUCK.SAV geladen ist, wie folgt vor:

- ▷ Wählen Sie „Daten", „Umstrukturieren" und im der sich öffnenden ersten Dialogbox des „Assistenten für die Datenumstrukturierung" „Umstrukturieren ausgewählter Variablen in Fälle". Bestätigen Sie mit „Weiter". Die zweite Dialogbox des Assistenten erscheint.
- ▷ Hier müssen Sie angeben, wie viele Variablengruppen umzustrukturieren sind. Zur Wahl stehen eine oder mehrere. Bei Auswahl der Option „Mehrere" wird die Anzahl in ein Eingabefeld eingetragen. (Im Beispiel ist „mehrere" zu wählen, da wir die beiden Gruppen für Blutdruck und Hämatokrit-Wert haben und als Anzahl die Voreinstellung 2 zu übernehmen.) Bestätigen Sie mit „Weiter". Es erscheint die dritte Dialogbox (⇨ Abb. 7.6).
- ▷ Dort ist anzugeben, wie die neu zu bildende(n) Variable(n) heißen sollen und aus welchen der bisherigen Variablen sie sich zusammensetzen. Im Feld Zielvariable ist der Name TRANS1 eingestellt. Ändern Sie ihn in „BLTDR". Geben Sie dann an, welche der bisherigen Variablen in der neuen BLTDR zusammengefasst werden. Dazu markieren Sie im Feld „Variablen in der aktuellen Datei" die zutreffenden Variablen (hier: BLTDR1, BLTDR2 und BLTDR3) und übertragen Sie diese durch Anklicken des Pfeils in das Feld „Zu transponierende Variablen".
- ▷ Wiederholden Sie dasselbe für die Variable „HÄM". Dazu öffnen Sie zunächst durch Klicken auf den Pfeil neben dem Feld „Zielvariable" eine Liste mit den Zielvariablen. Da wir im vorigen Fenster 2 angegeben haben, ist eine zweite Zielvariable „trans2" in der Liste enthalten. Markieren Sie diese und verfahren Sie für die bisherigen Variablen „HÄM1 bis 3" wie für BLTDR beschrieben. Andere nicht gruppierte Variablen werden in die neue Datei nur übernommen, wenn sie in das Feld „Variable(n) mit festem Format" übertragen werden. (Im Beispiel übertragen wir die Variablen PATIENT und GESCHL.)
- ▷ Außerdem kann mit Hilfe einer Auswahlliste im Feld „Angabe von Fallgruppen" noch festgelegt werden, wie die Fallgruppen bezeichnet werden sollen, mit der Fallnummer, mit dem Wert einer anzugebenden Variablen oder überhaupt nicht. Im Beispiel ist die Fallnummer adäquat. Wir ändern noch die Bezeichnung den Namen der Variablen, in der diese Bezeichnung ausgegeben wird von ID in FALL. (man kann auch noch ein Label für diese neue Variable vergeben) und bestätigen mit „Weiter".

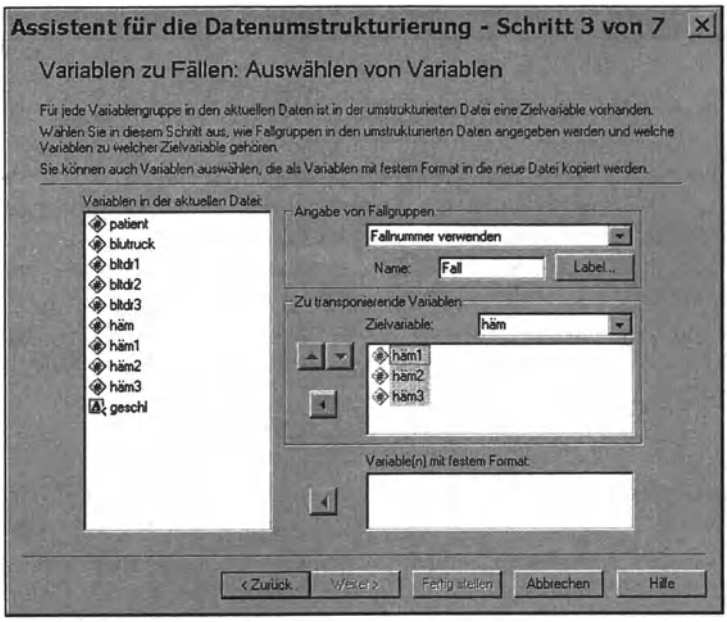


Abb. 7.6. Schritt 3 der Datenumstrukturierung bei Umstrukturieren von Variablen in Fälle

- ▷ In der nächsten Dialogbox kann festgelegt werden, ob eine mehrere oder keine Indexvariable erstellt werden soll. In einer solchen Indexvariablen wird die Information darüber erzeugt, welche Faktorstufe in der jeweiligen Zeile der Ergebnisdatei enthalten ist. Man sollt eine Indexvariable erstellen, wenn Sie nicht schon anderweitig durch die Gruppenvariablen erzeigt wird.
- ▷ Falls man bestimmt hat, dass eine Indexvariable gebildet werden soll, erscheint ein weiteres Dialogfenster. Hier kann man den Namen der Indexvariablen (Voreinstellung INDEX1 etc.) verändern und ein Label eingeben (im Beispiel ändern wir den Namen in Zeit und vergeben das Label Zeitpunkt). Weiter wird festgelegt, wie die Werte dieser Variablen gebildet werden. Zur Auswahl stehen „Fortlaufende Zahlen" und „Variablennamen". Wählt man eine dieser Optionsschaltfläche an, werden die verwendeten Indexwerte angezeigt (im Beispiel wäre dies bei „Fortlaufenden Zahlen" die Werte 1, 2 und 3, bei „Variablennamen" BLTDR1, BLTDR2 und BLTDR3 oder HÄM1, HÄM2 und HÄM3). Bei Verwendung mehrerer Gruppen kann bei „Variablennamen" über die Auswahlliste „Indexwerte" festgelegt werden, welcher der Variablennamen zur Bildung der Werte genutzt wird. (Im Beispiel wollen wir „Fortlaufende Zahlen" benutzen.)
- ▷ Es öffnet sich eine weitere Dialogbox mit verschiedenen „Optionen".
  - Falls bei der Auswahl noch nicht geschehen, kann man jetzt noch festlegen, dass nicht ausgewählte Variablen zu Variablem mit festem Format beibehalten werden (ansonsten werden sie aus der Datei entfernt).

- Was mit fehlenden Werte geschehen soll. Entweder wird aus ihnen ein Fall in der neuen Datei erstellt (Voreinstellung) oder sie werden daraus ganz entfernt.
- Anzahl neuer Fälle, die von einem Fall der aktuellen Datei erzeugt wurden. Wählt man diese Option aus, erstellt das Programm eine weitere Variable, die angibt wie viele Zeilen der neuen Matrix einem Fall der alten Matrix entsprechen (im Beispiel wird aus einer Zeile drei, weil für jeden Messzeitpunkt eine neue Zeile für denselben Fall erzeugt wird).

▷ In einer letzten Dialogbox bestimmt man schließlich, ob die Umstrukturierung fertig gestellt werden soll oder aber zur weiteren Bearbeitung oder späteren Nutzung in ein Syntaxfenster transferiert wird (falls man auf einige der Optionen verzichtet, kann die Umstrukturierung auch bereits in einem der vorherigen Fenster fertiggestellt werden).

7.2 Zusammenfügen von Dateien

Zwei Dateien können so zusammengeführt werden, dass an eine bestehende Datei aus einer zweiten neue Daten angefügt werden. Die Daten können sein:

❒ *Neue Fälle* mit Variablen gleichen Inhalts oder
❒ *Neue Variablen* für bereits erfasste Fälle.

7.2.1 Hinzufügen neuer Fälle

Beispiel. In einer Wahluntersuchung wurden die Wahlabsichten zweier Stichproben zu zwei nicht zu weit auseinander liegenden Zeitpunkten erfasst. Die Daten stehen in zwei SPSS-Dateien WAHLEN1.SAV und WAHLEN2.SAV. Die beiden Dateien sollen zu einer neuen Datei WAHLEN.SAV zusammengefasst werden. Die Variablen beider Dateien sind weitgehend identisch. Allerdings ist eine inhaltlich identische Variable, in der die aktuelle Wahlabsicht erfasst wurde, unterschiedlich benannt, in der ersten Datei als PART_AK2, in der zweiten als PARTAKT2. Außerdem sind einige Variablen der zweiten Datei in der ersten nicht enthalten. Eine davon, KOAL2, in der die Koalitionswünsche der Befragten erfasst wurden, soll in die gemeinsame Datei übernommen werden. Schließlich sind einige Variablen vorhanden, die in der neuen Datei ohne Interesse sind und daher gestrichen werden können. Sofern sie nicht in beiden Dateien enthalten sind (nicht gepaarte Variablen), geschieht das automatisch. Ansonsten muss eine entsprechende Auswahl erfolgen. Es soll zudem eine neue Variable erzeugt werden, die für die einzelnen Fälle festhält, aus welcher der Quelldateien die Daten kommen (Datei-Indikator).

Laden Sie dazu zunächst die Datei WAHLEN1.SAV als Arbeitsdatei in den Dateneditor. Um dieser Datei Fälle aus der anderen SPSS-Datei anzufügen, verfahren Sie wie folgt:

▷ Wählen Sie die Befehlsfolge „Daten“, „Dateien zusammenfügen ▷“.
▷ Wählen Sie das Untermenü „Fälle hinzufügen...“. Es öffnet sich die Dialogbox „Fälle hinzufügen: Datei lesen“.

- ▷ Geben Sie dort in das Eingabefeld „Dateiname:“ den Namen der Datei an, die sie mit der Arbeitsdatei verbinden wollen (hier: WAHLEN2). Sie können sie auch über die Verzeichnis- und Dateienlisten wählen.
- ▷ Klicken Sie auf die Schaltfläche „Öffnen“. Es öffnet sich die in Abb. 7.7 dargestellte Dialogbox „Fälle hinzufügen aus“, auf deren rechten Seite die Variablen der neuen Arbeitsdatei angeführt sind.

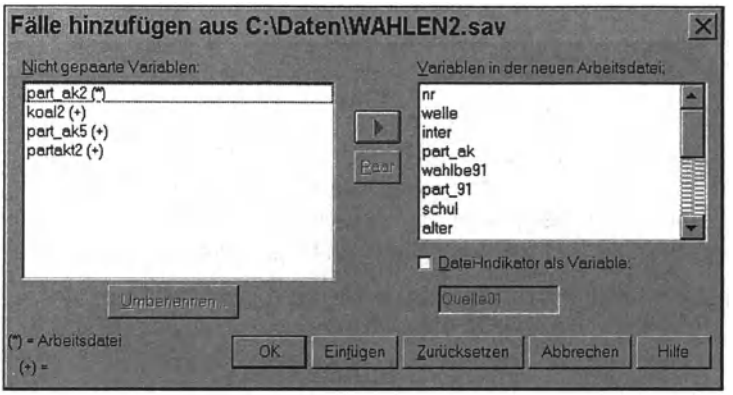


Abb. 7.7. Dialogbox „Fälle hinzufügen aus:“

In dem Feld „Nicht gepaarte Variablen“ auf der linken Seite werden zunächst die Variablen angezeigt, die kein Pendant in der anderen Datei besitzen: weil keine Variable gleichen Namens vorhanden ist oder weil bei Variablen gleichen Namens die eine numerisches, die andere Stringformat besitzt. Damit erkennbar ist, in welcher Datei die ungepaarte Variable enthalten ist, sind + oder * als Symbol hinzugefügt.

- * Bedeutet, dass eine Variable der Arbeitsdatei kein Pendant in der hinzugefügten Datei besitzt.
- \+ Bedeutet, dass eine Variable der hinzugefügten Datei kein Pendant in der Arbeitsdatei besitzt.

Zunächst enthält das Feld „Variablen in der neuen Arbeitsdatei:“ alle gepaarten Variablen. Man kann aber aus dieser Liste die nicht gewünschten Variablen entfernen. Bisher nicht gepaarte Variablen (mit gleicher Information, aber unterschiedlichem Namen) können gepaart werden. Variablen, die nur in einer der beiden Dateien enthalten sind, können nachträglich in die Auswahl aufgenommen werden. Bei den Fällen der anderen Datei werden dann System-Missings als Variablenwerte eingesetzt. Zusätzlich ist es möglich, Variablen umzubenennen.

Entfernen von Variablen. Zunächst sollen aus der Liste der ausgewählten Variablen die Variablen WELLE und FILTER_$ entfernt werden.

- ▷ Markieren Sie dazu jeweils die Variablen. Liegen sie nicht nebeneinander, muss bei der zweiten Variablen beim Klicken die <Ctrl>-Taste gedrückt sein.

▷ Klicken Sie auf . Die Variablen werden in das Feld „Nicht gepaarte Variablen“ verschoben.

Hinzufügen einer nicht gepaarten Variablen.

▷ Markieren Sie die Variable (hier: KOAL2).
▷ Klicken Sie auf . Die Variable wird in die Liste „Variablen in der neuen Arbeitsdatei:“ übertragen. Das Zeichen (+), aus dem zu entnehmen ist, dass dieser Variablen keine Variable in der Arbeitsdatei entspricht, bleibt erhalten. Bei den Fällen, für die kein Wert für diese Variable vorhanden ist, wird ein System-Missing-Wert eingesetzt.

Kombinieren zweier Variablen zu einem neuen Paar.

▷ Markieren Sie beide Variablen [hier: PART_AK2 (*) und PARTAKT2 (+)].
▷ Klicken Sie auf die Schaltfläche „Paar“. Das Paar erscheint im Feld „Variablen in der neuen Arbeitsdatei“. In der neuen Datei wird diese Variable unter dem Namen der Variablen der ursprünglichen Arbeitsdatei gespeichert.

Erzeugen eines Datei-Indikators. Durch Anklicken von „Datei-Indikator als Variable:“ erzeugt man eine Variable, in der festgehalten wird, aus welcher Datei der jeweilige Fall entstammt. Per Voreinstellung hat diese Variable den Namen QUELLE01. Der Namen kann durch Eintrag in das Feld geändert werden.

Umbenennen einer Variablen. Variablen der Liste „Nicht gepaarte Variablen:“ können umbenannt werden. Dies soll in unserem Beispiel verwendet werden, um die beiden Variablen PART_AK2 und PARTAKT2 mit gleichem Inhalt, aber unterschiedlichem Namen, gleich zu benennen. Um PARTAKT2 in PART_AK2 umzubenennen, gehen Sie folgt vor.

▷ Markieren Sie dazu den Variablennamen (hier: PARTAKT2).
▷ Klicken Sie auf die Schaltfläche „Umbenennen...“. Eine Dialogbox zum „Umbenennen“ erscheint (⇨ Abb. 7.8).

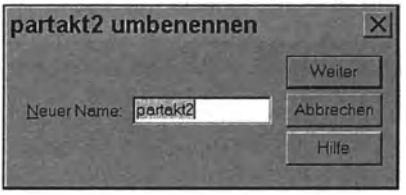


Abb. 7.8. Dialogbox „Umbenennen“ mit neuem Variablennamen

▷ Tragen Sie in das Eingabefeld „Neuer Name:“ den gewünschten Namen ein (hier: PART_AK2).
▷ Bestätigen Sie mit „Weiter“. Die Veränderung des Namens wird in der Liste dadurch kenntlich gemacht, dass alter und neuer Namen durch einen Pfeil verbunden angezeigt werden (hier: **partakt2 -> part_ak2 (+)**).

Auch wenn dadurch ein identischer Name zur komplementären Datei erzeugt wird, paart SPSS die beiden Variablen nicht automatisch nachträglich. Soll eine Paarbil-

dung erfolgen, muss diese ausdrücklich in der oben angegebenen Art durchgeführt werden. In unserem Beispiel würden zunächst die beiden Variablen gleichen Namens im Feld „Nicht gepaarte Variablen:" verbleiben, bis man sie ausdrücklich als Paar definiert. Dann allerdings wird für das Paar nur der gemeinsame Name in die Liste „Variablen in der neuen Arbeitsdatei:" übertragen.

- ▷ Mit „OK" führen Sie die Zusammenfügung aus. Es entsteht die zusammengeführte Datei unter dem Namen „UNBENANNT.SAV".
- ▷ Speichern Sie auf gewohnte Weise die neue Datei unter dem gewünschten Namen (hier: WAHLEN.SAV).

Informationen des Datenlexikons. Alle Informationen des Datenlexikons (Variablen- und Werte-Labels, benutzerdefinierte fehlende Werte und Anzeigeformate) werden aus der Arbeitsdatei übernommen. Nur wenn in der Arbeitsdatei für eine Variable keine solchen Informationen enthalten sind, werden sie aus der externen Datei übernommen. Zusätzliche Werte-Labels oder nutzerdefinierte fehlende Werte werden nicht aus der externen Datei übernommen, wenn entsprechende Werte schon in der Arbeitsdatei definiert sind. Deshalb kann es von Interesse sein, sich genau zu überlegen, welche der zu vereinenden Dateien als Arbeitsdatei benutzt wird, um möglichst viele bzw. die richtigen Informationen aus dem Datenlexikon zu übernehmen.

7.2.2 Hinzufügen neuer Variablen

Hier ist zu unterscheiden, ob für dieselben Fälle Dateien mit unterschiedlichen Variablen zusammengeführt werden (gleichwertige Dateien) oder ob eine Datei als Referenztabelle für die Zuordnung von Merkmalen für mehrere Fälle der anderen Datei dient (eine Datei ist Schlüsseltabelle). Beide Fälle sind unterschiedlich zu behandeln.

Gleichwertige Dateien. Es kann vorkommen, dass man für dieselben Fälle Variablen aus unterschiedlichen Dateien zusammenführen will. Das träfe z.B. zu, wenn Messwerte verschiedener Erhebungszeitpunkte zu Analysezwecken in einer Datei zusammengefasst enthalten sein sollen. Oder es wurden bestimmte Variablen für die Fälle nach erhoben oder sie entstammen unterschiedlichen Quellen. Außerdem kann es vorkommen, dass – aus Mangel an Speicherplatz, wegen Begrenzung der Verarbeitungskapazität des Programms oder aus Gründen der Übersichtlichkeit – Variablen auf mehrere Dateien verteilt wurden, die aber für bestimmte Analysen wieder vereint werden müssen. Man kann solche Dateien zusammenfassen, wenn beide entweder im Format SPSS für Windows oder im SPSS/PC+-Format vorliegen. Außerdem müssen die Fälle in beiden Dateien in der gleichen Reihenfolge sortiert sein. Ist dies nicht der Fall, sortiert man sie vorher. (Wird eine Schlüsselvariable verwendet, müssen sie nach der Schlüsselvariablen in aufsteigender Reihenfolge sortiert werden.)

Beispiel. Nehmen wir Daten des ALLBUS von 1990. Für dieselben Fälle sollen zwei Dateien existieren: In der ersten (ALL1.SAV) sind die Variablen enthalten, die wir für unsere Übungsdatei in Kapitel 2 verwendet haben. In einer zweiten Datei (ALL2.SAV) sind weitere Variablen enthalten, von denen wir jetzt einige

zusätzlich für Analysen benötigen. Dabei handelt es sich um die Variablen, die den Familienstand erfassen (FAMILIEN) sowie die Beurteilung verschiedener Arten kriminellen Verhaltens, nämlich von Alkohol am Steuer, Kaufhausdiebstahl, Schwarzfahren und Steuerhinterziehung (ALKOHOL, KAUFHAUS, SCHWARZ, STEUERA). Beide Dateien enthalten die Fallnummer, in der ersten lautet der Name dieser Variablen allerdings NR, in der zweiten LFD.NR. Weitere ebenfalls enthaltene Variablen sind nicht von Interesse. Gegenüber der ersten Datei fehlt in der zweiten ein Fall. Um die zwei Dateien zu verbinden, gehen Sie wie folgt vor:

- ▷ Öffnen Sie zuerst eine der beiden Dateien und machen Sie diese damit zur Arbeitsdatei (hier: ALL1.SAV).
- ▷ Wählen Sie die Befehlsfolge „Daten", „Dateien zusammenfügen ▷", „Variablen hinzufügen...". Es öffnet sich die Dialogbox „Variablen hinzufügen: Datei lesen".
- ▷ Wählen Sie auf die übliche Weise das gewünschte Verzeichnis und die gewünschte externe Datei, aus der Variablen in die Arbeitsdatei überführt werden sollen (hier: ALL2.SAV).
- ▷ Klicken Sie auf die Schaltfläche „Öffnen". Es öffnet sich die Dialogbox „Variablen hinzufügen aus" (⇨ Abb. 7.9).

Wird keine Schlüsselvariable verwendet, unterstellt das Programm automatisch, dass beide Dateien gleichwertig sind. Dies kann nur genutzt werden, wenn beide Dateien gleich viele Fälle umfassen (also nicht in unserem Beispiel).

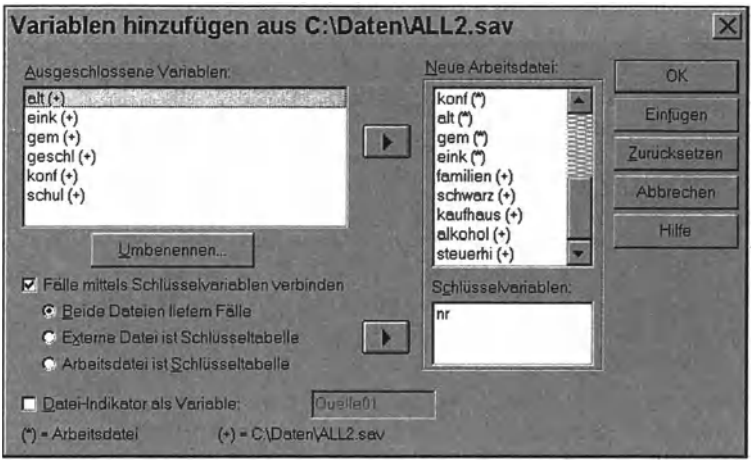


Abb. 7.9. Dialogbox „Variablen hinzufügen aus"

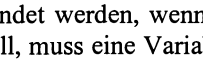
Jetzt gilt es, die Variablenlisten zu überarbeiten. Links findet sich die Liste „Ausgeschlossene Variablen:". Per Voreinstellung enthält sie alle Variablen der externen Datei, die in der Arbeitsdatei schon vorhanden sind. In der Liste „Neue Arbeitsdatei:" befinden sich alle Variablen, die in der neuen Datei vorhanden sind.

Per Voreinstellung sind das alle Variablen, die nur in einer der beiden Dateien vorhanden sind. Diese Listen gilt es nun den Wünschen entsprechend anzupassen.

Ausschließen von Variablen. Markieren Sie in der Liste „Neue Arbeitsdatei" eine Variable, bzw. mehrere Variablen, die ausgeschlossen werden sollen, und übertragen Sie diese mit  in die Liste „Ausgeschlossene Variablen:".

Umbenennen von Variablen. Sie können Variablen umbenennen. Das kann dazu dienen, einen ansprechenderen Namen zu wählen. Vor allem ist es aber nötig, wenn zwei Variablen gleichen Namens, aber unterschiedlichen Inhalts, in der neuen Arbeitsdatei enthalten sein sollen. Dies kann z.B. der Fall sein, wenn Variablen zu verschiedenen Erhebungszeitpunkten gleich benannt wurden, aber als Messzeitpunktsvariablen unterschieden werden sollen. Beide Variablen können dann nur in die Arbeitsdatei aufgenommen werden, wenn eine der beiden Variablen umbenannt wird. Dasselbe gilt, wenn eine Variable als Schlüsselvariable benutzt werden soll, die zwar in den beiden Ausgangsdateien denselben Inhalt hat, aber unterschiedliche Namen besitzt. Dann muss der Name vereinheitlicht werden. Umbenannt werden können nur Variablen aus der Liste der ausgeschlossenen Variablen. Deshalb müssten im letzteren Fall die Variablen zuerst aus der Liste „Variablen in der neuen Arbeitsdatei:" in die Liste „Ausgeschlossene Variablen:" übertragen werden (evtl. für beide durchführen !). Zur Umbenennung gehen Sie wie folgt vor:

- ▷ Markieren Sie die umzubenennende Variable in der Liste „Ausgeschlossene Variablen:".
- ▷ Klicken Sie auf die Schaltfläche „Umbenennen...". Die Dialogbox „Umbenennen" (⇨ Abb. 7.8) öffnet sich.
- ▷ Tragen Sie den neuen Namen in das Eingabefeld „Neuer Name:" ein.
- ▷ Bestätigen Sie mit „Weiter".

Der alte und der neue Name erscheinen im Feld „Ausgeschlossene Variable(n):" *Beispiel:* 

. Wenn gewünscht, kann jetzt die Variable in die Arbeitsdatei übertragen werden.

Verwenden einer Schlüsselvariablen. Eine Schlüsselvariable muss immer dann nicht verwendet werden, wenn beide Dateien gleich viele Fälle umfassen. Ist das nicht der Fall, muss eine Variable vorhanden sein, mit der es möglich ist, die Fälle der beiden Dateien einander zuzuordnen. Die Fallnummer dient in den meisten Fällen diesem Zweck, so auch in unserem Beispiel. Auch wenn eine Schlüsselvariable verwendet wird, müssen die Fälle beider Dateien zuvor nach dieser Variablen geordnet sein. Da die Schlüsselvariable in beiden Dateien vorhanden sein muss, steht sie automatisch im Feld „Ausgeschlossene Variablen:". (Haben Sie aber, wie in unserem Beispiel, unterschiedliche Namen, müssen beide zunächst in die Liste der ausgeschlossenen Namen übertragen werden. Dann erzeugt man den gleichen Namen durch Umbenennung einer der beiden Variablen. Auf diese Weise müsste im Beispiel etwa die Variable LFD.NR in NR umbenannt werden. Jetzt können die Variablen als Schlüsselvariablen verwendet werden.) Um eine Schlüsselvariable zu verwenden, verfahren Sie wie folgt:

- ▷ Klicken Sie auf das Kontrollkästchen „Fälle mittels Schlüsselvariablen verbinden".
- ▷ Damit die Dateien als gleichwertig behandelt werden, müssen Sie jetzt den Optionsschalter „Beide Dateien liefern Fälle" anklicken.
- ▷ Markieren Sie den Namen der Schlüsselvariablen, und übertragen Sie diese durch Anklicken von  in das Feld „Schlüsselvariablen:".

Fälle, die nur in einer der beiden Dateien vorhanden sind, bekommen automatisch für die Variablen, die nur in der anderen Datei vorhanden sind, einen System-Missing-Wert zugewiesen.

Datei-Indikator speichern. Durch Auswahl des Kontrollkästchens „Datei-Indikator als Variable:" kann man wiederum eine Variable erzeugen, die angibt, aus welcher Datei der jeweilige Fall entstammt. Der Name kann beliebig gewählt werden, Voreinstellung ist QUELLE01.

Eine der Dateien ist eine Schlüsseltabelle. Eine weitere interessante Möglichkeit besteht darin, dass man zwei Dateien miteinander verbinden kann, die nicht gleichwertig sind. Die Dateien enthalten unterschiedliche Typen von Fällen und Informationen. In einer der Dateien stehen jeweils bei einem Fall Informationen, die mehreren Fällen der anderen Datei zugeordnet werden. Die erstgenannte Datei dient dann als Referenztabelle für die Zuordnung von Werten zur anderen Datei. Diese Option ist vor allem deshalb interessant, weil es dadurch möglich ist, Daten aus verteilten Tabellen, wie sie dem modernen Datenmanagement entsprechen, zur statistischen Bearbeitung zusammenzufügen. Um Redundanz bei der Dateneingabe und Datenhaltung zu vermeiden, werden Daten in relationalen Datenbanken möglichst so auf mehrere verschiedene Tabellen verteilt, dass man den Aufwand für die Dateneingabe minimiert. So wird z.B. ein Betrieb eine getrennte Datei jeweils für Kundendaten, Bestellungen und Artikel halten, die aber für bestimmte Zwecke, z.B. der Rechnungsstellung, aber auch statistische Auswertungen kombiniert werden können. Ähnliches gilt für Mehrebenenanalysen. Sollen etwa in einer Wahluntersuchung einerseits individuelle Merkmale, andererseits Kollektivmerkmale, etwa Eigenschaften des Bundeslandes, verwendet werden, wird man die Merkmale der Bundesländer in einer Datei, die Individualdaten der befragten Wähler in einer anderen halten. Beide lassen sich aber bei relationalen Datenbanken über Schlüsselvariablen verknüpfen. In SPSS können solche Datenbanken zusammengeführt werden, aber nur in der Weise, dass die Informationen der Referenztabellen allen zutreffenden Fällen der anderen Tabelle zugeordnet werden.

Beispiel. Nehmen wir als Beispiel Daten der Schuldnerberatung (VZ.SAV). Dort hatten die meisten Schuldner mehr als einen Kredit aufgenommen, z.T. bei unterschiedlichen Banken und zu unterschiedlichen Zinskonditionen. Wir haben dies in der Originaldatenmatrix zunächst so erfasst, dass sieben Variablen für bis zu sieben Kredite vorgesehen waren. Jeweils auch eine entsprechende Variable für die Zinshöhe, den Namen der Bank usw.. Die moderne Datenhaltung wird normalerweise anders verfahren. Sie wird eine Datei mit den allgemeinen Daten der Schuldner, eine Kreditdatei mit den Kreditdaten und eine Bankendatei mit den Bankdaten erstellen. Jeweils zwei Dateien ist eine Schlüsselvariable gemeinsam, mit der sie verknüpft werden können, z.B. wird eine Klientennummer in der

Schuldnerdatei und in der Kreditdatei enthalten sein und eine Bankennummer sowohl in der Bankendatei als auch in der Kreditdatei. Entsprechend dieser Datenhaltung, wurde auch in unserem Falle eine zusätzliche Kreditdatei erstellt. Diese enthält alle kreditspezifischen Daten, in unserem Beispiel beschränkt auf Kredithöhe (KREDIT), Kreditzinsen (ZINS) und Bankennummer (BANKNR). Eine solche Datei erleichtert es, kreditbezogene Auswertungen vorzunehmen. Man kann z.B. ohne weiteres die durchschnittliche Kredithöhe, die durchschnittliche Zinsbelastung usw. berechnen. Das wäre in der Ausgangsdatei nur mit einigem Aufwand möglich, da ja solche Daten wie Kredithöhe über sieben Variablen verstreut sind. Will man jetzt auch Klientendaten, wie Geschlecht oder Einkommenshöhe, mit diesen Kreditdaten in Beziehung bringen, etwa um eine Korrelation zwischen Einkommenshöhe und Kredithöhe zu berechnen, müssen die Klientendaten der Kreditdatei hinzugefügt werden. So werden z.B. allen Krediten, die ein bestimmter Schuldner aufgenommen hat, dessen Geschlecht, Einkommenshöhe usw. zugeordnet. (Umgekehrt ist es allerdings nicht möglich, einem Fall die Daten mehrerer Kredite zuzuordnen, die ja dann in verschiedenen Variablen gespeichert werden müssen. Wenn man solche Daten benötigt, kann leider eine Mehrfacheingabe nicht verhindert werden.)

Eine Datei VZ.SAV mit den Klienten- und Kreditdaten haben Sie evtl. im Kapitel 3.1 mit dem Dateneditor selbst erstellt. Wenn Sie die folgenden Darstellungen anhand von Daten nachvollziehen wollen, können Sie daraus leicht die beiden hier zu kombinierenden Dateien erzeugen. (Falls Sie sich die Datendiskette beschafft haben, laden Sie die Dateien von dieser Diskette ⇨ Anhang C.) Die Datei mit den Klientendaten KLIENT.SAV erzeugen Sie, indem Sie alle kreditbezogenen Variablen aus der Ausgangsdatei löschen. Die Datei KREDITE.SAV zu erstellen, ist etwas komplizierter. Löschen Sie aus der Ausgangsdatei alle klientenbezogenen Daten, außer der laufenden Nummer. Kopieren Sie einmal die Fallnummern, so dass sie die Fallnummern zweimal unter einander stehen haben. Schneiden Sie die Daten der Variablen KREDIT2 aus und kopieren Sie diese hinter die dazugehörigen neuen Fallnummern in die Spalte KREDIT1. Genauso gehen Sie bei Übertragung der Daten aus ZINS2 in die Spalte ZINS1 vor. Löschen Sie die Spalten KREDIT2 und ZINS2. Jetzt haben Sie eine Datei, in der alle Kredite gleichwertig behandelt werden.

Grundsätzlich entspricht die Vorgehensweise der für gleichwertige Dateien geschilderten. Aber wichtig: Sie müssen auf jeden Fall mit einer Schlüsselvariablen arbeiten, die in beiden Dateien enthalten ist. Und beide Dateien müssen zuvor nach der Schlüsselvariablen in aufsteigender Ordnung sortiert sein. In unserem Beispiel ist die Schlüsselvariable die Klientennummer (NR). Öffnen Sie zuerst eine der beiden Dateien (etwa KLIENT.SAV), und wählen Sie die Befehlsfolge:

- ▷ „Daten", „Fälle sortieren...". Die Dialogbox „Fälle sortieren" erscheint (⇨ Abb. 7.10).
- ▷ Übertragen Sie den Namen der Sortiervariablen NR aus der Quellvariablenliste in das Feld „Sortieren nach", und bestätigen Sie mit „OK". Speichern Sie die sortierte Datei ab.

Wiederholen Sie dasselbe für die andere Datei.

Abb. 7.10. Dialogbox „Fälle sortieren"

- ▷ Öffnen Sie – falls noch nicht geschehen – die Datei, die Sie als Arbeitsdatei benutzen wollen (hier: KREDITE), und wählen Sie die Befehlsfolge „Daten", „Dateien zusammenfügen ▷", „Variablen hinzufügen...". Es öffnet sich die Dialogbox „Variablen hinzufügen: Datei lesen".
- ▷ Übertragen Sie den Namen der zu verbindenden Datei (hier: KLIENT.SAV) aus der Auswahlliste in das Eingabefeld „Dateiname:" (oder schreiben Sie ihn ein).
- ▷ Klicken Sie auf die Schaltfläche „Öffnen". Die Dialogbox „Variablen hinzufügen aus" erscheint (⇨ Abb. 7.11).

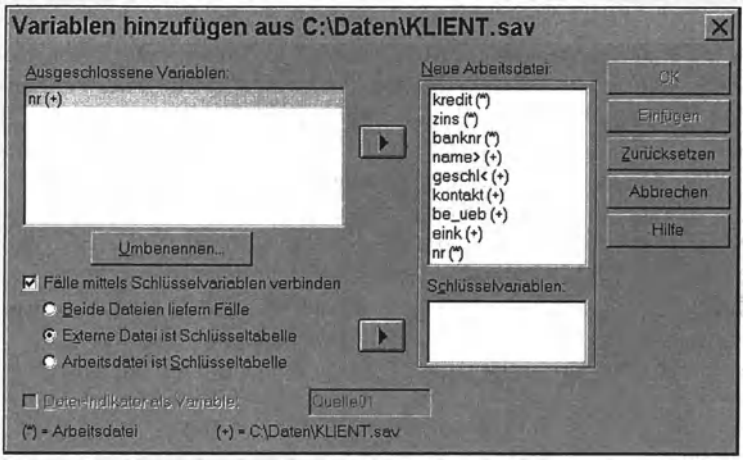


Abb. 7.11. Dialogbox „Variablen hinzufügen aus"

- ▷ Klicken Sie auf das Auswahlkästchen „Fälle anhand von Schlüsselvariablen verbinden".
- ▷ Markieren Sie die Schlüsselvariable NR in dem Feld „Ausgeschlossene Variablen:". Übertragen Sie diese mit [▸] in das Feld „Schlüsselvariablen:".

Jetzt müssen Sie noch angeben, in welcher der Dateien die Referenztabelle steht. Es kann sowohl die Arbeitsdatei als auch die externe Datei sein. (Es ist immer die Tabelle, in der ein Fall Informationen für mehrere Fälle der anderen enthält, in unserem Beispiel KLIENT.SAV. Beachten Sie das nicht, verweigert SPSS unter be-

stimmten Umständen mit einer Fehlermeldung die Ausführung oder sie führt zu einem unsinnigen Ergebnis.)

- ▷ Wählen Sie über Anklicken des Optionsschalters entweder die externe oder die Arbeitsdatei als Schlüsseltabelle (hier die externe).
- ▷ Bestätigen Sie mit „OK".

SPSS warnt, dass die Verbindung über Schlüsselvariablen misslingt, wenn die Datei nicht in aufsteigender Reihenfolge der Schlüsselvariablen sortiert ist.

- ▷ Bestätigen Sie mit „OK". Die erweiterte Datei wird gebildet und standardmäßig als UNBENANNT.SAV bezeichnet. Sie können sie speichern und umbenennen. (Andere Optionen, wie Umbenennen von Variablen, werden analog dem oben beschriebenen Vorgehen benutzt.)

7.3 Gewichten von Daten

SPSS bietet auch eine Möglichkeit, Daten zu gewichten. Das Vorgehen bei einer Gewichtung ist bereits in Kapitel 2.7 geschildert. Es wird hier nur in seinen Grundzügen dargestellt.

Eine Gewichtung von Daten wird vor allem benutzt, um Verzerrungen von Stichproben gegenüber der Grundgesamtheit, die sie repräsentieren sollen, zu korrigieren. Dazu muss zunächst eine Gewichtungsvariable (z.B. mit dem Namen GEWICHT) gebildet werden. In dieser wird jedem Fall in Abhängigkeit zu einem bestimmten Merkmal als Wert ein Gewicht zugewiesen, mit dem seine anderen Werte später bei statistischen Auswertungen multipliziert werden sollen (*Beispiel:* Männer bekommen den Wert 0,84, Frauen den Wert 1,21). Die Gewichte können eingetippt werden. Häufiger wird man die Variable aber durch eine Datentransformation bilden.

Um die Gewichtung für nachfolgende statistische Auswertungen wirksam werden zu lassen, wählen Sie dann die Befehlsfolge „Daten", „Fälle gewichten...". Es öffnet sich die Dialogbox „Fälle gewichten" (⇨ Abb. 2.26). Dort klicken Sie auf den Optionsschalter „Fälle gewichten mit" und übertragen die Gewichtungsvariable (hier: GEWICHT) aus der Liste der Quellvariablen in das Eingabefeld „Häufigkeitsvariable:". Bestätigen Sie mit „OK".

Die Gewichtung wirkt sich direkt auf alle bei einer Auswertung benutzten Variablen aus. Alle Daten werden so umgerechnet, als gäbe es entsprechend weniger Fälle in den schwächer gewichteten Gruppen und mehr in den stärker gewichteten (im Beispiel weniger Männer und mehr Frauen).

Wollen Sie die Gewichtung nicht mehr oder vorübergehend nicht verwenden, können Sie diese durch Auswählen des Optionsschalters „Fälle nicht gewichten" wieder ausschalten. Der aktuelle Status wird in der Statuszeile angezeigt.

Beachten Sie. Speichern Sie eine Datei mit dem Status „Fälle gewichten mit", so ist nach dem neuen Öffnen der Datei zwar der Optionsschalter „Fälle nicht gewichten" durch einen schwarzen Punkt als ausgewählt gekennzeichnet, in Wirklichkeit bleibt aber die Gewichtung erhalten, was auch die Statuszeile anzeigt. Wollen Sie die Gewichtung

ausschalten, müssen Sie ausdrücklich noch einmal „Fälle nicht gewichten" auswählen und mit „OK" bestätigen.

7.4 Aufteilen von Dateien und Verarbeiten von Teilmengen der Fälle

Manchmal kann es von Interesse sein, eine Datei aufzuteilen und die so gewonnenen Teilgruppen getrennt zu analysieren. Oder man wünscht, nur einen bestimmten Teil der Fälle zu betrachten. Zu diesem Zwecke bietet SPSS mehre Möglichkeiten an.

7.4.1 Aufteilen von Daten in Gruppen

Die Datei WAHLEN.SAV setzt sich aus den Angaben von zwei Wählerbefragungen zu unterschiedlichen Zeitpunkten zusammen. Für verschiedene Analysen kann es von Interesse sein, die Daten der beiden Zeitpunkte getrennt zu betrachten. Als diese Datei in Kap. 7.2.1 aus den Dateien WAHLEN1 und WAHLEN2 gebildet wurde, haben wir als Indikator für die Herkunft der Fälle die Variable QUELLE01 gebildet. Deshalb ist es möglich, die Datei WAHLEN auf Grundlage dieser Variablen nach den Erhebungszeitpunkten wieder in zwei Unterdateien aufzuteilen. Dann können Prozeduren, je nach Bedarf, entweder für alle Daten gemeinsam oder nur für jede Untergruppe getrennt durchgeführt werden. Bei Verwendung der Option „Gruppen vergleichen" werden die beiden Gruppen getrennt analysiert, die Ergebnisse für alle Gruppen aber in gemeinsamen Tabellen ausgegeben, bei Verwendung von „Ausgabe nach Gruppen aufteilen" entsteht für jede Gruppe eine eigene Ausgabe. Um eine Aufteilung vorzunehmen und getrennt Ausgaben für die Gruppen zu erhalten, gehen Sie wie folgt vor:

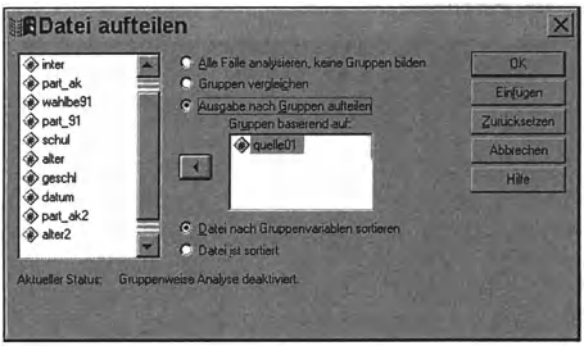


Abb. 7.12. Dialogbox „Datei aufteilen"

▷ Wählen Sie die Befehlsfolge „Daten", „Datei aufteilen...". Die Dialogbox „Datei aufteilen" erscheint (⇨ Abb. 7.12).

▷ Klicken Sie zuerst auf den Optionsschalter „Ausgabe nach Gruppen aufteilen".
▷ Übertragen Sie die zur Aufteilung verwendete Variable aus der Quellvariablenliste in das Feld „Gruppen basierend auf:".

Sie können mehrere Gruppierungsvariablen kombinieren. Es werden aber immer alle vorhandenen gültigen Werte der Variablen zur Gruppierung verwendet, so dass Sie auf dieser Ebene keine Umdefinition der Gruppen vornehmen können. Außerdem geht die Prozedur die Fälle in ihrer Reihenfolge durch und bildet jedesmal, wenn sie auf einen neuen Wert trifft, eine neue Gruppe. Deshalb müssen die Fälle vor Durchführung der Prozedur nach den Werten der Gruppierungsvariablen geordnet werden. Ist dies noch nicht geschehen oder sind Sie unsicher:

▷ Wählen Sie den Optionsschalter „Datei nach Gruppenvariablen sortieren". Ansonsten können Sie die Option „Datei ist sortiert" verwenden. Der Statusanzeige zeigt noch „Gruppenweise Analyse deaktiviert".
▷ Mit „OK" bestätigen Sie die Eingabe. Die Prozedur wird durchgeführt, die Statusanzeige verändert sich in „Ausgabe organisiert nach:" und zeigt die Gruppierungsvariable an.

Wurde eine Sortierung vorgenommen, sind die Daten im Dateneditorfenster in der neuen Anordnung zu sehen. Für Ihre weiteren Prozeduren können Sie wahlweise die Aufteilung der Daten ein- oder ausschalten.

7.4.2 Teilmengen von Fällen auswählen

Man kann auf vier verschiedene Weisen Teilmengen von Fällen für die Analyse auswählen:

❒ Fälle werden ausgewählt, wenn bestimmte Bedingungen zutreffen.
❒ Fälle werden aufgrund einer Filtervariablen ausgewählt.
❒ Ein bestimmter Zeit- oder Fallbereich wird ausgewählt.
❒ Es wird eine Zufallsstichprobe von Fällen ausgewählt.

Auswählen mit einem Bedingungsausdruck. Die Datei ALLBUS90.SAV entstammt einer Untersuchung, bei der bestimmte Fragen nur der Hälfte der Befragten gestellt wurden. Entsprechend wird zwischen dem Split 1 und dem Split 2 unterschieden. In Variable VN ist kodiert, ob ein Fall zu Split 1 oder Split 2 gehört. Wenn man eine Frage auswertet, die nur einem der Splits gestellt wurde, ist es sinnvoll, die Analyse auf die zutreffenden Fälle zu begrenzen. Das kann z.B. mit Hilfe eines Bedingungsausdruckes geschehen. Dazu gehen Sie wie folgt vor:

▷ Wählen Sie die Befehlsfolge „Daten", „Fälle auswählen...". Die Dialogbox „Fälle auswählen" erscheint (⇨ Abb. 7.13).
▷ Klicken Sie auf den Optionsschalter „Falls Bedingung zutrifft".
▷ Klicken Sie auf die Schaltfläche „Falls...". Die Dialogbox „Fälle auswählen: Falls" erscheint.

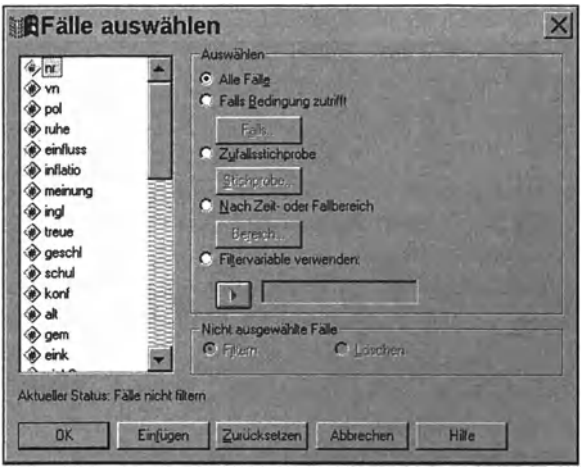


Abb. 7.13. Dialogbox „Fälle auswählen"

Hier können Sie dann in dem Eingabefeld den notwendigen Bedingungsausdruck zusammenstellen. Der Bedingungsausdruck muss zumindest einen Variablennamen enthalten. Ansonsten sind möglich:

- Werte bzw. Wertebereiche
- Arithmetische Ausdrücke
- Logische Ausdrücke
- Funktionen

Man kann auf diese Weise sehr komplexe Bedingungsausdrücke konstruieren. In unserem Beispiel wird lediglich der Wert 1 (entspricht Split 1) für die Variable „VN" (Versionsnummer) als Bedingung gesetzt (die Bedingung lautet „VN = 1").

- Bestätigen Sie die Eingabe mit „Weiter". Die Dialogbox „Fälle auswählen" öffnet sich erneut.
- Durch Anwahl einer der Optionen in der Gruppe „Nicht ausgewählte Fälle" kann weiter bestimmt werden, wie die nicht ausgewählten Fälle behandelt werden sollen:
  - *Filtern.* Die Fälle werden nicht für die weiteren Prozeduren verwendet, bleiben aber erhalten. Diese Option ist voreingestellt.
  - *Löschen.* Die Fälle werden gänzlich aus der Datei gelöscht. Man erhält dann eine verkleinerte Datei, die nur noch die ausgewählten Fälle umfasst. Diese Option sollte man mit Vorsicht verwenden. Leicht können damit Daten verloren gehen. Sicherheitshalber sollte man die neue, gekürzte Datei sofort unter neuem Namen speichern.
- Mit „OK" wird die Prozedur ausgeführt. Falls die Option „Filtern" gewählt wurde, zeigt die Statuszeile nach Ausführung die Meldung „Fälle anhand der Variablen ... filtern" und an die Daten im Dateneditor wird eine Filtervariable

FILTER_$ angehängt mit den Werten „1" (Label: „Ausgewählt") und „0" (Label: „Nicht ausgewählt"), die auch mit abgespeichert werden kann.

Die Filterung kann jederzeit wieder ausgeschaltet werden, wenn man in der Gruppe „Auswählen" die Option „Alle Fälle" auswählt.

Filtervariable verwenden. Diese Option dient im wesentlichen dazu, schon gebildete und mit abgespeicherte Filtervariablen anzuwenden. Die zur Analyse benötigten Fälle müssen auf einer numerischen Variablen einen von Null verschiedenen Wert, der kein Missing-Wert ist, besitzen, die auszusortierenden Fälle dagegen mit Null und/oder einem Missing-Wert verkodet sein. Dann kann man diese Variable als Filtervariable verwenden. Das sollte man evtl. schon bei der Verschlüsselung berücksichtigen und entsprechenden Fällen auf geeigneten Variablen den Wert 0 vergeben. (Häufig wird das bei der Verschlüsselung von „nicht zutreffenden Fragen" der Fall sein.) *Beispiel:* Wenn Sie, wie gerade geschildert, für ALLBUS90.SAV eine Variable FILTER_$ erzeugt haben, in der Split 1 mit 1 und Split 2 mit 0 kodiert ist und diese mit den Daten abspeichern, können Sie in Zukunft den Split 1 unter Verwendung dieser Filtervariablen auswählen. Um Fälle mit einer Filtervariablen auszuwählen, gehen Sie wie folgt vor.

- ▷ Wählen Sie die Befehlsfolge „Daten", „Fälle auswählen...". Die Dialogbox „Fälle auswählen" erscheint (⇨ Abb. 7.13).
- ▷ Klicken Sie auf den Optionsschalter „Filtervariable verwenden".
- ▷ Übertragen Sie die Filtervariable (hier: FILTER_$) aus der Variablenliste in das Feld „Filtervariable verwenden:".
- ▷ Bestimmen Sie durch Auswahl der zutreffenden Option der Gruppe „Nicht ausgewählte Fälle", ob die nicht ausgewählten Fälle nur ausgefiltert oder gelöscht werden sollen.
- ▷ Bestätigen Sie mit „OK". Die Statuszeile zeigt die Meldung „Fälle anhand der Variablen ... filtern".

Die Filterung kann auch hier jederzeit wieder ausgeschaltet werden, wenn man in der Gruppe „Auswählen" die Option „Alle Fälle" anwählt.

Auswählen von Zeit- oder Fallbereichen. Mit dieser Option kann ein abgegrenzter Teil der Fälle oder – in Zeitreihen – ein Zeitbereich ausgewählt werden. Dazu gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Daten", „Fälle auswählen...". Die Dialogbox „Fälle auswählen" erscheint (⇨ Abb. 7.13).
- ▷ Klicken Sie auf den Optionsschalter „Nach Zeit- oder Fallbereich" und die Schaltfläche „Bereich...". Die Dialogbox „Fälle auswählen: Bereich" öffnet sich.
- ▷ Legen Sie in der Gruppe „Beobachtung:" durch Eintrag in die Eingabefelder „Erster Fall" und „Letzter Fall" den Bereich fest, und bestätigen Sie mit „Weiter" und „OK".

Auswählen einer Zufallsstichprobe. Um Speicherplatz und/oder Rechenzeit zu sparen, wird man mitunter eine Zufallsstichprobe aus einem größeren Datensatz ziehen. Eine solche Stichprobe kann z.B. für Lehrzwecke ausreichen. Auch für die

Entwicklung und Erprobung von Programmen genügt zumeist eine kleine Fallzahl. Hat man sehr große Fallzahlen in einer Datei, kann es sogar sein, dass man auch eine ernsthafte Analyse nur mit einer Stichprobe durchführen kann. Unsere Übungsdatei ALLBUS90.SAV ist z.B. dadurch zustande gekommen, dass aus der Originaldatei des ALLBUS 1990 eine Stichprobe von ungefähr 10 % der Fälle ausgewählt wurde.

Startwert Zufallszahlen. SPSS wählt die Fälle für die Stichprobe mit Hilfe eines Pseudo-Zufallszahlengenerators aus. Das heißt, die Fallzahl der ausgewählten Fälle wird nicht wirklich ausgelost, sondern nach einem Algorithmus berechnet. Dabei werden fortlaufende Zufallszahlen, ausgehend von einer Startposition, verwendet. Beginnt man von derselben Startposition aus, kommt daher bei Verwendung derselben Auswahlalternativen immer genau die gleiche Stichprobe zustande. Um dies zu verhindern, verwendet SPSS für jede Zufallsstichprobe innerhalb einer Sitzung einen anderen Startwert, den es aus der internen Uhr des Rechners gewinnt. Es kann aber sein, dass man gerade eine Stichprobe reproduzieren will, vielleicht, um später den Fällen neue Variablen anzufügen, vielleicht, um bei einem unbeabsichtigten Datenverlust die Datenbasis wiederherstellen zu können. Will man das sichern, sollte man von vornherein einen festen Startwert benutzen. Erlaubt sind ganze Zahlen bis 2.000.000.000. Einen Startwert setzt man, mit folgender Befehlsfolge:

▷ „Transformieren", „Startwert für Zufallszahlen...". Es öffnet sich die Dialogbox „Startwert für Zufallszahlen" (⇨ Abb. 7.14).

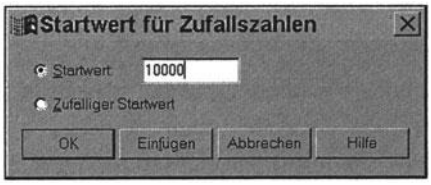


Abb. 7.14. Dialogbox „Startwert für Zufallszahlen" mit eingefügtem Startwert

▷ Geben Sie den Startwert in das Eingabefeld ein.
▷ Bestätigen Sie mit „OK".

Hinweis. Das Fenster neben dem Optionsschalter „Startwert:" den Schalter „Zufälliger Startwert". Letzterer ist (umgekehrt wie bei früheren Versionen) beim Beginn einer zunächst angewählt. Das Eingabefeld beim Optionsschalter „Startwert" enthält den Wert 2000000. Diese Voreinstellung wirkt auch bei der ersten Zufallsoperation, wenn „Zufälliger Startwert" gewählt ist. Erst danach werden zufällig andere Startwerte verwendet. Wählt man dagegen „Startwert:" aus und gibt einen beliebigen Startwert ein, ist zu beachten, dass dieser nur einmal bei der nächsten Zufallsoperation wirkt, auch wenn man diese Option angewählt lässt. Die nächste Operation beginnt mit einem anderen zufälligen Startwert. Will man dagegen denselben Startwert weiter benutzen, muss dieser vor jeder Zufallsoperation wieder mit „OK" ausdrücklich bestätigt werden.

Um eine Stichprobe zu ziehen, gehen Sie wie folgt vor :

- ▷ Wählen Sie die Befehlsfolge „Daten", „Fälle auswählen...". Die Dialogbox „Fälle auswählen" erscheint (⇨ Abb. 7.13).
- ▷ Klicken Sie auf den Optionsschalter „Zufallsstichprobe" und die Schaltfläche „Stichprobe...". Die Dialogbox „Fälle auswählen: Zufallsstichprobe" erscheint (⇨ Abb. 7.15).

Für die Bildung der Stichprobe stehen zwei Alternativen zur Verfügung:

- ❒ *Ungefähr* ein festzulegender Prozentsatz der Fälle (z.B. 10 %). Der Prozentsatz wird in das dafür vorgesehene Feld eingegeben.
- ❒ *Exakt* eine festgelegte Zahl von Fällen (z.B. 300) aus den ersten x Fällen (= einer festzulegenden Zahl von Fällen kleiner/gleich der Gesamtzahl). Will man aus sämtlichen Fällen der Ausgangsdatei auf diese Weise eine Stichprobe ziehen, muss der Wert im Eingabefeld „aus den ersten ... Fällen" gleich der Gesamtzahl der Fälle gesetzt werden.

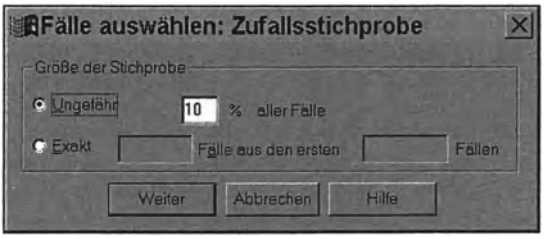


Abb. 7.15. Dialogbox „Fälle auswählen: Zufallsstichprobe"

Wie bei den anderen Auswahlverfahren auch, wird eine Filtervariable gebildet, die mit den Daten gespeichert werden kann. Für die Behandlung der nicht ausgewählten Daten kann zwischen „Filtern" und „Löschen" gewählt werden. Letzteres wird man wählen, wenn man auf Dauer mit dem verkleinerten Datensatz zu arbeiten beabsichtigt. Wählt man „Filtern", wird eine Filterung vorgenommen, wie oben bereits bei der Auswahl durch Bedingungsausdrücke beschrieben. SPSS fügt den Daten eine Filtervariable mit den Werten „Ausgewählt" und „Nicht ausgewählt" an. Die Statuszeile meldet nach Beendigung der Prozedur „Fälle anhand der Variablen ... filtern". Die Filterung kann durch Auswahl „Alle Fälle" ausgeschaltet werden.

Soll eine feste Zahl von Fällen ausgewählt werden, wählen Sie in der Dialogbox „Fälle auswählen: Zufallsstichprobe" die Option „Exakt" und setzen im ersten Eingabefeld den Wert für die Größe der gewünschten Stichprobe ein, im zweiten die Zahl der ersten Fälle der Datendatei, aus denen ausgewählt werden soll.

7.5 Erstellen einer Datei mit aggregierten Variablen

Aus den Variablen einer vorhandenen Datei kann man neue Variablen einer aggregierten Datei erzeugen. Hat man etwa eine Datei, deren Fälle Personen sind und in der als Variablen Bundesland und monatliches Einkommen enthalten sind, so kann man daraus eine neue aggregierte Datei gewinnen. Darin könnten Fälle die Bundesländer und die Variable das Durchschnittseinkommen der Bewohner sein. Man unterscheidet dabei zwei Variablentypen:

- *Break-Variable(n).* Es muss in der Ausgangsdatei mindestens eine Variable vorhanden sein, deren Ausprägungen jeweils einen Fall der neuen Variablen ergeben. In unserem Falle ist es die Variable Bundesland. Jedes Bundesland wird in der aggregierten Variablen ein Fall.
- *Aggregierungvariable(n).* Die Variablen, aus denen die Werte der neuen Fälle berechnet werden, sind die Aggregierungsvariablen. Ihre Werte kommen dadurch zustande, dass auf Basis einer geeigneten Aggregierungsfunktion sämtliche Werte der Fälle einer Kategorie der Break-Variablen zu einem einzigen Wert zusammengefasst werden. In unserem Beispiel werden u.a. sämtliche Einkommen der Befragten aus einem Bundesland (z.B. Bayern) zu einem Durchschnittswert zusammengefasst.

Sinnvoll ist eine solche Aggregierung nur, wenn die auf diese Weise neu gewonnenen Variablen Eigenschaften der neuen aggregierten Einheit messen. Ginge es nur um den Vergleich des Durchschnittseinkommens in den Bundesländern, würde man in unserem Beispiel besser die Statistikprozedur „Mittelwerte vergleichen" verwenden. Soll aber ein spezielles Merkmal des Bundeslandes (z.B. ein Indikator für seine ökonomische Kraft) ermittelt werden, das mit anderen Merkmalen (etwa Siedlungsdichte, geographischer Lage) in Beziehung gesetzt werden soll, dann ist die Aggregation angebracht.

Es kann auch sinnvoll sein, die Daten einer solchen aggregierten Datei für eine Mehrebenen- oder Kontextanalyse zu verwenden. *Beispiel:* Nehmen wir eine Frage aus der Wahlforschung: Man nimmt an, das Wahlverhalten einer Person hänge sowohl von seinen persönlichen sozialen Merkmalen als auch denen seines Wohnumfeldes ab. Arbeiter sein wäre z.B. ein persönliches Merkmal, in einem Arbeitergebiet zu wohnen ein Merkmal des Wohnumfeldes. In diesem Beispiel könnte man evtl. aus einer Personendatei durch Aggregation eine Datei mit Merkmalen von Wohnumfeldern gewinnen, etwa, indem man Wohnbezirke mit mehr als 50 % Arbeiteranteil als Arbeiterviertel klassifiziert. Diese Datei könnte wieder (wie weiter oben geschildert) als Referenztabelle benutzt werden, um der Personendatei die Merkmale des Wohnumfeldes anzufügen. Nach Vollzug des ganzen Prozesses wären dann für jede Person beide Arten von Variablen verfügbar, einerseits ihr persönliches Merkmal (Arbeiter), andererseits das Merkmal des Wohnumfeldes (Arbeitergebiet). (Ein Arbeiter muss keinesfalls in einem Arbeitergebiet wohnen.) Dadurch wird der Einfluss beider Merkmale, sowohl des persönlichen als auch des Kontextmerkmals, auf das Wahlverhalten untersuchbar.

Nehmen wir folgende Aufgabe: Aus den Daten des ALLBUS90.SAV soll eine aggregierte Datei für die Bundesländer gewonnen werden. Diese soll folgende ag-

gregierten Variablen enthalten: Durchschnittseinkommen der Erwerbstätigen, Streuung der Einkommen, durchschnittliche Arbeitszeit, Arbeitslosenanteil, Katholikenanteil, Protestantenanteil und Befragtenzahl. Sofern dies nötig erscheint, sollen die neu gebildeten Variablen sinnvolle Variablennamen erhalten. Die neue Datei soll unter dem Namen LAENDER.SAV gespeichert werden. Um die Fälle zu aggregieren, gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Daten", „Aggregieren...". Die Dialogbox „Daten aggregieren" öffnet sich (⇨ Abb. 7.16).
- ▷ Übertragen Sie die Break-Variable (BUNDL) aus der Quellvariablenliste in das Eingabefeld „Break-Variable(n):".
- ▷ Übertragen Sie die Aggregierungsvariablen (EINK,...) aus der Quellvariablenliste in das Eingabefeld „Variablen aggregieren:".

Dabei ist folgendes zu beachten:

- ❒ Standardmäßig wird als Aggregierungsfunktion das arithmetische Mittel benutzt. Die Aggregierungsfunktion kann aber über die Option „Funktion..." geändert werden. Welche Funktion benutzt wurde (gegebenenfalls mit welchen Werten), wird hinter dem neuen Namen der aggregierten Variablen angezeigt.
- ❒ Standardmäßig wird ein neuer Name für die aggregierte Variable vergeben, der aus dem alten Namen und dem Zusatz _1 (bei Mehrfachverwendung _2 usw.) besteht. Dieser kann über die Option „Name & Label..." geändert werden. Zusätzlich kann dort ein Variablen-Label bestimmt werden.
- ❒ Jede Variable der Auswahlliste kann mehrmals zur Bildung von Aggregatdaten verwendet werden. Dabei kann man unterschiedliche Aggregierungsfunktionen anwenden.

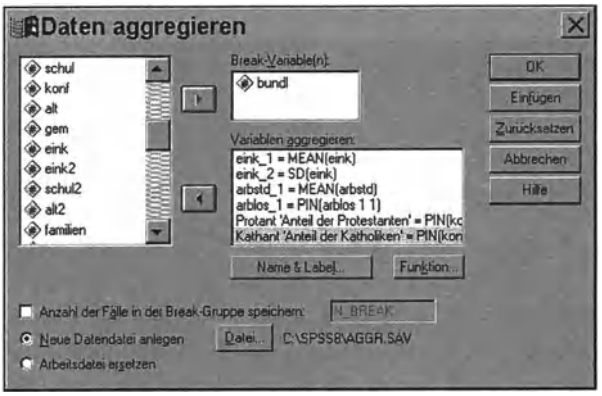


Abb. 7.16. Dialogbox „Daten aggregieren" mit Break- und Aggregierungsvariablen

- ❒ Es wird für jeden Fall ein Wert vergeben. Deshalb müssen qualitative Variablen mit mehr als zwei Ausprägungen mit Hilfe der Option „Funktionen" dichotomisiert werden, um zu sinnvollen Ergebnissen zu gelangen. Aus einer Variablen KONFESSION muss z.B. durch Auswahl einer geeigneten Aggregierungsfunk-

tion eine dichotomische Variable gemacht werden, etwa als Dichotomie „Katholiken" – „Nichtkatholiken". Sinnvoll ist es z.B., den Anteil oder den Prozentsatz einer der beiden Ausprägungen als Wert auf der aggregierten Variablen zu verwenden.

In Abb. 7.16 sehen Sie das Ergebnis der Eingaben unseres Beispiels. Zunächst wurde die Variable EINK zweimal als Aggregierungsvariable verwendet. Automatisch bekamen die Aggregierungsvariablen die Namen EINK_1 und EINK_2. Automatisch wurde bei beiden Variablen zunächst die Aggregierungsfunktion „Mittelwert" (Mean) angenommen. Die Variable EINK_2 sollte aber das Streuungsmaß Standardabweichung (SD) enthalten. Um das in Abb. 7.16 angezeigte Ergebnis für EINK_2 zu erreichen, müssen Sie wie folgt verfahren:

- ▷ Markieren Sie EINK_2, und klicken Sie auf die Schaltfläche „Funktion...". Die Dialogbox „Daten aggregieren: Aggregierungsfunktion" erscheint (⇨ Abb. 7.17).
- ▷ Klicken Sie auf den gewünschten „Optionsschalter" (hier: Standardabweichung), und bestätigen Sie mit „Weiter".

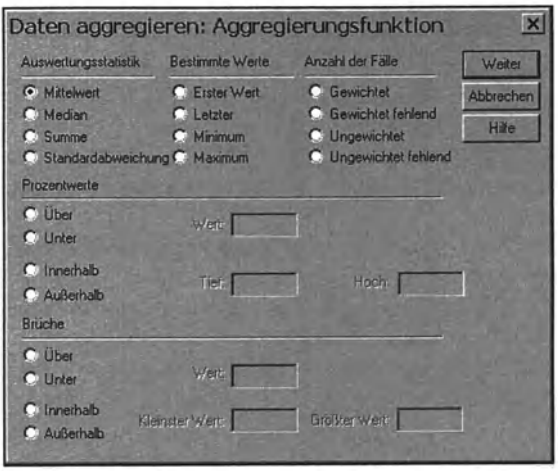


Abb. 7.17. Dialogbox „Daten aggregieren: Aggregierungsfunktion"

Entsprechend können Sie die Funktionen auf andere Variablen anwenden. Während die Funktionen im oberen Teil der Dialogbox sich für metrische Daten eignen, sind die im unteren Teil insbesondere für qualitative Daten von Bedeutung. Sie stellen verschiedene Möglichkeiten zur Dichotomisierung und zur Zusammenfassung der Werte zur Verfügung.

Man kann die Werte auf zwei Arten dichotomisieren. Im ersten Falle werden die Werte durch die Festlegung eines Wertes in einen oberen und unteren Bereich aufteilt. (Je nach Wunsch wird für die Aggregierung der obere oder untere Teil der Werte benutzt.) Im zweiten Falle unterteilt man den Wertebereich durch Festlegung einer Unter- und Obergrenze („Kleinster Wert:" bzw. „Größter Wert:") in ei-

nen Teil innerhalb und einen außerhalb dieser Grenzen. (Je nach Wunsch werden die Fälle innerhalb oder außerhalb des Bereichs zur Aggregierung benutzt.) Die Zusammenfassung in der Aggregatvariablen erfolgt als Anteilszahl (zwischen 0 und 1) oder als Prozentwert (zwischen 0 und 100).

In unserem Beispiel wurden diese Möglichkeiten zur Bildung der Variablen Arbeitlosenanteil, Katholikenanteil und Protestantenanteil benutzt.

Der Protestantenanteil in Prozent ergibt sich durch Zusammenfassung der Kategorien 1 = „evang. Kirche“ und 2 = „evang. Freikirche“ zu einem Bereich, dessen Anteil in Prozent angegeben wird (⇨ Abb. 7.18).

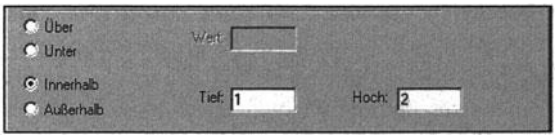


Abb. 7.18. Mittlerer Teil der Dialogbox „Daten aggregieren: Aggregierungsfunktion“. (Ausgewählt sind Prozentwerte innerhalb des Bereichs zwischen den Werten 1 und 2)

Im Feld „Variablen aggregieren:“ wird dies durch PIN(konf 1 2) gekennzeichnet. Dies bedeutet, dass Prozente innerhalb eines Bereichs auf der Variablen KONF mit den Grenzen 1 und 2 gebildet wurden.

Zur besseren Unterscheidung wurden anschließend für die Variablen Protestantenanteil und Katholikenanteil die neuen Namen PROTANT und KATHANT vergeben sowie ein ausführlicheres Variablen-Label. Um einen neuen Namen und/oder ein Variablenlabel für eine Aggregierungsvariable festzulegen, gehen Sie wie folgt vor:

- ▷ Klicken Sie auf die Schaltfläche „Name & Label...“. Die Dialogbox „Daten aggregieren: Variablenname und –label“ öffnet sich.
- ▷ Tragen Sie den gewünschten Namen in das Eingabefeld „Name:“ ein.
- ▷ Geben Sie das Variablenlabel in das Eingabefeld „Label:“ ein.
- ▷ Bestätigen Sie mit „Weiter“.

Aggregierungsfunktionen. Die Bezeichnungen der Aggregierungsfunktionen (⇨ Dialogbox „Daten aggregieren: Aggregierungsfunktion“ Abb. 7.17) sind weitgehend selbsterklärend. Zu beachten ist jedoch: Aggregiert wird unter Ausschluss der „fehlenden Werte“. Liegt eine gewichtete Datei vor, so werden die gewichteten Daten aggregiert. Die im oberen Teil der Box angezeigten Funktionen für metrische Daten (in Klammern ihre Kurzbezeichnung bei der Anzeige) sind: *Mittelwert* (MEAN), *Standardabweichung* (SD), *Minimalwert* (MIN) und *Maximalwert* (MAX) sowie *Summe* der Werte (SUM), jeweils bezogen auf die gültigen Werte der Breakgruppen. Außerdem kann der erste (FIRST) und der letzte (LAST) gültige Wert eine Variablen für die Breakgruppe angezeigt werden. Es handelt sich jeweils um die ersten und letzten Werte in der Reihenfolge der Matrix. Dafür wird sich selten eine sinnvolle Verwendung finden. Wichtig ist dagegen die Gruppe *„Anzahl Fälle“*. Man kann sich die gültigen Fälle der Breakgruppe gewichtet (N) oder ungewichtet (NU) ausgeben lassen. Auch die Zahl der fehlenden Werte pro

Breakgruppe kann ermittelt werden. Die Option „Gewichtet fehlend" aggregiert die Anzahl der fehlenden Werte in den Breakgruppen der gewichteten Datei (NMISS), „Ungewichte fehlend" dagegen ermittelt die Zahl der fehlenden Werte ohne Berücksichtigung der Gewichtung (NUMISS).

Im unteren Teil der Box finden sich für metrische und qualitative Daten geeignete Funktionen. Sie ist wiederum geteilt in die Bereiche „Prozentwerte" und „Brüche", (besser wäre die Bezeichnung Prozentanteile und Anteile). Beide verfügen über analoge Optionen: *Über, Unter, Innerhalb* und *Außerhalb.* Im mittleren Teil ergeben diese Funktionen Prozentsätze: *Prozentwert über* (PGT), *Prozentwert unter* (PLT) geben jeweils den Prozentanteil der gültigen Werte an allen gültigen Fällen der Breakgruppe an, die oberhalb oder unterhalb eines nutzerdefinierten Wertes liegen (der nutzerdefinierte Wert gehört nicht zur Aggregationsgruppe). *Prozentwert innerhalb* (PIN) und *Prozentwert außerhalb* (POUT) geben jeweils den Prozentanteil einer durch den Nutzer definierten niedrigster bzw. höchsten Werte eingeschlossenen bzw. ausgeschlossenen Gruppe an. Die nutzerdefinierten Grenzwerte gehören zur eingeschlossenen Gruppe, nicht zur ausgeschlossenen.

Die Optionen in der Gruppe „Brüche" führen zu analogen Ergebnissen. Anstelle von Prozentanteilen treten lediglich Anteilszahlen, die auf 1 statt auf 100 summieren (ein Prozentanteil von 50 entspricht also einem Anteil von 0,500 etc.). In der Syntax erscheinen sie mit der Abkürzung FGT *(Brüche/Anteil oberhalb)* und FLT *(Brüche/Anteil unterhalb)* bzw. FIN *(Brüche/Anteil innerhalb*) und FOUT (Brüche/*Anteil außerhalb*).

Eine weitere Möglichkeit zur Bildung einer aggregierten Variablen findet sich in der Dialogbox „Daten Aggregieren". Durch Anklicken des Kontrollkästchen „Anzahl der Fälle in der Break-Gruppe speichern:" erstellt man eine aggregierte Variable mit dem voreingestellten Namen N_Break. Der Name kann beliebig geändert werden. Die aggregierte Variable enthält die gesamte Fallzahl der Breakgruppe, also einschließlich der fehlenden Werte. Liegt eine gewichtete Datei vor, ist die Fallzahl ebenfalls gewichtet.

Name der aggregierten Datei. Sie können entweder die aktuelle Arbeitsdatei ersetzen oder durch Anwahl des entsprechenden Optionsschalters die Daten als neue Datei speichern. Für Letzteres gehen Sie wie folgt vor:

▷ Klicken Sie in der Dialogbox „Daten aggregieren" zunächst auf den Optionsschalter „Neue Datendatei anlegen". Per Voreinstellung wird die aggregierte Datei als neue Datei mit dem Namen „AGGR.SAV" gespeichert. Wollen Sie das ändern, klicken Sie auf die Schaltfläche „Datei...". Es öffnet sich die Dialogbox „Daten aggregieren: Ausgabedatei". Diese sieht wie jede Dialogbox zum Speichern aus.

In ihr können Sie für die neue Ausgabedatei einen neuen Namen und/oder ein neues Verzeichnis angeben. Laufwerk und Verzeichnis wechseln Sie in der bekannten Weise. Den Namen ändern Sie entweder durch Eingabe eines neuen Namens in das Eingabefeld „Dateiname:" oder durch Übertragen eines Namens aus den Auswahlfeld.

8 Häufigkeiten, deskriptive Statistiken und Verhältnis

8.1 Überblick über die Menüs „Deskriptive Statistiken", „Berichte" und „Mehrfachantworten"

Die Kapitel 8 bis 12 stellen Verfahren vor, die alle in den fünf Optionen des Menüs „Deskriptive Statistiken ▷" enthalten sind oder in den beiden in enger Beziehung zu deren Inhalten stehen die Menüs „Berichte" und „Mehrfachantworten". Die genannten Menüs versammeln ein Gemisch von Statistikverfahren, die keinesfalls alle nur der deskriptiven Statistik zuzuzählen sind. Vielfach überschneiden sich die Angebote. Ein kurzer Überblick soll die Orientierung erleichtern. Mit den verschiedenen Optionen können folgende statistische Auswertungen erstellt werden:

- ❒ Einfaches Auflisten von Fällen. Dafür benutzt man „Fälle zusammenfassen" oder „Bericht in Zeilen" bzw. „Bericht in Spalten" im Menü „Berichte".
- ❒ Beschreibung eindimensionaler Verteilungen.
  - ● Eindimensionale Häufigkeitstabellen. Diese erstellt man mit „Häufigkeiten" im Menü „Deskriptive Statistiken". Liegen Mehrfachantworten vor, ist es mit dem Menü „Mehrfachantworten" möglich.
  - ● Univariate statistische Maßzahlen. Für alle Messniveaus erstellt man sie im Programm „Häufigkeiten". Schneller, aber nur für intervallskalierte Daten geeignet, geht es mit „Deskriptive Statistiken". Im Untermenü „Explorative Datenanalyse" werden sie ebenfalls ausgegeben. Eine Besonderheit ist hier, dass auch robuste Lageparameter berechnet werden können. Schließlich liefern beide „Berichte"-Menüs diese Maßzahlen in besonderer Darstellungsweise.
  - ● Grafische Darstellung. Balkendiagramme, Kreisdiagramme und Histogramme kann man mit „Häufigkeiten" abrufen. Letzteres ist auch in „Explorative Datenanalyse" verfügbar, dazu „Stengel-Blatt"(Stem-and-Leaf-) Plots".
- ❒ Beschreibung zwei- und mehrdimensionaler Häufigkeitsverteilungen.
  - ● Zwei- und mehrdimensionale Kreuztabellen. Kreuztabellen gibt das Menü „Kreuztabellen" aus. Sind Mehrfachantworten vorhanden, muss man das Menü „Mehrfachantworten" verwenden. Verwendet man „Break-Variablen", erstellt das Programm OLAP-Würfel im Menü „Berichte" ebenfalls Kreuztabellen einer besonderen Form, allerdings wird die abhängige Variable überwiegend durch univariate Statistiken beschrieben.

- Zusammenhangsmaße. „Kreuztabellen" bietet eine Vielzahl von Zusammenhangsmaßen für jedes Messniveau an.
- Grafische Darstellungen. Boxplots, die im Menü „Explorative Datenanalyse" erstellt werden können, dienen dazu, Gruppen zu vergleichen. „Kreuztabellen" bietet „gruppierte Balkendiagramme" an.

❒ Schließende Statistik für eindimensionale Verteilungen. Der Standardfehler für Mittelwerte, aus dem sich das Konfidenzintervall errechnet, wird in den Menüs „Häufigkeiten", „Deskriptive Statistik" und „Explorative Datenanalyse" angeboten.

❒ Schließende Statistik für Zusammenhänge. „Kreuztabellen" bietet mit dem Chi-Quadrat-Test einen Signifikanztest.

❒ Prüfung der Anwendungsbedingungen für statistische Verfahren. Das Menü „Explorative Datenanalyse" bietet für die Prüfung der Normalverteilungsvoraussetzung zwei Normalverteilungsdiagramme und zwei Normalverteilungstests. Für die Überprüfung der Voraussetzung gleicher Varianzen in den Vergleichsgruppen kann man daraus „Boxplots" sowie den „Streuung gegen Zentralwert-Plot (Streubreite vs. mittleres Niveau)" und den „Levene-Test" verwenden. In „Häufigkeiten" kann man das Histogramm mit einer Normalverteilungskurve überlagern.

8.2 Durchführen einer Häufigkeitsauszählung

Mit der Option „Häufigkeiten..." des Menüs „Deskriptive Statistiken" kann eine eindimensionale Häufigkeitsverteilung mit absoluten Häufigkeiten, Prozentwerten und kumulierten Prozentwerten erstellt werden. Zusätzlich bietet diese Prozedur die ganze Palette statistischer Kennzahlen für eindimensionale Häufigkeitsverteilungen, also Lagemaße, Streuungsmaße, Schiefe- und Steilheitsmaße. Die Option „Deskriptive Statistiken..." bietet einen Teil dieses Angebotes ein zweites Mal, nämlich alle statistischen Maßzahlen, soweit sie für Daten gelten, die mindestens auf Intervallskalenniveau gemessen wurden. „Häufigkeiten..." ermöglicht weiter die grafische Darstellung eindimensionaler Häufigkeitsverteilungen in Form von Balkendiagrammen, Kreisdiagrammen und Histogrammen.

8.2.1 Erstellen einer Häufigkeitstabelle

Beispiel. Wir wollen aus den Daten des ALLBUS90.SAV eine Häufigkeitstabelle über die Einstellung der deutschen Bevölkerung zu einem außerehelichen Seitensprung erstellen. Um eine Häufigkeitstabelle zu erstellen, gehen Sie wie folgt vor:

▷ Wählen Sie die Befehlsfolge „Analysieren", „Deskriptive Statistiken ▷", „Häufigkeiten...". Es öffnet sich die Dialogbox „Häufigkeiten" (⇨ Abb. 2.11).
▷ Wählen Sie aus der Quellvariablenliste die Variable TREUE aus.
▷ Bestätigen Sie mit „OK".

Sie erhalten eine Standardhäufigkeitstabelle für diese Variable (⇨ Tab. 8.1).

In der Überschrift der Tabelle sind Variablennamen und die ersten 40 Zeichen des Variablen-Labels angezeigt.

Die Vorspalte unterscheidet zunächst die gültigen und fehlenden Werte und zeigt in der zweiten Hälfte – je nach Voreinstellung – Werte und/oder Wertelabels. Die eigentlichen Daten stehen im Tabellenkörper. Jede Zeile des Tabellenkörpers enthält jeweils Angaben für die Fälle, die dem entsprechenden Wert der Variablen zuzuordnen sind. In der ersten Zeile sind diejenigen enthalten, die einen Seitensprung für „sehr schlimm" erachten, in der zweiten, diejenigen, die ihn als „ziemlich schlimm" bewerten usw.. In der letzten Zeile ist die Zahl aller Fälle (es sind 301), in der vorletzten die Gesamtzahl der Fälle mit fehlenden Werte angegeben. In unserem Beispiel liegen sehr viele Fälle (148) mit fehlenden Werten vor. Als Zwischensumme der gültigen Werte (Gesamt) finden wir 153 Fälle. Das liegt vor allem daran, dass der Hälfte der Befragten diese Frage gar nicht gestellt wurde.

Tabelle 8.1. Häufigkeitstabelle für die Variable TREUE

TREUE VERHALTENSBEURTEILUNG: SEITENSPRUNG

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	SEHR SCHLIMM	39	13,0	25,5	25,5
	ZIEMLICH SCHLIMM	49	16,3	32,0	57,5
	WENIGER SCHLIMM	40	13,3	26,1	83,7
	GAR NICHT SCHLIMM	25	8,3	16,3	100,0
	Gesamt	153	50,8	100,0	
Fehlend	NICHT ERHOBEN	145	48,2		
	WEISS NICHT	2	,7		
	KEINE ANGABE	1	,3		
	Gesamt	148	49,2		
Gesamt		301	100,0		

Worum es sich im einzelnen handelt, ergibt sich aus den Spaltenüberschriften. Die zweite Spalte enthält die absoluten Häufigkeiten („Häufigkeit") der einzelnen Wertekategorien. So haben 39 Personen „sehr schlimm", 49 „ziemlich schlimm" geantwortet usw.. Da Absolutwerte häufig sehr schwer interpretierbar sind, rechnet man sie in der Regel in Anteilszahlen um. „Häufigkeiten" bietet automatisch Prozentwerte an. Dieses ist zunächst in der dritten Spalte („Prozent") der Fall. Man kann ihr entnehmen, dass die 39 Personen, die einen Seitensprung als „sehr schlimm" bezeichnen, 13 % aller Befragten ausmachen usw.. Bei dieser Prozentuierung sind hier allerdings auch die Fälle, für die kein gültiger Wert vorliegt, mit berücksichtigt. Dies kann für verschiedene Zwecke eine wichtige Information sein. Z.B. kann man daran erkennen, ob eine Frage durch zahlreiche Antwortverweigerungen in ihrer Brauchbarkeit beeinträchtigt ist. In unserem Beispiel sind z.B. nur 1 % Ausfälle durch Antwortverweigerungen „weiß nicht" und „keine Angabe" entstanden, der Löwenanteil dagegen dadurch, dass einem Teil der Befragten die Frage nicht gestellt wurde. Daher liegt wohl keine Beeinträchtigung vor.

Für die eigentliche Analyse sind aber nur die gültigen Werte von Interesse. Die Einbeziehung der ungültigen Werte würde zu einem völlig verzerrten Bild führen. In der vierten Spalte sind daher die Prozentwerte auf der Basis der gültigen Fälle

errechnet („Gültige Prozente"). Danach finden 25,5 % der Befragen einen Seitensprung „sehr schlimm", 32 % „ziemlich schlimm" usw..

Schließlich enthält die letzte Spalte die kumulierten Prozentwerte für die gültigen Fälle. Die Prozentwerte werden, vom ersten angeführten Variablenwert ausgehend, schrittweise aufaddiert. So kommt der zweite kumulierte Prozentwert 57,5 durch Addition von 25,5 und 32,0 der Kategorien „sehr schlimm" und „ziemlich schlimm" zustande. Er besagt also, dass 57 % der Befragten einen Seitensprung zumindest für „ziemlich schlimm" erachten. Solche kumulierten relativen Häufigkeiten können für viele Analysezwecke sinnvoll sein. Sie sind allerdings erst brauchbar, wenn zumindest Daten des Ordinalskalenniveaus vorliegen. Will man kumulierte Prozentwerte benutzen, muss man außerdem klären, von welcher Seite der Werteskala her aufaddiert werden soll. SPSS geht bei der Berechnung automatisch vom in der Tabelle zuerst angeführten Wert aus. Per Voreinstellung ist das der kleinste Wert. Man kann aber das Ende, von dem her kumuliert wird, dadurch bestimmen, dass man die Reihenfolge der Ausgabe der Werte mit der Formatierungsoption (⇨ Kap. 8.2.2) entsprechend festlegt.

Unterdrücken des Tabellenoutputs. Die Dialogbox „Häufigkeiten" enthält auch das Kontrollkästchen „Häufigkeitstabellen anzeigen". Per Voreinstellung ist dieses ausgewählt. Schaltet man es aus, so wird der Tabellenoutput unterdrückt. Sinnvollerweise unterdrückt man den Tabellenoutput, wenn man lediglich an einer Grafik bzw. an statistischen Maßzahlen interessiert ist.

8.2.2 Festlegen des Ausgabeformats von Tabellen

Um das Format der Ausgabe zu verändern, gehen Sie wie folgt vor:

▷ Wählen Sie in der Dialogbox „Häufigkeiten" die Schaltfläche „Format...". Es öffnet sich die Dialogbox „Häufigkeiten: Format" (⇨ Abb. 8.1).

Diese enthält zwei Gruppen für die Auswahl von Optionen. In der Gruppe *„Sortieren nach"* kann die Reihenfolge der Ausgabe der Variablenwerte beeinflusst werden:

- ❒ *Aufsteigende Werte.* Ordnet die Kategorien in aufsteigender Reihenfolge. Das ist die Voreinstellung.
- ❒ *Absteigende Werte.* Ordnet die Kategorien in fallender Reihenfolge.
- ❒ *Aufsteigende Häufigkeiten.* Hier werden die Kategorien nach der Zahl der in ihnen enthaltenen Fälle geordnet, und zwar ausgehend von der Kategorie mit den wenigsten Fällen. (Die Missing-Werte werden dabei nicht berücksichtigt.)
- ❒ *Absteigende Häufigkeiten.* Ordnet umgekehrt die Kategorie mit den meisten Fällen an die erste Stelle.

Die Anordnung wirkt sich auf die Berechnung der kumulierten Prozentwerte aus. Will man diese vom niedrigsten Wert ausgehend berechnen, behält man die Standardeinstellung bei. Sollen sie vom höchsten Wert ausgehend berechnet werden, muss „Absteigende Werte" gewählt werden. Die beiden anderen Einstellungen machen die kumulierten Prozentwerte dagegen praktisch unbrauchbar, weil sie in der Regel die sinnvolle Ordnung zerstören.

Abb. 8.1. Dialogbox „Häufigkeiten: Format"

Mit einem weiteren Kontrollkästchen kann man die Ausgabe von großen Tabellen unterdrücken.

- *Keine Tabelle mit mehr als ... Kategorien.* Zeigt Tabellen mit mehr als der eingegebenen Zahl von Kategorien nicht an. Voreingestellt ist 10. Man kann diesen Wert aber durch eine ganze Zahl größer 1 überschreiben. Das ist sinnvoll, wenn mehrere Variablen gleichzeitig ausgezählt werden und bei den Variablen mit vielen Werten nur die Maßzahlen oder die Grafik interessiert.
- *Mehrere Variablen.* Diese Gruppe enthält Optionen, die nur die Ausgabe von Statistiken betreffen. Die Häufigkeitstabellen werden immer für jede Variable einzeln ausgegeben. Werden dagegen Statistiken für mehrere Variablen angefordert, sind zwei Alternativen möglich:
  - *Variablen vergleichen.* Die Statistiken für alle Variablen werden in einer einzigen Tabelle ausgegeben.
  - *Ausgabe nach Variablen ordnen.* Die Statistiken für jede Variable werden in einer eigenen Tabelle ausgegeben.

8.2.3 Grafische Darstellung von Häufigkeitsverteilungen

Im Rahmen von „Häufigkeiten" bietet SPSS drei Arten von Grafiken zur Visualisierung von Häufigkeitsverteilungen an.

- *Balkendiagramme.* Bei einem Balkendiagramm wird die absolute oder relative Häufigkeit jeder Variablenkategorie durch die Höhe eines isoliert stehenden Balkens dargestellt. Diese Form der Darstellung ist geeignet für jede Art von Daten , insbesondere aber Kategorialdaten.
- *Kreisdiagramme.* In einem Kreisdiagramm wird die absolute oder relative Häufigkeit jeder Variablenkategorie durch die Größe eines Kreissegments dargestellt. Geeignet für jede Art von Daten mit nicht zu großer Zahl der Ausprägungen.
- *Histogramme.* Sie stellen Daten in Form von direkt aneinander anschließenden Flächen dar. Sinnvoll ist die Darstellung von Verteilungen durch ein Histogramm bei Vorliegen kontinuierlicher oder quasi-kontinuierlicher Daten. Es ist mindestens Ordinalskalenniveau, besser Intervallskalenniveau erforderlich. Im Gegensatz zum Balkendiagramm müssen die Kategorien eine sinnvolle Ordnung bilden. Anders als beim Balkendiagramm werden auch Klassen, in denen keine Fälle vorhanden sind, angezeigt. Die Option „Histogramme" ist gedacht für die automatische Generierung eines Histogramms aus differenziert erhobe-

nen Daten. Es werden automatisch per Voreinstellung gleich breite Klassen gebildet. Als Richtwert für die Zahl der Klassen dient 21, aber insgesamt wird, ausgehend von der Gesamtskalenbreite, die sich aus höchstem und niedrigstem Wert ergibt, eine Unterteilung mit glatten Klassenbreiten vorgenommen. Daher kann auch die Verwendung bei schon vorher klassifizierten Daten zu einer unkorrekten Darstellung führen. (Sie müssen gegebenenfalls die Klassengrenzen und -breiten im Grafikeditor, Befehlsfolge „Diagramme", „Achse", „Intervall", „Anpassen", „Definieren" und Eingabe der richtigen Werte ändern. Zur Darstellung von Verteilungen mit ungleicher Klassenbreite sind die Grafikmöglichkeiten von SPSS generell ungeeignet. Hier müssen Sie gegebenenfalls andere Programme heranziehen.) Zusätzlich steht in einem Kontrollkästchen die Möglichkeit zur Verfügung, das Histogramm durch eine *Normalverteilung* zu überlagern, die anzeigt, wie eine Normalverteilung bei Daten gleichen Mittelwerts und gleicher Streuung aussehen würde. Dies kann als Hilfsmittel für die Beurteilung der Verteilung dienen, insbesondere auch zur Überprüfung der Normalverteilungsvoraussetzung, die für viele Signifikanztests im Rahmen der multivariaten Analyse gilt.

Um Häufigkeitsverteilungen als Balkendiagramm, Kreisdiagramm oder Histogramm darzustellen, wird in der Dialogbox „Häufigkeiten" (⇨ Abb. 2.11) auf die Schaltfläche „Diagramme..." geklickt. Es öffnet sich die Dialogbox „Häufigkeiten: Diagramme" (⇨ Abb. 8.2). In der Dialogbox wird der Diagrammtyp durch Anklicken des entsprechenden Optionsschalters gewählt. Für ein Balken- oder Kreisdiagramm bestimmt man weiter durch Anklicken des entsprechenden Optionsschalters, ob die Höhe der Balken bzw. die Größe des Kreissegments in absoluten oder prozentualen Häufigkeiten dargestellt werden soll. Klickt man auf das Kontrollkästchen „Mit Normalverteilungskurve", wird ein Histogramm mit einer Normalverteilungskurve überlagert. Falls man nur an den Diagrammen interessiert ist, kann man die Ausgabe von Tabellenoutput durch Anklicken des Kontrollkästchens „Häufigkeitstabelle anzeigen" in der Dialogbox „Häufigkeiten" unterdrücken.

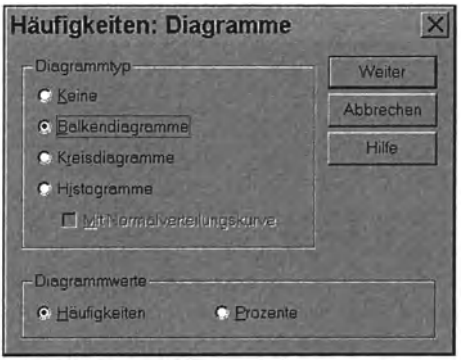


Abb. 8.2. Dialogbox „Häufigkeiten: Diagramme"

Wurde eine Grafik erstellt, erscheint diese im Ausgabefenster. Durch Doppelklicken auf die Grafik kann man den Diagramm-Editor öffnen. Dort kann sie mit verschiedenen Gestaltungsmöglichkeiten überarbeitet werden. Insbesondere ist diese Möglichkeit für Histogramme zu empfehlen, wenn auf den höchsten und niedrigsten angezeigten Wert und Klassenbreite (Intervall) Einfluss genommen werden muss. (Mit den Befehlen der Befehlssyntax lässt sich die Intervallbreite nicht steuern.) Die drei Grafiktypen können auch im Menü „Grafiken" erstellt werden (⇨ Kap. 26).

8.3 Statistische Maßzahlen

8.3.1 Definition und Aussagekraft

Überblick. Mit Hilfe statistischer Maßzahlen kann man wesentliche Eigenschaften eindimensionaler Verteilungen noch knapper erfassen. Dazu stehen in SPSS die vier gebräuchlichen Typen von Maßzahlen zur Verfügung (⇨ Abb. 8.4).

- *Lagemaße.* Sie geben auf unterschiedliche Weise in etwa die Mitte der Verteilung wieder.
- *Streuungsmaße.* Sie geben an, wie weit die einzelnen Werte um die Mitte der Verteilung herum streuen.
- *Verteilungsmaße* (Schiefe- und Steilheitsmaße). Schiefemaße geben Hinweise darauf, ob eine Verteilung symmetrisch ist oder nach der einen oder anderen Seite schief, Steilheitsmaße dagegen, ob eine Verteilung im Vergleich zu einer Normalverteilung von Daten gleichen Mittelwerts und gleicher Streuung im Bereich des Mittelwertes eher enger oder weiter streut.
- *Perzentilwerte.* Sie geben den Wert einer Verteilung an, unterhalb dessen ein festgelegter Prozentsatz der Fälle mit einem geringeren Wert liegt. Es sind ebenfalls Lagemaße, die aber nur in einem Spezialfall (dem Medianwert) die Mitte einer Verteilung kennzeichnen. Die Distanz zwischen zwei Perzentilen kann als Streuungsmaß Anwendung finden. Gebräuchlich ist die Distanz zwischen dem 25. Perzentil (unteres Quartil) und dem 75. (oberes Quartil), der Quartilsabstand oder dessen Hälfte, der Mittlere Quartilsabstand.

Abhängigkeit der Statistiken vom Messniveau. Welche statistische Maßzahl im konkreten Fall geeignet ist, hängt nicht nur vom Zweck, sondern auch vom Messniveau der Daten ab. Die in den vier Optionsgruppen angebotenen Ma0zahlen unterscheiden sich z.T. hinsichtlich des vorausgesetzten Messniveaus. Deshalb soll darauf etwas näher eingegangen werden.

Daraus, dass wir Messwerten bestimmte Zahlen zugeordnet haben, ist nicht zu schließen, dass diese etwa wie reelle Zahlen behandelt werden können. Vielmehr muss dem empirischen Relativ ein äquivalentes numerisches Relativ zugeordnet werden. Das heißt, man darf Zahlen nur Eigenschaften unterstellen, die sie auch tatsächlich abbilden, und es dürfen nur Rechenoperationen durchgeführt werden, die lediglich auf den abgebildeten Eigenschaften beruhen. Statistische Maßzahlen dürfen deshalb ebenfalls jeweils nur mathematische Operationen verwenden, die dem Messniveau der Daten angemessen sind. So sind bei rationalskalierten Daten

alle geläufigen Rechenoperationen erlaubt. Dagegen dürfen etwa bei intervallskalierten Daten keine Quotienten aus den Messwerten gebildet werden.

Tabelle 8.2 führt die vier in der Statistik bedeutsamen Messniveaus und die dazu gehörigen Unterscheidungskriterien an. Diese vier Kriterien bauen hierarchisch aufeinander auf, so dass ein höheres immer die Existenz des niedrigeren Kriteriums voraussetzt. Es liegt eine Hierarchie von Messniveaus von niedrigerem zu höherem vor. Wir unterscheiden Nominal-, Ordinal-, Intervall- und Verhältnis- (oder Rational-)skalenniveau. Für viele Zwecke reicht eine Unterscheidung in qualitative bzw. kategoriale Daten (nominal- und ordinalskalierte) und metrische (intervall- oder rationalskalierte).

Tabelle 8.2. Überblick über Messniveaus und ihre Bedeutung für die Statistik

Messniveau	Mögliche empirische Aussagen	Beispiele
Nominal	1. Gleichheit und Ungleichheit	Automarken, Geschlecht, Schulform, Fächer
Ordinal	1. Gleichheit und Ungleichheit 2. Ordnung	Schulnoten, Hackordnung, Soziale Schichtung
Intervall	1. Gleichheit und Ungleichheit 2. Ordnung 3. Gleichheit von Differenzen	Celsiustemperaturskala, Intelligenzpunktwerte, Leistungspunktwerte
Verhältnis	1. Gleichheit und Ungleichheit 2. Ordnung 3. Gleichheit von Differenzen 4. Gleichheit von Quotienten	Gewicht, Körpergröße, Alter, Zahl der Kinder pro Familie, Reaktionszeit

Quelle: in Anlehnung an Wolf, W. (1974), S. 58.

Aus Tabelle 8.3 kann man ablesen, welche Verfahren aus jeder der drei oben genannten Gruppen von Maßzahlen je nach Messniveau prinzipiell in Frage kommen. Dabei ist das Messniveau vom Nominal- bis zum Verhältnisskalenniveau als hierarchische Ordnung von niedrigerem zu höherem zu verstehen. Auf Daten des höheren Niveaus sind prinzipiell auch alle Verfahren anwendbar, die für niedrigere Messniveaus geeignet sind. Diese auch bei höherem Messniveau zu verwenden, ist oft sinnvoll, weil sich die Art der Information der verschiedenen statistischen Maßzahlen auch in derselben Gruppe etwas unterscheidet. Insbesondere ist es immer angebracht, die Häufigkeitsverteilung mit zu betrachten. Andererseits wird ein Teil der vorhandenen Information verschenkt, wenn man bei höherem Messniveau nicht die dafür angepassten Verfahren verwendet, so dass man normalerweise die Verfahren für das erreichte höhere Messniveau auch nutzen sollte.

Neben dem Messniveau der Daten, sind für die Auswahl der geeigneten Statistiken zwei weitere Kriterien wichtig:

- Der Anspruch an die *Robustheit* der Messung. So geht in die Berechnung des arithmetischen Mittels jeder einzelne Wert ein. Es kann daher durch Extremwerte verzerrt werden. Dagegen ergibt sich der Medianwert aus dem Wert eines

einzigen Falles. Er ist sehr robust. (Im Untermenü „Explorative Datenanalyse..." werden andere robuste Lageparameter angeboten, die eine größere Zahl von Werten einbeziehen, aber dennoch Extremwerte nicht oder mit geringem Gewicht beachten.)

- Die *Eigenschaften der Parameter.* So fallen arithmetisches Mittel, Modalwert und Medianwert bei symmetrischen Verteilungen zusammen, unterscheiden sich aber bei schiefen Verteilungen charakteristisch.

Tabelle 8.3. Sinnvolle Parameter in Abhängigkeit zum Messniveau

	sinnvolle Parameter	
Messniveau	**Lageparameter**	**Streuungsparameter**
Nominal	Modalwert	Häufigkeitsverteilung
Ordinal	Median (Perzentile)	Quartilsabstand
Intervall	arithmetisches Mittel	Varianz Standardabweichung Spannweite
Verhältnis	geometrisches Mittel	Variationskoeffizient

Lagemaße (zentrale Tendenz).

Modalwert (Modus). Der am häufigsten auftretende Wert. Bei klassifizierten Daten ist der Modalwert die Klassenmitte der Klasse mit den meisten Fällen. (Achtung! Bei ungleicher Klassenbreite oder bei vielen wenig besetzten Klassen nicht aussagefähig.)

Median. Ordnet man die Fälle nach ihrem Wert, so ist es der Wert, unter und über dem jeweils die Hälfte der Fälle liegt. Bei nicht klassifizierten Daten ist es bei einer ungeraden Zahl von Fällen der Wert des mittleren Falles. Bei einer geraden Zahl von Fällen gibt es keinen mittleren Fall, sondern zwei. Es wird das arithmetische Mittel der Werte dieser beiden Fälle verwendet. Sind die Daten klassifiziert, fällt der mittlere Fall in eine Klasse mit bestimmter Klassenbreitee. Es wird daher unterstellt, dass alle Fälle, die in dieser Einfallsklasse liegen, ein gleich großes Stück dieser Spannweite abdecken. Daraus wird der Wert innerhalb der Klasse ermittelt, an dem genau der mittlere Fall liegen würde.

Arithmetisches Mittel (Mittelwert). Ist die Summe der Werte aller Fälle, dividiert durch die Zahl der Fälle. Bei klassifizierten Daten wird jeweils der Klassenmittelwert als Wert verwendet.

$$\bar{x} = \frac{\sum x}{n} \tag{8.1}$$

Summe. Ebenfalls angeboten wird die Gesamtsumme der Werte. Hierbei handelt es nicht um ein Lagemaß. Jedoch kann die Summe eine interessante Information ergeben. (*Beispiel:* Die Gesamtsumme der Schulden aller von einer Schuldenbera-

tungsstelle regulierten Fälle.) Außerdem wird die Summe für viele andere Berechnungen als Zwischengröße benötigt (z.B. arithmetisches Mittel).

Streuungsmaße (Dispersionsparameter). Verteilungen mit dem gleichen Messwert für die zentrale Tendenz können sich in anderer Hinsicht unterscheiden. So kann sich in einer Untersuchung von 2.000 Personen in einem Extremfall ein Durchschnittseinkommen von 2.500 DM ergeben, wenn alle Personen 2.500 DM verdienen, in dem entgegengesetzten Falle, wenn 1.000 Personen je 0 DM und die anderen 1.000 je 5.000 DM verdienen. Dispersionsparameter geben an, wie stark die Einzelwerte mit dem Mittelwert übereinstimmen oder von ihm abweichen.

Spannweite. Dieser Wert wird einfach aus der Differenz zwischen höchstem und niedrigstem Wert ermittelt. Sie ist ein sehr simples Streuungsmaß. Es ist extrem sensitiv für Extremwerte und daher häufig unbrauchbar. Die gebräuchlichsten Streuungsmaße sind Varianz und Standardabweichung.

Varianz. Ist die Summe der quadrierten Abweichungen der Einzelwerte vom arithmetischen Mittel, geteilt durch die Zahl der Werte. In SPSS wird die Varianz als Stichprobenvarianz (= Schätzwert für die Varianz der Grundgesamtheit) berechnet. Daher wird durch $n-1$ dividiert.

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} \tag{8.2}$$

Die Varianz ist 0, wenn alle Werte mit dem Mittelwert identisch sind und wird um so größer, je größer die Streuung ist. Die Varianz wird häufig als Zwischenergebnis für weitere Berechnungen benutzt.

Standardabweichung. Ist die Quadratwurzel aus der Varianz. Die Standardabweichung s ist leichter zu interpretieren als die Varianz, weil sie dieselben Maßeinheiten wie die Originaldaten verwendet. Auch sie wird 0 bei völliger Übereinstimmung aller Daten mit dem arithmetischen Mittel und wird umso größer, je größer die Streuung.

Standardfehler für das arithmetische Mittel (Mittelwert Standardfehler). Die Auswahlbox für die Streuungsparameter bietet diesen Parameter an, der eigentlich eher dem Bereich der schließenden Statistik zuzurechnen ist. Er dient bei Stichprobendaten zur Bestimmung des Konfidenzintervalls (Fehlerspielraums, Standardirrtums, Mutungsbereichs), in dem das „wahre" arithmetische Mittel mit einer festgelegten Wahrscheinlichkeit liegt. Üblicherweise benutzt man ein Sicherheitsniveau von 95 oder 99 %. Dann muss der Standardfehler zur Bestimmung des Konfidenzintervalls mit 1,96 bzw. 2,58 multipliziert werden (⇨ Kap. 8.4).

Formmaße. Die Auswahlgruppe *„Verteilung"* bietet zwei Maßzahlen zur Form der Verteilung an. Lage- und Streuungsmaße kennzeichnen Verteilungen gut, wenn sie symmetrisch um einen Mittelpunkt herum aufgebaut sind. Noch besser ist es, wenn zudem auch der Gipfel der Verteilung in der Mitte liegt und die Verteilung eingipflig ist. Ideal ist es, wenn die Werte normalverteilt sind.

Bei der Beurteilung der Form einer Verteilung gehen die von SPSS angebotenen Maße von einem Vergleich mit einer Normalverteilung mit demselben arithmetischen Mittel und derselben Streuung aus. Für die Normalverteilung gelten einige

charakteristische Merkmale. Die Normalverteilung ist glockenförmig und symmetrisch. Der Abstand zwischen dem arithmetischen Mittel und dem zu einem Wendepunkt gehörenden x-Wert beträgt genau eine Standardabweichung. In den Bereich von ± einer Standardabweichung um das arithmetische Mittel fallen immer ca. 68 % der Fälle der Verteilung. Auch für jeden anderen Bereich der Verteilung ist der Anteil der Fälle bekannt.

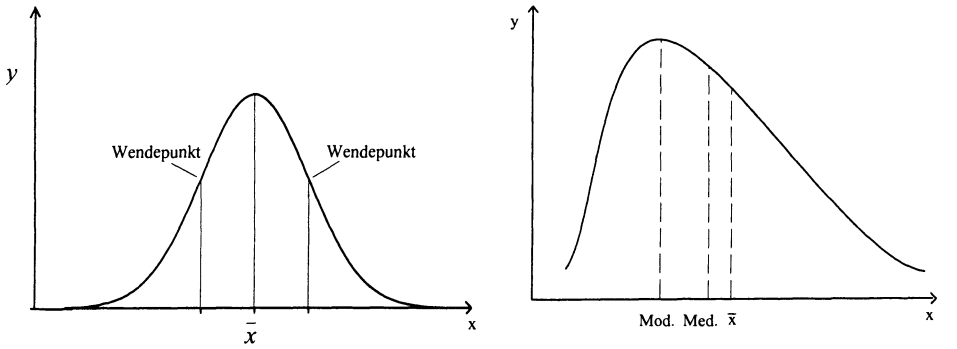


Abb. 8.3. Normalverteilung, rechtsschiefe (linkssteile) Verteilung

Eingipflige Verteilungen müssen aber nicht symmetrisch aufgebaut sein. Der Gipfel kann mehr zu dem einen oder anderen Ende der Verteilung verschoben sein. Dann handelt es sich um eine schiefe Verteilung. Ist der Gipfel mehr zu den niederen Werten hin verschoben, liegt er also links vom Mittelwert, müssen rechts vom Mittelwert die meisten extremen Werte liegen. Eine solche Verteilung nennt man linksgipflig (linkssteil) oder rechts (positiv) schief. Kommen dagegen höhere Werte häufiger vor, liegt der Gipfel also rechts vom arithmetischen Mittel, während die meisten extremen Werte links davon liegen, heißt die Verteilung rechtsgipflig (rechtssteil) oder links (negativ) schief.

Die Schiefe einer Verteilung kann man bereits aus dem Vergleich der drei Lagemaße arithmetisches Mittel, Medianwert und Modalwert erkennen. Es gilt: Im Falle einer symmetrischen Verteilung fallen die drei Werte zusammen. Bei einer linksgipfligen bzw. rechtsschiefen Verteilung ist: Modalwert < Median < arithmetisches Mittel. Bei einer rechtsgipfligen bzw. linksschiefen gilt umgekehrt: Modus > Median > arithmetisches Mittel.

Schiefe. SPSS verwendet als Schiefemaß das sogenannte dritte Moment. Es ist definiert als:

$$\text{Schiefe} = \frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3}{n} \tag{8.3}$$

Es nimmt den Wert 0 an, wenn die Verteilung total symmetrisch ist. Je unsymmetrischer die Verteilung, desto größer der Wert. Der Wert wird positiv bei linksgipfligen Verteilungen und negativ bei rechtsgipfligen.

Kurtosis (Steilheit, Wölbung, Exzess). Es ist ein Maß dafür, ob die Verteilungskurve im Vergleich zu einer Normalverteilung bei gleichem Mittelwert und gleicher Streuung spitzer oder flacher verläuft. Bei spitzem Verlauf drängen sich die Fälle im Zentrum der Verteilung stärker um den Mittelwert als bei einer Normalverteilung, während dann im Randbereich weniger Fälle auftreten. Eine im Vergleich zur Normalverteilung flachere Verteilung hat im Bereich des Mittelwertes weniger Fälle aufzuweisen, fällt dann dafür aber zunächst nur langsam ab und enthält dort mehr Fälle. Erst ganz am Rand fällt sie schneller ab.

SPSS benutzt das vierte Moment als Steilheitsmaß. Die Definitionsgleichung lautet:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4}{n} - 3 \qquad (8.4)$$

Nimmt Kurtosis einen Wert von 0 an, entspricht die Form genau einer Normalverteilung. Ein positiver Wert zeigt eine spitzere Form an, ein negativer eine flachere.

Zu beiden Formmaßen wird auch der zugehörige Standardfehler berechnet. Er kann auf dieselbe Weise, wie beim Standardfehler für Mittelwerte beschrieben, zur Berechnung eines Konfidenzintervalls benutzt werden.

Perzentilwerte. Die Auswahlgruppe „Perzentilwerte" (⇨ Abb. 8.4) ermöglicht es, auf verschiedene Weise Perzentile zu berechnen. Ein Perzentilwert P einer Verteilung ist der Wert auf der Messskala, unter dem P % und über dem (100-P) % der Messwerte liegen, z.B. liegen unterhalb des 10. Perzentilwert 10 %, darüber 90 % der Werte.

- ❒ Durch Anklicken des Auswahlkästchens „Perzentile", Eingabe eines Wertes in das Eingabefeld und Anklicken der Schaltfläche „Hinzufügen" kann man beliebige Perzentile anfordern. Dieses kann man mehrfach wiederholen. Die Liste der eingegebenen Werte wird im entsprechenden Feld angezeigt.
- ❒ Das Auswahlkästchen „Trennen ... gleiche Gruppen" vereinfacht die Auswahl mehrerer gleich großer Perzentilgruppen. Wählt man es an und gibt den Wert 10 ein, so wird das 10., 20., 30. bis 90. Perzentil gebildet. Es handelt sich um 10 gleiche Gruppen, denn die ersten 10 % der Fälle haben einen Wert von unter dem angegebenen Perzentilwert bis zu ihm hin, die zweiten 10 % liegen zwischen diesem Wert und dem des 20. Perzentils usw.. Letztlich haben die Glieder der 10. Gruppe, die letzten 10 %, Werte, die größer sind als der des 90. Perzentils. Dieser Perzentilwert wird nicht angegeben, da er automatisch der größte auftretende Wert sein muss. Gibt man als Wert 5 ein, werden das 20., 40. usw. Perzentil ermittelt.
- ❒ Das Auswahlkästchen „Quartile" wählt vereinfacht die gebräuchlichsten Perzentile aus, das 25. (unteres Quartil) das 50. und das 75. (oberes Quartil).

Anmerkung. Auch der Medianwert, der in der Gruppe „Lagemaße" angeboten wird, ist ein besonderer Perzentilwert. Er kann – wie auch die Quartile – in den anderen Auswahlkästchen ebenfalls gewählt werden.

Anwendung auf klassifizierte Daten. Im Fall von gruppierten (klassifizierten) Daten ist für die Berechnung aller Perzentilwerte (d.h. bei *allen* Optionen der Gruppe „Perzentile" und der Option „Median" in der Gruppe „Lagemaße") das Auswahlkästchen *„Werte sind Gruppenmittelpunkte"* einzuschalten, sonst wird lediglich der (nicht aussagefähige) Wert der Einfallsklasse als Perzentilwert angegeben.

Anmerkung. Bei gruppierten (klassifizierten) Daten muss dann allerdings auch wirklich der Klassenmittelwert als Gruppenwert verschlüsselt sein und nicht etwa ein beliebiger anderer Wert. Eine Einkommensklasse von 0 bis 500 DM darf also nicht als Klasse 1, sondern muss als 250 kodiert werden. Das gilt auch für die Berechnung anderer Maßzahlen wie arithmetisches Mittel, Varianz und Standardabweichung. Sollen sie aus klassifizierten Werten berechnet werden, muss der Klassenmittelwert als Wert angegeben sein. Allerdings muss bei diesen Maßzahlen das Kästchen „Werte sind Gruppenmittelpunkte" nicht angekreuzt werden, eine zutreffende Berechnung erfolgt bei entsprechender Vorkehrung ohnehin. Immer aber ist die Berechnung von statistischen Kennzahlen aus nicht klassifizierten Daten genauer. Deshalb sollte man bei Zusammenfassung von Daten zu Klassen immer die Variable mit den unklassifizierten Daten erhalten und sie zur Berechnung der statistischen Kennzahlen benutzen.

8.3.2 Berechnen statistischer Maßzahlen

Es sollen jetzt zur Tabelle 8.1 über die Einstellung zur ehelichen Treue die sinnvollen statistischen Kennzahlen berechnet werden. Die Tabelle soll nicht mehr angezeigt werden. Dazu gehen Sie wie folgt vor:

- ▷ Wählen Sie „Analysieren", „Deskriptive Statistiken ▷", „Häufigkeiten ...".
- ▷ Wählen Sie die Variable TREUE.
- ▷ Schalten Sie „Häufigkeitstabellen anzeigen" aus.
- ▷ Klicken Sie auf die Schaltfläche „Statistik...". Die in Abb. 8.4 angezeigte Dialogbox öffnet sich.

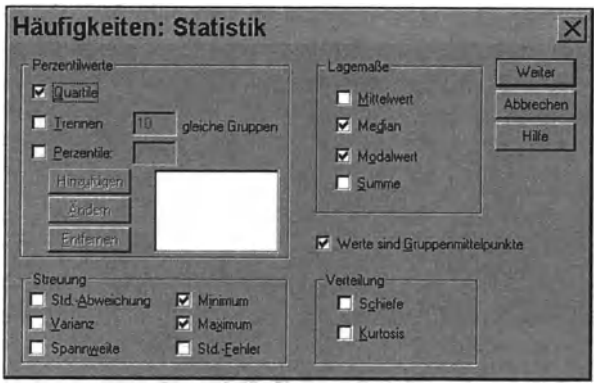


Abb. 8.4. Dialogbox „Häufigkeiten: Statistik"

Wir wählen jetzt die geeigneten statistischen Maßzahlen aus. Die Messung der Einstellung zu ehelicher Treue hat Ordinalskalenniveau. Die Kategorien „sehr schlimm", „ziemlich schlimm" usw. zeigen Unterschiede an und haben eine eindeutige Ordnung. Gleiche Abstände können dagegen kaum unterstellt werden. Man sollte daher nur Maßzahlen auswählen, die höchstens Ordinalskalenniveau verlangen.

- ▷ Wählen Sie: „Quartile", „Median", „Modalwert", „Minimum" und „Maximum".
  Außerdem müssen wir davon ausgehen, dass es sich um gruppierte Daten handelt. Wir haben mit der Einstellung ein kontinuierliches Merkmal. Die Klassen müssen also unmittelbar aneinander anschließen. Deshalb dürfen auch die Werte 1 „sehr schlimm", 2 „ziemlich schlimm" nicht als klar unterschiedene Werte auf der Zahlengerade interpretiert werden, sondern als Repräsentanten von Klassen. Die erste geht von 0,5 bis 1,5, die zweite von 1,5 bis 2,5 usw..
- ▷ Wählen Sie daher das Kontrollkästchen „Werte sind Gruppenmittelpunkte".
- ▷ Bestätigen Sie mit „Weiter" und „OK". Es erscheint die hier in Tabelle 8.4 pivotiert wiedergegebene Ausgabe.

Der niedrigste Wert („Minimum") beträgt 1, der höchste („Maximum") 4. Das ist in diesem Falle wenig informativ. Es sind jeweils die höchste und niedrigste angebotene Kategorie. Der häufigste Wert („Modus") beträgt 2. Das ist uns schon aus der Tabelle 8.1 bekannt. Es ist die Kategorie, in der die meisten gültigen Werte (nämlich 49 Fälle) stehen. Der Medianwert („Median") beträgt nach Tabelle 8.4 2,29. Schon aus der Tabelle 8.1 können wir in der Spalte der kumulierten Prozentwerte gut erkennen, dass der mittlere Fall in der Einfallsklasse 2 „ziemlich schlimm" liegt. Hätten wir nicht angegeben, dass 2 der Gruppenmittelwert einer Klasse ist, wäre als Medianwert einfach die 2 angegeben worden. Denn von allen Werten in dieser Klasse wäre angenommen worden, dass sie denselben Wert 2 hätten. Da wir aber gruppierte Daten haben, wird angenommen, dass sich die 49 Fälle der Klasse 2 gleichmäßig über den Bereich 1,5 bis 2,5 verteilen. Der insgesamt mittlere Fall (der 76,5te von 153 gültigen) wäre der 37,5 von 49 in der Einfallsklasse, liegt also im dritten Viertel dieser Einfallsklasse. Das gibt genau das Ergebnis an. Bei der Berechnung der Quartile wird der Medianwert ein zweites Mal angegeben, zusätzlich die Werte für das untere Quartil („Perzentile 25") 1,43 und das obere Quartil („Percentile 75") 3,21.

Tabelle 8.4. Statistische Maßzahlen zur Kennzeichnung der Verteilung der Einstellung zur ehelichen Treue

Statistiken

TREUE VERHALTENSBEURTEILUNG: SEITENSPRUNG

N						Perzentile		
Gültig	Fehlend	Median	Modus	Minimum	Maximum	25	50	75
153	148	2,29[a]	2	1	4	1,43[b]	2,29	3,21

a. Aus gruppierten Daten berechnet

b. Perzentile werden aus gruppierten Daten berechnet.

In einem zweiten Anwendungsbeispiel sollen geeigneten statistischen Maßzahlen für die Variable EINK (Monatseinkommen) berechnet werden.

- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Deskriptive Statistiken ▷", „Häufigkeiten...".
- ▷ Wählen Sie die Variable EINK.
- ▷ Schalten Sie „Häufigkeitstabelle anzeigen" aus.
- ▷ Klicken Sie auf die Schaltfläche „Statistik...". Die in Abb. 8.4 angezeigte Dialogbox öffnet sich. Da das Einkommen auf Verhältnisskalenniveau gemessen ist (neben Unterschied und Ordnung liegen auch gleiche Abstände und ein absoluter Nullpunkt vor), können wir alle statistischen Maßzahlen benutzen. Klicken Sie (außer in der Gruppe „Perzentilewrte") alle an.
- ▷ Wählen Sie außerdem in der Gruppe „Perzentilwerte" folgende Optionen aus: Markieren Sie das Kontrollkästchen „Quartile". Markieren Sie ebenfalls das Kontrollkästchen „Trennen ... gleiche Gruppen", und ändern Sie im Eingabefeld den Wert in „5".
- ▷ Schalten Sie – falls eingeschaltet – im entsprechenden Kontrolllkästchen die Option „Werte sind Gruppenmittelpunkte" aus.
- ▷ Bestätigen Sie mit „Weiter" und „OK".

Sie erhalten eine umfangreiche Ausgabe (⇨ Tabelle 8.5).

Die wichtigsten Informationen des Outputs sollen kurz besprochen werden. Der Output enthält zunächst die drei Lagemaße, das arithmetisches Mittel („Mittelwert") = 2096,78 DM, den häufigsten Wert („Modus") = 2100 DM und den Zentralwert („Median") 1900 DM. Man erkennt, dass um die 2000 DM (je nach Maßzahl etwas höher oder geringer) in etwa die Mitte der Verteilung liegt. Der häufigste Wert ist nicht besonders aussagekräftig, da wir eine sehr differenziert erhobene Verteilung haben. In einem solchen Falle kann es relativ zufällig sein, welche Kategorie nun gerade am stärksten besetzt ist. Er wird daher auch bei der Interpretation der Schiefe der Verteilung außer acht gelassen.

Die Verteilung ist nicht ganz symmetrisch. Das kann man schon daran erkennen, dass arithmetisches Mittel und Median auseinanderfallen. Das arithmetische Mittel ist größer als der Median. Demnach ist die Verteilung linksgipflig. Dasselbe besagt auch das Schiefemaß („Schiefe"). Es beträgt 1,186. Es ist positiv, zeigt also eine linksgipflige Verteilung an. Das Steilheitsmaß („Kurtosis") beträgt 2,0. Als positiver Wert zeigt es eine Verteilung an, die spitzer ist als eine Normalverteilung. Dies alles können wir auch durch Betrachtung des Histogramms bestätigen.

Darüber hinaus enthält die Ausgabe die Streuungsmaße Varianz und Standardabweichung. Letztere beträgt ±1133,80 DM. Das ist bei einem Mittelwert von 2096 DM eine recht beträchtliche Streuung. Die Spannweite ist ebenfalls ein einfaches Streuungsmaß. Sie beträgt 6871 DM. Aus der Differenz zwischen oberem und unterem Quartil lässt sich ebenfalls ein Streuungsmaß, der Quartilsabstand ermitteln. Er beträgt 2500 – 1300 DM = 1200 DM. Für all diese Maße gilt: Je größer der Wert, desto größer die Streuung. Ein Wert von 0 bedeutet keinerlei Streuung. Am aussagefähigsten sind diese Werte im Vergleich mit anderen Verteilungen.

Es sind weiter die Werte für das 20., 40. usw. Perzentil angezeigt, zusammen damit auch die Quartile und der Median. Der Wert 1200 für das 20. Perzentil bedeutet z.B., dass 20 Prozent der Befragten weniger als 1200 DM verdienen und 80 Prozent 1200 DM und mehr.

Außerdem sind die Standardfehler für das arithmetische Mittel, Schiefe und Kurtosis angegeben. Beispielhaft soll dieser für das arithmetische Mittel interpretiert werden. Die Interpretation setzt voraus, dass die Daten einer Zufallsstichprobe entstammen. Dann kann man das Konfidenzintervall bestimmen. Der Standardfehler beträgt ± 94,81. In diesem Bereich um das arithmetische Mittel der Stichprobe liegt mit 68prozentiger Sicherheit der „wahre Wert". Da man gewöhnlich aber 95prozentige Sicherheit wünscht, muss man den Wert mit 1,96 multiplizieren. 94,81 * 1,96 = 185,833. Mit 95prozentiger Sicherheit liegt daher der „wahre Mittelwert" im Bereich von 2096,783 ± 185,833 DM, d.h. im Bereich: 1910,95 bis 2282,61 DM.

Tabelle 8.5. Statistische Maßzahlen zur Variablen Einkommen

Statistiken

EINK BEFR.: MONATLICHES NETTOEINKOMMEN

N	Gültig	143
	Fehlend	158
Mittelwert		2096,78
Standardfehler des Mittelwertes		94,81
Median		1900,00
Modus		2100
Standardabweichung		1133,80
Varianz		1285506
Schiefe		1,186
Standardfehler der Schiefe		,203
Kurtosis		2,000
Standardfehler der Kurtosis		,403
Spannweite		6871
Minimum		129
Maximum		7000
Summe		299840
Perzentile	20	1200,00
	25	1300,00
	40	1700,00
	50	1900,00
	60	2100,00
	75	2500,00
	80	2820,00

Weitere Möglichkeiten bei Verwenden der Befehlssyntax.

- Bestimmte Datenbereiche können aus der Analyse ausgeschlossen werden (mit dem VARIABLES-Unterkommando).
- Zusätzliche Formatierungsoptionen ermöglichen es, jede Tabelle auf einer neuen Seite beginnen zu lassen, Tabellen mit doppeltem Zeilenabstand zu erstellen und die Tabelle in eine Datei zu schreiben (mit dem FORMAT-Unterkommando).
- Weitere Optionen für die Behandlung gruppierter (klassifizierter) Daten stehen im GROUPED-Unterkommando zur Verfügung. Damit können die Klassengrenzen bei Bedarf genauer definiert werden.
- Das MISSING-Unterkommando ermöglicht es, benutzerdefinierte fehlende Werte in die Berechnung einzuschließen.
- Mit den Unterkommandos BARCHART, HISTOGRAMM und PIECHART kann man Grafiken näher spezifizieren. Da aber auch im Grafikfenster dieselben Bearbeitungsmöglichkeiten bestehen, werden diese empfohlen (⇨ Kap. 26).

8.4 Bestimmen von Konfidenzintervallen

Einführung. Will man in der beschreibenden Statistik eine statistische Maßzahl oder Parameter einer Variablen, z.B. ein Lage-, Streuungs- oder Formmaß für eine Grundgesamtheit bestimmen, so ist das bei einer Vollerhebung ohne weiteres möglich. Dasselbe gilt für deskriptive Maße für Zusammenhänge, also z.B. Zusammenhangsmaße, Regressionskoeffizienten. Stammen statistische Maßzahlen aus Stichproben, können sie von den Maßzahlen der Grundgesamtheit (= Parameter) mehr oder weniger stark abweichen, und dies ist bei der Interpretation von Stichprobenergebnissen mit zu berücksichtigen. Statistische Maßzahlen aus Stichproben dienen deshalb nur als *Schätzwerte* für die Parameter der Grundgesamtheit, für die wahren Werte. Das arithmetische Mittel der Werte in der Stichprobe kann z.B. als Schätzwert für das arithmetische Mittel derselben Variablen in der Grundgesamtheit dienen.

Wenn der Stichprobe eine Zufallsauswahl zugrunde liegt, sind Abweichungen der aus der Stichprobe gewonnenen statistischen Maßzahlen vom Parameter der Grundgesamtheit als Ergebnis des Zufalls zu interpretieren. In diesem Falle können wahrscheinlichkeitstheoretische Überlegungen zum Tragen kommen. Auf deren Basis ist es möglich, einen Bereich abzuschätzen, in dem mit angebbarer Wahrscheinlichkeit der wahre Wert der Grundgesamtheit liegt. Der wahre Wert wird mit einer festlegbaren Wahrscheinlichkeit in einem bestimmten Bereich um den Stichprobenwert liegen. Diesen Bereich nennt man *Konfidenzintervall* (Schätzintervall, Fehlerspielraum, Sicherheitsspielraum, Vertrauensbereich). Zur Ermittlung des Konfidenzintervalls benötigt man die Streuung der (gedanklichen) Verteilung (= Stichprobenverteilung), die durch wiederholte Ziehungen einer großen Anzahl von Stichproben entsteht. Die Standardabweichung dieser Stichprobenverteilung wird auch als *Standardfehler* bezeichnet.

SPSS gibt den Standardfehler und/oder die Ober- und Untergrenze des Konfidenzintervalls (...%-Konfidenzintervall) für arithmetisches Mittel, Schiefe- und Wölbungsmaß (Kurtosis) sowie Regressionskoeffizienten z.T. auf Anforderung, z.T. automatisch in sehr vielen Prozeduren aus. Im Menü Grafiken werden zusätzlich bei den Regelkartendiagrammen auch Konfidenzintervalle für Spannweite und Standardabweichung ausgewiesen. (Eine Besonderheit liegt bei Verwendung des Moduls „Exact Tests" vor. Wendet man dort die Monte-Carlo-Simulation an, werden für die Wahrscheinlichkeiten P eines Stichprobenergebnisses Konfidenzintervalle angegeben (⇨ Kap. 27)).

Konfidenzintervall für das arithmetische Mittel. Der Gedanke, der zur Bestimmung von Konfidenzintervallen führt, soll hier am Beispiel des Konfidenzintervalls für das arithmetische Mittel kurz geschildert werden. Angenommen, man möchte den durchschnittlichen Verdienst von Männern (Variable x) durch eine Stichprobenerhebung in Erfahrung bringen. Unter der Voraussetzung, dass die Erhebungsdaten als eine Zufallsstichprobe aus einer definierten Grundgesamtheit (diese habe den Mittelwert μ und die Standardabweichung σ_x) interpretiert werden können, ist der aus der Stichprobe gewonnene Mittelwert $\bar{x}$ eine Punktschätzung für den unbekannten Mittelwert μ der Grundgesamtheit. Da ein Punktschätzwert wegen der Zufallsauswahl der Stichprobe nur selten dem Parameter entspricht, wird häufig eine Intervallschätzung vorgenommen. Bei einer Intervallschätzung wird ein Bereich berechnet – angegeben durch einen unteren und oberen Grenzwert – in dem man das unbekannte μ mit einer Wahrscheinlichkeit von z.B. 95 % (= 0,95 oder allgemein: $1-\alpha$) erwarten kann. Die Wahrscheinlichkeit α kann als Irrtumswahrscheinlichkeit interpretiert werden: Bei einem z.B. 95 %-Konfidenzintervall besteht eine Wahrscheinlichkeit von 5 %, dass der unbekannte Wert nicht in dem zu berechnenden Konfidenzintervall liegt.

Dabei geht man im *direkten Schluss* zunächst von folgender Grundüberlegung aus: Würden aus einer Grundgesamtheit mit normalverteilten Werten unendlich viele Stichproben gezogen, so würde die Verteilung von $\bar{x}$ dieser Stichproben selbst wieder eine Normalverteilung sein, wobei deren Mittelwert dem wahren Wert μ entspricht und deren Standardabweichung (= Standardfehler) $\sigma_{\bar{x}}$ aus der Standardabweichung der Grundgesamtheit σ_x und dem Stichprobenumfang n ableitbar ist: $\sigma_{\bar{x}} = \sigma_x / \sqrt{n}$. Glücklicherweise führt eine Verletzung der Voraussetzung normalverteilter Werte in der Grundgesamtheit in den meisten Fällen zu keinen großen Problemen. So ist z.B. auch die Stichprobenverteilung von $\bar{x}$ aus einer Grundgesamtheit mit uniform verteilten Werten bei genügend großem Stichprobenumfang nahezu normalverteilt mit einem Mittelwert von μ und einer Standardabweichung von $\sigma_{\bar{x}} = \sigma_x / \sqrt{n}$. Demgemäß kann man z.B. erwarten, dass ein aus einer Zufallsstichprobe gewonnenes $\bar{x}$ mit einer Wahrscheinlichkeit P = $1-\alpha$ = 0,95 (= 95 %) in den zu μ symmetrischen Bereich mit der Untergrenze $\mu - 1{,}96\sigma_{\bar{x}}$ und Obergrenze $\mu + 1{,}96\sigma_{\bar{x}}$ fällt. Der Wert 1,96 entspricht der Standardnormalverteilungsvariable z für eine Wahrscheinlichkeit von $\frac{\alpha}{2} = 0{,}025$. Ganz allgemein lässt sich formulieren:

$$P(\mu - z_{\frac{\alpha}{2}}\sigma_{\bar{x}} \le \bar{x} \le \mu + z_{\frac{\alpha}{2}}\sigma_{\bar{x}}) = P(\mu - z_{\frac{\alpha}{2}}\frac{\sigma_x}{\sqrt{n}} \le \bar{x} \le \mu + z_{\frac{\alpha}{2}}\frac{\sigma_x}{\sqrt{n}}) = 1-\alpha \qquad (8.5)$$

In Abb. 8.5 links ist dieses dargestellt: die Variable $\bar{x}$ fällt mit einer Wahrscheinlichkeit von $1-\alpha$ (schraffierter Bereich) in die Grenzen $\mu \pm z_{\alpha/2}\sigma_{\bar{x}}$. Auch für andere symmetrisch um μ liegende Intervalle lassen sich Wahrscheinlichkeiten bestimmen. So liegen z.B. im Bereich ± 2,57 Standardabweichungen um das arithmetische Mittel 99 % der Stichprobenmittelwerte. Soweit der *direkte* Schluss.

Wird im *Umkehrschluss* ein solcher Bereich zur Bestimmung eines Schätzintervalles für μ benutzt, so spricht man von einem Konfidenzintervall.
Im Umkehrschluss kann man bei Kenntnis des arithmetischen Mittels $\bar{x}$ nur *einer* Stichprobe sagen, dass ein gesuchtes z.B. 95 %- Konfidenzintervall für das unbekannte arithmetische Mittel μ in den Grenzen $\bar{x} - 1{,}96\sigma_{\bar{x}}$ bzw. $\bar{x} + 1{,}96\sigma_{\bar{x}}$ um das Mittel $\bar{x}$ der Stichprobe liegt. Weil gemäß direkten Schlusses $\bar{x}$ mit einer Wahrscheinlichkeit von P = 0,95 im Intervall $\mu \pm 1{,}96 \cdot \sigma_{\bar{x}}$ liegt, muss umgekehrt in dem Konfidenzintervall $\bar{x} \pm 1{,}96 \cdot \sigma_{\bar{x}}$ das unbekannte μ mit einer Wahrscheinlichkeit von P liegen. In allgemeiner Formulierung gilt:

$$P(\bar{x} - z_{\frac{\alpha}{2}}\sigma_{\bar{x}} \le \mu \le \bar{x} + z_{\frac{\alpha}{2}}\sigma_{\bar{x}}) = P(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma_x}{\sqrt{n}} \le \mu \le \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma_x}{\sqrt{n}}) = 1-\alpha \qquad (8.6)$$

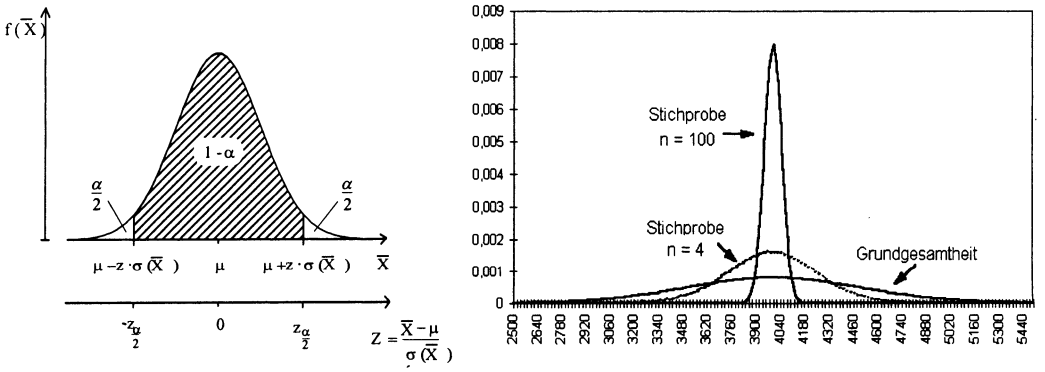


Abb. 8.5. links: Stichprobenverteilung von $\bar{x}$; rechts: Streuung der Stichprobenmittelwerte $\bar{x}$ bei den Stichprobengrößen n = 4 und n = 100

In der Regel wird eine 95prozentige oder 99prozentige Sicherheit angestrebt und entsprechend ein Konfidenzintervall von ± 1,96 Standardabweichungen bzw. ± 2,57 Standardabweichungen um den gefundenen Wert gelegt.

In der Realität kennen wir i.d.R. aber nicht die Streuung σ_x der Grundgesamtheit und damit auch nicht die Standardabweichung der Stichprobenverteilung. Bekannt sind nur die statistischen Maßzahlen *einer* Stichprobe. Deshalb ersetzt man σ_x durch seinen aus der Stichprobe gewonnenen unverzerrten Schätzwert $s = \sqrt{\frac{1}{n-1} \cdot (x - \bar{x})^2}$. Dann wird die standardnormalverteilte Variable (z

$= \frac{\bar{x}-\mu}{\sigma_x/\sqrt{n}}$) zur t-verteilten Variablen ($t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$) mit $n-1$ Freiheitsgraden. In Gleichung 8.6 muss demgemäß σ_x durch s und z durch t der t-Verteilung mit $n-1$ Freiheitsgraden (FG) ersetzt werden. Es gilt dann

$$P(\bar{x} - t_{\frac{\alpha}{2},FG}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\frac{\alpha}{2},FG}\frac{s}{\sqrt{n}}) = 1-\alpha \qquad (8.7)$$

Die Größe der Standardabweichung $\sigma_{\bar{x}}$ und damit das Konfidenzintervall hängt wegen $\sigma_{\bar{x}} = \sigma_x/\sqrt{n}$ erstens von der Streuung σ_x der Grundgesamtheit (bzw. s der Stichprobe) ab. Er wird umso größer, je größer die Streuung in der Grundgesamtheit ist. Zweitens ist er vom Stichprobenumfang n abhängig. Er wird umso geringer, je größer der Umfang der Stichprobe ist. In Abb. 8.5 rechts ist diese Gesetzmäßigkeit demonstriert: mit größerem Stichprobenumfang n (hier: n = 4 und n = 100) aus der Grundgesamtheit wird die Standardabweichung der Stichprobenverteilung kleiner.

Durch den Multiplikator z bzw. t, den Sicherheitsfaktor, der das Vielfache des Standardfehlers angibt, wird festgelegt, mit welcher Sicherheit der wahre Wert in das Konfidenzintervall fällt. Üblich sind die Sicherheitsniveaus 95 % (Multiplikator 1,96 bei Normalverteilung) und 99 % (Multiplikator 2,576 bei Normalverteilung). Bei der t-Verteilung gilt für dieselbe Wahrscheinlichkeit je nach Stichprobengröße (genauer Zahl der Freiheitsgrade: FG = $n-1$) ein anderes t. Den zu einer Wahrscheinlichkeit gehörigen t-Wert müssen Sie gegebenenfalls in einer Tabelle der t-Verteilung nachschlagen.

Angenommen, man hat eine Stichprobe mit $n = 30$, $\bar{x} = 2500$, $s = 850$ erhoben und möchte einen 95 %-Konfidenzbereich für den unbekannten Mittelwert μ berechnen. Für FG = 29 und $\frac{\alpha}{2} = 0{,}025$ ergibt sich aus einer t-Tabelle $t = 2{,}045$.

Als Grenzwerte für den 95 %-Konfidenzbereich ergeben sich: $2500 - 2{,}045 \cdot \frac{850}{\sqrt{30}} = 2182{,}64$ und $2500 + 2{,}045 \cdot \frac{850}{\sqrt{30}} = 2817{,}36$. Bei einem höheren Stichprobenumfang n kann die t-Verteilung durch die Normalverteilung approximiert werden, so dass dann zur Vereinfachung mit z-Werten der Standardnormalverteilung gerechnet werden darf.

Wenn SPSS Konfidenzintervalle berechnet, fordert man überwiegend nur das gewünschte Sicherheitsniveau in Prozent an. Die SPSS-Prozeduren benutzen dann automatisch die richtigen zu dieser Wahrscheinlichkeit gehörenden t-Werte aus der t-Verteilung. Ausnahmen gelten bei Regelkartendiagrammen (dort muss der gewünschte t-Wert eingegeben werden) und beim Fehlerbalkendiagramm (dort kann dieser alternativ eingegeben werden).

Hinweise zu Einschränkungen der Anwendbarkeit von Konfidenzintervallen und Anwendung bei anderen Wahrscheinlichkeitsauswahlen.

- Die Konfidenzintervallberechnung ist natürlich nur geeignet, die durch Zufallsauswahl entstandenen Fehlerschwankungen zu berücksichtigen. Voraussetzung ist also, dass überhaupt eine solche Auswahl vorliegt. Das ist bei sehr vielen sozialwissen-

schaftlichen Untersuchungen (Quotenauswahl, Auswahl typischer Fälle, Auswahl auf Geratewohl) nicht der Fall. Selbst bei einer Zufallsstichprobe aber werden andere, systematische Auswahlverzerrungen nicht berücksichtigt.

- Das wahrscheinlichkeitstheoretische Modell der Ziehung einer einfachen uneingeschränkten Zufallsauswahl *mit Zurücklegen* muss zutreffen. SPSS geht grundsätzlich bei der Berechnung von Standardfehler von diesem Modell aus. Die Ergebnisse können aber auch bei einer einfachen uneingeschränkten Zufallsauswahl *ohne Zurücklegen* verwendet werden, wenn der Anteil der Stichprobe an der Grundgesamtheit relativ gering ist. Gewöhnlich setzt man das voraus, wenn der Stichprobenumfang weniger als 10 % des Umfanges der Grundgesamtheit ausmacht.
- Für einige Parameter wie Prozentwerte, Perzentilwerte, Zusammenhangsmaße bietet SPSS keine Berechnung von Standardfehler bzw. Konfidenzintervall an.

Es sollen nun einige Hinweise auf den Unterschied der Konfidenzintervalle bei Anwendung anderer wahrscheinlichkeitstheoretischer Auswahlverfahren gegeben werden. Wenn, wie in den Sozialwissenschaften kaum vermeidbar, wegen weiterer systematischer Auswahlfehler, die berechneten Konfidenzintervalle ohnehin nur als Anhaltspunkte gewertet werden können, kann es ausreichen, mit den Formeln von SPSS zu arbeiten und grobe Korrekturen im Hinblick auf das tatsächlich verwendete Verfahren vorzunehmen. (Die Überlegungen gelten nur, wenn die Auswahl *nicht disproportional* erfolgt.)

- *Großer Anteil der Stichprobe an der Grundgesamtheit (ca. ab 10 %).* Beim Ziehen ohne Zurücklegen ist die Endlichkeitskorrektur vorzunehmen. Der Standardfehler ist mit dem Faktor $\sqrt{\frac{N-n}{N-1}}$ zu multiplizieren.

  (N = Anzahl der Untersuchungseinheiten in der Grundgesamtheit).
- *Klumpenauswahl.* Wenn die Klumpen per Zufall gezogen werden, kann die einfache Formel benutzt werden. Fälle sind aber die Klumpen, nicht die Einzelfälle. Gegebenenfalls muss also – nur zur Berechnung des Standardfehlers – durch Aggregation eine neue Datei mit den Klumpen als Fällen erstellt werden.
- *Geschichtete Zufallsauswahl.* Sie führt zu geringeren Auswahlfehlern als eine einfache Zufallsauswahl. Der Grad der Verbesserung hängt allerdings sehr von der Heterogenität zwischen den Schichten und der Homogenität innerhalb der Schichten ab. Bei sozialwissenschaftlichen Untersuchungen ist der positive Schichtungseffekt nicht allzu hoch zu veranschlagen. Es mag genügen, mit den Formeln für einfache Zufallsauswahl zu arbeiten und sich zu vergegenwärtigen, dass man den Fehlerspielraum etwas überschätzt.
- *Mehrstufige Auswahl* (wenn auf jeder Ebene per Zufall ausgewählt wird). Der Standardfehler ist gegenüber der einfachen Zufallsauswahl höher. Die Berechnung kann je nach Zahl der Ebenen und der auf diesen jeweils angewendeten Auswahlmethode überaus komplex sein. Für eine zweistufige Auswahl kann der Standardfehler recht gut auf zwei verschiedene Arten näherungsweise berechnet werden. Erstes Verfahren: Man vernachlässigt den Effekt der zweiten Auswahlstufe und betrachtet die erste Stufe als Klumpenauswahl. Dann kann man den Standardfehler wie unter Klumpenauswahl beschrieben berechnen. Zweites Verfahren: Man berechnet den Standardfehler so, als läge eine einfache Zufallsauswahl vor und multipliziert das Ergebnis mit $\sqrt{2}$. Dies hat sich als grobe Annäherung bewährt (⇨ Böltken, S. 370).

8.5 Das Menü „Deskriptive Statistiken“

Das Menü „Deskriptive Statistiken“ enthält als Option ein gleichnamiges Untermenü. Dieses Untermenü bietet statistische Maßzahlen für zumindest auf dem Intervallskalenniveau gemessene (metrische) Daten an. Gegenüber dem Angebot von „Häufigkeiten...“ fehlen daher die Perzentilwerte und der Modalwert. Ansonsten handelt es sich um dieselben statistischen Maßzahlen wie im Untermenü „Häufigkeiten“. Es werden allerdings lediglich die statistischen Maßzahlen berechnet, also keine Tabellen oder Grafiken erstellt. Zusätzlich zu „Häufigkeiten...“ bietet „Deskriptive Statistiken...“ die Möglichkeit an, die Rohdaten in standardisierte z-Werte zu transformieren und diese als neue Variable zu speichern.

Z-Transformation. Eine Transformation der Rohdaten in standardisierte z-Werte kann aus zwei Gründen erfolgen:

- Erstens sind die Rohdaten verschiedener Variablen aufgrund der unterschiedlichen Messskalen in vielen Fällen kaum vergleichbar. Durch die z-Transformation werden dagegen Daten beliebiger metrischer Variablen auf einer vergleichbaren Messskala dargestellt.
- Zweitens wird die z-Transformation oft quasi als ein Mittel verwendet, auf Ordinalskalenniveau gemessene Daten auf Intervallskalenniveau zu heben. Man unterstellt dabei, dass die zugrundeliegende Verteilung einer Normalverteilung entspricht und die Bestimmung der relativen Position eines Falles innerhalb einer solchen Verteilung einer Messung auf einer Intervallskala gleich kommt.

Der z-Wert gibt nun die relative Position in einer solchen Verteilung an, indem er die Differenz des Rohwertes zum arithmetischen Mittel in Standardabweichungen ausdrückt.

$$z_i = \frac{x_i - \bar{x}}{s} \tag{8.8}$$

Das arithmetische Mittel der z-Werte ist 0 und die Standardabweichung 1.

So lässt sich etwa der z-Wert für eine Person mit einem Einkommen von 1500 DM in unserer oben dargestellten Einkommensverteilung berechnen. Das arithmetische Mittel beträgt 2096,78 DM, die Standardabweichung 1133,80:

$$Z_{1500} = (1500 - 2096{,}78) : 1133{,}80 = -0{,}526$$

Ein Einkommen von 1500 DM weicht demnach ca. eine halbe Standardabweichung vom durchschnittlichen Einkommen nach unten ab. Aus Tabellen für die Standardnormalverteilung kann man für einen so ermittelten Wert auch entnehmen, wieviel Prozent der Einkommensbezieher ein geringeres, wieviel ein höheres Einkommen beziehen.

Die so ermittelten z-Werte werden häufig für die Berechnung multivariater Statistiken benutzt. Nur nach einer solchen Standardisierung lässt sich z.B. die relative Bedeutung verschiedener Variablen beurteilen.

Nach Auswahl von:

▷ „Analysieren“, „Deskriptive Statistiken ▷“, „Deskriptive Statistiken...“ öffnet sich die Dialogbox „Deskriptive Statistik“ (⇨ Abb. 8.6). Hier können Sie aus der Variablenliste die Variablen auswählen. Außerdem steht ein Kontrollkästchen *.Standardisierte Werte als Variable speichern* zur Verfügung. Damit bestimmen Sie, ob z-Werte als neue Variable gesichert werden.

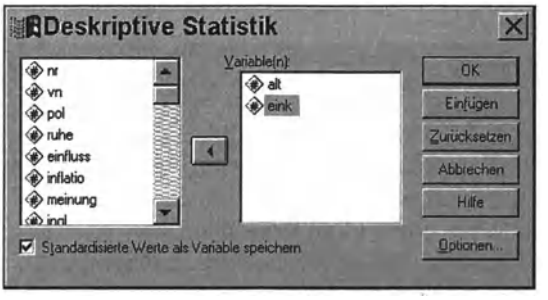


Abb. 8.6. Dialogbox „Deskriptive Statistik“

Das Anklicken der Schaltfläche „Optionen...“ öffnet die Dialogbox „Deskriptive Statistik: Optionen“ (⇨ Abb. 8.7).

Abb. 8.7. Dialogbox „Deskriptive Statistik: Optionen“

Hier können die gewünschten Statistiken durch Anklicken von Kontrollkästchen ausgewählt werden. Voreingestellt sind: arithmetisches Mittel („Mittelwert“), Standardabweichung, Minimum und Maximum.

Eine weitere Gruppe „Anzeigereihenfolge“ ermöglicht es, wenn gleichzeitig mehrere Variablen bearbeitet werden, durch Anklicken des entsprechenden Optionsschalters die Reihenfolge der Ausgabe zu bestimmen:

- ❒ *Variablenliste.* Ordnet sie in der Reihenfolge ihrer Auswahl (Voreinstellung).
- ❒ *Alphabetisch (Variablennamen).* Ordnet die Variablen nach ihrem Namen in alphabetischer Ordnung.
- ❒ *Aufsteigende Mittelwerte.* Ordnet die Variablen nach ihrem arithmetischen Mittel in ansteigender Reihenfolge, ausgehend vom kleinsten Mittelwert.
- ❒ *Absteigende Mittelwerte.* Ordnet umgekehrt nach absteigender Größe des arithmetischen Mittels.

▷ Wählen Sie die gewünschten Optionen aus und bestätigen Sie mit „Weiter" und „OK".

Die vorgeschlagenen Einstellungen ergeben die Ausgabe in Tabelle 8.6. Die Variablen „Einkommen" und „Alter" sind in umgekehrter Reihenfolge geordnet, weil sich für Alter eine kleineres arithmetisches Mittel ergibt als für Einkommen. Für beide Variablen werden alle ausgewählten statistischen Kennzahlen angezeigt.

Außerdem werden für EINK und ALT z-Werte in zwei neuen Variablen ZEINK und ZALT gespeichert wurden. Als Variablenlabel wird das alte Label mit vorangestelltem „Z-Wert:" übernommen. Mit diesen neuen Variablen können in Zukunft beliebige statistische Operationen ausgeführt werden.

Tabelle 8.6. Einige Deskriptive Statistiken für die Variablen „Alter" und „Einkommen"

Deskriptive Statistik

	N	Minimum	Maximum	Mittelwert	Standardabweichung
ALT	298	18	89	47,67	18,12
EINK	143	129	7000	2096,78	1133,80
Gültige Werte (Listenweise)	143				

Weitere Möglichkeiten bei Verwenden der Befehlssyntax.

- ❒ Z-Werte können nur für ein Subset der Variablen berechnet werden (mit dem VARIABLES-Unterkommando).
- ❒ Benutzerspezifische Namen können für die z-Wert-Variablen definiert werden (mit dem VARIABLES-Unterkommando).
- ❒ Mit dem MISSING-Unterkommando können Fälle, die bei irgendeiner der benutzten Variablen einen fehlenden Wert haben, generell (LISTWISE) von der Analyse ausgeschlossen werden. Benutzerdefinierte fehlende Werte können aber auch in die Analyse einbezogen werden (INCLUDE). Voreinstellung ist, dass nur die Fälle mit fehlenden Werten bei der betreffenden Variablen ausgeschlossen werden.
- ❒ Die Variablen können bei der Ausgabe auch nach der Größe einer beliebigen der berechneten statistischen Maßzahlen geordnet werden (mit dem SORT-Unterkommando).

8.6 Das Menü „Verhältnis“

Das Menü „Verhältnis“ dient dem Vergleich von Gruppen (unabhängige Variablen), wenn die abhängige Variable eine zusammengesetzte Variable ist, deren Wert sich aus dem Verhältnis der Werte zweier Ausgangsvariablen ergibt. (*Beispiel:* Stundenkilometer, Stundenlohn, Umsatz zu Verkaufsfläche etc.). Man könnte diese abhängige Variable auch aus den Ausgangsvariablen mit dem Menü „Berechnen“ bilden und für die Analyse z.B. das Menü „Mittelwerte vergleichen verwenden“. Das Menü „Verhältnis“ erspart aber diesen Umweg und bietet darüber hinaus einige Statistiken (Lage-, Streuungs- und Konzentrationsmaße) an, die in den anderen Menüs nicht zur Verfügung stehen.

Beispiel. Für die Daten von ALLBUS90.SAV soll der Stundenlohn von Männern und Frauen verglichen werden. Eine Variable Stundenlohn existiert nicht, sie ergibt sich vielmehr aus dem Verhältnis von EINK (Einkommen im Monat) und STDMON (Arbeitsstunden im Monat). Es sollen das arithmetische Mittel (mitsamt Konfidenzintervall) und die Standardabweichung verglichen werden. Außerdem soll festegestellt werden, ob sich der Grad der Konzentration in einem Bereich mit Untergrenze Mittelwert – 50% des Mittelwertes und der Obergrenze Mittelwert + 50% des Mittelwerts bei den beiden Gruppen unterscheidet.

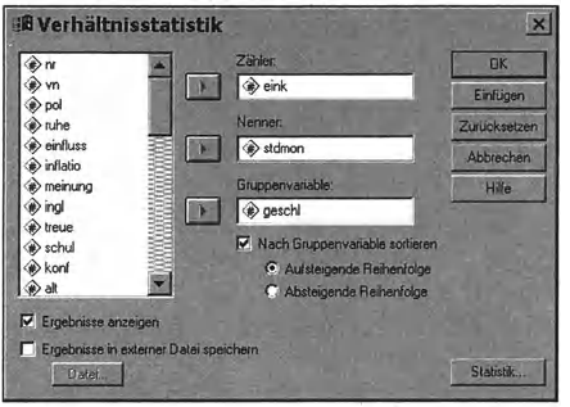


Abb. 8.8. Dialogbox “Verhältnisstatistik“

Um diese Analyse durchzuführen, gehen Sie wie folgt vor:

- ▷ Wählen Sie „Analysieren“, „Deskriptive Statistiken“ und „Verhältnis“. Die Dialogbox „Verhältnisstatistik“ erscheint (⇨ Abb. 8.8).
- ▷ Bilden Sie die abhängige Variable, indem Sie EINK aus der Variablenliste in das Feld „Zähler“ übertragen und STDMON in das Feld „Nenner“.
- ▷ Geben Sie an, für welche Gruppen der Vergleich durchgeführt werden soll, indem Sie die unabhängige Variable GESCHL in das Feld „Gruppenvariable“ übertragen. Die Voreinstellungen hinsichtlich Sortierung der Ausgabe behalten wir bei.

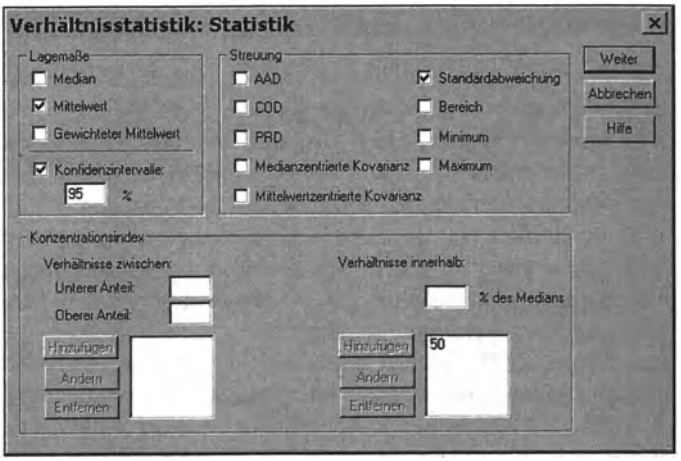


Abb. 8.9. Dialogbox „Verhältnisstatistik: Statistik"

Jetzt muss festgelegt werden, welche Statistiken zum Vergleich herangezogen werden sollen.

- ▷ Dazu öffnen Sie durch Anklicken der Schaltfläche „Statistik" die Dialogbox „Verhältnisstatistik: Statistik" (⇨ Abb. 8.9).
- ▷ Markieren Sie in der Gruppe „Lagemaße" die Option „Mittelwert", in der Gruppe „Streuung" die Option „Standardabweichung".
- ▷ Zur Definition des Konzentrationsmaßes tragen Sie in das Eingabefeld „Verhältnis innerhalb" „Prozent des Medians" den Wert 50 ein und übertragen dies durch „Hinzufügen" in das Auswahlfeld. Bestätigen Sie mit „Weiter" und „OK".

Die Ausgabe sehen Sie in Tabelle. 8.7.

Tabelle 8.7. Einige Statistiken für die Variable EINK/STDMO

Verhältnistatistik für EINK / STDMON

Gruppe	Mittelwert	Konfidenzintervall 95% für Mittelwert		Std.-Abweichung	Konzentrationskoeffizient
		Untergrenze	Obergrenze		Innerhalb 50% des Medians
MAENNLICH	17,061	15,326	18,796	5,975	81,3%
WEIBLICH	13,476	11,768	15,184	4,405	82,1%
Insgesamt	15,740	14,440	17,041	5,691	78,9%

Beim Erstellen der Konfidenzintervalle wird von einer Normalverteilung der Verhältnisse ausgegangen.

Man kann dieser u.a. entnehmen, dass der Stundenlohn im Mittel bei den Männern höher liegt als bei den Frauen (ca. 17 gegenüber ca. 13,47 DM). Die Löhne der Männer Streuen mit einer Standardabweichung von 5,975 etwas stärker als die der Frauen mit 4,405. Dabei konzentriert sich bei beiden Gruppen ungefähr ein gleich

starker Anteil von 81 bis 82% in einer mittleren Einkommensgruppe Median ± 50% des Medians.

Optionen. In der Dialogbox „Verhältnisstatistik“ können sie die Sortierung der Ausgabe bestimmen. Wählen Sie „Nach Gruppenvariable sortieren“, werden die Gruppen in der Ausgabetabelle in der Reihenfolge ihrer Werte ausgegeben, je nach weiter gewählter Option in aufsteigender oder absteigender Folge. (*B.:* Aufsteigende Folge sortiert 1 = männlich, 2 = weiblich, absteigende die umgekehrte Folge.) Ist die Option ausgeschaltet, werden die Gruppen in der Reihenfolge ausgegeben, in der sie bei den ersten Fällen erscheinen.

Statistiken.

Lagemaße. Als Lagemaße werden Mittelwert (arithmetisches Mittel), Median und Gewichteter Mittelwert angeboten. Letzterer wird als Quotient aus dem Mittelwert der Zählervariable und dem Mittelwert der Nennervariablen gebildet (im Gegensatz zum einfachen Mittelwert, der aus den Quotienten gebildet wird). Für alle drei Lagemaße können Konfidenzintervalle angefordert werden (allerdings mit dem gleichen Sicherheitsniveau für alle Lagemaße). Wird ein Konfidenzintervall angefordert, kann man das Sicherheitsniveau selbst im Feld „Konfidenzintervalle“ festlegen (Voreinstellung 95%).

Streuungsmaße. Neben den bekannten Streuungsmaßen Standardabweichung und Bereich (Spannweite), letzteres errechnet sich als Differenz aus Maximum und Minimum, zwei ebenfalls angebotenen Maßen, stehen 5 weitere Streuungsmaße zur Auswahl. Sie werden in der Ausgabe teilweise anders – und z.T. irreführend – beschriftet als in der Auswahlliste. Deshalb wird diese Beschriftung in Klammern angeführt.

- ❒ *AAD* (Mittlere absolute Abweichungen). Summe der absoluten Abweichungen von Mittelwert durch Zahl der Fälle.
- ❒ *COD* (Streuungskoeffizient). Ist AAD geteilt durch Mittelwert (das Gegenstück zum Variationskoeffizienten, der sich aus der Standardabweichung errechnet).
- ❒ *PRD* (Preisgebundene Differenz). Ist der Quotient aus Mittelwert und gewichtetem Mittelwert.
- ❒ *Mittelwertzentrierte Kovarianz* (Variationskoeffizient / zentrierter Mittelwert). Es handelt sich um den bekannten Variationskoeffizienten: Standardabweichung durch Mittelwert.
- ❒ *Medianzentrierte Kovarianz* (Variationskoeffizient / zentrierter Median). Standardabweichung geteilt durch Median.

Konzentrationsindex. Das Ergebnis der Konzentrationsmaße ist immer der Anteil der Gruppe, deren Wert in einen bestimmten Bereich fällt (*B.:* Anteil der Männer bzw. der Frauen, die einen Stundenlohn zwischen 10 und 20 DM erreichen). (Ein solcher Konzentrationsindex kann immer nur in Kombination mit einer Statistik gewählt werden, z.B. dem arithmetischen Mittel, obwohl das für das Ergebnis keine Bedeutung hat.) Der Bereich, für den der Anteil der Fälle ermittelt werden soll, kann auf zweierlei Weise bestimmt werden:

- ❒ *Verhältnisse zwischen.* Hier werden feste Ober- und Untergrenzen des Bereichs angegeben, z.B. zwischen 10 und 20 (DM für Stundenlohn).

- *Verhältnisse innerhalb.* Auch hier wird ermittelt, wie viel Prozent einer Gruppe mit ihren Werten zwischen zwei Grenzen liegen. Nur werden diese Grenzen implizit ermittelt aus einer bestimmten prozentualen Abweichung vom Medianwert nach oben und unten. Die hoch die prozentuale Abweichung sein soll, gibt man im Feld „% des Medians" an (B.: Der Median beträgt DM 16. Gewünscht ist 50% des Medians. Dies wären DM 8. Also liegt die Untergrenze des Bereichs, für den der Anteil der Gruppe berechnet wird, bei 8, die Obergrenze bei 24 DM.

9 Explorative Datenanalyse

Das Untermenü „Explorative Datenanalyse“ (im Syntaxhandbuch und im Algorithmenhandbuch wird es als „Examine“ geführt) vereinigt zwei unterschiedliche Arten von Optionen:

- Zunächst bietet es Ergänzungen der deskriptiven – zumeist eindimensionalen – Statistik. Das sind zum einen die *robusten Lageparameter*. Hierbei handelt es sich um auf besondere Weise berechnete Mittelwerte, bei denen der Einfluss von Extremwerten ausgeschaltet oder reduziert wird. Zum anderen handelt es sich um zwei besondere Formen der grafischen Aufbereitung, den *Stengel Blatt(Stem-und-Leaf-)Plot* und den *Boxplot*. Beide dienen dazu, Verteilungen genauer bzw. unter speziellen Aspekten aussagekräftig darzustellen. Diese Hilfsmittel können zur normalen deskriptiven Analyse gebraucht werden, aber auch – was für andere deskriptive Statistiken gleichfalls zutrifft – zur Prüfung der Daten auf Fehler und zur Vorbereitung weiterer Analysen. Auch das Vorliegen der Anwendungsvoraussetzungen statistischer Prüfmodelle kann damit teilweise untersucht werden. Die Fehlersuche, aber auch die Hypothesengenerierung, wird zusätzlich durch Optionen zur Identifikation von Extremfällen unterstützt.
- Es kann das Vorliegen einer Normalverteilung oder von homogenen Streuungen in Untergruppen geprüft werden. Dies sind Anwendungsvoraussetzungen verschiedener statistischer Testmodelle.

9.1 Robuste Lageparameter

Das gebräuchlichste Lagemaß (Lokationsparameter) für metrische Daten ist das arithmetische Mittel. Es besitzt eine Reihe von Vorteilen gegenüber anderen Parametern, unter anderem den, dass alle Werte einer Untersuchungspopulation in die Berechnung eingehen. Andererseits aber hat es den Nachteil, dass es durch Extremwerte (Ausreißer) u.U. stark beeinflusst werden kann und dann ein unrealistisches Bild ergibt. Ausreißer wirken sich insbesondere bei kleinen Populationen störend aus. Diesen Nachteil hat z.B. der Medianwert nicht. Dafür besteht bei ihm aber der umgekehrte Nachteil, dass – insbesondere bei metrisch gemessenen Daten – die verfügbaren Informationen nur rudimentär genutzt werden. Um einerseits möglichst viele Werte zur Berechnung des Lagemaßes zu benutzen, andererseits aber die störenden Einflüsse von Extremwerten auszuschließen, wurden sogenannte *robuste* Lagemaße entwickelt. Allgemein gesprochen, handelt es sich um

gewogene arithmetische Mittel, bei deren Berechnung die Werte, je nach Grad der Abweichung vom Zentrum, mit ungleichem Gewicht eingehen, im Extremfalle mit dem Gewicht 0. Allgemein gilt für die robusten Lokationsparameter die Formel:

$$\bar{x} = \frac{\sum w_i \cdot x_i}{\sum w_i} \tag{9.1}$$

Wobei w_i das jeweilige Gewicht des Wertes angibt.

Getrimmte Mittelwerte (Trimmed Mean). Die einfachste Form sind sogenannte „getrimmte Mittelwerte". Sie werden als normales arithmetisches Mittel unter Ausschluss von Extremwerten berechnet. Die Extremwerte erhalten (formal gesprochen) das Gewicht 0, alle anderen das Gewicht 1. Als Extremwerte wird ein bestimmter Prozentanteil der Werte an jedem Ende der geordneten Rangreihe der Fälle bestimmt. So bedeutet eine 5 % Trimmung, dass die 5 % niedrigsten und die 5 % höchsten Werte nicht in die Berechnung des arithmetischen Mittels einbezogen werden.

M(aximum-Likelihood)-Schätzer (M-Estimators). SPSS bietet vier verschiedene M-Schätzer. Im Unterschied zu getrimmten Mittelwerten, teilen sie die Werte nicht nur in zwei Kategorien – benutzte und nicht benutzte – ein, sondern vergeben unterschiedliche Gewichte: extremeren Werten geringere, Werten nahe dem Zentrum höhere. Der Unterschied zwischen den verschiedenen Schätzern besteht in den verwendeten Gewichtungsschemata.

Allen gemeinsam ist, dass die Berechnung nicht aus den Rohdaten (x_i), sondern aus einer standardisierten Abweichung u_i des jeweiligen Wertes von dem geschätzten Lageparameter (z.B. Mittelwert oder Median) erfolgt.

$$u_i = \frac{|x_i - \text{Lageschätzer}|}{\text{Streuungsschätzer}} \tag{9.2}$$

Die absolute Abweichung des Rohwertes vom (zunächst unbekannten!) robusten Mittelwert (Lageschätzer) wird also durch einen Streuungsparameter geteilt. Da in die Formel der Lageschätzer eingeht, der ja selbst erst Ergebnis der Berechnung sein soll, muss die Berechnung iterativ erfolgen. Als Streuungsschätzer wird gewöhnlich der Median der absoluten Abweichungen vom Stichprobenmedian verwendet. Die Formel für MAD (Median der Abweichungsdifferenzen) lautet:

$$MAD = Md \text{ von allen } |x_i - Md| \tag{9.3}$$

Die Gewichtungsschemata der vier angebotenen M-Schätzer unterscheiden sich nun wie folgt:

- *M-Schätzer nach Hampel.* Hier wird ein kompliziertes Wägungsschema benutzt, das von drei Grenzwerten von u abhängt. Es sind die Grenzen a = 1,7, b = 3,4 und c = 8,5. Werte unterhalb der Grenze a bekommen ein Gewicht von 1, Werte zwischen a und b, ein Gewicht a : u und Werte zwischen b und c ein Ge-

wicht von: $\frac{a}{u} \cdot \frac{c-u}{c-b}$. Alle Werte oberhalb von c erhalten das Gewicht 0. Abb. 9.1 zeigt das Wägungsschema.

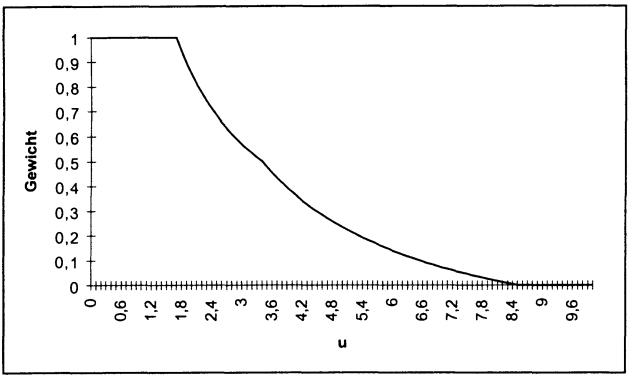


Abb. 9.1. Wägungsschema für „M-Schätzer nach Hampel"

Die anderen Verfahren arbeiten nur mit einer kritischen Grenze c.

- *M-Schätzer nach Huber.* Das Gewicht bleibt bis zur kritischen Grenze c = 1,339 gleich hoch und sinkt dann kontinuierlich.
- *Tukey-Biweight.* Das Gewicht sinkt langsam von 1 auf 0, bis zur kritischen Grenze c = 4,685. Bei größeren Werten ist das Gewicht 0.
- *Andrews-Welle.* Die Gewichte sinken ohne abrupten Übergang von 1 auf 0. Die kritische Grenze ist c = 1,339π. Höhere Werte erhalten das Gewicht 0.

Um die robusten Lokationsparameter zu berechnen, gehen Sie wie folgt vor (Beispiel aus ALLBUS90.SAV):

▷ Wählen Sie die Befehlsfolge „Analysieren", „Deskriptive Statistiken ▷", „Explorative Datenanalyse...". Es öffnet sich die Dialogbox „Explorative Datenanalyse" (⇨Abb. 9.2).

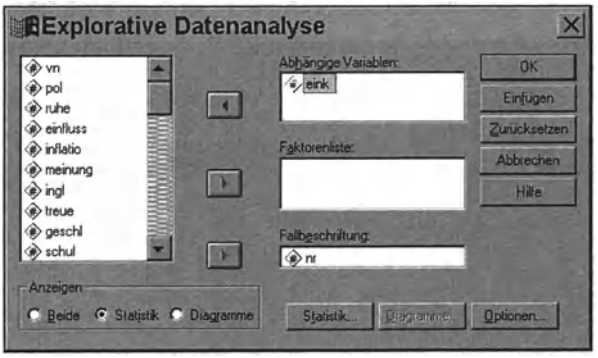


Abb. 9.2. Dialogbox „Explorative Datenanalyse"

- ▷ Übertragen Sie die gewünschte Variable aus der Quellvariablenliste in das Eingabefeld „Abhängige Variablen:“ (hier: EINK).
- ▷ Sollten Sie auch an der Identifikation von Extremwerten interessiert sein, übertragen Sie die Identifikationsvariable aus der Quellvariablenliste in das Eingabefeld „Fallbeschriftung:“ (hier: NR, mit den Fallnummern als Werten).
- ▷ Interessieren ausschließlich die Statistiken, klicken sie in der Gruppe „Anzeigen“ auf die Optionsschaltfläche „Statistik“.
- ▷ Klicken Sie auf die Schaltfläche „Statistik...“. Die Dialogbox „Explorative Datenanalyse: Statistik“ erscheint (⇨ Abb. 9.3).

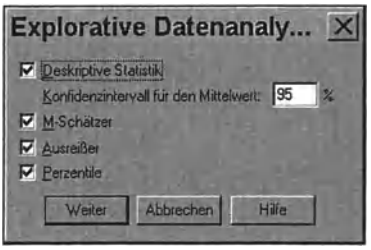


Abb. 9.3. Dialogbox „Explorative Datenanalyse: Statistik“

- ▷ Klicken Sie auf die Kontrollkästchen „Deskriptive Statistik“ und „M-Schätzer“.
- ▷ Bestätigen Sie mit „Weiter“ und „OK“.

Im Beispiel ergibt sich die in Tabelle 9.1 dargestellte Ausgabe. Die erste Tabelle ergibt sich aus der Option „Deskriptive Statistik“. Sie enthält die typischen Lage-, Streuungs- und Form-Maße, wie sie schon bei der Besprechung der Menüs „Häufigkeiten“ und „Deskriptive Statistiken“ dargestellt wurden. Ergänzend sind zwei Maße zu erwähnen. Das Maß „Interquartilbereich“. Es gibt die Distanz zwischen oberem Quartil (75. Perzentil) und unterem Quartil (25. Perzentil) an. Es ist ein gebräuchliches Streuungsmaß. „5 % getrimmtes Mittel“ ist ein getrimmtes arithmetisches Mittel, das unter Auslassung der 5 % Fälle mit den höchsten und der 5 % Fälle mit den niedrigsten Werten berechnet wird. Der Wert liegt mit 2025 etwas unter dem normalen $\bar{x}$-Wert von 2096,78 DM. Offensichtlich haben die Extremwerte des oberen Bereiches $\bar{x}$ etwas stärker bestimmt als die des unteren. Außerdem ist noch die Ober- und Untergrenze des 95%-Konfidenzintervall für das arithmetische Mittel angegeben. (Die Voreinstellung des Sicherheitsniveaus von 95% kann in der Dialogbox geändert werden.)

Die eigentlichen M-Schätzer sind in der unteren Tabelle enthalten. Diese Tabelle enthält als Fußnoten auch die verwendeten Gewichtungskonstanten. Die mit „M-Schätzer nach Huber“ überschriebene Ausgabe 1903,83 gibt den nach diesem Verfahren berechneten robusten Mittelwert von 1903,83 DM an. Die Fußnote (mit der völlig irreführenden Beschriftung „Die Gewichtungskonstante ist“) teilt mit, dass mit einer kritischen Grenze von 1,339 gerechnet wurde. Nach Hampel beträgt das robuste arithmetischen Mittel 1897,7930. Die verwendeten kritischen Grenzen waren laut Fußnote 1,700; 3,400; 8,500. Wie man sieht, liegen die Werte der robu-

sten Lageparameter alle deutlich unter dem des gewöhnlichen arithmetischen Mittels. Es wurde bei allen mehr oder weniger stark der Einfluss der nach oben abweichenden Extremwerte ausgeschaltet. Gleichzeitig schwanken aber auch die robusten Mittelwerte deutlich untereinander. Am niedrigsten fällt Andrews M-Schätzer mit 1796,61 DM aus, am höchsten der M-Schätzer nach Huber mit 1903,83 DM.

Tabelle 9.1. Ausgabe von deskriptiven Statistiken und M-Schätzern

Univariate Statistiken

			Statistik	Standardfehler
BEFR.: MONATLICHES NETTOEINKOMMEN	Mittelwert		2096,78	94,813
	95% Konfidenzintervall des Mittelwerts	Untergrenze	1909,36	
		Obergrenze	2284,21	
	5% getrimmtes Mittel		2025,45	
	Median		1900,00	
	Varianz		1285506	
	Standardabweichung		1133,801	
	Minimum		129	
	Maximum		7000	
	Spannweite		6871	
	Interquartilbereich		1200,00	
	Schiefe		1,186	,203
	Kurtosis		2,000	,403

M-Schätzer

	M-Schätzer nach Huber[a]	Tukey-Biweight[b]	M-Schätzer nach Hampel[c]	Andrews-Welle[d]
BEFR.: MONATLICHES NETTOEINKOMMEN	1903,83	1800,22	1897,79	1796,61

a. Die Gewichtungskonstante ist 1,339.

b. Die Gewichtungskonstante ist 4,685.

c. Die Gewichtungskonstanten sind 1,700, 3,400 und 8,500

d. Die Gewichtungskonstante ist 1,340*pi.

Weitere Statistikoptionen. Die Dialogbox „Explorative Datenanalyse: Statistiken" bietet weitere Statistikoptionen an:

- *Perzentile.* Gibt verschiedene wichtige Perzentilwerte aus (⇨ Tabelle 9.2). Diese werden nach etwas anderen Methoden als üblich berechnet (siehe unten). Die Verfahren „Weighted Average" und „Tukey Angelpunkte" sind voreingestellt. Weitere können mit der Befehlssyntax angefordert werden.
- *Ausreißer.* Gibt die fünf Fälle mit den höchsten und den niedrigsten Werten aus (⇨ Tabelle 9.3).

Tabelle 9.2. Ausgabe bei Nutzung der Option „Perzentile"

Perzentile

		Perzentile						
		5	10	25	50	75	90	95
Gewichtetes Mittel (Definition 1)	BEFR.: MONATLICHES NETTOEINKOMMEN	720,00	894,00	1300,00	1900,00	2500,00	3952,00	4300,00
Tukey-Angelpunkte	BEFR.: MONATLICHES NETTOEINKOMMEN			1300,00	1900,00	2500,00		

Tabelle 9.3. Ausgabe bei Verwendung der Option „Ausreißer"

Extremwerte

		BEFR.: MONATLICHES NETTOEINKOMMEN		
		Fallnummer	IDENTIFIKATIONSNUMMER DER BEFRAGTEN	Wert
Größte Werte	1	289	4959	7000
	2	136	2666	5300
	3	249	4329	4800
	4	6	83	4800
	5	192	3527	,[a]
Kleinste Werte	1	168	3090	129
	2	1	38	150
	3	231	4014	370
	4	141	2742	520
	5	287	4911	650

a. Nur eine partielle Liste von Fällen mit dem Wert 4500 wird in der Tabelle der oberen Extremwerte angezeigt.

Angegeben werden die Werte der Fälle mit den fünf größten und den fünf kleinsten Werten, außerdem die automatisch vergebene SPSS-Fallnummer (und/oder der Wert einer selbst gewählten Identifikationsvariablen [wie hier: NR]). Haben mehrere Fälle denselben Wert, wird nur der erste Fall ausgegeben. Eine Fußnote gibt – wie hier für den Wert 4500 – an, dass noch mehr Fälle mit diesem Wert existieren. Die Identifikation von Extremwerten dient in erster Linie der Suche nach Datenfehlern, aber auch der Prüfung der Frage, inwieweit normale Lokationsparameter angewendet werden können.

Berechnen von Perzentilwerten. Da die Explorative Datenanalyse verschiedene Berechnungsarten für Perzentilwerte anbietet und diese sich etwas von der üblichen Berechnung unterscheiden, sollen diese etwas näher erläutert werden: Ein Perzentilwert ist bekanntlich derjenige Wert, den genau der Fall in einer geordneten Rangreihe hat, unter dem ein bestimmter (durch das gewünschte Perzentil festgelegter) Anteil der Fälle liegt. Nun ist das aber häufig kein bestimmter Fall, sondern die Grenze liegt zwischen zwei Fällen. Beim Medianwert gilt das z.B. immer, wenn er aus einer geraden Anzahl von Fällen zu ermitteln ist. Bei anderen Perzentilwerten tritt diese Situation noch häufiger ein. Die verschiedenen Arten der Per-

zentilberechnung unterscheiden sich darin, wie sie in einer solchen Situation den Perzentilwert bestimmen. Im Prinzip sind zwei Vorgehensweisen geläufig:

- Es wird (auf unterschiedliche Weise) durch Interpolation ein Zwischenwert zwischen den beiden Werten der Fälle ermittelt, zwischen denen die Grenze verläuft.
- Es wird der Wert einer dieser beiden Fälle (welcher, wird wiederum unterschiedlich festgelegt) als Perzentilwert bestimmt.

Explorative Datenanalyse benutzt per Voreinstellung folgende Berechnungsarten:

① *Weighted Average (HAVERAGE).* Wird in der Ausgabe als „Gewichtetes Mittel (Definition 1)" bezeichnet. Diese Berechnungsart entspricht der üblichen Berechnung bei nicht klassifizierten Werten. Es handelt sich um einen gewogenen Mittelwert bei $x_{(n+1)*p}$. Es wird ein gewogener Mittelwert von x_i und x_{i+1} gebildet nach der Formel:

$$(1-f)\cdot x_i + f\cdot x_{i+1} \qquad (9.4)$$

Dabei wird $(n+1)\cdot p$ in einen ganzzahligen Anteil i und einen Nachkommaanteil f zerlegt.

Dabei gilt:

n = Zahl der Fälle.
p = Perzentil, angegeben als Anteilszahl.
i = der Rangplatz des unteren der beiden Fälle, zwischen denen die Grenze liegt, i+1 der Rangplatz des oberen.

Beispiel:

Fall	1	2	3	4	5	6	7	8	9	10
Wert	10	20	30	40	50	50	50	60	70	70

Aus den angegebenen Werten von zehn Fällen soll der untere Quartilswert oder das 25. Perzentil berechnet werden. Die Zahl der Fälle n = 10. Das Perzentil p = 0,25. Entsprechend ergibt $(n+1)\cdot p = (10+1)\cdot 0{,}25 = 2{,}75$. Dies ist der Rangplatz, für den der Wert zu errechnen ist. Da es sich hier aber um keinen ganzzahligen Wert handelt, muss ein Mittelwert zwischen dem zweiten Fall (dessen Wert ist $x_i = 20$) und dem dritten (dessen Wert ist $x_{i+1} = 30$) gebildet werden. Dazu wird zunächst der Wert 2,75 in den ganzzahligen Anteil i = 2 und den gebrochenen Anteil f = 0,75 zerlegt. Der gebrochene Anteil gibt praktisch den Anteil der Spanne zwischen dem Fall i und dem Fall i+1 an, der noch zu den unterhalb der Grenze liegenden Fällen zu zählen ist. f wird daher zur Gewichtung bei der Mittelwertbildung benutzt.

$$(1-f)\cdot x_i + f\cdot x_{i+1} = (1-0{,}75)\cdot 20 + 0{,}75\cdot 30 = 27{,}5$$

② *Tukey-Angelpunkte (Tukey's Hinges).* Wird zusammen mit irgendeiner der Berechnungsarten das 25., das 50. oder das 75. Perzentil aufgerufen, gibt SPSS automatisch auch das Ergebnis der Berechnung nach der Methode „Tukey-An-

gelpunkte" aus. In diesem Fall werden diese drei Werte nach einem komplexen Verfahren ermittelt, das hier nicht näher erläutert werden kann.

Über die Befehlssyntax sind weitere Berechnungsarten verfügbar:

③ *WAVERAGE.* Gewogener Mittelwert bei x_{n*p}. Dieser Wert wird im Prinzip auf dieselbe Weise gebildet. Jedoch wird der mittlere Rangplatz nicht von n+1, sondern von w ausgehend gebildet. Entsprechend verändert sich die Berechnung unseres Beispiels: $n \cdot p = 10 \cdot 0{,}25 = 2{,}5$. Mit diesem veränderten Wert weiter berechnet ist: i= 2 und f = 0,5. Daraus folgt:
$(1-f) \cdot x_i + f \cdot x_{i+1} = (1-0{,}5) \cdot 20 + 0{,}5 \cdot 30 = 25$

④ *ROUND.* Es wird der Wert x_i genommen. Dabei ist i der ganzzahlige Teil von $n \cdot p + 0{,}5$. Im Beispiel wäre $n \cdot p + 0{,}5 = (10 \cdot 0{,}25) + 0{,}5 = 3$. Da nur ein ganzzahliger Teil vorhanden ist, ist i = 3. Der Wert des dritten Falles ist der untere Quartilswert, also 30.

⑤ *EMPIRICAL.* Der Wert von x_i wird verwendet, wenn der gebrochene Teil von $n \cdot p = 0$. Sonst wird x_{i+1} genommen. Im Beispiel ist $n \cdot p = 10 \cdot 0{,}25 = 2{,}5$. Es ist ein nicht ganzzahliger Rest vorhanden. Also wird $x_{i+1} = 30$ verwendet.

⑥ *AEMPIRICAL.* Wenn der gebrochene Teil von $n*p = 0$ ist, wird als Wert ein nicht gewogenes arithmetisches Mittel zwischen x_i und x_{i+1} verwendet, ansonsten der Wert x_{i+1}. Da im Beispiel ein gebrochener Teil vorliegt, wird wieder der Wert des dritten Falles, also 30 verwendet.

Bei großen Fallzahlen, wo meist mehrere Fälle denselben Wert haben, unterscheiden sich die Ergebnisse in der Regel nicht voneinander. Das gilt vor allem auch deshalb, weil unter bestimmten Bedingungen – wenn festgelegte Grenzwerte überschritten sind – auch bei den mit gewichteten Mitteln arbeitenden Verfahren auf eine Mittelwertbildung verzichtet und der Wert des Falles i+1 verwendet wird (⇨ SPSS Statistical Algorithms). Liegen kleine Fallzahlen vor, können dagegen deutliche Unterschiede zwischen den Ergebnissen der verschiedenen Berechnungsarten auftreten.

Anmerkung. Alle Berechnungsarten gehen vom Vorliegen nicht klassifizierter Daten aus. Nur die Option „Perzentile" des Menüs „Häufigkeiten" ermöglicht es, für klassifizierte Daten exakte Perzentilwerte zu berechnen.

9.2 Grafische Darstellung von Daten

Das Menü „Explorative Datenanalyse" bietet verschiedene Formen der grafischen Darstellung von Daten. Einerseits ergänzen sie die beschreibende Statistik, zum anderen sind sie z.T. mit besonderen Features zur Identifikation von Extremwerten ausgestattet. Dies unterstützt die Suche nach Datenfehlern und u.U. die Generierung neuer Hypothesen. Schließlich werden sie auch zur Prüfung der Voraussetzungen statistischer Prüfverfahren benutzt: Geprüft werden können die Vorausset-

zung der Normalverteilung und die Voraussetzung gleicher Varianz in Vergleichsgruppen.

- *Histogramm.* Es ist für kontinuierliche metrische Daten geeignet. SPSS teilt den Bereich der Daten automatisch in Klassen gleicher Breite. Die Punkte auf der x-Achse repräsentieren jeweils den Mittelpunkt einer Klasse (bei SPSS fehlerhaft) (⇨ Kap. 8.2.3). Außer zur üblichen deskriptiven Analyse kann man ein Histogramm auch zur Beurteilung der Anwendbarkeit statistischer Testverfahren nutzen. Insbesondere ist es möglich, die Verteilung auf Eingipfligkeit und Annäherung an die Normalverteilung zu prüfen. Auch Lücken und Extremwerte kann man durch Analyse des Histogramms aufdecken.
- *Stengel-Blatt (Stem-and-Leaf) Plot.* Sind histogrammähnliche Darstellungen. Allerdings werden die Säulen durch Zahlen dargestellt, die einzelne Untersuchungsfälle repräsentieren und nähere Angaben über deren genauen Wert machen. Dadurch sind detaillierte Informationen über die Verteilung innerhalb der Klassen gegeben, die bei der Verwendung des Histogramms verloren gehen. Die *Stengel-Blatt-Diagramme* werden durch die besondere Aufbereitung der Extremwerte insbesondere zur Fehlersuche verwendet.
- *Boxplots.* Sie geben keine Auskunft über Einzelwerte, sondern über zusammenfassende Statistiken (die Lage von Median, oberem und unterem Quartil) und Extremwerte. Sie sind besonders geeignet zur Identifikation von Extremwerten. Der Vergleich von Boxplots verschiedener Gruppen wird verwendet, um die für viele statistische Tests gültige Voraussetzung gleicher Streuung in den Vergleichsgruppen zu prüfen.
- *Normalverteilungsdiagramme.* Sind spezielle Darstellungsweisen zur Überprüfung der Voraussetzung der Normalverteilung.

9.2.1 Univariate Diagramme: Histogramm und Stengel-Blatt-Diagramm

Um ein Histogramm und/oder ein Stengel-Blatt Diagramm für die Variable EINK (Monatseinkommen) zu erstellen, gehen Sie wie folgt vor:

▷ Wählen Sie in der Dialogbox „Explorative Datenanalyse“ (⇨ Abb. 9.2) die gewünschte abhängige Variable (hier: EINK).

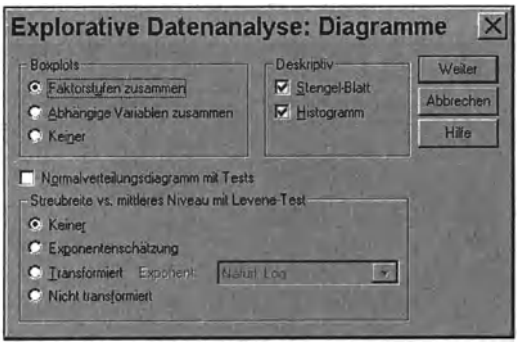


Abb. 9.4. Dialogbox „Explorative Datenanalyse: Diagramme“

- ▷ Falls ausschließlich das Diagramm gewünscht wird, wählen Sie in der Gruppe „Anzeigen" die Option „Diagramme" (⇨ Abb. 9.2).
- ▷ Klicken Sie auf die Schaltfläche „Diagramme...". Die Dialogbox „Explorative Datenanalyse: Diagramme" erscheint (⇨ Abb. 9.4).
- ▷ Klicken Sie in der Gruppe „Deskriptiv" auf die beiden Kontrollkästchen „Stengel-Blatt" und „Histogramm".
- ▷ Klicken Sie in den beiden anderen Gruppen jeweils auf die Optionsschaltfläche „Keine".
- ▷ Bestätigen Sie mit „Weiter" und „OK".

Für die Variable Einkommen (EINK) wird ein Histogramm im Ausgabefenster dargestellt. Da dieses nicht mit einer Normalverteilungskurve überlagert werden kann, empfiehlt es sich, das Histogramm besser im Menü „Häufigkeiten" (⇨ Kap. 8.2.3) oder im Menü „Grafiken" zu erstellen (⇨ Kap. 26.12) bzw. im „Diagramm-Editor" entsprechend zu bearbeiten (⇨ Kap. 27).

Tabelle 9.4. Stengel-Blatt Diagramm für die Variable Einkommen

BEFR.: MONATLICHES NETTOEINKOMMEN Stem-and-Leaf Plot

Frequency	Stem & Leaf
3,00	0 . 113
14,00	0 . 56678888888999
28,00	1 . 0000000111222222233333334444
28,00	1 . 55555566666777777778888888999
30,00	2 . 000000011111111122222233344444
14,00	2 . 55555566788899
6,00	3 . 002344
6,00	3 . 555688
6,00	4 . 000002
8,00	Extremes (>=4300)

Stem width: 1000
Each leaf: 1 case(s)

Im Ausgabefenster erscheint außerdem das in Tabelle 9.4 dargestellte Stengel-Blatt-Diagramm. Im Gegensatz zum Histogramm ist das Stengel-Blatt Diagramm ein besonderes Angebot des Programms „Explorative Datenanalyse". Diese Grafikart soll näher erläutert werden. In einem Stengel-Blatt-Diagramm wird die Häufigkeit der einzelnen Kategorien – wie im Histogramm – als Säulenhöhe dargestellt. Die Säulen werden aber aus Zahlen gebildet, aus denen man – kombiniert mit den Zahlen am Fuß der Säule – die Werte jedes Einzelfalles – zumindest näherungsweise – entnehmen kann. Dazu werden die Werte in zwei Teile zerlegt, die führenden Ziffern (Stengel, Stems) und die Folgeziffern (Blätter, Leafs). Die führenden Ziffern werden jeweils am Fuß der Säule angegeben, die Leafs als Werte in der Säule. Sind die Werte klein (bis 100), wird so der exakte Wert mitgeteilt. Ein Wert 56 würde z.B. in die führende Zahl 5 und die folgende Zahl 6 aufgeteilt. Ein

Fall mit dem Wert 56 würde in einer Säule mit der Beschriftung 5 (=Stem) mit dem Wert 6 (=Leaf) eingetragen. Der Stem gibt dann die Zehnerwerte, der Leaf die Einer an.

Das Stem-und-Leaf Diagramm in Tabelle 9.4 bezieht sich auf die Einkommen der Befragten. Es ist etwas schwerer zu lesen und gibt die Daten etwas ungenauer an, weil die Werte wesentlich höher sind, nämlich von 0 bis 7000 DM reichen. Deshalb werden als Stem-Werte nur ganze Tausender verwendet. Man entnimmt das der Angabe „Stem width: 1000" am Fuß der Tabelle. Jeweils am Fuß einer Säule stehen dann die Stem-Werte in der Spalte „Stem". Der erste ist 0, d.h. in dieser Säule stehen Werte mit 0 Tausendern im Wert. Da am Anfang zwei Säulen mit der Beschriftung 0 bei „Stem" stehen, sind in beiden Säulen Werte mit einer 0 auf der Tausenderstelle. Die erste enthält aber die erste Hälfte dieses Bereiches – also von 0 bis unter 500 –, die zweite die folgende – von 500 bis unter 1000. Die nächsten zwei Säulen sind mit 1 beschriftet, hier stehen die Werte von 1000 bis unter 2000 DM usw.. Jede Säule ist praktisch eine Doppelsäule. Das liegt daran, dass zumindest in einer Säule zu viele Fälle existieren, um sie der Höhe nach in einer Einzelsäule darzustellen. Je nach Bedarf wird daher von SPSS die Säulenzahl innerhalb der Stem-Weite vergrößert. Wird die Zahl der Fälle zu groß, kann auch jeder Leaf-Wert für mehrere Fälle stehen. In unserem Beispiel ist das nicht der Fall. Die Anmerkung „Each leaf: 1 case(s)" am Fuß der Säule gibt an, dass jeder Fall durch eine eigene Zahl repräsentiert ist.

Die Zahlen innerhalb der Säule, in der Spalte „Leaf", geben nun für je einen Fall die Folgezahl an. Es wird immer nur eine Ziffer angegeben. Diese hat den Wert der Stelle, die nach der dem Wert der Stelle von „Stem" folgt. Da unsere Führungszahl (Stem) Tausenderwerte angibt, sind es bei der Folgezahl (Leaf) Hunderterwerte. Betrachten wir jetzt die erste Säule mit dem Stem 0, so geben die ersten zwei Ziffern 1 an, dass jeweils ein Fall mit einem Einkommen von 100 DM existiert (Zehner und Einer werden nicht ausgewiesen, daher kann der wahre Wert zwischen 100 und unter 200 DM liegen). Es folgt ein Fall mit einem Einkommen von DM 300. Die Zahl der Fälle ist zusätzlich in der Spalte „Frequency" mit 3 angegeben. So ist jeder Fall rekonstruierbar enthalten. Die letzte Säule z.B. enthält 6 Fälle. Davon haben 5 den Wert 4000 DM und einer den Wert 4200 DM.

Extremwerte werden in diesem Diagramm gesondert behandelt. Ihr Wert wird in einer letzten Reihe in Klammern in Klarform (nicht in Stem-und-Leaf-Aufgliederung) angegeben. Im Beispiel sind es acht Fälle, mit Werte >=4300 DM. Das Kriterium für die Klassifikation als Extremwert entspricht der des Boxplots (⇨ unten).

Bei der Verwendung von Stem-and-Leaf Plots sollte man weiter beachten, dass Kategorien ohne Fälle nicht angezeigt werden. Die Verteilung muss also zunächst sorgfältig nach möglichen Lücken inspiziert werden. Dieses Diagramm eignet sich besonders für kontinuierliche metrische Daten. Liegen diskontinuierliche Daten vor, steht eine Säule für den jeweils vorhandenen Wert. Eine Reihe von leeren Säulen, die zusätzlich beschriftet sind, geben den leeren Bereich zwischen den einzelnen Säulen wieder. Handelt es sich um nicht metrische Daten, kann man ein solches Diagramm zwar auch verwenden, sinnvoller ist in diesem Falle aber das Erstellen eines Balkendiagrammes.

9.2.2 Boxplot

Boxplots werden im Kap. 26.9 ausführlich erläutert. Deshalb geben wir hier nur einen kurzen Überblick über ihre Anwendung.

Der Boxplot jeder Gruppe enthält in der Mitte einen schwarz oder farbig ausgefüllten Kasten (Box). Er gibt den Bereich zwischen dem ersten und dem dritten Quartil an (also den Bereich, in dem die mittleren 50 % der Fälle der Verteilung liegen). Die Breite dieses Kästchens (entspricht dem Interquartilbereich) gibt einen Hinweis auf die Streuung der Werte dieser Gruppe. Außerdem zeigt ein schwarzer Strich in der Mitte dieses Kästchens die Lage des Medianwertes an. Seine Lage innerhalb des Kästchens gibt einen Hinweis auf Symmetrie oder Schiefe. Liegt er in der Mitte, ist die Verteilung symmetrisch, liegt er zu einer Seite verschoben, ist sie schief.

Zusätzlich geben die Querstriche am Ende der jeweiligen Längsachse die höchsten bzw. niedrigsten beobachteten Werte an, die keine „Extremwerte" bzw. „Ausreißer" sind. Auch hier kann man gewisse Informationen über die Spannweite und über die Schiefe der Verteilung gewinnen.

Boxplots eignen sich besonders für die Identifikation von Ausreißern und Extremwerten:

- *Ausreißer* (Outliers) sind Werte, die zwischen 1,5 und 3 Boxenlängen vom oberen Quartilswert nach oben bzw. vom unteren Quartilswert nach unten abweichen. Sie werden durch einen kleinen Kreis ○ gekennzeichnet.
- *Extremwerte* sind Werte, die mehr als drei Boxenlängen vom oberen Quartilswert nach oben bzw. vom unteren Quartilswert nach unten abweichen. Sie werden mit ✶ gekennzeichnet.

9.3 Überprüfen von Verteilungsannahmen

Viele statistische Tests beruhen auf Modellen, die gewisse Annahmen über die Verteilung(en) in der Grundgesamtheit voraussetzen. Darunter sind die wichtigsten die Annahme einer Normalverteilung der Werte in der Grundgesamtheit und der Homogenität (Gleichheit) der Varianzen in Vergleichsgruppen. SPSS stellt in mehreren Programmteilen Tests für diese beiden Annahmen zur Verfügung. Im Menü „Explorative Datenanalyse" werden zusätzlich für beide Zwecke Grafiken und Tests angeboten. Eine Überprüfung sollte vor Anwendung statistischer Verfahren, die auf solchen Voraussetzungen basieren, durchgeführt werden. Allerdings geben diese Hilfsmittel nur ungefähre Orientierungen, denn die Tests erweisen sich als in unterschiedlichem Maße robust gegenüber Verletzungen der Annahmen. Darüber, welches Ausmaß der Abweichung noch hinzunehmen ist, gibt es aber nur vage Vorstellungen, die angeführten Hilfsmittel können allenfalls entscheidungsunterstützend wirken.

9.3.1 Überprüfen der Voraussetzung homogener Varianzen

Um die Voraussetzung der Homogenität (Gleichheit) der Varianzen von Vergleichsgruppen zu überprüfen, kann man im Menü „Explorative Datenanalyse" zweierlei benutzen:

- *Levene-Test.* Es handelt sich um eine besondere Variante des F-Tests zur Überprüfung der Homogenität von Varianzen. Er wird von SPSS im Rahmen mehrerer Menüs angeboten (⇨ u.a. Kap. 13.4.2 und Kap. 14.2).
- *Streuung über Zentralwertdiagramm* (Streubreite vs. mittleres Niveau). Es handelt sich um zwei Grafikarten, die es erlauben zu überprüfen, inwieweit die Varianz einer Variablen von der Größe der betrachteten Werte abhängt.

Ist die Voraussetzung der Homogenität der Varianz verletzt, kann dies durch Datentransformation evtl. geheilt werden. Streuung gegen Zentralwert-Plots unterstützen auch die Auswahl von Transformationsformeln. Diese können innerhalb des Menüs „Explorative Datenanalyse" auf ihre Wirkung geprüft werden.

Der Levene-Test. Untersucht man den Zusammenhang zwischen einer kategorialen unabhängigen Variablen und einer metrischen abhängigen, wird bei vielen statistischen Tests vorausgesetzt, dass die Varianz der Werte der metrischen Skala in den Gruppen der unabhängigen Variablen in etwa gleich ist. Der Levene-Test ist ein Test auf Homogenität der Varianzen, der gegenüber anderen Tests den Vorteil hat, nicht selbst von der Voraussetzung einer Normalverteilung in der Grundgesamtheit abzuhängen. Bei Durchführung des Levene-Test wird für jeden einzelnen Fall die absolute Abweichung vom Gruppenmittelwert gebildet. Dann wird eine Einweg-Varianzanalyse der Varianz dieser Differenzen durchgeführt. Sollte die Nullhypothese gelten, dürfte sich die Variation innerhalb der Gruppen von der zwischen den Gruppen nicht signifikant unterscheiden. Der klassische Levene-Test geht von der Abweichung der einzelnen Fälle vom arithmetischen Mittel aus. SPSS bietet jetzt auch drei weitere Varianten an: „Basiert auf dem Median", „Basierend auf dem Median und mit angepassten df", „Basiert auf dem getrimmten Mittel". Diese Levene-Tests sind robuster, da die zugrunde liegenden Lagemaße selbst robuster, also weniger anfällig für die Wirkung von Ausreißern und Extremwerten sind als das arithmetische Mittel.

Beispiel. Eine solche Analyse soll für das Einkommen nach Schulabschlüssen (Datei: ALLBUS90.SAV) durchgeführt werden. Zur Vorbereitung ist die Ursprungsvariable SCHUL etwas verändert und als SCHUL2 abgespeichert worden. Der Wert für Personen, die noch Schüler sind, wurden als Missing-Wert deklariert. Personen ohne Hauptschulabschluss wurden durch „Umkodieren" mit den Personen mit Hauptschulabschluss zusammengefasst, ebenso Fachoberschulabsolventen und Abiturienten.

Um einen „Levene-Test" und ein „Streuung über Zentralwertdiagramm" aufzurufen, gehen Sie wie folgt vor:

▷ Übertragen Sie in der Dialogbox „Explorative Datenanalyse" die abhängige Variable (hier: EINK) in das Feld „Abhängige Variablen:" und die unabhängige (hier: SCHUL2) in das Feld „Faktorenliste:".

- ▷ Wenn nur eine Grafik gewünscht wird: Klicken Sie in der Gruppe „Anzeigen" auf „Diagramme".
- ▷ Klicken Sie auf die Schaltfläche „Diagramme...". Die Dialogbox „Explorative Datenanalyse: Diagramme" öffnet sich (⇨ Abb. 9.4).
- ▷ Klicken Sie in der Gruppe „Streuungsbreite vs. mittleres Niveau mit Levene-Test" auf die Optionsschaltfläche „Nicht transformiert".
- ▷ Bestätigen Sie mit „Weiter" und „OK".

Tabelle 9.5 zeigt den Output des Levene-Tests für unser Beispiel. Sein Ergebnis ist in allen vier Varianten, dass sich die Varianz der Gruppen nicht signifikant unterscheidet. Die Wahrscheinlichkeit dafür, dass beide Gruppen aus ein und derselben Grundgesamtheit stammen könnten, ist z.B. mit 0,1970 (Spalte „Signifikanz") beim klassischen Test noch so hoch (noch höher bei den anderen Varianten), dass man die Annahme gleicher Varianz nicht verwerfen kann. Je nach vorher festgelegtem Signifikanzniveau würde man erst ab einem Wert von 0,05 und niedriger bzw. 0,01 und niedriger die Annahme ablehnen, dass beide Gruppen dieselbe Varianz haben. Demnach könnte man also statistische Verfahren anwenden, die Homogenität der Varianz voraussetzen.

Tabelle 9.5. Ausgabe des Levene-Tests auf Homogenität der Varianz von Schulbildungsgruppen

Test auf Homogenität der Varianz

		Levene-Statistik	df1	df2	Signifikanz
BEFR.: MONATLICHES NETTOEINKOMMEN	Basiert auf dem Mittelwert	1,643	2	139	,197
	Basiert auf dem Median	1,244	2	139	,291
	Basierend auf dem Median und mit angepaßten df	1,244	2	136,819	,291
	Basiert auf dem getrimmten Mittel	1,540	2	139	,218

Streubreite vs. mittleres Niveau (Streuung über Zentralwertdiagramm). Ergänzend betrachten wir das „Streuung über Zentralwertdiagramm" (⇨ Abb. 9.5). Auf der Abszisse ist der Zentralwert abgetragen, auf der Ordinate die Streuung (ermittelt als Interquartilsbereich). Mit den drei Punkten werden die Einkommen der drei Schulbildungsgruppen abgebildet. So liegt bei der ersten Gruppe, den „Hauptschülern", der Zentralwert etwa bei 1700, die Streuung bei ca. 1000, bei der zweiten Gruppe, den „Abiturienten/Fachoberschulabsolventen", ist sowohl der Medianwert mit 2100 als auch die Streuung mit 2300 deutlich höher. Bei den „Mittelschulabsolventen" ist der Medianwert am höchsten, die Streuung liegt im mittleren Bereich. Ideal wäre es, wenn die Streuungen gleich wären. Dann würden die Linien auf einer Geraden, parallel zur x-Achse liegen. Dies ist ersichtlich nicht der Fall. Trotzdem besteht, wie oben festgestellt, keine signifikante Differenz zwischen den Streuungen der Gruppen.

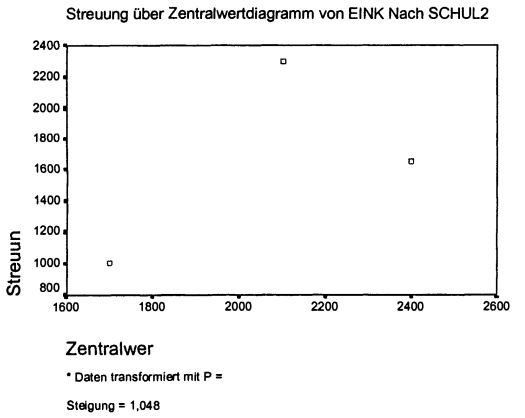


Abb. 9.5. „Streuung über Zentralwertdiagramm" für die Variable Einkommen, gruppiert nach Schulbildung"

Datentransformation. Man könnte versuchen, durch Datentransformation das Kriterium der Homogenität der Varianzen noch besser zu erreichen. Wir haben bisher für den Levene-Test und das „Streuung über Zentralwertdiagramm" die nicht transformierten Daten verwendet. Das Programm bietet aber Möglichkeiten zur Datentransformation an:

- ❒ *„Exponentenschätzung"* (Power Estimation). Trägt den natürlichen Logarithmus des Medianwertes gegen den natürlichen Logarithmus des Interquartilsbereichs ab.
- ❒ *„Transformiert"*. Es können unterschiedliche Transformationsformeln benutzt werden.

▷ Klicken Sie zuerst die Optionsschaltfläche „Transformiert" an (⇨ Abb. 9.4).
▷ Klicken Sie dann auf den Pfeil am rechten Rand des Auswahlkästchens. Es öffnet sich eine Drop-Down-Auswahlliste mit den verfügbaren Transformationsfunktionen.

Bei der Auswahl einer geeigneten Transformationsfunktion kann man sich nach folgender Formel richten:

$$\text{Power} = 1 - \text{Steigung}(\text{"Slope"}) \tag{9.5}$$

Dabei ist Power der Exponent der Transformationsfunktion und Steigung die Steigung einer durch die Punkte des „Streuung über Zentralwertdiagramms" (aus nicht transformierten Daten) gelegten Regressionsgerade. Die Angabe dieser Steigung finden wir unter der Bezeichnung „Steigung" in der letzten Zeile des Streuung über Zentralwertdiagramm. In unserem Beispiel beträgt die Steigung 1,048. Entsprechend können wir als geeignete Power berechnen:

$$\text{Power} = 1 - 1{,}048 = -0{,}048$$

Man verwendet den nächstgelegenen Wert aus der Auswahlliste. Dabei gelten folgende Entsprechungen:

Exponent (Power)	Transformation
3	Kubisch
2	Quadratisch
1	Untransformiert
1/2	Quadratwurzel
0	Natürl. Log.
−1/2	1 / Quadratwurzel
−1	Reziprok

Der im Beispiel gefundene Wert liegt nahe 0. Eine geeignete Transformation wäre daher die Bildung des natürlichen Logarithmus.

Tabelle 9.6. Ausgabe des Levene-Tests mit transformierten Daten

Test auf Homogenität der Varianz

		Levene-Statistik	df1	df2	Signifikanz
BEFR.: MONATLICHES NETTOEINKOMMEN	Basiert auf dem Mittelwert	,751	2	139	,474
	Basiert auf dem Median	,593	2	139	,554
	Basierend auf dem Median und mit angepaßten df	,593	2	112,776	,554
	Basiert auf dem getrimmten Mittel	,666	2	139	,516

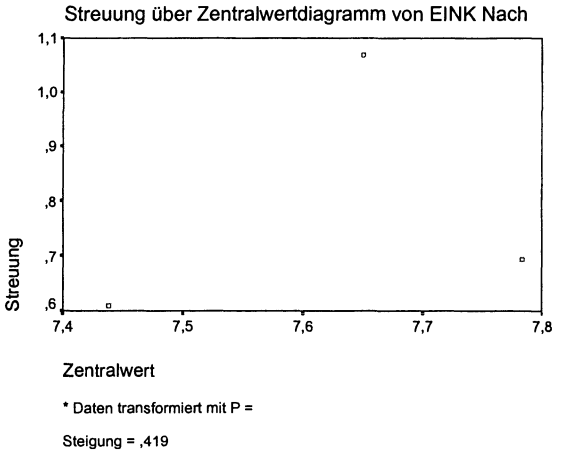


Abb. 9.6. Streuung über Zentralwertdiagramm für „Einkommen" nach „Schulbildung" mit transformierten Daten

▷ Markieren Sie die gewünschte Transformationsfunktion.
▷ Bestätigen Sie mit „Weiter" und „OK".

Das Ergebnis sind ein veränderter Levene-Test und ein verändertes „Streuung über Zentralwertdiagramm" (⇨ Tabelle 9.6 und Abb. 9.6).

Beide Ergebnisse zeigen jetzt verbesserte Befunde. Der Levene-Tests erweisen jetzt die Unterschiede der Varianzen als noch weniger signifikant. Die Steigung der Regressionsgerade im „Streuung über Zentralwertdiagramm" ist geringer. Das kann man gut an dem Wert Steigung von 0,419 sehen. In der Grafik ist es auf den ersten Blick weniger ersichtlich, wird aber deutlich, wenn man beachtet, dass jetzt die Achsen anders skaliert sind als in der ersten Grafik mit den nicht transformierten Werten.

Hinweis. Würde man dieselbe Prozedur für die Gruppen Männer und Frauen durchführen, wären die Ergebnisse anders. Die Varianz der Einkommen dieser beiden Gruppen unterscheidet sich signifikant. Das würde dort ebenfalls durch Logarithmierung geheilt. Trotzdem werden wir im folgenden weiter mit den Rohdaten arbeiten. Dafür spricht, dass die Ergebnisse dann anschaulicher bleiben. Außerdem hat eine Überprüfung ergeben, dass die meisten Ergebnisse kaum von denjenigen abweichen, die bei Verwendung transformierter Daten entstünden. Trotz Verletzung der Voraussetzung der Homogenität der Varianzen, sind die Verfahren insgesamt robust genug, dass sich dies nicht entscheidend auf die Ergebnisse auswirkt.

9.3.2 Überprüfen der Voraussetzung der Normalverteilung

Für viele statistische Tests ist auch die Normalverteilung der Daten in der Grundgesamtheit vorauszusetzen. Deshalb muss dieses vor Anwendung solcher Tests überprüft werden. Glücklicherweise sind die meisten Tests relativ robust, so dass mehr oder weniger große Abweichungen von der Normalverteilungsannahme hingenommen werden können. Von zentraler Bedeutung ist meistens nicht die Normalverteilung der Werte in der Grundgesamtheit, sondern die Normalverteilung der Stichprobenverteilung, also derjenigen Verteilung, die entstünde, wenn unendlich viele Stichproben gezogen würden. Diese ist zumindest näherungsweise auch bei relativ groben Abweichungen der Grundgesamtheitswerte von der Normalverteilung noch gegeben. Normalverteilung der Stichprobenwerte ist z.B. auch dann noch gegeben, wenn bei nicht zu kleinem Stichprobenumfang eine uniforme Verteilung der Werte in der Grundgesamtheit vorliegt, also in alle Kategorien gleich viele Werte fallen. Sehr grobe Abweichungen, insbesondere mehrgipflige und extrem schiefe Verteilungen können dagegen nicht mehr akzeptiert werden.

Das Menü „Explorative Datenanalyse" stellt zur Überprüfung der Voraussetzung der Normalverteilung der Werte in der Grundgesamtheit zwei Hilfsmittel zur Verfügung:

- *Normalverteilungsdiagramm* (Q-Q-Diagramm) und *Trendbereinigtes Normalverteilungsdiagramm* (Trendbereinigtes Q-Q-Diagramm).
- *Kolmogorov-Smirnov* und *Shapiro-Wilk-Test.*

Auch in anderen Menüs sind Prüfungshilfsmittel verfügbar. Zu denken ist insbesondere an das durch eine Normalverteilung überlagerte Histogramm, das man im Menü „Häufigkeiten" bzw. im Menü „Grafiken" erstellen kann. Leider gibt es keine eindeutigen Kriterien dafür, ab wann die Anwendungsvoraussetzungen für einen Test nicht mehr gegeben sind, der eine Normalverteilung voraussetzt. Der Anwender ist daher stark auf sein eigenes Urteil und seine Erfahrung angewiesen.

Hilfreich sind insbesondere die Grafiken. Die beiden Tests dagegen sind kaum brauchbar.

Normalverteilungsplots und trendbereinigte Normalverteilungsplots. Der „Normalverteilungsplot" ist eine Grafik, bei der die beobachteten Werte gegen die bei einer Normalverteilung zu erwartenden Werte in einem Achsenkreuz abgetragen werden. Die Skala für die beobachteten Werte ist auf der Abszisse, die der erwarteten Werte auf der Ordinate abgetragen. Ist eine Normalverteilung gegeben, müssen die Punkte dieser Verteilung auf einer Geraden liegen, die diagonal vom Nullpunkt ausgehend nach oben verläuft.

Beim „Trendbereinigten Normalverteilungsplot" werden dagegen die Abweichung der beobachteten Werte von der Normalverteilungslinie grafisch dargestellt. Auf der Abszisse ist die Skala der beobachteten Werte, auf der Ordinate diejenige der Abweichungen abgetragen. Die Punktewolke sollte zufällig um eine horizontale Gerade durch den Nullpunkt streuen. Zufällig heißt, dass keine Struktur erkennbar ist. Um die beiden Diagramme zu erstellen, gehen Sie wie folgt vor:

- ▷ Übertragen Sie in der Dialogbox „Explorative Datenanalyse" die gewünschte Variable in das Eingabefeld „Abhängige Variablen:" (hier: EINK).
- ▷ Klicken Sie in der Gruppe „Anzeigen" auf die Optionsschaltfläche „Diagramme".
- ▷ Klicken Sie auf die Schaltfläche „Diagramme...". Die Dialogbox „Explorative Datenanalyse: Diagramme" öffnet sich (⇨ Abb. 9.4).
- ▷ Klicken Sie auf das Auswahlkästchen „Normalverteilungsdiagramm mit Tests".
- ▷ Schalten Sie gegebenenfalls alle anderen ausgewählten Plots aus.
- ▷ Bestätigen Sie mit „Weiter" und „OK".

Da Normalverteilungs-Plots im Menü „Grafiken" erläutert werden, kann hier auf eine Darstellung und Erläuterung der zwei erzeugten Grafiken „Q-Q Diagramm" und „Trendbereinigtes Q-Q Diagramm" verzichtet werden (⇨ Kap. 26.13).

Normalverteilungstests. Wenn Sie die Option „Normalverteilungsdiagramm mit Tests" verwenden, werden zusammen mit den beiden Grafiken auch zwei Normalverteilungstests ausgegeben:

- ❒ *Shapiro-Wilk-Test.* Er sollte bei Stichprobengrößen unter 50 verwendet werden. Gegenüber vergleichbaren Tests zeichnet er sich durch gute Teststärke aus.
- ❒ *Kolmogorov-Smirnov.* Ist eine Kolgomorov-Smirnov Statistik, die für den Test der Normalitätsvoraussetzung spezielle Signifikanzlevels nach Lilliefors benutzt.

Da in unserer Beispielstichprobe mehr als 50 Fälle enthalten sind, sind hier die Ergebnis des Kolmogorov-Smirnov-Tests adäquat.

Tests auf Normalverteilung

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
EINK	,128	143	,000	,923	143	,000

a. Signifikanzkorrektur nach Lilliefors

Für die Interpretation entscheidend ist die Spalte „Signifikanz". Da hier nur Nullen enthalten sind, ist klar, dass die beobachtete Verteilung mit an Sicherheit grenzender Wahrscheinlichkeit *nicht* aus einer normalverteilten Grundgesamtheit stammt. Die Normalverteilungsannahme kann also nicht bestätigt werden. Dies würde dafür sprechen, dass Tests, die Normalverteilung der Werte in der Grundgesamtheit voraussetzen, nicht angewendet werden sollen.

Man muss bei der Entscheidung aber bedenken, dass ein Normalverteilungstest wenig hilfreich ist. Dies liegt daran, dass eine Nullhypothese überprüft wird. Man müsste hier nicht α, sondern β zur Bestimmung des Signifikanzniveaus benutzen. Tut man das nicht – und dies ist bei einem Test von Punkt- gegen Bereichshypothesen nicht möglich – führt das zu dem paradoxen Ergebnis, dass die zu prüfende Hypothese umso eher bestätigt wird, je kleiner die Stichprobengröße n ist (⇨ Kap. 13.3). Es muss daher von allzu schematischer Anwendung der Normalverteilungstests abgeraten werden. Im Prinzip wären für die Klärung der Fragestellung Zusammenhangsmaße, die den Grad der Übereinstimmung mit einer Normalverteilung ausdrücken, geeigneter. Noch günstiger wäre es, wenn Maßzahlen entwickelt werden könnten, die Grenzfälle noch akzeptabler Verteilungen zugrunde legten. Dies steht aber bislang nicht zur Verfügung.

Optionen. Beim Anklicken der Schaltfläche „Optionen..." in der Dialogbox „Explorative Datenanalyse" öffnet sich die Dialogbox „Explorative Datenanalyse: Optionen". Hier kann die Behandlung fehlender Daten beeinflusst werden. Diese können entweder listenweise oder paarweise aus der Berechnung ausgeschlossen werden. Beim Befehl „Werte einbeziehen" werden Berechnungen und Diagramme auch für die Gruppen der fehlenden Werte der Faktorvariablen (unabhängigen Variablen) erstellt. Diese Gruppen wird (nicht immer) mit der Beschriftung „Fehlend" gekennzeichnet.

Weitere Möglichkeiten bei Verwenden der Befehlssyntax.

- Es können mehrere kategoriale unabhängige Variablen kombiniert werden.
- Mit dem Unterkommando STATISTICS kann die Zahl der ausgewiesenen Extremwerte verändert werden.
- Mit dem Unterkommando PERCENTILES können alternative Berechnungsarten für die Berechnung der Perzentile gewählt werden.
- Mit dem Unterkommando PLOT können beliebige Transformationen der Werte für die Streuung gegen Zentralwert-Plots festgelegt werden.
- Mit dem Unterbefehl MSTIMATOR können die kritischen Punkte der verschiedenen Maximum-Likelihood-Schätzer verändert werden.
- Mit dem Unterkommando MISSING können mit dem Befehl INCLUDE die nutzerdefinierten fehlenden Werte in die Berechnung einbezogen werden (die systemdefinierten bleiben ausgeschlossen).

10 Kreuztabellen und Zusammenhangsmaße

Zusammenhänge zwischen zwei kategorialen Variablen können am einfachsten in Form einer Kreuztabelle dargestellt werden. Durch die Einführung von Kontrollvariablen ist es möglich, dies auf drei- und mehrdimensionale Zusammenhänge auszudehnen. SPSS bietet dazu das Untermenü „Kreuztabellen" an. Bei einer größeren Zahl von Variablenwerten werden Kreuztabellen leicht unübersichtlich. Deshalb bevorzugt man oft die Darstellung von Zusammenhängen durch ein einziges Zusammenhangsmaß. Das Menü „Kreuztabellen" ermöglicht die Berechnung einer Reihe von Zusammenhangsmaßen für Daten unterschiedlichen Messniveaus. Zudem bietet es verschiedene Varianten des Chi-Quadrat-Tests für die Überprüfung der Signifikanz von Zusammenhängen zwischen zwei Variablen an.

10.1 Erstellen einer Kreuztabelle

Im folgenden Beispiel soll festgestellt werden, ob die Einstellung auf der Inglehartschen „Materialismus-Postmaterialismus"-Skala von der Schulbildung der Befragten abhängt (Datei: ALLBUS90.SAV). Dazu muss eine Kreuztabelle mit der in Kap. 9.3.1 gebildeten Schulbildungsvariablen (SCHUL2) als unabhängiger und der Variablen Inglehartindex (INGL) als abhängiger Variable gebildet werden. (Die Bildung der Variablen INGL aus den Variablen RUHE, EINFLUSS, INFLATIO und MEINUNG wurde in Kap. 2.6 geschildert.)

Zum Erstellen einer Kreuztabelle gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Deskriptive Statistiken ▷", „Kreuztabellen...". Es öffnet sich die Dialogbox „Kreuztabellen" (⇨ Abb. 10.1).
- ▷ Wählen Sie aus der Variablenliste die Zeilenvariable aus, und übertragen Sie diese in das Feld „Zeilen:".
- ▷ Übertragen Sie aus der Quellvariablenliste die Spaltenvariable in das Feld „Spalten:".

In der so erzeugten Tabelle wird die in das Feld „Zeilen:" ausgewählte Variable in der Vorspalte stehen und ihre Werte werden die Zeilen bilden. Die im Feld „Spalten:" ausgewählte Variable wird im Kopf der Tabelle stehen, ihre Werte werden die Spalten bilden. (Es können mehrere Zeilen- und Spaltenvariablen ausgewählt werden. Zwischen allen ausgewählten Zeilen- und Spaltenvariablen werden dann zweidimensionale Tabellen gebildet.)

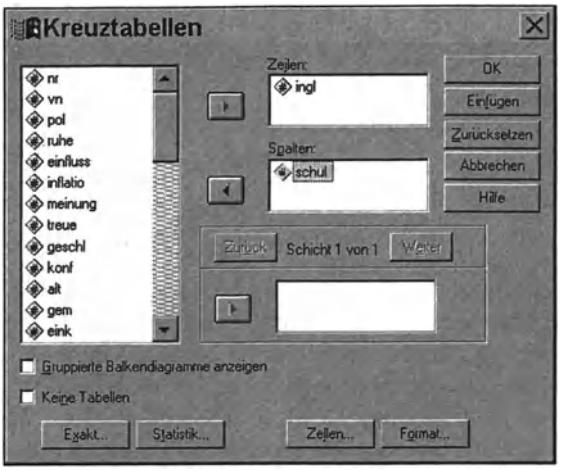


Abb. 10.1. Dialogbox „Kreuztabellen"

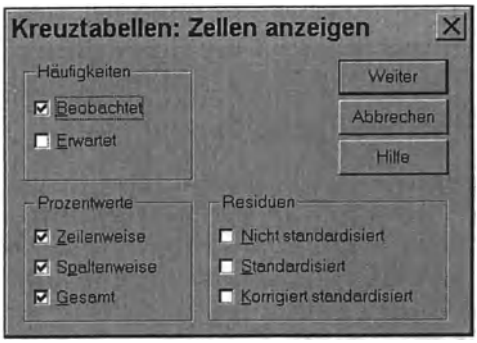


Abb. 10.2. Dialogbox „Kreuztabellen: Zellen anzeigen"

▷ Klicken Sie auf die Schaltfläche „Zellen...". Es öffnet sich die Dialogbox „Kreuztabellen: Zellen anzeigen" (⇨ Abb. 10.2). Diese enthält drei Auswahlgruppen:

❒ *Häufigkeiten.*
- *Beobachtet* (Voreinstellung). Gibt in der Kreuztabelle die Anzahl der tatsächlich beobachteten Fälle an.
- *Erwartet.* Gibt in der Kreuztabelle die Anzahl der Werte an, die erwartet würden, wenn kein Zusammenhang zwischen den beiden Variablen bestünde, wenn sie also voneinander unabhängig wären. Das ist interessant im Zusammenhang mit dem Chi-Quadrat-Test (⇨ Kap. 10.2).

❒ *Prozentwerte.* In dieser Gruppe wird festgelegt, ob in der Kreuztabelle eine Prozentuierung vorgenommen und in welcher Weise diese durchgeführt wird:

- *Zeilenweise.* Zeilenweise Prozentuierung. Die Fälle in den Zellen werden als Prozentanteile an den Fällen der zugehörigen Zeile ausgedrückt.
- *Spaltenweise.* Spaltenweise Prozentuierung. Die Fälle in den Zellen werden als Prozentanteile an den Fällen der zugehörigen Spalte ausgedrückt.
- *Gesamt.* Die Fälle in den Zellen werden als Prozentanteile an allen Fällen ausgedrückt.

Die Richtung der Prozentuierung muss je nach Fragestellung und Art der Aufbereitung der Daten bestimmt werden. In der Regel setzt das eine Entscheidung darüber voraus, welche Variable die „unabhängige Variable" sein soll und welche die „abhängige". Von der unabhängigen wird angenommen, dass sie einen ursächlichen Effekt auf die abhängige hat. Ist das der Fall, sollen die verschiedenen Ausprägungen der unabhängigen Variablen hinsichtlich der Verteilung der Werte auf der abhängigen verglichen werden. Entsprechend wird die Gesamtzahl der Fälle jedes Wertes der unabhängigen Variablen gleich 100 % gesetzt. Dementsprechend prozentuiert man spaltenweise, wenn die unabhängige Variable die Spaltenvariable und zeilenweise, wenn sie die Zeilenvariable ist. Prozentuierung auf Basis der Gesamtzahl der Fälle kommt nur für spezielle Zwecke in Frage, etwa, wenn zweidimensionale Typen gebildet werden sollen oder wenn es um Veränderungen zwischen zwei Zeitpunkten geht.

❒ *Residuen.* Diese Auswahlbox betrifft wiederum Zwischenergebnisse des Chi-Quadrat-Tests.

- *Nicht standardisiert.* Die Differenzen zwischen beobachteten Werten und Erwartungswerten werden als Absolutbeträge angegeben.
- *Standardisiert.* Diese Differenzen werden als standardisierte Werte angegeben.
- *Korrigiert standardisiert.* Diese Differenzen werden in der Tabellenausgabe als korrigierte Residuen bezeichnet.

▷ Wählen Sie die gewünschte(n) Prozentuierung(en).
▷ Wählen Sie gegebenenfalls „Häufigkeiten" und „Residuen" aus.
▷ Bestätigen Sie die Auswahl mit „Weiter" und „OK".

Die in Abb. 10.1 und 10.2 dargestellten Einstellungen führen bei den Beispielsdaten zu Tabelle 10.1.

Die Tabelle enthält in ihrem Kopf die unabhängige Variable Schulbildung. Sie ist hier als Spaltenvariable benutzt. Ihre Werte bilden die Spaltenüberschriften. Die abhängige Variable „Inglehart-Index" bildet die Zeilenvariable. Ihre vier Werte stehen zur Beschriftung der Zeilen in der Vorspalte. Da die unabhängige Variable drei und die abhängige vier Kategorien besitzt, ergibt die Kombination eine 3*4-Tabelle. Die Tabelle hat zwölf Zellen. In jeder stehen die Werte für eine der Wertekombinationen beider Variablen.

Da als Eintrag die beobachteten Werte und alle drei Prozentuierungsarten gewählt wurden, stehen in jeder Zelle vier Werte. Um welche es sich handelt, zeigen die Eintragungen am Ende der Vorspalte. Der erste Wert („Anzahl") gibt die Zahl

der Fälle mit dieser Wertekombination an. So gilt für 16 Befragte die Kombination Hauptschulabschluss/Postmaterialisten.

Tabelle 10.1. Kreuztabelle „Inglehart-Index" nach „Schulbildung"

INGL * SCHUL2 Kreuztabelle

			SCHUL2			
			Hauptschule	Mittelschule	Fachh/Abi	Gesamt
INGL	POSTMATERIALISTEN	Anzahl	16	24	40	80
		% von INGL	20,0%	30,0%	50,0%	100,0%
		% von SCHUL2	11,2%	32,0%	58,0%	27,9%
		% der Gesamtzahl	5,6%	8,4%	13,9%	27,9%
	PM-MISCHTYP	Anzahl	36	23	15	74
		% von INGL	48,6%	31,1%	20,3%	100,0%
		% von SCHUL2	25,2%	30,7%	21,7%	25,8%
		% der Gesamtzahl	12,5%	8,0%	5,2%	25,8%
	M-MISCHTYP	Anzahl	56	22	13	91
		% von INGL	61,5%	24,2%	14,3%	100,0%
		% von SCHUL2	39,2%	29,3%	18,8%	31,7%
		% der Gesamtzahl	19,5%	7,7%	4,5%	31,7%
	MATERIALISTEN	Anzahl	35	6	1	42
		% von INGL	83,3%	14,3%	2,4%	100,0%
		% von SCHUL2	24,5%	8,0%	1,4%	14,6%
		% der Gesamtzahl	12,2%	2,1%	,3%	14,6%
Gesamt		Anzahl	143	75	69	287
		% von INGL	49,8%	26,1%	24,0%	100,0%
		% von SCHUL2	100,0%	100,0%	100,0%	100,0%
		% der Gesamtzahl	49,8%	26,1%	24,0%	100,0%

Die zweite Zahl („% von INGL") ist ein Reihenprozentwert. Er gibt an, wieviel Prozent die Fälle dieser Zelle an allen Fällen der dazugehörigen Reihe ausmachen. Die 16 Hauptschüler/Postmaterialisten sind z.B. 20 % der insgesamt 80 Postmaterialisten.

Die Spaltenprozente (hier: % von SCHUL2) folgen als Drittes. Sie geben an, wieviel Prozent die Fälle dieser Zelle an allen Fällen der dazugehörigen Spalte ausmachen. Die 16 genannten Fälle sind z.B. 11,2 % aller 143 Hauptschüler.

Schließlich geben die Gesamtprozentwerte („% der Gesamtzahl") an, welchen Prozentanteil die Fälle dieser Zelle an allen Fällen ausmachen. Die 16 postmaterialistisch eingestellten Hauptschüler sind 5,6 % aller 287 gültigen Fälle.

Angebracht ist in unserem Beispiel lediglich die spaltenweise Prozentuierung. Es geht ja darum festzustellen, ob unterschiedliche Schulbildung auch unterschiedliche Einstellung auf der „Materialismus-Postmaterialismus"-Dimension nach sich zieht. Das ist nur ersichtlich, wenn die verschiedenen Bildungsgruppen vergleichbar gemacht werden. Vergleichen wir entsprechend nur die dritten Zahlen in den jeweiligen Zellen. Dann zeigen sich recht eindeutige Trends. Von den Hauptschülern sind 11,2 % als Postmaterialisten eingestuft, von den Mittelschülern dagegen 32,0 % und von den Personen mit Abitur/Fachhochschulreife sogar 58,0 % usw.. Solche Unterschiede sprechen deutlich dafür, dass die unabhängige Variable einen Einfluss auf die abhängige Variable besitzt.

Die Tabelle zeigt weiter am rechten und am unteren Rand sowohl die absoluten Häufigkeiten als auch die Prozentwerte an, die sich ergeben würden, wenn die beiden Variablen für sich alleine ausgezählt würden. Am rechten Rand ist die Verteilung auf der abhängigen Variablen „Inglehart-Index" angegeben, am unteren die Verteilung nach Schulbildung. Man spricht hier auch von den Randverteilungen der Tabelle oder Marginals. Sie kann für verschiedene Zwecke interessant sein, u.a. ist sie Ausgangspunkt zur Kalkulation der Erwartungswerte für den Chi-Quadrat-Test.

Hinzufügen einer Kontrollvariablen. In den Sozialwissenschaften haben wir es in der Regel mit wesentlich komplexeren als zweidimensionalen Beziehungen zu tun. Es wird auch nur in Ausnahmefällen gelingen, den Einfluss weiterer Variablen von vornherein auszuschalten oder unter Kontrolle zu halten. Ist das nicht der Fall, kann das Ergebnis einer zweidimensionalen Tabelle möglicherweise in die Irre führen. Die Einflüsse weiterer Variablen können die wirkliche Beziehung zwischen den beiden untersuchten Variablen durch Vermischung verschleiern. Ein einfacher Weg, möglichen Fehlinterpretationen vorzubeugen, aber auch die komplexere Beziehung zwischen drei und mehr Variablen zu studieren, ist die Ausweitung der Tabellenanalyse auf drei- und mehrdimensionale Tabellen. Dabei wird/werden eine oder mehrere weitere mögliche „unabhängige Variable(n)" als „Kontrollvariable(n)" in die Tabelle eingeführt. Diese Variable(n) steht/stehen dann noch oberhalb der unabhängigen Variablen. Der zweidimensionale Zusammenhang wird für die durch die Werte der Kontrollvariablen bestimmten Gruppen getrennt analysiert.

Unser Beispiel soll jetzt um die Kontrollvariable Geschlecht erweitert werden. Man kann von der Variablen Geschlecht durchaus erwarten, dass sie die Einstellung auf der Dimension „Materialismus-Postmaterialismus" beeinflusst, also eine weitere unabhängige Variable darstellt. (Eine entsprechende Tabelle bestätigt das auch, wenn auch nicht so deutlich wie bei der Schulbildung.) Außerdem besteht zwischen Geschlecht und Schulbildung ein deutlicher Zusammenhang. Deshalb wäre es z.B. durchaus denkbar, dass sich im oben festgestellten Zusammenhang zwischen Schulbildung und der Einstellung nach dem Inglehart-Index etwas anderes verbirgt, nämlich ein Zusammenhang zwischen Geschlecht und der Einstellung nach dem Inglehart-Index.

Um eine dreidimensionale Tabelle zu erstellen, gehen Sie wie folgt vor:

- Verfahren Sie zunächst wie bei der Erstellung einer zweidimensionalen Tabelle.
- Wählen Sie aber in der Dialogbox „Kreuztabellen" zusätzlich die Kontrollvariable aus, und übertragen Sie diese in das Auswahlfeld „Schicht 1 von 1". (Sie können mehrere Variablen als jeweils dritte Variable einführen. Mit jeder dieser Variablen wird dann eine dreidimensionale Tabelle erstellt. Sie können aber auch eine vierte usw. Dimension einführen, indem Sie die Schaltfläche „Weiter" anklicken. Es öffnet sich dann ein Feld zur Definition der nächsten Kontrollebene „Schicht 2 von 2" usw.. Auf eine niedrigere Ebene kann man durch Anklicken der Schaltfläche „Zurück" zurückgehen.)

▷ Ändern Sie in der Dialogbox „Kreuztabellen: Zellen anzeigen" die Einstellung so, dass nur die angemessene Prozentuierung ausgewiesen wird (hier: Spaltenprozente). Bestätigen Sie mit „Weiter" und „OK".

Wurden die angegebenen Einstellungen vorgenommen, ergibt das die in Tabelle 10.2 dargestellte Ausgabe.

Tabelle 10.2. Kreuztabelle „Inglehart-Index" nach „Schulbildung" und „Geschlecht"

INGL * SCHUL2 * GESCHL Kreuztabelle

GESCHL				SCHUL2 Hauptschule	SCHUL2 Mittelschule	SCHUL2 Fachh/Abi	Gesamt
MAENNLICH	INGL	POSTMATERIALISTEN	Anzahl	10	12	18	40
			% von SCHUL2	14,7%	35,3%	52,9%	29,4%
		PM-MISCHTYP	Anzahl	20	10	10	40
			% von SCHUL2	29,4%	29,4%	29,4%	29,4%
		M-MISCHTYP	Anzahl	26	10	6	42
			% von SCHUL2	38,2%	29,4%	17,6%	30,9%
		MATERIALISTEN	Anzahl	12	2		14
			% von SCHUL2	17,6%	5,9%		10,3%
	Gesamt		Anzahl	68	34	34	136
			% von SCHUL2	100,0%	100,0%	100,0%	100,0%
WEIBLICH	INGL	POSTMATERIALISTEN	Anzahl	6	12	22	40
			% von SCHUL2	8,0%	29,3%	62,9%	26,5%
		PM-MISCHTYP	Anzahl	16	13	5	34
			% von SCHUL2	21,3%	31,7%	14,3%	22,5%
		M-MISCHTYP	Anzahl	30	12	7	49
			% von SCHUL2	40,0%	29,3%	20,0%	32,5%
		MATERIALISTEN	Anzahl	23	4	1	28
			% von SCHUL2	30,7%	9,8%	2,9%	18,5%
	Gesamt		Anzahl	75	41	35	151
			% von SCHUL2	100,0%	100,0%	100,0%	100,0%

Wie wir sehen, wurden zwei Teiltabellen für den Zusammenhang zwischen Schulbildung und Einstellung auf der „Materialismus-Postmaterialismus"-Dimension erstellt, zuerst für die Männer, dann für die Frauen. In beiden Teiltabellen bestätigt sich der Zusammenhang zwischen Schulbildung und Einstellung. Dabei scheint dieser Zusammenhang bei Frauen noch ein wenig stärker zu sein. Generell kann man sagen:

❒ Zeigen die neuen Teiltabellen nahezu dieselben Zusammenhänge wie die alte, spricht man von Bestätigung. Verschwindet dagegen der ursprüngliche Zusammenhang, wurde eine Scheinkorrelation aufgedeckt oder es besteht eine Intervention (d.h. der direkte Einflussfaktor ist nur die Kontrollvariable. Sie wird aber selbst durch die zunächst als unabhängige Variable angenommene Variable beeinflusst). Häufig wird der Zusammenhang nicht verschwinden, aber sich in seiner Stärke verändern. Hat der Ursache-Wirkungs-Zusammenhang in allen Untertabellen die gleiche Richtung, spricht man von Multikausalität (beide unabhängigen Variablen haben eine unabhängige Wirkung), hat er dagegen in den Untertabellen unterschiedliche Richtung, spricht man von Interaktion, denn die

jeweilige Kombination der Werte der unabhängigen Variablen haben eine besondere Wirkung.

- Auch wenn eine zweidimensionale Tabelle zunächst keinen Zusammenhang erkennen lässt, kann es sein, dass bei Einführung einer Kontrollvariablen sich in den Teiltabellen ein Zusammenhang zeigt. Es wurde dann eine scheinbare Non-Korrelation aufgedeckt. Tatsächlich liegt entweder Multikausalität oder Interaktion vor.

Bestimmen des Tabellenformats. In der Dialogbox „Kreuztabellen" kann durch Anklicken der Schaltfläche „Format..." die Dialogbox „Kreuztabellen: Tabellenformat" geöffnet werden. In diesem können zwei Formatierungsoptionen für die „Zeilenfolge" gewählt werden (⇨ Abb. 10.3).

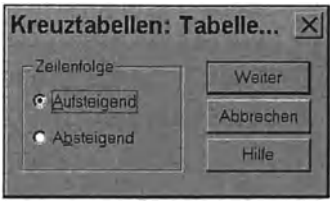


Abb. 10.3. Dialogbox „Kreuztabellen: Tabellenformat"

- *Aufsteigend* (Voreinstellung). Die Variablenwerte werden vom kleinsten Wert ausgehend nach ansteigenden Werten geordnet.
- *Absteigend.* Die Variablenwerte werden vom größten Wert ausgehend nach fallenden Werten geordnet.

Anzeigen eines Balkendiagramms. Die Dialogbox „Kreuztabellen" enthält zwei weitere Kontrollkästchen.

- *Gruppierte Balkendiagramme anzeigen.* Es wird ein Balkendiagramm für den Zusammenhang der Untersuchungsvariablen erstellt. In ihm erscheinen Kategorienkombinationen der unabhängigen und abhängigen Variablen als Balken. Deren Höhe entspricht der Anzahl der Fälle. Bei Verwendung von Kontrollvariablen, wird für jede Kategorie jeder Kontrollvariable ein eigenes Diagramm erstellt. Ein entsprechendes Diagramm können Sie auch im Menü „Grafiken" erstellen (⇨ Kap. 26.2.2)
- *Keine Tabellen.* Es werden nur zusätzlich angeforderte Statistiken und/oder Diagramme ausgegeben.

Weitere Möglichkeiten bei Verwenden der Befehlssyntax.

- Mit dem VARIABLES-Unterkommando können Variablen im Integer-Modus (nur ganzzahlige Werte) benutzt werden. Das spart Speicherplatz und erlaubt es, auch fehlende Werte in der Tabelle anzuzeigen. Es müssen dann die Variablen zuerst in einer Variablenliste angegeben und Grenzen für die einzubeziehenden Werte definiert werden.

- Die Art der Benutzung der fehlenden Werte kann mit dem „MISSING"-Unterkommando beeinflusst werden. Per Voreinstellung werden in jeder Tabelle die Fälle ausgeschlossen (Schlüsselwort: TABLE), die bei einer der in dieser Tabelle benutzten Variablen einen fehlenden Wert aufweisen. Man kann sie aber auch in die Berechnung der Prozentwerte einschließen (INCLUDE) oder sie in der Tabelle berichten lassen (REPORT), ohne dass sie bei der Berechnung der Prozentwerte und der statistischen Maßzahlen berücksichtigt werden. Letzteres ist nur im Integer-Modus möglich. Fallen die fehlenden Werte nicht in die definierten Grenzen der Integer-Variablen, sind sie unabhängig vom Schlüsselwort von der Berechnung ausgeschlossen.
- Mit dem Unterbefehl WRITE können Tabellen im ASCII-Format geschrieben werden.

10.2 Der Chi-Quadrat-Unabhängigkeitstest

Theoretische Grundlagen. In den meisten Fällen entstammen unsere Daten keiner Vollerhebung, sondern nur ein Teil der Zielpopulation wurde untersucht (Teilerhebung). Ist das der Fall, kann nicht ohne weitere Prüfung ein in einer Tabelle erkannter Zusammenhang zwischen zwei Variablen als gesichert gelten. Er könnte in der Grundgesamtheit gar nicht existieren und lediglich durch Auswahlverzerrungen vorgetäuscht werden. Falls die Teilpopulation durch Zufallsauswahl zustande gekommen ist (Zufallsstichprobe), kann eine weitgehende Absicherung vor zufallsbedingten Ergebnissen mit Hilfe von Signifikanztests erfolgen (⇨ Kap. 13.3). Das Menü „Kreuztabellen" bietet dazu den Chi-Quadrat-Test an, der geeignet ist, wenn zwei oder mehr unabhängige Stichproben vorliegen und die abhängige Variable auf Nominalskalenniveau gemessen wurde.

Ein Chi-Quadrat-Test zur Überprüfung der statistischen Signifikanz von Zusammenhängen zwischen zwei Variablen geht im Prinzip wie folgt vor:

- Die Nullhypothese (H_0, die Annahme es bestehe keine Beziehung zwischen den untersuchten Variablen) wird einer Gegenhypothese (H_1, mit der Annahme, dass ein solcher Zusammenhang bestehe) gegenübergestellt. Es soll entschieden werden, ob die Hypothese H_1 als weitgehend gesichert angenommen werden kann oder H_0 (vorläufig) beibehalten werden muss.
- Die statistische Prüfgröße Chi-Quadrat wird ermittelt, für die eine Wahrscheinlichkeitsverteilung (hier: Chi-Quadrat-Verteilung) bekannt ist (asymptotischer Test) oder berechnet werden kann (exakter Test). In diesem Kapitel wird der asymptotische Test besprochen (zum exakten Test ⇨ Kap. 29).
- Es wird ein Signifikanzniveau festgelegt, d.h. die Wahrscheinlichkeit, ab der H_1 angenommen werden soll. Üblich sind das 5 %-Niveau (ist dieses erreicht, spricht man von einem signifikanten Ergebnis) und das 1 %-Niveau (ist dieses erreicht, spricht man von einem hoch signifikanten Ergebnis).
- Feststellen der Freiheitsgrade (df = degrees of freedom) für die Verteilung.
- Aus diesen Festlegungen ergibt sich der „kritische Bereich", d.h. der Bereich der Werte der Prüfgröße, in dem H_1 angenommen wird.

❒ Die Prüfgröße wird daraufhin überprüft, ob sie in den kritischen Bereich fällt oder nicht. Ist ersteres der Fall, wird H_1 angenommen, ansonsten H_0 vorläufig beibehalten.

Der Chi-Quadrat-Test ist ein Test, der prüft, ob nach ihrer empirischen Verteilung zwei in einer Stichprobe erhobenen Variablen voneinander unabhängig sind oder nicht. Sind sie unabhängig, wird die Nullhypothese beibehalten, ansonsten die Hypothese H_1 angenommen. Der Chi-Quadrat-Test hat breite Anwendungsmöglichkeiten, da er als Messniveau lediglich Nominalskalenniveau voraussetzt. Zur Hypothesenprüfung wird nicht ein Parameter, sondern die ganze Verteilung verwendet. Deshalb spricht man von einem nicht-parametrischen Test (⇨Kap. 22). Außerdem macht er keine Voraussetzungen hinsichtlich der Verteilung der Werte in der Grundgesamtheit. Auch dies ist ein Merkmal von verteilungsfreien oder nichtparametrischen Tests. Allerdings sollte man beachten, dass die Daten aus einer Zufallsstichprobe stammen müssen. Weil die gesamte Verteilung geprüft wird, ergibt sich aber aus einem signifikanten Ergebnis nicht, an welcher Stelle der Verteilung die signifikanten Abweichungen auftreten. Dazu bedarf es weiterer Prüfungen.

Im Chi-Quadrat-Test wird die empirisch beobachtete Verteilung mit einer erwarteten Verteilung verglichen. Die erwartete Verteilung ist diejenige, die auftreten würde, wenn zwischen den beiden Variablen keine Beziehung bestünde, wenn sie also voneinander unabhängig wären. Die erwarteten Häufigkeiten (Erwartungswerte) für die einzelnen Zellen ij einer Tabelle (i = Zeile, j = Spalte) können aus den Randverteilungen ermittelt werden:

$$e_{ij} = \frac{(\text{Fallzahl in Zeile i}) \cdot (\text{Fallzahl in Spalte j})}{n} \tag{10.1}$$

Die Prüfgröße Chi-Quadrat (χ^2) ist ein Messwert für die Stärke der Abweichung der beobachteten Verteilung von der erwarteten Verteilung in einer Kreuztabelle:

$$\chi^2 = \sum_i \sum_j \frac{\left(f_{ij} - e_{ij}\right)^2}{e_{ij}} \tag{10.2}$$

f_{ij} = beobachtete Fälle in der Zelle der iten Reihe und jten Spalte

e_{ij} = unter H_0 erwartete Fälle in der Zelle der iten Reihe und jten Spalte

Die Prüfgröße χ^2 folgt asymptotisch einer Chi-Quadrat-Verteilung mit folgenden Freiheitsgraden:

$$df = (\textit{Zahl der Spalten} - 1) \cdot (\textit{Zahl der Zeilen} - 1) \tag{10.3}$$

SPSS führt die Berechnungen des Chi-Quadrat-Tests auf Anforderung durch. Es gibt den χ^2-Wert, die Freiheitsgrade und die Wahrscheinlichkeit des χ^2-Wertes unter der Annahme, dass H_0 gilt, an. Das Signifikanzniveau müssen Sie selbst festlegen und auf dieser Basis feststellen, ob Ihr Ergebnis signifikant ist oder nicht. Außerdem kann man sich die Erwartungswerte und die Differenz zwischen beob-

achteten Werten und Erwartungswert (die sogenannten Residuen) als Zwischenprodukte der Berechnung ausgeben lassen.

Ein Anwendungsbeispiel. Es soll ein (asymptotischer) Chi-Quadrat-Test für die Kreuztabelle zwischen den Variablen Einstellung nach dem „Materialismus-Postmaterialismus"-Index und Schulbildung durchgeführt werden. Die Nullhypothese besagt, dass zwischen beiden Variablen kein Zusammenhang bestehe, die Gegenhypothese dagegen, dass die Einstellung nach dem Index von der Schulbildung abhängig sei.

Um den Chi-Quadrat-Test durchzuführen und sich die Erwartungswerte sowie die Residuen zusätzlich anzeigen zu lassen, gehen Sie wie folgt vor:

- ▷ Wählen Sie zunächst in der Dialogbox „Kreuztabellen" die Zeilen- und die Spaltenvariable aus (⇨ Abb. 10.1).
- ▷ Klicken Sie auf die Schaltfläche „Statistik...". Es öffnet sich die Dialogbox „Kreuztabellen: Statistik" (⇨ Abb. 10.4).
- ▷ Wählen Sie das Kontrollkästchen „Chi-Quadrat", und bestätigen Sie mit „Weiter".
- ▷ Klicken Sie in der Dialogbox „Kreuztabellen" auf die Schaltfläche „Zellen...". Es öffnet sich die Dialogbox „Kreuztabellen: Zellen anzeigen" (⇨ Abb. 10.2).
- ▷ Wählen Sie dort alle gewünschten Kontrollkästchen an.
- ▷ Bestätigen Sie mit „Weiter" und „OK".

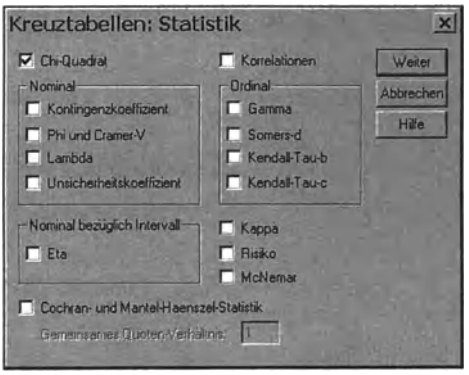


Abb. 10.4. Dialogbox „Kreuztabellen: Statistik"

In unserem Beispiel wurden neben den beobachteten Werten die Erwartungswerte, die nicht standardisierten sowie die standardisierten und die korrigierten standardisierten Residuen angefordert. Das Ergebnis steht in Tabelle 10.3.

Der Output enthält zunächst die angeforderte Tabelle mit den beobachteten Werten („Anzahl"), den erwarteten Werten („Erwartete Anzahl"), den nicht standardisierten Residuen („Residuen"), den standardisierten Residuen und den korrigierten standardisierten Residuen („Korrigierte Residuen"). Sie stehen untereinander in der genannten Reihenfolge. Für den einfachen Pearsonschen Chi-Quadrat-

Test sind nur die ersten drei Werte relevant. Betrachten wir die linke obere Zelle der Postmaterialisten, die einen Hauptschulabschluss oder weniger haben. Es sind 16 beobachtete Fälle. Der Erwartungswert beträgt 39,9. Er berechnet sich nach Formel 10.1 als $(80 \cdot 143) : 287 = 39{,}9$. Das dazugehörige Residuum, die Differenz zwischen beobachtetem und erwartetem Wert, beträgt $16 - 39{,}9 = -23{,}9$.

Tabelle 10.3. Chi-Quadrat-Test und Residuen für die Kreuztabelle „Inglehart-Index" nach „Schulbildung"

INGL * SCHUL2 Kreuztabelle

			SCHUL2			
			Hauptschule	Mittelschule	Fachh/Abi	Gesamt
INGL	POSTMATERIALISTEN	Anzahl	16	24	40	80
		Erwartete Anzahl	39,9	20,9	19,2	80,0
		% von SCHUL2	11,2%	32,0%	58,0%	27,9%
		Residuen	-23,9	3,1	20,8	
		Standardisierte Residuen	-3,8	,7	4,7	
		Korrigierte Residuen	-6,3	,9	6,4	
	PM-MISCHTYP	Anzahl	36	23	15	74
		Erwartete Anzahl	36,9	19,3	17,8	74,0
		% von SCHUL2	25,2%	30,7%	21,7%	25,8%
		Residuen	-,9	3,7	-2,8	
		Standardisierte Residuen	-,1	,8	-,7	
		Korrigierte Residuen	-,2	1,1	-,9	
	M-MISCHTYP	Anzahl	56	22	13	91
		Erwartete Anzahl	45,3	23,8	21,9	91,0
		% von SCHUL2	39,2%	29,3%	18,8%	31,7%
		Residuen	10,7	-1,8	-8,9	
		Standardisierte Residuen	1,6	-,4	-1,9	
		Korrigierte Residuen	2,7	-,5	-2,6	
	MATERIALISTEN	Anzahl	35	6	1	42
		Erwartete Anzahl	20,9	11,0	10,1	42,0
		% von SCHUL2	24,5%	8,0%	1,4%	14,6%
		Residuen	14,1	-5,0	-9,1	
		Standardisierte Residuen	3,1	-1,5	-2,9	
		Korrigierte Residuen	4,7	-1,9	-3,6	
Gesamt		Anzahl	143	75	69	287
		Erwartete Anzahl	143,0	75,0	69,0	287,0
		% von SCHUL2	100,0%	100,0%	100,0%	100,0%

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	64,473[a]	6	,000
Likelihood-Quotient	67,950	6	,000
Zusammenhang linear-mit-linear	58,862	1	,000
Anzahl der gültigen Fälle	287		

a. 0 Zellen (,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 10,10.

In der unteren Tabelle stehen die Ergebnisse verschiedener Varianten des Chi-Quadrat-Tests. Uns interessiert nur die erste Reihe, die die Werte des klassischen Pearsonschen Chi-Quadrat-Tests enthält. Für die Prüfgröße χ^2 wurde der Wert 64,473 ermittelt. Die Tabelle hat sechs Freiheitsgrade (df). Das errechnet sich gemäß Formel 10.3 aus $(3-1)\cdot(4-1) = 6$. Der Wert unter „Asymptotische Signifikanz (2-seitig)" gibt an, wie wahrscheinlich ein solcher Chi-Quadrat-Wert in einer Tabelle mit sechs Freiheitsgraden bei Geltung von H_0 ist. Es ist so unwahrscheinlich, dass der Wert 0,000 angegeben wird. Wir müssen uns daher keine weiteren Gedanken über das Signifikanzniveau machen, sondern können H_1 als statistisch signifikant annehmen. Nicht immer sind die Ergebnisse so eindeutig wie in diesem Beispiel. Allgemein gilt: Würde das Signifikanzniveau auf 5 %-Irrtumswahrscheinlichkeit festgelegt ($\alpha = 0{,}05$) werden, würden „Signifikanz"-Werte, die kleiner als α sind, als signifikant angesehen werden. Setzt man die Grenze bei 1 %-Irrtumswahrscheinlichkeit ($\alpha = 0{,}01$), so gilt entsprechendes (⇨ Kap. 13.3).

Der asymptotische Chi-Quadrat-Test bringt gute Ergebnisse, wenn die Daten einer Zufallsauswahl aus multinominalen Verteilungen entspringen. Außerdem dürfen für den asymptotischen Test die Erwartungswerte nicht zu klein sein. Als Faustregel gilt, dass diese in jeder Zelle mindestens fünf betragen sollen. Sind irgendwelche Erwartungswerte geringer als fünf, gibt SPSS unter der Tabelle an, in wie vielen Zellen solche Erwartungswerte auftreten. Sind die Anwendungsbedingungen für einen asymptotischen Test nicht gegeben, sollte ein exakter Test durchgeführt werden (⇨ Kap. 29).

Für 2*2-Tabellen führt SPSS *Fisher's exact Test* durch, wenn der Erwartungswert für irgendeine Zelle unter fünf liegt. Er ist besonders nützlich, wenn die Sample-Größe gering und die Erwartungswerte klein sind. Dieser Test berechnet exakte Werte für die Wahrscheinlichkeit, die beobachteten Resultate zu erhalten, wenn die Variablen unabhängig voneinander sind und die Randverteilung als fest angenommen werden kann.

Für 2*2-Tabelle wird häufig die *Yates Korrektur* (Continuity Correction) benutzt, die berücksichtigt, dass wir es mit diskontinuierlichen Merkmalen zu tun haben, die asymptotische Verteilung aber auf kontinuierlichen beruht. Dabei werden die Residuen korrigiert. Von positiven Residuen wird in Gleichung 10.2 der Wert 0,5 subtrahiert, zu negativen Werten der Wert 0,5 addiert, bevor man die Quadrierung vornimmt.

Um diese Berechnungen zu simulieren, wurde für ein kleineres Sample von 32 Personen (Datei: ALLBUS.SAV) der Zusammenhang zwischen politischem Interesse (POL1) und Schulbildung (SCHUL2) ermittelt. Außerdem wurden beide Variablen dichotomisiert, so dass eine 2*2- Tabelle entsteht. Die Ergebnisse finden Sie in Tabelle 10.4. In Fußnote b. ist angegeben, dass in zwei der vier Zellen der Erwartungswert unter fünf liegt. In der Tabelle sind beobachtete Werte „Anzahl", Erwartungswerte („Erwartete Anzahl") und Residuen aufgeführt. In der unteren Tabelle finden wir zunächst die Ergebnisse des Chi-Quadrat-Tests nach Pearson, danach die Ergebnisse nach Anwendung der Yates Korrektur („Kontinuitätskorrektur"). Der Chi-Quadrat-Wert wird mit 1,094 deutlich geringer als der Wert nach Pearson. Entsprechend ist auch die Wahrscheinlichkeit, dass unter der Nullhypo-

these $\chi^2 \geq \chi^2$ – Prüfwert gilt, mit 0,296 größer. (In beiden Fällen aber ist das Ergebnis nicht signifikant, die Nullhypothese wird beibehalten.)

Tabelle 10.4. Ergebnisse der Chi-Quadrat-Statistik bei einer 2 * 2-Tabelle mit kleinen Erwartungswerten

POL1 * SCHUL2 Kreuztabelle

			SCHUL2		
			eimfach	höher	Gesamt
POL1	vorhanden	Anzahl	18	8	26
		Erwartete Anzahl	19,5	6,5	26,0
		Residuen	-1,5	1,5	
	nicht vorhanden	Anzahl	6	0	6
		Erwartete Anzahl	4,5	1,5	6,0
		Residuen	1,5	-1,5	
Gesamt		Anzahl	24	8	32
		Erwartete Anzahl	24,0	8,0	32,0

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	2,462[b]	1	,117		
Kontinuitätskorrektur[a]	1,094	1	,296		
Likelihood-Quotient	3,893	1	,048		
Exakter Test nach Fisher				,296	,149
Zusammenhang linear-mit-linear	2,385	1	,123		
Anzahl der gültigen Fälle	32				

a. Wird nur für eine 2x2-Tabelle berechnet

b. 2 Zellen (50,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 1,50.

Für Fisher's Exact Test („Exakter Test nach Fisher „) sind für den einseitigen [„Exakte Signifikanz (1-seitig)"] und den zweiseitigen [„Exakte Signifikanz (2-seitig)"] Test jeweils die Wahrscheinlichkeiten angegeben, dass unter H_0 $\chi^2 \geq \chi^2$ – Prüfwert gilt. (Zum Unterschied zwischen einseitigen und zweiseitigen Tests ⇨ Kap 13.3.) Da beide Werte größer als 0,05 sind, liegen keine signifikanten Differenzen zwischen Erwartungswerten und beobachteten Werten vor. Die Nullhypothese wird auch danach beibehalten.

Kommen wir auf das erste Beispiel mit der größeren Stichprobe zurück. In Tabelle 10.3 sind weitere Residuen („Standardisierte Residuen", „Korrigierte Residuen") enthalten. Sie stehen in Verbindung mit den ebenfalls angebotenen alternativen Chi-Quadrat-Tests. Deren Ergebnisse finden sich ebenfalls in den der unteren Tabelle. Es sind der *„Likelihood-Quotient Chi-Quadrat"*-Test und der *„Man-*

tel-Haenszel"-Test („Zusammenhang linear-mit-linear"). Ersterer beruht auf der Maximum-Likelihood-Theorie und wird häufig für kategoriale Daten benutzt. Bei großen Stichproben bringt er dasselbe Ergebnis wie der Pearson-Test. Der Mantel-Haenszel-Test wird später genauer besprochen. Er ist ausschließlich für Ordinaldaten gedacht, hat *immer* einen Freiheitsgrad und kann auch als Zusammenhangsmaß benutzt werden. Im Prinzip werden die Ergebnisse dieser Tests auf die gleiche Weise interpretiert wie der Chi-Quadrat-Test nach Pearson. Wie man sieht, erbringen alle drei Tests in der Tabelle 10.3 auch in etwa die gleichen Ergebnisse. In Tabelle 10.4, der 2 * 2-Tabelle, dagegen führt der „Likelihood-Quotient" zu einem anderen Ergebnis. Nach ihm sind die Differenzen zwischen den Absolventen verschiedener Schularten signifikant, während alle anderen Tests ein nicht signifikantes Ergebnis anzeigen.

Exakte Tests. Wenn die Anwendungsbedingungen für den Chi-Quadrat-Test (erwartete Häufigkeiten > 5) nicht erfüllt sind, sollte ein exakter Test durchgeführt werden (⇨ Kap. 29).

10.3 Zusammenhangsmaße

Signifikanztests geben an, ob ein beobachteter Zusammenhang zwischen zwei Variablen statistisch abgesichert ist oder nicht, ihnen ist aber keine direkte Information zu entnehmen, wie eng dieser Zusammenhang ist. Das Ergebnis eines Signifikanztestes hängt nämlich vor allem auch von der Größe der Stichprobe ab. Je größer die Stichprobe, desto eher lässt sich die Signifikanz auch schwacher Zusammenhänge nachweisen. Maßzahlen, die die Stärke eines Zusammenhanges zwischen zwei Variablen ausdrücken, nennt man Zusammenhangs- oder Assoziationsmaße. Da es, insbesondere vom Messniveau der Variablen abhängig, unterschiedliche Arten von Assoziationen gibt, wurden auch verschiedene Zusammenhangsmaße entwickelt. Die Zusammenhangsmaße unterscheiden sich aber auch in anderer Hinsicht. Z.B. berücksichtigen manche neben dem Grad der Assoziation auch Informationen über die Randverteilung (Margin-sensitive-Maße). Außerdem definieren sie unterschiedlich, was ein perfekter Zusammenhang ist und geben unterschiedliche Zwischenwerte an. Die Mehrzahl der Zusammenhangsmaße ist allerdings so angelegt, dass der Wert 0 anzeigt, dass kein Zusammenhang zwischen den beiden Variablen vorliegt. Der Wert 1 dagegen indiziert einen perfekten Zusammenhang. Werte zwischen 1 und 0 stehen für mehr oder weniger starke Zusammenhänge. Sind beide Variablen zumindest auf dem Ordinalskalenniveau gemessen, kann auch die Richtung des Zusammenhangs angegeben werden. Ein positiver Wert des Zusammenhangsmaßes zeigt an, dass ein größerer Wert auf der unabhängigen Variablen auch einen größeren auf der abhängigen nach sich zieht. Dagegen indiziert ein negatives Vorzeichen, dass mit Vergrößerung des Wertes der unabhängigen Variablen der Wert der abhängigen sinkt.

Tabelle 10.5. Beziehung zwischen Messniveau und Zusammenhangsmaß

Messniveau	Maßzahlen	Bemerkungen
	Chi-Quadrat-basierte Messungen	
	Phi Koeffizient ϕ	Für 2 * 2-Tabellen geeignet. Ansonsten beträgt das Maximum nicht 1, z.T. auch größer 1.
	Cramers V	Auch für größere Tabellen. Der maximale Wert beträgt immer 1.
	Kontingenzkoeffizient C	Auch für größere Tabellen. Der maximale Wert beträgt nicht immer 1.
Nominal	*Relative (proportionale) Irrtumsreduktion*	
	Lambda λ *)	Erreicht nur 1, wenn jede Reihe mindestens eine nicht leere Zelle enthält.
	Kruskals und Goodmans tau	Beruht auf der Randverteilung. Wird mit dem Kontrollkästchen Lambda mit ausgewählt.
	Unsicherheitskoeffizient	Beruht auf den Randverteilungen.
	Sonstiges (Zustimmungsmessung)	Speziell für Übereinstimmungsmessungen bei Überprüfung von Zuverlässigkeit und Gültigkeit.
	kappa	Kann nur für quadratische Tabellen mit gleicher Zeilen- und Spaltenzahl berechnet werden.
	Spearmans Rangkorrelationskoeffizient r.	Bereich zwischen -1 und +1.
	Mantel-Haenszel Chi-Quadrat Zusammenhang linear-mit-linear	Nur für ordinal skalierte Daten.
	auf paarweisen Vergleich beruhende Maßzahlen	
Ordinal	Kendalls tau-b	Berücksichtigt Bindungen auf einer der Variablen, kann nicht immer die Werte -1 und +1 erreichen.
	Kendalls tau-c	Kann näherungsweise bei jeder Tabellenform die Werte −1 und +1 erreichen.
	Gamma	0 ist nur bei 2 * 2-Tabelle ein sicheres Indiz für Unabhängigkeit der Variablen. Für 3- bis 10-dimensionale Tabellen werden bedingte Koeffizienten berechnet.
	Somers d	Asymmetrische Variante von Gamma.
Intervall	Pearsonscher Produkt-Moment Korrelations-Koeffizient r	Gilt für lineare Beziehungen. Wertebereich zwischen -1 und +1.
Mischform/ Sonderauf-gaben	Eta	Unabhängige Variable nominal, abhängige mindestens intervallskaliert.
	Risk und odds-ratio*)	Speziell für Kohorten- bzw. Fall-Kontroll-Studien. Kann nur für 2 * 2-Tabellen berechnet werden.

* keine exakten Tests verfügbar

Im folgenden geben wir eine Übersicht über die von SPSS verwendeten Zusammenhangsmaße in Abhängigkeit vom Messniveau der Variablen. Entscheidend ist das Messniveau der auf dem geringsten Niveau gemessenen Variablen.

10.3.1 Zusammenhangsmaße für nominalskalierte Variablen

Wenn beide Variablen auf Nominalskalenniveau gemessen sind, ist es lediglich möglich, Aussagen über die Stärke des Zusammenhanges zu machen. Da keine eindeutige Ordnung existiert, sind Aussagen über Art und Richtung des Zusammenhanges sinnlos. Zusammenhangsmaße für nominalskalierte Daten können auf zwei unterschiedlichen Logiken aufbauen. Die einen gründen sich auf die Chi-Quadrat-Statistik, die anderen auf der Logik der proportionalen Irrtumsreduktion.

Auf der Chi-Quadrat-Statistik basierende Zusammenhangsmaße. Der Chi-Quadrat Wert selbst ist als Zusammenhangsmaß nicht geeignet. Er ist nämlich außer von der Stärke des Zusammenhangs auch von der Stichprobengröße und der Zahl der Freiheitsgrade abhängig. Auf seiner Basis können aber Zusammenhangsmaße errechnet werden, wenn man den Einfluss der Stichprobengröße und der Freiheitsgrade berücksichtigt und dafür sorgt, dass die Maßzahl Werte im Bereich zwischen 0 und 1 annimmt. Dafür sind verschiedene Verfahren entwickelt worden:

- *Phi Koeffizient.* Er ist hauptsächlich für 2 * 2-Tabellen geeignet. Zur Ermittlung des Phi Koeffizienten wird Pearsons Chi-Quadrat durch die Stichprobengröße dividiert und die Quadratwurzel daraus gezogen. Phi ergibt für 2 * 2-Tabellen denselben Wert wie Pearsons Produkt-Moment-Korrelationskoeffizient. Bei größeren Tabellen kann es sein, dass die Werte nicht zwischen 0 und 1 liegen, weil Chi-Quadrat größer als die Stichprobengröße ausfallen kann. Deshalb sollte dann lieber der Kontingenzkoeffizient berechnet werden. Koeffizienten unterschiedlicher Tabellenformen sind nicht vergleichbar.

$$\phi = \sqrt{\frac{\chi^2}{n}} \tag{10.4}$$

- *Cramers V.* Ist eine weitere Variation, die auch bei größeren Tabellen immer einen Wert zwischen 0 und 1 erbringt. In jeder Tabelle beträgt der maximal erreichbare Wert 1. Die Formel lautet:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} \tag{10.5}$$

  Dabei ist k der kleinere Wert der Anzahl der Reihen oder der Spalten.

- *Kontingenzkoeffizient.* Zu seiner Berechnung wird der Chi-Quadrat-Wert nicht durch n, sondern durch $\chi^2 + n$ geteilt. Dadurch liegen die Werte immer zwischen 0 und 1.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{10.6}$$

Bei der Interpretation ist allerdings zu beachten, dass der maximal erreichbare Wert nicht immer 1 beträgt. Der maximal erreichbare Wert hängt von der Zahl der Reihen und Spalten der Tabelle ab. Er ist nur für quadratische Tabellen nach der folgenden Formel ermittelbar:

$$C_{Max} = \sqrt{\frac{r-1}{r}}, \text{ wobei r = Zahl der Reihen bzw. Spalten} \qquad (10.7)$$

Er beträgt beispielsweise bei einer 2 * 2-Tabelle 0,707 und bei einer 4 * 4-Tabelle nur 0,87. Man sollte daher nur Kontingenzkoeffizienten für Tabellen gleicher Größe vergleichen.

Zum Vergleich unterschiedlich großer quadratischer Tabellen kann der Kontingenzkoeffizient nach der folgenden Formel korrigiert werden:

$$C_{korr} = \frac{C}{C_{Max}} \qquad (10.8)$$

Beispiel. Es soll überprüft werden, wie eng der Zusammenhang zwischen der Beurteilung ehelicher Untreue und Geschlecht des/der Befragten ist. Da Geschlecht ein nominal gemessenes Merkmal ist, kommen nur Zusammenhangsmaße für nominal skalierte Merkmale in Frage. Um die entsprechenden Statistiken zu erstellen, gehen Sie wie folgt vor:

- ▷ Wählen Sie zunächst in der Dialogbox „Kreuztabellen" Spalten- und Zeilenvariable aus.
- ▷ Schalten Sie in die Dialogbox „Zellen...".
- ▷ Wählen Sie dort die gewünschte Prozentuierung aus.
- ▷ Wählen Sie in der Dialogbox „Kreuztabellen" die Schaltfläche „Statistik...". Es öffnet sich die Dialogbox „Kreuztabellen: Statistik" (⇨ Abb. 10.4). Für Sie ist jetzt die Auswahlgruppe „Nominal" relevant.
- ▷ Wählen Sie durch Anklicken der Kontrollkästchen die gewünschten Statistiken.
- ▷ Bestätigen Sie mit „Weiter" und „OK".

Wenn Sie die Kästchen „Kontingenzkoeffizient" und „Phi und Cramers V" angewählt haben, enthält die Ausgabe den in Tabelle 10.6 dargestellten Output. Er enthält unter der Kreuztabelle die drei angeforderten Zusammenhangsmaße. Da es sich nicht um eine 2 * 2-Tabelle handelt, kommt eigentlich Phi nicht in Frage. Wie man sieht, weichen aber die drei Maße nur minimal voneinander ab.

Alle drei liegen bei ca. 0,22. Das zeigt einen schwachen Zusammenhang zwischen Geschlecht und Einstellung zur ehelichen Treue an. Für alle drei Zusammenhangsmaße wird zugleich als Signifikanztest der Pearsonsche Chi-Quadrat-Test durchgeführt. Sein Hauptergebnis wird in der Spalte „Näherungsweise Signifikanz" mitgeteilt. Der Wert 0,057 besagt, dass eine etwa 5,6 %-ige Wahrscheinlichkeit besteht, dass das Ergebnis auch dann per Zufall zustande gekommen sein kann, wenn in Wirklichkeit die Nullhypothese gilt. Nach der üblichen Konvention sind deshalb die Zusammenhangsmaße nicht signifikant. Es ist also nicht statistisch abgesichert, dass überhaupt ein Zusammenhang existiert (⇨ Kap. 13.3).

Tabelle 10.6. Kreuztabelle mit Chi-Quadrat-basierten Zusammenhangsmaßen

TREUE * GESCHL Kreuztabelle

			GESCHL		
			MAENNLICH	WEIBLICH	Gesamt
TREUE	SEHR SCHLIMM	Anzahl	12	27	39
		% von GESCHL	16,2%	34,2%	25,5%
	ZIEMLICH SCHLIMM	Anzahl	24	25	49
		% von GESCHL	32,4%	31,6%	32,0%
	WENIGER SCHLIMM	Anzahl	23	17	40
		% von GESCHL	31,1%	21,5%	26,1%
	GAR NICHT SCHLIMM	Anzahl	15	10	25
		% von GESCHL	20,3%	12,7%	16,3%
Gesamt		Anzahl	74	79	153
		% von GESCHL	100,0%	100,0%	100,0%

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	,222	,057
	Cramer-V	,222	,057
	Kontingenzkoeffizient	,217	,057
Anzahl der gültigen Fälle		153	

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

Auf der Logik der relativen Reduktion des Irrtums beruhende Zusammenhangsmaße. Alle diese Messungen beruhen auf demselben Grundgedanken. Sie gehen zunächst von der Annahme aus, man wolle Werte der abhängigen Variablen für einzelne Fälle auf Basis vorhandener Kenntnisse voraussagen. Eine gewisse Trefferquote würde man schon durch reines Raten erreichen. Ausgangspunkt der Überlegungen ist aber die Trefferquote, die man wahrscheinlich erreicht, wenn man die Verteilung der vorauszusagenden Variablen selbst kennt. Genauer geht man von deren Kehrwert aus, nämlich von der Wahrscheinlichkeit, dass man sich dabei irrt. Man muss dann mit Irrtümern in einer bestimmten Größenordnung rechnen. Hat man eine weitere unabhängige Variable zur Verfügung, die das Ergebnis der abhängigen Variablen mit beeinflusst, kann man aufgrund der Kenntnis der Werte auf der unabhängigen Variablen verbesserte Aussagen machen, die allerdings immer noch mit Irrtümern bestimmter Größe behaftet sind. Jedoch werden die Irrtümer aufgrund der zusätzlichen Kenntnis geringer ausfallen, und zwar wird sich die Größe des Irrtums umso mehr verringern, je enger die Beziehung zwischen der unabhängigen und der abhängigen Variablen ist. Die Maße, die sich auf die Logik der proportionalen Reduktion der Irrtumswahrscheinlichkeit stützen, basieren alle darauf, dass sie zwei Irrtumsmaße ins Verhältnis zueinander setzen. Das erste misst die Größe des Irrtums, der bei einer Voraussage ohne die zusätzliche Kenntnis der unabhängigen Variablen auftritt, die andere die Größe des Irrtums,

der bei der Voraussage auftritt, wenn man dabei die Kenntnis der unabhängigen Variablen nutzt.

Gehen wir zur Tabelle 10.6, dem angeführten Beispiel über die Einstellung bundesdeutscher Bürger zur ehelichen Untreue. Wollte man bei einzelnen Bürgern die Einstellung zur Untreue voraussagen und wüsste nur die Verteilung der Einstellungen insgesamt, so hätten wir die Information, die uns die Verteilung am rechten Rand der Tabelle gibt. Demnach wird am häufigsten – von 32 % der Befragten – der Wert 2 „ziemlich schlimm" gewählt. Die sicherste Voraussage machen wir, wenn wir diese häufigste Kategorie für die Voraussage verwenden. Allerdings wird man sich dann bei 68 % der Voraussagen irren. Die Wahrscheinlichkeit, sich zu irren ist demnach 1 minus dem Anteil der Fälle in der am stärksten besetzten Kategorie, $P_i(1) = 1 - 0{,}32 = 0{,}68$.

Wissen wir jetzt noch das Geschlecht, so können wir die Voraussage verbessern, indem wir bei der Voraussage der Werte für die Männer die Kategorie benutzen, die bei Männern am häufigsten auftritt, bei der Voraussage des Wertes für die Frauen dagegen die, die bei diesen am häufigsten auftritt. Bei den Männern ist das die Kategorie 2 „ziemlich schlimm", die 32,4 % der Männer wählen, bei den Frauen die Kategorie 1 „sehr schlimm", die diese zu 34,2 % wählen. Um die Irrtumswahrscheinlichkeit zu ermitteln, benutzen wir die Gesamtprozente (TOTAL). Wir haben uns bei allen geirrt, die nicht in die beiden Zellen „Männer/ziemlich schlimm" bzw. „Frauen/sehr schlimm" fallen.

$P_i(2) = 7{,}8 + 15 + 9{,}8 + 16{,}3 + 11{,}1 + 6{,}5 = 66{,}5\ \%$

Diese Irrtumswahrscheinlichkeit ist also geringer als 68 %, in unserem Falle aber so minimal, dass man kaum von einem Gewinn sprechen kann. Die verschiedenen, auf der Logik der relativen Reduktion der Irrtumswahrscheinlichkeit beruhenden, Maßzahlen berechnen das Verhältnis der Irrtumswahrscheinlichkeiten auf verschiedene Weise (wir lassen im folgenden das Subskript i für Irrtum in den Formeln weg):

① *Goodmans und Kruskals Lambda.* Wird nach folgender Formel berechnet:

$$\lambda = \frac{P(1) - P(2)}{P(1)} \tag{10.9}$$

In unserem Beispiel ergibt das $(68 - 66{,}5) : 68 = 0{,}02$. Damit sind nur 2 % der Irrtumswahrscheinlichkeit reduziert.

Lambda ergibt Ergebnisse zwischen 0 und 1. Ein Wert 0 bedeutet, dass die unabhängige Variable die Voraussage überhaupt nicht verbessert, ein Wert 1, dass sie eine perfekte Voraussage ermöglicht. Bei der Interpretation ist allerdings zu berücksichtigen, dass der Wert 1 nur erreicht werden kann, wenn in jeder Reihe mindesten eine nicht leere Zelle existiert. Außerdem kann man zwar sagen, dass bei statistischer Unabhängigkeit Lambda den Wert 0 annimmt, aber nicht umgekehrt, dass 0 unbedingt völlige statistische Unabhängigkeit anzeigt. Lambda bezieht sich ausschließlich auf die besondere statistische Beziehung, dass aus einem Wert der unabhängigen Variablen einer der abhängigen vorausgesagt werden soll.

Je nachdem, welche Variable in einer Beziehung die unabhängige, welche die abhängige ist, kann Lambda unterschiedlich ausfallen. SPSS bietet daher für beide mögliche Beziehungsrichtungen ein asymmetrisches Lambda an. Der Benutzer muss selbst entscheiden, welches in seinem Falle zutrifft. Für den Fall, dass keine der Variablen eindeutig die unabhängige bzw. abhängige ist, wird darüber hinaus eine symmetrische Version von Lambda angezeigt, die die Zeilen- und Spaltenvariable gleich gut voraussagt. Ein Nachteil von Lambda ist, dass die Voraussage der Werte der abhängigen Variablen lediglich auf der Zelle mit dem häufigsten Wert beruht. Bei größeren Tabellen muss daher zwangsläufig eine große Irrtumswahrscheinlichkeit auftreten, wenn nicht ganz extreme Verteilungen vorliegen. Außerdem werden unter bestimmten Bedingungen selbst klare Zusammenhänge nicht ausgewiesen. Wenn z.B. die verschiedenen Gruppen der unabhängigen Variablen den häufigsten Wert in derselben Kategorie der abhängigen Variablen haben, wird auch dann kein Zusammenhang ausgewiesen, wenn sich die relativen Häufigkeiten in diesen Kategorien klar unterscheiden.

② *Goodmans und Kruskals Tau.* Dieser Wert wird beim Aufruf von Lambda mit ausgegeben. Bei der Berechnung von Lambda wird auf Basis des häufigsten Wertes für alle Werte einer Spalte oder Zeile die gleiche Voraussage gemacht. Die Berechnung von Tau beruht auf einer anderen Art von Voraussage. Hier wird die Voraussage stochastisch auf Basis der Randverteilung getroffen. Man würde deshalb in unserem Beispiel (vor Einbeziehung der Variablen „Geschlecht") nicht für alle Fälle den Wert 2 „ziemlich schlimm" voraussagen, sondern durch Zufallsziehung mit unterschiedlich gewichteten Chancen für die Kategorien 1 bis 4 gemäß der Randverteilung für die Einstellung gegenüber ehelicher Untreue 25,6 % der Fälle den Wert 1, 32,0 % den Wert 2, 26,1 % den Wert 3 und 16,3 % den Wert 4 zuordnen. Man kann auf dieser Basis ermitteln, dass 27,436 % aller Fälle richtig vorausgesagt würden oder umgekehrt in 72,564 % der Fälle eine falsche Voraussage getroffen würde. Tau wird ansonsten parallel zu Lambda berechnet.

Tau = (73,7 – 72,564) : 73,7 = 0,0154.

Auch hier kann in unserem Beispiel nur eine geringfügige Verbesserung der Voraussage mit etwa 1,6 %iger Reduktion der Irrtumswahrscheinlichkeit errechnet werden.

Für Tau kann näherungsweise ein Signifikanztest auf Basis der Chi-Quadrat-Verteilung durchgeführt werden. Das Ergebnis wird in der Spalte „Näherungsweise Signifikanz" mitgeteilt. Außerdem kann ein „Asymptotischer Standardfehler" berechnet werden. Aufbauend auf ihm, kann man ein Konfidenzintervall ermitteln.

Um die gewünschten Statistiken zu erhalten, gehen Sie wie folgt vor:

- ▷ Wählen Sie in der Dialogbox „Kreuztabellen" die Zeilen und die Spaltenvariable aus (⇨ Abb. 10.1).
- ▷ Wenn Sie lediglich die Statistiken und nicht die Tabelle angezeigt wünschen, wählen Sie das Kontrollkästchen „Keine Tabellen".
- ▷ Klicken Sie auf die Schaltfläche „Statistik...".

▷ Wählen Sie in der Dialogbox „Kreuztabellen: Statistik“ (⇨ Abb. 10.4) die gewünschten Statistiken.

Tabelle 10.7. Zusammenhangsmaße nach der Logik der relativen Irrtumsreduktion

Richtungsmaße

			Wert	Asymptotischer Standardfehler[a]	Näherungsweises T[b]	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Lambda	Symmetrisch	,073	,058	1,208	,227
		VERHALTENSBEURTEILUNG: SEITENSPRUNG abhängig	,019	,069	,277	,781
		GESCHLECHT, BEFRAGTE<R> abhängig	,149	,101	1,373	,170
	Goodman-und-Kruskal-Tau	VERHALTENSBEURTEILUNG: SEITENSPRUNG abhängig	,016	,011		,063[c]
		GESCHLECHT, BEFRAGTE<R> abhängig	,049	,034		,058[c]
	Unsicherheitskoeffizient	Symmetrisch	,024	,017	1,416	,053[d]
		VERHALTENSBEURTEILUNG: SEITENSPRUNG abhängig	,018	,013	1,416	,053[d]
		GESCHLECHT, BEFRAGTE<R> abhängig	,036	,026	1,416	,053[d]

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

c. Basierend auf Chi-Quadrat-Näherung

d. Chi-Quadrat-Wahrscheinlichkeit für Likelihood-Quotienten.

Haben Sie die Kästchen „Lambda“ und „Unsicherheitskoeffizient“ ausgewählt und auf die Ausgabe der Tabelle verzichtet, ergibt sich der in Tabelle 10.7 angezeigte Output.

In unserem Falle ist „Geschlecht“ eindeutig die unabhängige und „Einstellung zur ehelichen Untreue“ die abhängige Variable. Deshalb sind bei allen Koeffizienten die Angaben zur asymmetrischen Version mit „VERHALTENSBEURTEILUNG: SEITENSPRUNG“ als abhängiger Variablen die richtigen. Lambda zeigt den Wert 0,019, gibt also an, dass ungefähr 1,9 % der Fehlerwahrscheinlichkeit bei einer Voraussage durch Einbeziehung der Information über das Geschlecht reduziert werden. Ganz ähnlich ist das Ergebnis für Tau. Der Wert beträgt 0,016. Beides kommt den oben berechneten Werten nahe. Die Werte 0,781 und 0,063 für „Näherungsweise Signifikanz“ in den Reihen für Lambda und Tau zeigen darüber hinaus – allerdings sehr unterschiedlich deutlich –, dass das Ergebnis nicht signifikant ist. Es ist auch denkbar, dass die Variable Geschlecht gar keine Erklärungskraft hat.

Im ersten Falle ist der Signifikanztest auf einem näherungsweisen t-Wert aufgebaut, im zweiten Falle auf einer Chi-Quadrat-Näherung.

③ *Unsicherheitskoeffizient.* Er hat dieselbe Funktion wie die beiden besprochenen Koeffizienten und ist auf die gleiche Weise zu interpretieren. Er ähnelt in seiner Berechnung ebenfalls Lambda. Aber auch hier wird die ganze Verteilung, nicht nur der häufigste Wert für die Voraussage genutzt. Es existiert eine symmetrische und eine asymmetrische Version. Bei der asymmetrischen muss bekannt sein, welche Variable die unabhängige ist. Wenn x die unabhängige Variable ist, wird der Unsicherheitskoeffizient nach folgender Formel berechnet:

$$\text{Unsicherheitskoeffizient} = \frac{U(y) - U(y/x)}{U(y)} \quad (10.10)$$

Dabei ist U(y) die Unsicherheit, die besteht, wenn nur die Verteilung der abhängigen Variablen bekannt ist, U(x/y) ist die bedingte Unsicherheit, wenn auch die Werte der unabhängigen Variablen bekannt sind.

U(y) repräsentiert die durchschnittliche Unsicherheit in der Randverteilung von y. Es wird berechnet:

$$U(y) = -\sum_j p(y_j) \log p(y_j) \quad (10.11)$$

Dabei ist $p(y_j)$ die Wahrscheinlichkeit dafür, dass eine bestimmte Kategorie von y auftritt. U(y/x) wird berechnet:

$$U_{(y/x)} = -\sum_{kj} \sum (y_j, x_k) \log p(y_j / x_k) \quad (10.12)$$

10.3.2 Zusammenhangsmaße für ordinalskalierte Variablen

Allgemein gilt, dass Maßzahlen, die ein niedriges Messniveau voraussetzen, auch für Daten höheren Messniveaus Verwendung finden können. Man verschenkt dabei aber einen Teil der verfügbaren Information. Zusätzlich zur Information über einen Unterschied von Werten, die auch bei nominalskalierten Daten vorliegt, kann man ordinalskalierte Daten in eine eindeutige Rangfolge ordnen. Anders als bei reinen Kontingenztabellen und den dazugehörigen Zusammenhangsmaßen (Kontingenzkoeffizienten), kann man daher Zusammenhangsmaße bilden (Assoziationskoeffizienten), die auch Auskunft über die Richtung des Zusammenhanges geben und dem Konzept der Korrelation entsprechen. Nach diesem sind Variablen positiv korreliert, wenn niedrige Werte auf einer Variablen tendenziell auch niedrige auf der anderen nach sich ziehen und hohe Werte auf der ersten, hohe auf der zweiten. Umgekehrt sind sie negativ korreliert, wenn niedrige Werte auf der einen tendenziell mit hohen Werten auf der anderen verbunden sind.

Rangkorrelationsmaße.

Spearmans Rangkorrelationskoeffizient r_s. Er basiert auf dem später besprochenen Pearsonschen Produkt-Moment-Korrelationskoeffizienten r. Dieser verlangt aber Intervallskalenniveau der korrelierten Variablen. Der Spearmansche Rangkorrelationskoeffizient umgeht dieses Problem, indem er anstelle der Werte der Variablen die Rangplätze der Fälle bezüglich dieser Variablen verwendet. Die Fälle werden zuerst auf jeder Variablen nach ihrer Position angeordnet. Entsprechend kann man für jeden Fall auf diesen Variablen den Rangplatz ermitteln (liegt bei mehreren Fällen derselbe Wert vor, bekommen sie alle denselben mittleren Rangplatz). Wenn Rangplätze verwendet werden, kann die Formel für den Pearsonschen Produkt-Moment-Korrelationskoeffizienten gemäß Gleichung 10.19 umgeformt werden in:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{10.13}$$

Dabei ist n die Zahl der Fälle und d jeweils für jeden Fall die Differenz zwischen dem Rangplatz auf der ersten und dem Rangplatz auf der zweiten Variablen.

Mantel-Haenszel Chi-Quadrat (Zusammenhang linear-mit-linear). Ist ein Signifikanztest für Rangkorrelationsmaße. Er beruht ebenfalls auf dem Pearsonschen Korrelationskoeffizienten. Die Zahl der Freiheitsgrade ist immer 1. Dieser Koeffizient wird von SPSS immer zusammen mit der Chi-Quadrat-Statistik ausgegeben, sollte aber nur bei ordinalskalierten Daten Verwendung finden. Die Formel lautet:

$$\text{Mantel-Haenszel} = r_s^2 \cdot (n - 1) \tag{10.14}$$

Auf paarweisem Vergleich beruhende Maßzahlen. Alle anderen Zusammenhangsmaße für Ordinaldaten beruhen auf dem paarweisen Vergleich aller Fälle hinsichtlich ihrer Werte auf beiden Variablen. Das Grundprinzip ist wie folgt: Alle möglichen Paare zwischen den Fällen werden verglichen. Dabei wird bei jedem Paar festgestellt, in welcher Beziehung die Werte stehen. Sind beide Werte des ersten Falles höher als beide Werte des zweiten Falles oder sind sie umgekehrt beide niedriger, so spricht man davon, dass dieses Paar *konkordant* ist. Ist dagegen der eine Wert des ersten Falles niedriger als der Wert des zweiten Falles auf dieser Variablen, bei der anderen Variablen dagegen das Umgekehrte der Fall, ist das Paar *diskordant*. Schließlich ist das Paar *gebunden* (tied), wenn wenigstens einer der Werte gleich ist. Es gibt drei Arten von Bindungen, erstens: beide Werte sind gleich, zweitens: der eine ist gleich, der andere bei Fall zwei geringer oder drittens: ein Wert ist gleich, der andere bei Fall zwei höher.

Aus einer Kreuztabelle lässt sich leicht entnehmen, wieviel konkordante, wieviel diskordante und wieviel gebundene Paare existieren. Die Zusammenhangsmaße beruhen nun auf dem Anteil der verschiedenen konkordanten und diskordanten Paare. Überwiegen die konkordanten Paare, dann ist der Zusammenhang positiv, überwiegen die diskordanten, ist er negativ. Existieren gleich viele konkordante und diskordante, besteht kein Zusammenhang. Alle gehen von der Differenz aus: konkordante Paare (P) minus diskordante (Q). Sie unterscheiden sich in der Art, wie diese Differenz normalisiert wird:

① *Kendalls tau-b*. Berücksichtigt bei der Berechnung Bindungen auf einer der beiden Variablen, nicht aber Bindungen auf beiden. Die Formel lautet:

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + T_x)(P + Q + T_y)}} \tag{10.15}$$

Dabei ist T_x die Zahl der Paare, bei denen auf der ersten Variablen (x) eine Bindung vorliegt und T_y die Zahl der Paare, bei denen auf der zweiten Variablen (y) eine Bindung vorliegt.

Tau-b kann nicht immer die Werte −1 und +1 erreichen. Wenn kein Randwert Null vorliegt, ist das nur bei quadratischen Tabellen (mit gleicher Zahl der Reihen und Spalten) und symmetrischen Randhäufigkeiten möglich.

② *Kendalls tau-c.* Ist eine Maßzahl, die auch bei n * m-Tabellen die Werte −1 und +1 näherungsweise erreichen kann. Dies wird durch Berücksichtigung von m = Minimum von Spalten bzw. Reihen erreicht. Die Formel lautet:

$$\tau_c = \frac{2m(P-Q)}{n^2(m-1)} \tag{10.16}$$

Dabei ist m die kleinere Zahl der Reihen oder Spalten. In Abhängigkeit von m erreicht der Maximalwert aber auch nicht in jedem Falle 1.

Tau-b und tau-c ergeben in etwa den gleichen Wert, wenn die Randverteilungen in etwa gleiche Häufigkeiten aufweisen.

③ *Goodmans und Kruskals Gamma.* Es ist der tau-Statistik verwandt. Die Formel lautet:

$$G = \frac{P-Q}{P+Q} \tag{10.17}$$

Es ist die Wahrscheinlichkeit dafür, dass ein Paar konkordant ist minus der Wahrscheinlichkeit, dass es diskordant ist, wenn man die Bindungen vernachlässigt. Gamma wird 1, wenn alle Fälle in den Zellen der Diagonalen einer Tabelle liegen. Sind die Variablen unabhängig, nimmt es den Wert 0 an. Aber umgekehrt ist der Wert 0 kein sicheres Zeichen, dass Unabhängigkeit vorliegt. Sicher ist es bei 2 * 2-Tabellen. Durch Zusammenlegen von Kategorien kann Gamma leicht künstlich angehoben werden, deshalb sollte es vornehmlich für die Analyse der Originaldaten verwendet werden.

④ *Somers d.* Ist eine Variante von Gamma. Bei der Berechnung von Gamma wird allerdings eine symmetrische Beziehung zwischen den beiden Variablen angenommen. Dagegen bietet Somers d eine asymmetrische Variante. Es wird zwischen unabhängiger und abhängiger Variablen unterschieden. Im Nenner steht daher die Zahl aller Paare, die nicht auf der unabhängigen Variablen gebunden sind, also auch die Bindungen auf der abhängigen Variablen. Die Formel lautet:

$$d_y = \frac{P-Q}{P+Q+T_y} \tag{10.18}$$

d gibt also den Anteil an, um den die konkordanten die diskordanten Paare übersteigen, bezogen auf alle Paare, die nicht auf x gebunden sind. Die symmetrische Variante von d benutzt als Nenner das arithmetische Mittel der Nenner der beiden asymmetrischen Varianten.

Zwischen den auf paarweisem Vergleich beruhenden Maßzahlen besteht folgende generelle Beziehung:

$$|\tau_b| \leq |\gamma| \text{ und } |d_y| \leq |\gamma|$$

Beispiel. Es sollen für die Beziehung zwischen Schulabschluss und Einstellung auf der Dimension „Materialismus-Postmaterialismus" Zusammenhangsmaße ermittelt werden. Dabei wird die rekodierte Variable SCHUL2 verwendet. Die entsprechende Kreuztabelle findet sich am Anfang dieses Kapitels. Beide Variablen „Schulbildung" und „Einstellung nach dem Inglehart-Index" sind ordinalskaliert. Es gibt eine eindeutige Ordnung von geringer zur höheren Schulbildung und von postmaterialistischer zu materialistischer Einstellung. Daher kommen Koeffizienten für ordinalskalierte Daten in Frage.

Um diese zu ermitteln, gehen Sie wie oben beschrieben vor mit dem Unterschied, dass in der Dialogbox „Kreuztabellen: Statistik" die gewünschten Statistiken gewählt werden.

Wenn Sie sämtliche Statistiken der Gruppe „Ordinal" und zusätzlich das Kästchen „Korrelationen" ausgewählt haben, ergibt das die in Tabelle 10.8 dargestellte Ausgabe.

Tabelle 10.8. Zusammenhangsmaße für ordinalskalierte Daten

Richtungsmaße

			Wert	Asymptotischer Standardfehler[a]	Näherungsweises T[b]	Näherungsweise Signifikanz
Ordinal- bzgl. Ordinalmaß	Somers-d	Symmetrisch	-,397	,043	-9,089	,000
		INGLEHART-INDEX abhängig	-,432	,047	-9,089	,000
		Schulbildung umkodiert abhängig	-,368	,040	-9,089	,000

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

Symmetrische Maße

		Wert	Asymptotischer Standardfehler[a]	Näherungsweises T[b]	Näherungsweise Signifikanz
Ordinal- bzgl. Ordinalmaß	Kendall-Tau-b	-,399	,043	-9,089	,000
	Kendall-Tau-c	-,405	,045	-9,089	,000
	Gamma	-,568	,056	-9,089	,000
	Korrelation nach Spearman	-,453	,048	-8,590	,000[c]
Intervall- bzgl. Intervallmaß	Pearson-R	-,454	,047	-8,594	,000[c]
Anzahl der gültigen Fälle		287			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

c. Basierend auf normaler Näherung

Es werden alle in der Auswahlbox „Ordinal" angezeigten Maßzahlen ausgegeben. Durch Anklicken des Kontrollkästchens „Korrelationen" kann man zusätzlich den Spearmansche Rangkorrelationskoeffizienten („Korrelation nach Spearman") und den Pearsonsche Produkt-Moment-Korrelationskoeffizienten („Pearson-R") anfordern. Letzterer verlangt Intervallskalenniveau und ist hier unangebracht. Alle Koeffizienten weisen einen negativen Wert aus. Es besteht also eine negative Korrelation zwischen Bildungshöhe und Materialismus. Höhere Werte für Bildung ergeben niedrigere Werte (die postmateralistische Einstellung anzeigen) auf dem Inglehart-Index. Es handelt sich um einen Zusammenhang mittlerer Stärke. Wie man sieht, variieren die Maßzahlen in der Größenordnung etwas. Die Koeffizienten

schwanken zwischen etwa 0,4 und 0,45. Nur Gamma weist einen deutlich höheren Wert aus. Weiter ist tau-c tau-b gegenüber vorzuziehen, da wir es nicht mit einer quadratischen Tabelle zu tun haben. Da mit Sicherheit eine Reihe von Bindungen vorliegt, ist auch an Somers d zu denken. Hier ist die asymmetrische Variante mit INGLEHART-INDEX als abhängiger Variablen angebracht, da eindeutig ist, welche Variable die unabhängige und welche die abhängige ist. Für alle Koeffizienten ist auch ein „Asymptotischer Standardfehler" ausgewiesen, so dass man Konfidenzintervalle berechnen kann. Für alle Koeffizienten wurde weiter ein auf einem näherungsweisen T aufbauender Signifikanztest durchgeführt. Wie wir sehen, sind die Werte hoch signifikant. Es ist also so gut wie ausgeschlossen, dass in Wirklichkeit kein Zusammenhang zwischen den beiden Variablen besteht. Der ebenfalls erwähnte Mantel-Haenszel Chi-Quadrat-Wert wird nur bei Anforderung der Chi-Quadrat-Statistik in der Reihe „Zusammenhang linear-mit-linear" angegeben. Es handelt sich um einen Signifikanztest für ordinalskalierte Daten. Das Ergebnis zeigt die Tabelle 10.9. Wie man sieht, ist das Ergebnis auch nach diesem Test hoch signifikant.

Tabelle 10.9. Ergebnis des Mantel-Haenszel-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Zusammenhang linear-mit-linear	58,862	1	,000

10.3.3 Zusammenhangsmaße für intervallskalierte Variablen

Wenn beide Variablen auf Intervallskalenniveau gemessen werden, steht als zusätzliche Information der Abstand zwischen den Werten zur Verfügung. Eine Reihe von Maßen nutzt diese Information. Das bekannteste ist der *Pearsonsche Produkt-Moment-Korrelations-Koeffizient* r. Es ist ein Maß für Richtung und Stärke einer *linearen* Beziehung zwischen zwei Variablen. Die Definitionsformel lautet:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y - \bar{y})^2}} \qquad (10.19)$$

Ein negatives Vorzeichen zeigt eine negative Beziehung, ein positives Vorzeichen eine positive Beziehung zwischen zwei Variablen an. 1 steht für eine vollkommene Beziehung, 0 für das Fehlen einer Beziehung. Bei der Interpretation ist die Voraussetzung der Linearität zu beachten. Für nichtlineare Beziehungen ergibt r ein falsches Bild (⇨ Kap. 16).

Beispiel. Es soll der Zusammenhang zwischen Alter und Einkommen untersucht werden (Datei: ALLBUS90.SAV). Beide Variablen sind auf Rationalskalenniveau gemessen. Als Zusammenhangsmaß bietet sich daher Pearsons r an.

Um die gewünschte Statistik zu berechnen, gehen Sie wie oben beschrieben vor und wählen im Unterschied dazu nun in der Dialogbox „Kreuztabellen: Statistiken" das Kontrollkästchen „Korrelationen". Für die genannten Variablen ergibt sich der in Tabelle 10.10 gekürzt dargestellte Output.

Tabelle 10.10. Ausgabe von Korrelationskoeffizienten

Symmetrische Maße

		Wert	Asymptotischer Standardfehler[a]	Näherungsweises T[b]	Näherungsweise Signifikanz
Intervall- bzgl. Intervallmaß	Pearson-R	-,122	,067	-1,464	,145[c]
Anzahl der gültigen Fälle		143			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

c. Basierend auf normaler Näherung

Der Pearsonsche Korrelationskoeffizient r = – 0,12234 zeigt eine leichte negative Beziehung zwischen Alter und Einkommen an. Allerdings ist diese nach der Angabe in der Spalte „Näherungsweise Signifikanz" nicht signifikant. Ebenso kann man aus der Angabe des „Asymptotischer Standardfehlers" leicht ein Konfidenzintervall etwa für das Signifikanzniveau 95 % errechnen. Man sieht, dass es den Wert Null einschließt.

Eta ist ein spezieller Koeffizient für den Fall, dass die unabhängige Variable auf Nominalskalenniveau gemessen wurde, die abhängige aber mindestens auf Intervallskalenniveau. Er zeigt an, wie sehr sich die Mittelwerte für die abhängige Variable zwischen den verschiedenen Kategorien der unabhängigen unterscheiden. Unterscheiden sie sich gar nicht, wird eta 0. Unterscheiden sie sich dagegen stark und ist zudem die Varianz innerhalb der Kategorien der unabhängigen Variablen gering, tendiert er gegen 1. Wenn die abhängige Variable (die intervallskalierte) die Spalten definiert, lautet die Formel:

$$\text{Eta} = \sqrt{1{,}0 - \frac{\sum_{i=\text{niedr}}^{\text{höchst}} \left\{ \sum_{j=\text{niedr}}^{\text{höchst}} f_{ij}j^2 - \left[\left(\sum_{j=\text{niedr}}^{\text{höchst}} f_{ij}j \right) \Big/ \left(\sum_{j=\text{niedr}}^{\text{höchst}} f_{ij} \right) \right] \right\}}{\sum_{i=\text{niedr}}^{\text{höchst}} \sum_{j=\text{niedr}}^{\text{höchst}} f_{ij}j^2 - \left[\left(\sum_{i=\text{niedr}}^{\text{höchst}} \sum_{j=\text{niedr}}^{\text{höchst}} f_{ij}j \right)^2 \Big/ n \right]}} \qquad (10.20)$$

Dabei ist f_{ij} die Zahl der Fälle in der Reihe i und der Spalte j, *niedr* ist der niedrigste Wert, *höchst* der höchste Wert der betreffenden Variablen.

Eta-Quadrat gibt den Anteil der Varianz der abhängigen Variablen an, der durch die unabhängige Variable erklärt wird.

Betrachten wir die Abhängigkeit des Einkommens vom Geschlecht. Da Geschlecht auf Nominalskalenniveau gemessen wird, kommt eta als Zusammenhangsmaß infrage. Um eta zu berechnen, gehen Sie wie oben beschrieben vor. Im Unterschied dazu wählen Sie nun aber in der Dialogbox „Kreuztabellen: Statistik"

in der Gruppe mit der Bezeichnung „Nominal bezüglich Intervall" das Kontrollkästchen „Eta". Tabelle 10.11 zeigt den Output für das Beispiel.

Tabelle 10.11. Ausgabe bei Auswahl von Eta

Richtungsmaße

			Wert
Nominal- bzgl. Intervallmaß	Eta	EINK abhängig	,414
		GESCHL abhängig	,687

Da „Geschlecht" die unabhängige Variable ist, ist der Wert für eta mit Einkommen als abhängige Variable relevant. Er zeigt mit 0,414 einen mittelstarken Zusammenhang an. Eta^2 ist $(0{,}414)^2 = 0{,}171$, d.h. etwa 17 % der Varianz des Einkommens wird durch das Geschlecht erklärt.

10.3.4 Spezielle Maße

Kappa-Koeffizient (Übereinstimmungsmaß kappa). Um die Gültigkeit und/oder Zuverlässigkeit von Messinstrumenten zu überprüfen, wird häufig die Übereinstimmung von zwei oder mehr Messungen desselben Sachverhaltes ermittelt. Es kann sich dabei z.B. um die Übereinstimmung von zwei Beobachtern handeln oder von zwei verschiedenen Personen, die dieselben Daten kodieren. Es kann auch um die Übereinstimmung zu verschiedenen Zeitpunkten gemachter Angaben zu einem invarianten Sachverhalt gehen oder um den Vergleich der Ergebnisse zweier verschiedener Messverfahren.

Beispiel. In ihrem Buch „Autoritarismus und politische Apathie" (1971) gibt Michaela von Freyhold auf S. 47 eine Tabelle an, aus der hervorgeht, wie dieselben Untersuchungspersonen auf der Dimension Autoritarismus von den Interviewern (denen diese persönlich bekannt waren) zunächst vor dem Interview und später aufgrund der Autoritarismus(A)-Skala eingestuft wurden. Die Übereinstimmung dieser beiden Messungen soll als Nachweis der Gültigkeit der A-Skala dienen. Wir haben Tabelle 10.12 aus diesen Angaben errechnet (Datei A-SKALA.SAV). Dazu wurden zunächst aus den Prozentzahlen und der Fallzahl in den Spalten die Absolutwerte für die einzelnen Zellen ermittelt und außerdem die „tendenziell Autoritären" mit „ausgesprochen Autoritären" sowie „tendenziell Liberale" mit „absolut Liberalen" jeweils zu einer Kategorie zusammengefasst.

Als einfaches Maß kann man einfach den Anteil der beobachteten übereinstimmenden Einstufungen an allen Einstufungen verwenden.

$\ddot{U} = \frac{M}{N}$, wobei M = Zahl der Übereinstimmungen und N = Zahl der Vergleiche.

Im Beispiel stimmen 88 + 102 = 190 Einstufungen von 252 überein. Der Anteil der richtigen Einstufungen beträgt also: 190 : 252 = 0,75.

Tabelle 10.12. Einstufungen nach der A-Skala und dem Interviewereindruck

SKALA * INTERV Kreuztabelle

			INTERV		
			Autoritär	Liberal	Gesamt
SKALA	autoritär	Anzahl	88	36	124
		% von INTERV	77,2%	26,1%	49,2%
	liberla	Anzahl	26	102	128
		% von INTERV	22,8%	73,9%	50,8%
Gesamt		Anzahl	114	138	252
		% von INTERV	100,0%	100,0%	100,0%

Kappa ist ein etwas komplizierteres Übereinstimmungsmaß. Es stellt in Rechnung, dass auch bei zufälliger Zuordnung ein bestimmter Anteil an Übereinstimmungen zu erwarten ist. Deshalb ist auf die Qualität des Messverfahrens nur der darüber hinausgehende Anteil der Übereinstimmungen zurückzuführen. Dieser darf allerdings nur auf den nicht schon per Zufall erreichbaren Übereinstimmungsanteil bezogen werden. Die Formel lautet entsprechend:

$$\text{kappa} = \frac{\ddot{U} - \ddot{U}_E}{1 - \ddot{U}_E} \tag{10.21}$$

Dabei ist $\ddot{U}$ der Anteil der tatsächlich beobachteten Übereinstimmungen, $\ddot{U}_E$ der der erwarteten Übereinstimmungen. Der Anteil der erwarteten Übereinstimmungen errechnet sich:

$$\ddot{U}_E = \sum_{i=1}^{k} (p_i)^2 \tag{10.22}$$

Dabei ist p_i der relative Anteil der einzelnen Ausprägungen an der Gesamtzahl der Fälle und k die Zahl der Ausprägungen.

Um kappa zu berechnen, gehen Sie wie oben beschrieben vor, wählen aber im Unterschied dazu nun in der Dialogbox „Kreuztabellen: Statistik" das Kontrollkästchen „Kappa".

Tabelle 10.13. Ausgabe bei Auswahl des Kappa-Koeffizienten

Symmetrische Maße

		Wert	Asymptotischer Standardfehler[a]	Näherungsweises T[b]	Näherungsweise Signifikanz
Zustimmungsmaß	Kappa	,507	,054	8,077	,000
Anzahl der gültigen Fälle		252			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

Für unser Beispiel sehen Sie das Ergebnis in Tabelle 10.13. Kappa beträgt 0,507. Das ist zwar eine mittlere Korrelation, für den Nachweis der Gültigkeit einer Messung reicht diese aber kaum aus. Bei der Interpretation ist zu berücksichtigen, dass

kappa nur für Nominaldaten sinnvoll ist, weil nur die vollständige Übereinstimmung zweier Messungen verwendet wird, nicht aber eine mehr oder weniger große Annäherung der Werte. Für höher skalierte Daten sollte man entsprechend andere Zusammenhangsmaße wählen. Auch ist zu beachten, dass die gemessene Übereinstimmung stark von der Kategorienbildung abhängig ist. Hätten wir z.B. die vier Kategorien von Freyholds beibehalten, wäre zunächst einmal der Anteil der beobachteten Übereinstimmungen wesentlich kleiner ausgefallen. Aber auch kappa hätte einen viel geringeren Wert angenommen. Der asymptotische Standardfehler beträgt 0,54. Man kann daraus ein Konfidenzintervall für kappa berechnen. Da bei Überprüfungen von Gültigkeit und Zuverlässigkeit immer ein sehr hoher Zusammenhang gewünscht wird, ist der Nachweis, dass ein Wert signifikant der Nullhypothese widerspricht, aber wenig aussagekräftig. Das hält der Forscher normalerweise für selbstverständlich. Deshalb ist die Warnung des SPSS-Handbuchs davor, den Standardfehler hier nur mit Vorsicht für einen Signifikanztest zu verwerten zwar richtig, man sollte aber eher ganz darauf verzichten. Wenn man etwas prüfen sollte, dann, ob die Messgenauigkeit mit hoher Sicherheit ein sinnvolles unteres Niveau nicht unterschreitet.

Risikoeinschätzung in Kohortenstudien. Letztlich kann man mit den Kreuztabellen-Statistiken einen Risikokoeffizienten *(Relatives Risiko)* berechnen. Er gibt an, um das Wievielfache höher oder geringer gegenüber dem Durchschnitt das relative Risiko für eine bestimmte Gruppe ist, dass ein bestimmtes Ereignis eintritt. Dieses Maß ist sowohl für prospektive oder Kohortenstudien als auch für retrospektive oder Fall-Kontrollstudien gedacht, muss jedoch jeweils dem Design der Studie entsprechend verwendet werden. Auf jeden Fall ist es nur auf 2*2-Tabellen anwendbar.

Kohortenstudien sind Studien, die eine bestimmte, durch ein kohortendefinierendes Ereignis festgelegte Gruppe über einen längeren Zeitraum hinweg verfolgen. Dabei kann u.a. untersucht werden, bei welchen Fällen in diesem Zeitraum ein bestimmtes Ereignis (Risiko) eintritt. Das könnte eine bestimmte Krankheit, aber ebenso eine Heirat, die Geburt eines Kindes, Arbeitslosigkeit o.ä. sein. Das Interesse gilt der Frage, ob dieses Risiko sich zwischen verschiedenen Kategorien einer unabhängigen Variablen unterscheidet.

Beispiel. Als Beispiel entnehmen wir dem ALLBUS von 1990 eine Kohorte. Das kohortendefinierende Ereignis ist die Geburt zwischen den Jahren 1955 und 1960. Die Kohorte wurde durch ihr bisheriges Leben, also 30-35 Jahre lang, verfolgt, und es wurde festgestellt, wer in diesem Zeitraum der Versuchung, einen Kaufhausdiebstahl zu begehen, mindestens einmal unterlegen ist (Datei DIEB1.SAV). Es sollte zunächst untersucht werden, ob – wie allgemein in den Sozialwissenschaften angenommen – das Risiko, dass dies passiert, bei Personen aus niedrigeren Herkunftsschichten größer ist. Die Herkunftsschicht wird durch die Schulbildung des Vaters operationalisiert. Die folgende Tabelle (10.14) enthält die vermutete unabhängige Variable „Soziale Herkunft" als Zeilenvariable und das untersuchte Risiko „Diebstahl mindestens einmal begangen" als zweiten Wert der Spaltenvariablen. SPSS erwartet für die Berechnung des Risikokoeffizienten diese Anordnung der Variablen. Außerdem muss die Gruppe mit dem höheren Risiko als erste Zeile er-

scheinen (!). Dies hat mit dem benutzen Algorithmus zu tun. Da der erste Testlauf im Gegensatz zur Hypothese ergab, dass nicht die Kinder von Vätern mit geringerer, sondern die mit Vätern höherer Schulbildung eher einmal einen Kaufhausdiebstahl begehen, musste eine entsprechende Vorkehrung getroffen werden. Die Kinder mit Vätern höherer Schulbildung mussten in die erste, diejenigen mit niedrigerer Schulbildung in die zweite Zeile eingetragen werden.

Man kann nun für den 30 bis 35-jährigen Zeitraum die Vorkommensrate für das untersuchte Ereignis errechnen. Sie beträgt bei Personen, deren Vater eine Hauptschul- oder geringere Ausbildung erfahren hat, 48 von 251, also 0,19. Dagegen beträgt sie für Kinder eines Vaters mit erweiterter Schulbildung 30 von 97, also 0,31. Aufgrund dieser Daten muss zunächst einmal die Hypothese revidiert werden, denn nicht bei den Kindern von Eltern mit geringerer Schulbildung, sondern bei denen mit höherer liegt das größere Risiko. Vergleicht man nun die beiden Gruppen, so dass geprüft werden kann, um wieviel höher das Risiko der Kinder aus den besser gebildeten Schichten gegenüber denjenigen aus den geringer gebildeten ist, errechnet sich 0,31:0,19 = 1,6316. Das Risiko der Kinder aus den höher gebildeten Familien, einen Kaufhausdiebstahl zu begehen, ist 1,63 mal so hoch wie das der Kinder, deren Väter geringere formale Bildung haben.

Tabelle 10.14. Häufigkeit eines Kaufhausdiebstahls nach sozialer Herkunft

Schulabschluß Vater * Kaufhausdiebstahl Kreuztabelle

Anzahl

		Kaufhausdiebstahl		
		Nein	Ja	Gesamt
Schulabschluß Vater	höherer Schulabschluß	67	30	97
	Hauptschule und weniger	203	48	251
Gesamt		270	78	348

SPSS dividiert bei der Berechnung immer die Risikowahrscheinlichkeit der ersten Reihe durch die der zweiten. Deshalb sollte die Gruppe mit dem höheren Risiko in der ersten Reihe stehen. Dagegen ist es nicht vorgeschrieben, in welcher Spalte das interessierende Risikoereignis steht. Da es dem Programm nicht bekannt ist, berechnet es für alle Spalten die entsprechenden Werte. Der Nutzer muss sich den richtigen Wert heraussuchen (⇨ Tab. 10.15). In unserem Falle steht das Risikoereignis „Diebstahl" in der zweiten Spalte. Deshalb ist unter den beiden Zeilen, die die Ergebnisse für eine Kohortenstudie ausgeben (Beschriftung „Für Kohorten-Analyse Kaufhausdiebstahl") die untere Zeile „Für Kohorten-Analyse Kaufhausdiebstahl = Ja" die zutreffende. Hier wurde die Kategorie 2 der Variablen DIEB zur Berechnung benutzt, also die, in der diejenigen stehen, die tatsächlich einmal einen Diebstahl begangen haben. Wie man sieht, entspricht der ausgewiesene Wert (bis auf Rundungsungenauigkeiten) dem oben berechneten. Dazu wird das Konfidenzintervall angegeben (in diesem Falle schon für ein 95 %-Sicherheitsniveau) und auf die untere und obere Grenze des Intervalls umgerechnet. Mit 95 %iger Si-

cherheit ist demnach das Risiko, einen Kaufhausdiebstahl zu begehen, bei Kindern von Vätern mit besserer Schulbildung zwischen 1,093 und 2,392 mal größer als das anderer Kinder. Da der erste Wert über 1 liegt, ist es ziemlich sicher, dass ein Unterschied zwischen den beiden Gruppen tatsächlich besteht. Allerdings ist der Unterschied möglicherweise nur minimal. (Die zweite Zeile stellt die Fragestellung quasi um, errechnet das Risiko, keinen Kaufhausdiebstahl zu begehen. Dies ist bei Kindern von Vätern mit höherem Schulabschluss geringer. Die Quote beträgt 0,854. Der obere Wert Quotenverhältnis für Schulabschluss des Vaters setzt diese beiden Quoten in Beziehung. 0,054 : 1,617 = 0,528. D.h. das relative Risiko der Kinder von Vätern mit höherer Schulbildung, keinen Diebstahl zu begehen ist nur etwa halb so groß wie ihr relatives Risiko, einen solchen zu begehen.

Tabelle 10.15. Relatives Risiko für einen Kaufhausdiebstahl nach sozialer Herkunft

Risikoschätzer

	Wert	95%-Konfidenzintervall	
		Untere	Obere
Quotenverhältnis für Schulabschluß Vater (höherer Schulabschluß / Hauptschule und weniger)	,528	,310	,900
Für Kohorten-Analyse Kaufhausdiebstahl = Nein	,854	,738	,988
Für Kohorten-Analyse Kaufhausdiebstahl = Ja	1,617	1,093	2,392
Anzahl der gültigen Fälle	348		

Um *Relatives Risiko* (risk) für eine Kohortenanalyse zu berechnen, gehen Sie wie folgt vor:

- Vorarbeit: Prüfen Sie, welche Gruppe das größere Risiko trägt. Sorgen Sie gegebenenfalls durch Umkodierung dafür, dass diese die erste der beiden Gruppen wird und damit in der ersten Zeile steht. Ein Änderung der Ausgabereihenfolge über den Format-Befehl reicht nicht.
- Ist das gewährleistet: Wählen Sie in der Dialogbox „Kreuztabellen" die unabhängige Variable (!) als Zeilenvariable und die abhängige Variable (!) als Spaltenvariable aus.
- Wenn Sie lediglich die Statistiken und nicht die Tabelle angezeigt wünschen, wählen Sie das Kontrollkästchen „Keine Tabellen".
- Klicken Sie auf die Schaltfläche „Statistik...".
- Wählen Sie in der Dialogbox „Kreuztabellen: Statistik" das Kontrollkästchen „Risiko".
- Bestätigen Sie mit „Weiter" und „OK".

Risikoabschätzung in Fall-Kontrollstudien. Kohortenstudien verfolgen eine bestimmte Fallgruppe, bei denen das untersuchte Ereignis noch nicht eingetreten ist, über einen gewissen Zeitraum hinweg und stellen fest, bei welchen Fällen das Ereignis eintritt, bei welchen nicht, gegebenenfalls auch den Zeitpunkt. Fall(Case)-Kontrollstudien gehen umgekehrt vor. Sie nehmen eine Gruppe, bei denen das Ereignis eingetreten ist und vergleichen sie mit einer – mit Ausnahme des kritischen

Ereignisses – im wesentlichen gleich zusammengesetzten Kontrollgruppe. Normalerweise ist diese Kontrollgruppe in etwa gleich groß wie die Fallgruppe. Es soll festgestellt werden, ob sich diese beiden Gruppen auch hinsichtlich weiterer Variablen, die als Ursachen für das kritische Ereignis in Frage kommen, unterscheiden. Retrospektiv sind sie, da man zurückblickend mögliche Ursachenfaktoren untersucht. Handelt es sich um konstante Faktoren, kann man auch den aktuellen Wert verwenden.

Beispiel. Als Beispiel soll aus den ALLBUS von 1990 mit Daten in Form einer Fall-Kontrollstudie die Frage geklärt werden, ob die Wahrscheinlichkeit, einmal einen Kaufhausdiebstahl zu begehen, von der Herkunftsschicht abhängt. Die Herkunftsschicht wird, wie oben, über die Schulbildung des Vaters operationalisiert (Datei DIEB2.SAV). Unser Beispiel entstammt keiner tatsächlichen Kontrollstudie, sondern einer normalen Umfrage, könnte deshalb auch wie üblich ausgewertet werden. Es soll hier aber nach der Art einer Fall-Kontrollstudie geschehen. Wir betrachten also diejenigen, die schon einmal einen Kaufhausdiebstahl begangen haben, als die Gruppe, bei der das interessierende Risiko eingetreten ist. Alle anderen, bei denen das nicht der Fall war, werden als Kontrollgruppe verwendet. Eine entsprechende Tabelle sieht wie folgt aus:

Tabelle 10.16. Tabelle nach Art einer Fall-Kontrollstudie

Kaufhausdiebstahl * Schulabschluß Vater Kreuztabelle

			Schulabschluß Vater		
			1,00 höhere Schule	2,00 Hauptschule und weniger	Gesamt
Kaufhausdiebstahl	1,00 Ja	Anzahl	180	278	458
		% von Kaufhausdiebstahl	39,3%	60,7%	100,0%
	2,00 Nein	Anzahl	545	1685	2230
		% von Kaufhausdiebstahl	24,4%	75,6%	100,0%
Gesamt		Anzahl	725	1963	2688
		% von Kaufhausdiebstahl	27,0%	73,0%	100,0%

SPSS erwartet, dass die interessierende Risikovariable (die abhängige Variable) bei einer Fall-Kontrollstudie als Zeilenvariable benutzt wird, die mögliche ursächliche Variable als Spaltenvariable. Außerdem muss das untersuchte Ereignis (der „Fall") in der ersten Zeile stehen. Deshalb wird hier bei Diebstahl der Wert 1 (=Ja) vor dem Wert 2 (=Nein) ausgewiesen. Es reicht in diesem Falle auch nicht aus, nur mit dem Formatbefehl die Ausgabereihenfolge zu ändern, das interessierende Ereignis muss wirklich mit dem Wert 1 verkodet werden. Ebenso muss das ursächliche Ereignis (die Ausprägung, die das Risiko wahrscheinlich erhöht), in der ersten Spalte stehen. Deshalb wird die Gruppe der Personen, deren Väter höheren Schulabschluss haben, in der ersten Spalte ausgewiesen, die anderen in der zweiten. In einer normalen Studie würde man die Daten spaltenweise prozentuieren, in der Kontrollstudie geschieht das dagegen zeilenweise. Man behandelt also die eigentliche Wirkungsvariable anders als sonst wie eine unabhängige Variable, die Ursa-

chenvariable dagegen wie eine abhängige. Man untersucht ja von der Wirkung ausgehend, ob es evtl. Unterschiede hinsichtlich möglicher unabhängiger Variablen gibt. Man kann der Tabelle etwa entnehmen, dass von denjenigen, die schon einmal einen Kaufhausdiebstahl begangen haben, 60,7 % Väter mit Hauptschul- oder geringerem Abschluss haben und 39,3 % Väter mit höherem Schulabschluss. Bei denjenigen, die keinen Diebstahl begangen haben, sind 75,6 % aus der Schicht mit geringerer Bildung und 24,4 % aus der mit höherer. Bei der Interpretation so gewonnener Daten muss man sehr vorsichtig sein. So bedeutet die Tatsache, dass ein größerer Anteil der „Diebe" aus der Schicht mit geringerer Bildung stammt, keinesfalls, dass Kinder aus dieser Schicht relativ häufiger Diebstähle begehen. Sie sind ja auch in der Gruppe, die keine Diebstähle begangen hat, stärker vertreten als die anderen. Das liegt ganz einfach daran, dass diese Schicht insgesamt zahlreicher ist. Man muss ihren Anteil vielmehr mit dem Anteil an der Untersuchungsgruppe insgesamt vergleichen. Dieser ist bei den Kindern aus der niedrigeren Bildungsschicht 73,0 %. An denjenigen, die einmal gestohlen haben, ist der Anteil dagegen nur 60,7 %, also geringer. Solche Interpretationsprobleme ergeben sich bei der normalen Prozentuierung nicht, aber Fall-Kontrollstudien lassen sie eben häufig nicht zu. Dazu kommt oft noch das weitere Problem, dass der Anteil der Gruppen, die die Werte der unabhängigen Variablen repräsentieren, an der Grundgesamtheit nicht bekannt ist. Dann ist eine sinnvolle Interpretation oft gar nicht möglich.

Eine leichtere Interpretation erlaubt wiederum ein Risiko-Koeffizient. Wir können aber die relative Risikorate nicht auf dieselbe Weise berechnen wie bei der Kohortenstudie. Stattdessen verwenden wir die sogenannte *odds-ratio*, das Verhältnis der Anteile der Gruppen der unabhängigen Variablen an den Gruppen der Untersuchungsvariable. So ist das Verhältnis der Diebe aus der höheren Bildungsschicht zu den Dieben aus der Schicht mit geringerer Bildung der Väter 180 : 278 = 0,6475, das Verhältnis der „Nicht-Diebe" aus dieser Schicht zu den „Nicht-Dieben" unter den Personen mit niedrigerer Bildung des Vaters 545 : 1685 = 0,3234. Die odds-ratio ist entsprechend 0,6475 : 0,3234 = 2,002. Der Anteil der Personen mit besser gebildeten Vätern an den Kaufhausdieben ist als ca. zweimal so hoch wie ihr Anteil an den Personen, die noch keinen Kaufhausdiebstahl begangen haben.

Um *Relatives Risiko* für eine Fall-Kontrollstudie zu berechnen, gehen Sie wie folgt vor:

- Vorarbeit: Prüfen Sie, welche Gruppe das größere Risiko trägt. Gegebenenfalls kodieren Sie die unabhängige Variable so um, dass diese Gruppe an erster Stelle steht (und später die erste Zeile bildet).
- Kodieren Sie gegebenenfalls die abhängige Variable so um, dass das interessierende Merkmal an erster Stelle steht.
- Ist das gewährleistet: Wählen Sie in der Dialogbox „Kreuztabellen" die unabhängige (!) Variable als Spaltenvariable und die abhängige (!) Variable als Zeilenvariable aus.
- Klicken Sie die Schaltfläche „Zellen..." an, und wählen Sie in der sich öffnenden Dialogbox „Kreuztabellen: Zellen anzeigen" in der Auswahlbox „Prozentwerte" für die Prozentuierung „Zeilenweise" an.

- ▷ Klicken Sie auf die Schaltfläche „Statistik...".
- ▷ Wählen Sie in der Dialogbox „Kreuztabellen: Statistik" das Kontrollkästchen „Risiko".
- ▷ Bestätigen Sie mit „Weiter" und „OK".

Für das gewählte Beispiel ergibt der Output, neben der bereits angeführten Tabelle 10.16, die Tabelle 10.17.

Tabelle 10.17. Ausgabe bei einer Fall-Kontrollstudie

Risikoschätzer

	Wert	95%-Konfidenzintervall	
		Untere	Obere
Quotenverhältnis für Kaufhausdiebstahl (Ja / Nein)	2,002	1,621	2,472
Für Kohorten-Analyse Schulabschluß Vater = höhere Schule	1,608	1,405	1,841
Für Kohorten-Analyse Schulabschluß Vater = Hauptschule und weniger	,803	,743	,868
Anzahl der gültigen Fälle	2688		

Relevant ist die Zeile „Quotenverhältnis für Kaufhausdiebstahl (Ja / Nein)". Sie enthält die odds-ratio. Sie beträgt, wie berechnet, 2,002. Weiter sind die Grenzen des Konfidenzintervalls bei 95 %igem Sicherheitsniveau angegeben. Die untere Grenze beträgt 1,621, die obere 2,472. Selbst wenn man die untere Grenze annimmt, liegt also die relative Häufigkeit von Kaufhausdiebstählen durch Personen, deren Väter der höheren Bildungsschicht zugehören, deutlich über der von Personen aus niederer Bildungsschicht.

McNemar. Schließlich wird in „Kreuztabellen: Statistik" auch der McNemar-Test angeboten. Er fällt insofern etwas aus der Reihe, als es sich nicht um ein Zusammenhangsmaß, sondern einen Signifikanztest handelt. Gedacht ist er für Vorher-Nachher-Designs mit dichotomen Variablen. Anwendbar ist er aber auch auf quadratische Tabellen mit gleichen Ausprägungen auf den gekreuzten Variable. (⇨ Kap. 22.5.3)

10.3.5 Statistiken in drei- und mehrdimensionalen Tabellen

Das Menü „Kreuztabellen" ermittelt immer nur statistische Maßzahlen für zweidimensionale Tabellen. Werden zusätzliche Kontrollvariablen eingeführt, können ebenfalls die statistischen Maßzahlen angefordert werden. Diese gelten aber nicht für die gesamte mehrdimensionale Tabelle, sondern die Gesamttabelle wird in zweidimensionale Untertabellen zerlegt. Bei dreidimensionalen Tabellen entsteht z.B. für jede Ausprägung der Kontrollvariablen eine eigene Untertabelle. Für jede dieser Untertabellen werden die statistischen Maßzahlen getrennt ermittelt.

Eine Ausnahme bilden die mit der Option „Cochran- und Mantel-Haenszel-Statistik" anzufordernden Verfahren. Die dort ausgegebenen Signifikanztest prüfen

bei mehr als zweidimensionalen Tabellen die Signifikanz des Zusammenhangs zwischen zwei Variablen insgesamt unter Beachtung der Kontrollvariablen. Unabhängige und abhängige Variablen müssen aber dichotomisiert vorliegen, d.h. es wird die Signifikanz des Zusammenhangs dieser beiden Variablen in einer Schar zweidimensionaler Tabellen überprüft. Häufig wird dieser Test im Zusammenhang mit Kohortenstudien oder Fall-Kontrollstudien verwendet. Das relative Risiko spielt hier eine zentrale Rolle.

Beispiel: Wir greifen auf die Datei DIEB1 zurück, aus der eine Risikoabschätzung von Personen aus zwei sozialen Herkunftsschichten, einen Kaufhausdiebstahl zu begehen abgeschätzt wurde. Der Risikoquotient für den Schulabschluss des Vaters betrug 0,528. Wir könnten diese Analyse nun durch Auswahl von „Cochran- und Mantel-Haenszel-Statistik" um einen Signifikanztest ergänzen, wollen aber zusätzlich noch eine dritte Variable (Schichtvariable) GESCHL einführen.

Tabelle 10.18. Ausgabe bei Cochran- und Mantel-Haenszel-Statistik

Tests auf Homogenität des Quotenverhältnisses

Statistik		Chi-Quadrat	df	Asymptotische Signifikanz (zweiseitig)
Bedingte Unabhängigkeit	Cochran	5,431	1	,020
	Mantel-Haenszel	4,755	1	,029
Homogenität	Breslow-Day	,449	1	,503
	Tarone	,449	1	,503

Schätzung des gemeinsamen Quotenverhältnisses nach Mantel-Haenszel

Schätzung			,534
ln(Schätzung)			-,628
Standardfehler von ln(Schätzung)			,272
Asymptotische Signifikanz (zweiseitig)			,021
Asymptotisches 95% Konfidenzintervall	Gemeinsames Quotenverhältnis	Untergrenze	,313
		Obergrenze	,910
	ln(gemeinsames Quotenverhältnis)	Untergrenze	-1,161
		Obergrenze	-,094

Die Schätzung des gemeinsamen Quotenverhältnisses nach Mantel-Haenszel ist unter der Annahme des gemeinsamen Quotenverhältnisses von 1,000 asymptotisch normalverteilt. Dasselbe gilt für den natürlichen Logarithmus der Schätzung.

- Laden Sie die Datei DIEB1 und wählen Sie „Analysieren", „Deskriptive Statistiken" und „Kreuztabellen".
- Übertragen Sie in der Dialogbox „Kreuztabellen" die Variable „VATER in das Feld „Zeilen" und DIEB in das Feld „Spalten" sowie GESCHL in das Feld „Schicht 1 von1".

▷ Öffnen Sie durch Anklicken von „Statistik" die Dialogbox „Kreuztabellen: Statistik" und klicken Sie das Auswahlkästchen „Cochran- und Mantel-Haenszel-Statistik" an. Es erscheint die in Tabelle 10.18 etwas gekürzt dargestellt Ausgabe.

In den ersten beiden Zeilen der oberen Tabelle finden Sie die Ausgabe der Cochran und Mantel-Haenszel-Statistik. Es handelt sich um zwei gleichwertige Signifikanztests, wobei letzterer für kleinere Stichproben Korrekturen vornimmt. Beide zeigen, dass auf dem 5%Niveau (auch bei Beachtung der Kontrollvariablen GESCHL) der Zusammenhang zwischen DIEB und VATER signifikant ist.
Die beiden unteren Zeilen prüfen die Homogenität der Quotenverhältnisse und ergeben beide, dass die Hypothese der Homogenität nicht zu verwerfen ist, d.h. die Quotenverhältnisse könnten bei den beiden Gruppen der Schichtungsvariable GESCHL, den Männern und den Frauen gleich sein.

Die zweite Teiltabelle schließlich schätzt das Quotenverhältnis nach Mantel-Haenszel und kommt mit 0,534 zu einem von dem Ergebnis der einfachen Berechnung leicht abweichenden Wert. Zudem wird für diesen Wert ein 95%-Konfidenzintervall angegeben. Der wahre Wert liegt mit 95%-Wahrscheinlichkeit zwischen 0,313 und 0,910. (Ergänzend erscheint der Natürliche Logarithmus des Quotenverhältnisses und die Grenzen des dazu gehörigen Konfidenzintervalls.)

Aus dem Wert „Asymptotische Signifikanz (zweiseitig)" von 0,021 kann man schließen, dass das gefundene Quotenverhältnis signifikant von einem vorgegebenen Quotenverhältnis von 1 abweicht. (Der vorgegebene Wert kann in der Dialogbox „Kreuztabellen: Statistiken" geändert werden. Hätten wir ihn aufgrund von Vorkenntnissen z.B. auf 0,5 gesetzt, wäre das Ergebnis, dass der gefundene Wert nicht signifikant von erwarteten vorgegebenen abweicht.)

11 Fälle auflisten und Berichte erstellen

Das Untermenü „Berichte" enthält vier Menüs, mit denen Datenlisten und Berichte erstellt werden können. Die Auswertungsmöglichkeiten, die diese Menüpunkte bieten, sind weitgehend schon durch andere Optionen abgedeckt. Optisch durch einen Querstrich erkennbar abgegrenzt ist zunächst das Menü „OLAP-Würfel" (Online Analytical Processing) und darauf sind die drei Menüs „Fälle zusammenfassen", „Bericht in Zeilen" und „Bericht in Spalten" zu einer Gruppe zusammengefasst. Diese Menüs erlauben es, interaktive Tabellen zu erstellen, Listen zusammenzustellen und Berichte zu verfassen.

- *OLAP-Würfel.* Mit diesem Menü werden Pivot-Tabellen erstellt, in denen der Nutzer interaktiv zwischen den zu betrachtenden Schichten wählen kann. Der OLAP-Würfel eignet sich gut zur Weitergabe komplexer Datenstrukturen auch an externe Nutzer. SPSS bietet dafür geeignete Zusatzsoftware an.
- *Listen.* Die Menüs „Fälle zusammenfassen..." als „Bericht in Zeilen" ermöglichen es, Datenlisten zu erstellen. Darunter versteht man eine Aufstellung der Variablenwerte für die einzelnen Fälle einer Untersuchung. Über eine Datenliste verfügt man bereits im Editorfenster. Jedoch können mit den besprochenen Befehlen einzelne Variablen für die Liste ausgewählt werden. Ebenso kann man die Liste auf eine Auswahl der Fälle beschränken. Unterschiedliche Formatierungsmöglichkeiten stehen zur Verfügung. Listen wird man für die Datendokumentation und zur Überprüfung der Korrektheit der Datenübernahme aus externen Programmen verwenden. Auch zur Fehlersuche sind sie geeignet.
- *Zusammenfassende Berichte.* Darunter versteht man die Darstellung zusammenfassende Maßzahlen für Subgruppen in einer Tabelle. Solche Berichte können mit allen vier Menüs erstellt werden. Dabei werden Maßzahlen berechnet, wie sie in den Unterprogrammen „Deskriptive Statistiken", „Häufigkeiten" und „Mittelwerte vergleichen" ebenfalls geboten werden. Gegenüber diesen Programmen haben die hier besprochenen Unterprogramme den Vorteil, dass die Maßzahlen für mehrere Variablen gleichzeitig in einer zusammenfassenden Tabelle dargestellt werden können. Man kann sich dadurch einen leichten Überblick über mehrere charakteristische Variablen für jede interessierende Untergruppe verschaffen. Daneben stehen zahlreiche Formatierungsmöglichkeiten zur Verfügung, die es erlauben, eine präsentationsfähige Ausgabe zu gestalten.
- *Kombinierte Berichte.* In ihnen werden sowohl Datenlisten als auch zusammenfassende Maßzahlen für Gruppen präsentiert. Dies ist möglich mit den Menüs „Fälle zusammenfassen" und „Bericht in Zeilen".

11.1 Erstellen eines OLAP-Würfels

Das Menü „OLAP-Würfel" ist relativ einfach aufgebaut und dient dazu, in Schichten gegliederte Tabellen zu erstellen. Die abhängige Variable(n) (Auswertungsvariablen) müssen auf Intervall- oder Rationalskalenniveau gemessen sein. Für sie werden zusammenfassende Statistiken wie Mittelwerte, Standardabweichung etc. ausgegeben. Die unabhängige(n) Variable(n) (Gruppenvariablen) dagegen muss/müssen kategorialer Art sein, also entweder auf Nominal- oder Ordinalskalenniveau gemessen oder aber durch Klassenbildung in eine begrenzte Zahl von Gruppen aufgeteilt. Die Werte der unabhängigen Variable(n) ergeben die Schichten der Tabelle. Ergebnis ist eine Pivot-Tabelle, die überwiegend dieselben Informationen anbietet wie eine mit dem Menü „Mittelwerte" (⇨ Kap. 13.2) erstellte, allerdings ist per Grundeinstellung immer nur eine Schicht im Vordergrund zu sehen, also nur die Daten einer Gruppe, während per Grundeinstellung im Menü „Mittelwerte" die gesamten Informationen in der Datei zu sehen sind (durch Pivotieren kann diese wechselseitig ineinander übergeführt werden). Die verfügbaren Statistiken sind in beiden Menüs identisch. Der OLAP-Würfel bietet zusätzlich die Möglichkeit der Bildung von Differenzen zwischen Vergleichsgruppen oder Vergleichsvariablen.

Beispiel. Für die Daten von ALLBUS90 solle das Durchschnittseinkommen gegliedert nach Geschlecht ausgegeben werden. Zusätzlich wird die Differenz des Einkommens von Männern und Frauen ermittelt.

- ▷ Wählen Sie „Analysieren", „Berichte" und „OLAP-Würfel". Die Dialogbox „OLAP-Würfel" öffnet sich. Sie ermöglicht lediglich die Auswahl der „Auswertungsvariablen" und der „Gruppenvariablen" (ohne Gruppenvariable ist die Schaltfläche „OK" inaktiv).
- ▷ Übertragen Sie EINK in das Feld „Auswertungsvariablen" und GESCHL in das Feld „Gruppenvariable(n)".
- ▷ Öffnen Sie durch Anklicken der Schaltfläche „Statistiken" die Dialogbox "OLAP-Würfel: Statistiken". Zur Verfügung stehen zahlreiche Lage-, Streuungs-, Schiefe- und Formmaße. Dort können durch Übertragen aus dem Feld „Statistik" in das Feld „Zellenstatistiken" die statistischen Kennzahlen ausgewählt werden, die für die Berichtsvariable berechnet werden sollen. In umgekehrter Richtung wählt man die bereits voreingestellten Kennzahlen ab. Zur Verfügung stehen sind dieselben Statistiken wie im Menü „Mittelwerte", Dialogbox „Mittelwerte: Optionen" (⇨ Abb. 13.2) oder in der Dialogbox „Statistik" von „Fälle zusammenfassen". Zusätzlich dazu findet man hier die Möglichkeit zur Berechnung von „Prozent der Summe in" und „Prozent der Fälle in". Dies wird jeweils ergänzt durch den Namen der Gruppierungsvariablen und gibt die Prozentwerte der ausgewählten Schicht innerhalb der Fälle mit dieser Gruppierungsvariablen an. Außerdem sind per Voreinstellung wesentlich mehr Statistiken ausgewählt als in den Menüs „Mittelwerte vergleichen" und „Fälle zusammenfassen", nämlich „Summe", „Anzahl der Fälle", „Mittelwert", „Standardabweichung", „Prozent der Gesamtsumme", „Prozent der Gesamtzahl". Im Beispiel soll nur „Mittelwert", „Standardabweichung" und „Anzahl der Fälle" als Zellenstatistik ausgewählt werden.

- Durch Klicken auf „Differenzen" öffnet sich die Dialogbox „OLAP-Würfel: Differences" (⇨ Abb. 11.1). Hier können wir festlegen, zwischen welchen Berichtsvariablen oder zwischen welchen Gruppen einer Gruppenvariablen (mehrere Gruppenvariablen können hier nicht gleichzeitig verwendet werden) eine Differenz gebildet werden soll (im Beispiel zwischen den Einkommen der Männer und denen der Frauen). Wenn nur eine Berichtsvariable ausgewählt wurde, ist die Optionsschalter „Differenzen zwischen den Variablen inaktiv", wie in unserem Beispiel. Wir wählen „Differenzen zwischen den Gruppen". Damit wird der untere Teil der Dialogbox „Differenzen zwischen Fallgruppen" aktiv. Sind mehrere Gruppenvariablen angegeben, wäre jetzt die Gruppenvariable auszuwählen, für deren Gruppen Differenzen gebildet werden sollen, in unserem Beispiel ist es GESCHL. In das Feld „Kategorie" wird der Wert der Kategorie eingetragen, von deren Statistik die Statistik der anderen Kategorie „Minus Kategorie" abgezogen werden soll. Im Beispiel soll von Mittelwert des Einkommens der Männer derjenige der Frauen abgezogen werden. Entsprechend ist 1 (für männlich) in das obere, 2 (für weiblich) in das untere Feld einzutragen. In das Feld „Prozentbeschriftung" geben wir ein „Männer minus Frauen in Prozent" und in das Feld „Arithmetische Differenz" „Männer minus Frauen absolut".
- Durch Klicken auf den Pfeil übertragen wir dieses Wertepaar in das Feld „Paare". In der Gruppe „Art der Differenz" kann man auswählen, welche Differenz gebildet werden soll. Zur Verfügung stehen „prozentuale Differenz" und „Arithmetische Differenz". Im Beispiel wählen wir beide.

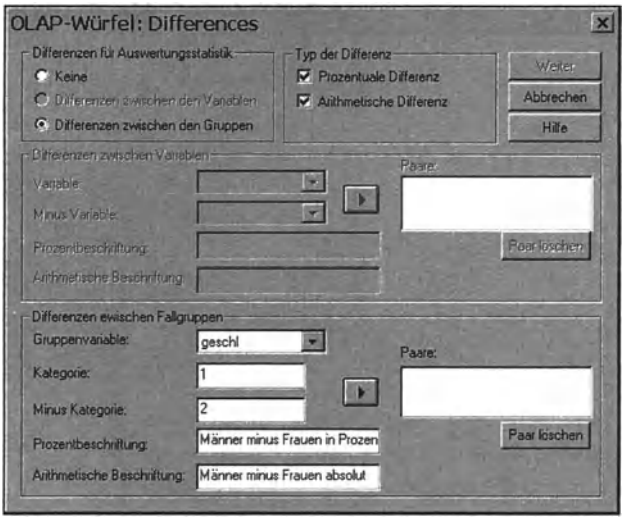


Abb. 11.1. Dialogbox „OLAP-Würfel: Differences"

- Klickt man auf die Schaltfläche „Titel", öffnet sich die Dialogbox „OLAP-Würfel: Titel". Hier kann man im Eingabefeld „Titel" einen Titel für die Ta-

belle eintragen (Voreinstellung „OLAP-Würfel"). Das Feld „Erklärung" dient dazu, einen Text für eine Fußnote der Tabelle zu erstellen.

▷ Bestätigen Sie mit „Weiter" und „OK".[1]

Die Art der Ausgabe macht den Hauptunterschied zu den anderen Menüs aus. Im Menü „OLAP-Würfel" wird immer eine geschichtete Tabelle ausgegeben. D.h. man sieht immer nur die Tabelle für eine Schicht der Gliederungsvariablen, zunächst für die Schicht „insgesamt". Man kann dann nacheinander die verschiedenen Schichten in der Pivot-Tabelle aufrufen (⇨ Kap. 4.1.4).

Tabelle 11.1 zeigt einen Ausschnitt aus der Schicht „insgesamt" eines Berichts mit den Auswertungsvariablen EINK und der Gruppenvariablen GESCHL sowie den Statistiken „Mittelwert", „Standardabweichung" und „Anzahl der Fälle (N)". Die Tabelle ist bereits durch Doppelklicken zum Pivotieren aktiviert. Beim Klicken auf den Pfeil neben „Insgesamt" öffnet sich eine Auswahlliste mit den Namen der Schichten (hier: „Männlich" und „Weiblich" sowie „Männer minus Frauen in Prozent" und „Männer minus Frauen absolut"). Durch Anklicken eines dieser Namen wechselt man in die Schicht der so bezeichneten Gruppe. Betrachten wir die Ergebnis für den Mittelwert des Einkommen in den verschiedenen Schichten, beträgt dieser in der Schicht insgesamt 2096,78, in der Schicht „Männlich" 2506,30 und in der Schicht „Weiblich" 1561,77, „Männer minus Frauen absolut" ergibt 944,42 und in Prozent 60,4%. Bei der Interpretation des letzten Wertes ist zu beachten, dass immer der zweite der eingegebenen Werte als Prozentuierungsbasis benutzt wird, die Differenz beträgt also ca. 60% mittleren Einkommens der Frauen.

Tabelle 11.1. Erste Schicht eines OLAP-Würfels, zum Pivotieren ausgewählt

OLAP-Würfel

GESCHL	Insgesamt

	Mittelwert	Standardabweichung	N
EINK	2096,78	1133,801	143

11.2 Das Menü „Fälle zusammenfassen"

11.2.1 Listen erstellen

Mit dem Menü „Fälle zusammenfassen" können ausgewählte Variablen für alle Fälle oder die ersten x Fälle aufgelistet werden.

Beispiel. Es sollen für die Überprüfung einer Datenübernahme Fälle der Datei ALLBUS90.SAV aufgelistet werden. Dafür soll es ausreichen, die ersten 10 Fälle auszugeben. Außerdem interessiert in einem ersten Durchgang nur eine kleine Zahl von Variablen. Um eine Liste zu erstellen, gehen Sie wie folgt vor:

[1] Prüfen Sie, ob die Beschriftung nach Dialogbox Abb. 11.1. in der Tabelle falsch ausgeführt wird. U.U. landet die Beschriftung für die Prozente bei den Absolutzahlen und umgekehrt.

▷ Wählen Sie die Befehlsfolge „Analysieren“, „Berichte“ und „Fälle zusammenfassen...“. Die Dialogbox „Fälle zusammenfassen“ öffnet sich (⇨ Abb. 11.2).

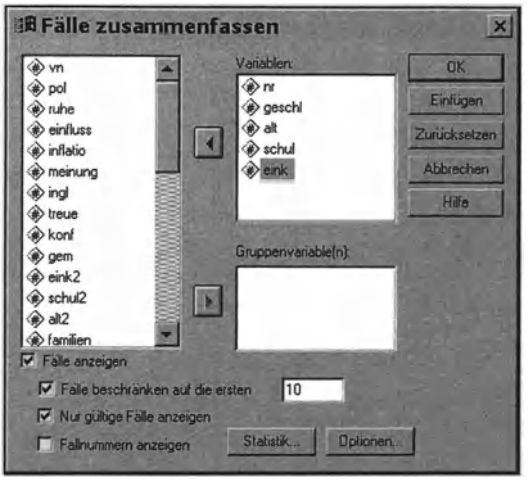


Abb. 11.2. Dialogbox „Fälle zusammenfassen“

▷ Übertragen Sie die interessierenden Variablen aus der Quellvariablenliste in das Feld „Variablen:“. Die Variablen werden später in der Reihenfolge angezeigt, in der Sie sie übertragen.
▷ Sollen nur die ersten x Fälle angezeigt werden, wählen Sie die Option „Fälle beschränken auf die ersten“, und tragen Sie in das Eingabefeld die Nummer des letzten Falles ein (hier: 10).
▷ Markieren Sie das Auswahlkästchen „Fälle anzeigen“. Damit werden die Daten für alle Fälle angezeigt. Ansonsten würden nur Auswertungen für Gruppen angezeigt.

❒ *Fallnummer anzeigen*. Das Anklicken dieses Kontrollkästchens bewirkt, dass eine weitere Variable mit der SPSS-internen Fallnummer ausgegeben wird. (Das wird man nutzen, wenn keine Fallnummern durch den Nutzer vergeben wurden oder diese aus irgendwelchen Gründen weniger übersichtlich sind.)
❒ *Nur gültige Fälle anzeigen*. Es werden nur Fälle ohne fehlende Werte angezeigt.

Das dargestellte Beispiel führt zu dem in Tabelle 11.2 wiedergegebenen Ergebnis. In der ersten Spalte befindet sich die SPSS-interne Nummer, in den folgenden stehen die ausgewählten Variablen in der Auswahlreihenfolge. Jede Spalte ist mit dem Variablennamen überschrieben.

Tabelle 11.2. Ausgabe einer Datenliste

Zusammenfassung von Fällen[a]

		NR	GESCHL	ALT	SCHUL	EINK
1		38	2	61	2	150
2		39	1	42	2	680
3		72	2	89	2	1450
4		76	2	26	3	99997
5		77	2	53	2	99997
6		83	1	34	3	4800
7		87	1	71	2	99997
8		100	1	80	2	2100
9		137	1	36	3	3200
10		138	2	30	3	1700
Insgesamt	N	10	10	10	10	7

a. Begrenzt auf die ersten 10 Fälle.

11.2.2 Kombinierte Berichte erstellen

Man kann mit dem Untermenü „Fälle zusammenfassen" auch eine gruppierte Liste erstellen lassen (kombinierter Bericht). Für die Gruppen können zusammenfassende Statistiken gewählt werden.

Beispiel: Wir benutzen dieselben Daten, gruppieren sie aber nach Geschlecht. Für die Gruppen sollen die Fallzahlen N und das arithmetische Mittel ausgegeben werden.

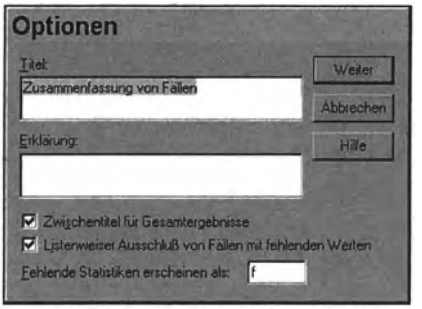


Abb. 11.3. Dialogbox „Optionen"

Dazu gehen sie zunächst wie oben vor:

- ▷ Übertragen Sie aber zusätzlich die Variable „GESCHL" in das Auswahlfeld „Gruppenvariable(n)".
- ▷ Klicken Sie auf die Schaltfläche „Statistik". Die Dialogbox „Zusammenfassung: Statistik" öffnet sich. Übertragen Sie die gewünschten Statistiken aus der Liste „Statistik:" in das Auswahlfeld „Zellenstatistik:" (hier: „Anzahl der Fälle und Mittelwert"). Zur Verfügung stehen zahlreiche Lage-, Streuungs-, Schiefe- und

Formmaße. Es sind dieselben Statistiken wie im Menü „Mittelwerte“, Dialogbox „Mittelwerte: Optionen“ (⇨ Abb. 13.2).

▷ Bestätigen Sie mit „Weiter“.

Außerdem können noch einige Optionen gewählt werden.

▷ Klicken Sie dazu auf die Schaltfläche „Optionen“. Die Dialogbox „Optionen“ öffnet sich (⇨ Abb. 11.3).

Tabelle 11.3. Kombinierter Bericht mit „Fälle zusammenfassen“

Zusammenfassung von Fällen[a]

					NR	ALT	SCHUL	EINK
GESCHL	1	1			39	42	2	680
		2			83	34	3	4800
		3			87	71	2	f
		4			100	80	2	2100
		5			137	36	3	3200
		Insgesamt	N		5	5	5	4
			Mittelwert		89,20	52,60	2,40	2695,00
	2	1			38	61	2	150
		2			72	89	2	1450
		3			76	26	3	f
		4			77	53	2	f
		5			138	30	3	1700
		Insgesamt	N		5	5	5	3
			Mittelwert		80,20	51,80	2,40	1100,00
	Insgesamt	N			10	10	10	7
		Mittelwert			84,70	52,20	2,40	2011,43

a. Begrenzt auf die ersten 10 Fälle.

Im Feld „Titel“ können Sie eine Überschrift für die Tabelle eintragen“ (Voreinstellung: Zusammenfassung von Fällen). Im Feld „Erklärung“ kann der Text einer Fußnote für die Tabelle eingetragen werden. Markiert man das Auswahlkästchen „Zwischentitel für Gesamtergebnisse“ (Voreinstellung), werden die zusammenfassenden Werte für Gruppen durch den Zwischentitel „insgesamt“ markiert, sonst nicht. Markiert man „Listenweiser Ausschluss von Fällen mit fehlenden Werten“, werden in die Berechnung der zusammenfassenden Statistiken nur die Fälle einbezogen, die in keiner der ausgewählten Variablen einen fehlenden Wert ausweisen. Im Feld „Fehlende Statistik erscheint als“ kann eine Zeichenkette eingetragen werden, die bei fehlenden Werten in der Tabelle erscheint (anstelle des nutzerdefinierten Wertes oder des Symbols für systemdefinierten fehlende Werte). In unserem Beispiel wurde „f“ als Symbol verwendet. Das Ergebnis sehen Sie in Tabelle 11.3.

Im Unterschied zu Tabelle 11.2. sind die Fälle nach Geschlecht geordnet. Für Männer und Frauen werden jeweils die Zahl der Fälle N und der Mittelwert als zusammenfassende Statistiken ausgegeben (und zwar für alle Variablen, auch wenn dies bei einigen sachlich unsinnig ist).

Ein einfacher Bericht, der nur die Statistiken der Gruppen enthielte, würde entstehen, wenn man das Auswahlkästchen „Fälle anzeigen" ausschalten würde.

11.3 Erstellen von Berichten in Zeilen oder Spalten

Das Menü „Berichte" enthält weiter die beiden Reportmenüs:

- *„Bericht in Zeilen"* und
- *„Berichte in Spalten"*.

Beide erlauben es, gegliedert nach einer oder mehreren Gliederungsvariablen (Break-Variablen), zusammenfassende Statistiken zu erstellen. Es handelt sich um Statistiken, die für die Beschreibung mindestens intervallskalierter Daten geeignet sind: das arithmetische Mittel, die Streuungsmaße Varianz und Standardabweichung, Schiefe und Steilheitsmaß sowie ergänzende Angaben wie kleinster und größter Wert. Außerdem können Prozentwerte unter oder über bzw. zwischen zwei Grenzwerten ermittelt werden. Der Hauptvorteil der Reports liegt darin, dass es möglich ist, diese Maße für mehrere Variablen in einer Übersichtstabelle parallel auszugeben und mit zahlreichen Formatierungsoptionen sowie durch Beschriftung optisch ansprechend zu gestalten. Die beiden Unterprogramme unterscheiden sich zunächst in der Art der Ausgabe der zusammenfassenden Statistiken. Verwendet man „Bericht in Zeilen", werden die verschiedenen zusammenfassenden Statistiken einer Teilgruppe in untereinanderliegenden Zeilen ausgedruckt. Wählt man dagegen „Bericht in Spalten", werden sämtliche Statistiken für die Gliederungsgruppe nebeneinander in Spalten angezeigt. Darüber hinaus kann nur mit „Bericht in Zeilen" eine Auflistung der Fälle angefordert werden. Mit „Bericht in Zeilen" kann man auch Listen und gemischte Berichte erstellen. „Bericht in Spalten" bietet dagegen die Möglichkeit, mit Hilfe einfacher Rechenoperationen aus zwei Ausgangsvariablen eine neue zu berechnen.

11.3.1 Berichte in Zeilen

Zeilenweise Berichte können als zusammenfassende Berichte angelegt sein, als Fallauflistungen oder eine Kombination von beiden. Wir beginnen mit den zusammenfassenden Berichten.

11.3.1.1 Zusammenfassende Berichte

Beispiel. Für die Datei ALLBUS90.SAV sollen, gegliedert nach Geschlecht, das arithmetische Mittel, die Standardabweichung sowie der höchste und niedrigste Wert für die Variablen Alter (ALT), Schulbildung (SCHUL), Einkommen (EINK) und monatliche Arbeitsstunden (STDMON) ausgegeben werden. (STDMON ist eine neue, vom Autoren aus den Variablen EINK und ARBSTD errechnete Variable.) Der Bericht soll mit einer Datumsangabe, einer Überschrift in einer Kopfzeile und der Seitenangabe in einer Fußzeile versehen werden. Die Überschriften über den Spalten sollen nicht nur die Angaben der Variablendefinition enthalten, sondern ein ausführliches Label. Sie sollen zudem zentriert sein. Auch die Spalte für

die Break-Variable wird mit einer ausführlichen Überschrift versehen werden. (Notwendig wären lediglich die Angabe von Datenvariablen und Break-Variablen sowie der gewünschten Statistiken. Alle weiteren Spezifikationen sind optional.) Um einen entsprechenden Report zu erstellen, gehen Sie wie folgt vor:

▷ Wählen Sie die Befehlsfolge „Analysieren", „Berichte" und „Bericht in Zeilen...". Die Dialogbox „Bericht in Zeilen" erscheint (⇨ Abb. 11.4).

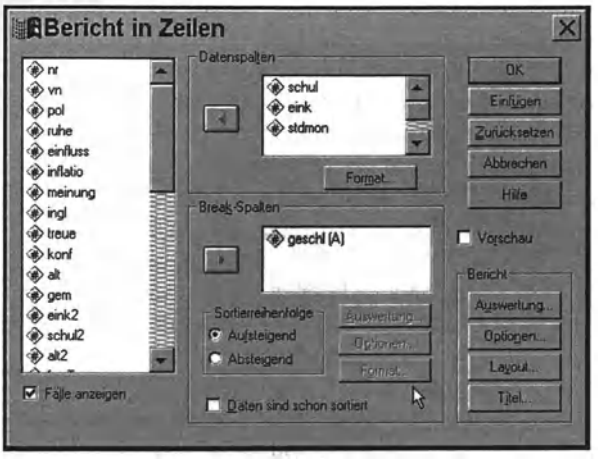


Abb. 11. 4. Dialogbox „Bericht in Zeilen"

▷ Übertragen Sie alle Berichtsvariablen aus der Quellvariablenliste in das Feld „Datenspalten".
▷ Übertragen Sie die Gliederungsvariable(n) das Feld „Break-Spalten".
▷ Falls Ihre Fälle bereits nach der Break-Variablen sortiert sind, kreuzen Sie das Kontrollkästchen „Daten sind schon sortiert" an. Sie sparen damit Rechenzeit.

Format Spaltenvariablen. Wenn Sie das Layout der Ausgabe für die einzelnen Berichtsvariablen beeinflussen wollen, können Sie das über die Schaltfläche „Format..." der Gruppe „Datenspalten" tun.

Abb. 11.5. Dialogbox „Bericht: Datenspaltenformat für"

▷ Markieren Sie den Namen der Variablen, deren Format sie gestalten wollen, und klicken Sie auf die Schaltfläche „Format...". Die Dialogbox „Bericht: Datenspaltenformat für" öffnet sich (⇨ Abb. 11.5).

❒ Im Feld *Spaltentitel* können Sie eine Überschrift für die Variablenspalte eingeben. (Voreinstellung: Falls vorhanden, wird die Variablenetikette verwendet, wenn nicht, der Variablennamen.)
❒ Im Feld *Ausrichtung der Spaltentitel* können Sie durch Anklicken des Pfeils eine Auswahlliste öffnen und zwischen den Optionen „Linksbündig", „Mitte" und „Rechtsbündig" wählen (Voreinstellung: Rechtsbündig).
❒ Durch Eingabe eines Wertes in das Feld *Spaltenbreite* können Sie die Spaltenbreite festlegen. Allerdings kann dadurch die Spaltenbreite für eine vollständige Ausgabe der Spaltenüberschrift zu eng werden. Reicht sie für die Werte nicht aus, werden Dezimalstellen gerundet, ansonsten auf wissenschaftliche Notation umgestellt. Reicht sie auch dafür nicht aus, wird durch * ein zu langer Wert angezeigt. Zu lange Stringwerte werden abgeschnitten. Das gilt auch für Spaltenüberschriften.
❒ In der Gruppe *Position des Wertes in der Spalte* können Sie die Ausrichtung des Wertes bestimmen. Voreingestellt ist „rechts" für numerische und „links" für Stringvariablen. Sie können alternativ „Zentriert in der Spalte" oder durch Anklicken des Optionsschalters „Einzug von rechts" (betrifft aber je nach Variablenart auch Einzug von links!) und die Eingabe eines Wertes in das Kästchen „Anzahl der Stellen:" einen Einzug von rechts/links wählen.
❒ Eine letzte Wahlmöglichkeit bietet die Gruppe *Spalteninhalt*. In ihr bestimmt man, ob für eine Variable in der Tabelle zur Bezeichnung der Ausprägungen die „Werte" oder die „Werte Labels" angezeigt werden (Voreinstellung „Werte"). Dies wird aber nur wirksam, wenn auflistende Berichte erstellt werden.

Wiederholen Sie die Prozedur gegebenenfalls für alle Spaltenvariablen.

Format (Layout) für die Break-Spalten. Wollen Sie das Layout der Break-Spalte(n) beeinflussen, gehen Sie auf die gleiche Weise vor:

▷ Markieren Sie in der Dialogbox „Bericht in Zeilen" eine Break-Spalte, und klicken Sie auf die Option „Format" zur Gruppe „Break-Spalte". Es erscheint die Dialogbox „Bericht: Break-Format für...".

Diese entspricht vollständig der Dialogbox „Bericht: Datenspaltenformat für..". In der Dialogbox „Bericht: Datenspaltenformat für.." legen Sie lediglich das Format für die Datenspalten auf der rechten Seite des Reports fest, in „Bericht: Break-Format für.." dagegen das Format für die Break-Spalten auf der linken Seite. Füllen Sie die Dialogbox entsprechend aus, und wiederholen Sie gegebenenfalls den Vorgang für weitere Break-Variablen. (Voreinstellung für den Namen: Falls vorhanden, wird die Variablenetikette verwendet, wenn nicht der Variablennamen.)

Zusammenfassende Statistiken. In der Gruppe „Break-Spalten" der Dialogbox „Bericht in Zeilen" legt man weiter fest, welche zusammenfassenden Statistiken für die Untergruppen dieser Variablen angefordert werden. Um dies festzulegen:

- ▷ Markieren Sie eine Break-Variable, und klicken Sie auf die Schaltfläche „Auswertung...“. Die in Abb. 11.6 dargestellte Dialogbox öffnet sich.
- ▷ Klicken Sie auf die Auswahlkästchen für die gewünschten Statistiken. Sollten Sie kumulierte Prozentwerte für bestimmte Bereiche anfordern, müssen Sie zusätzlich die entsprechenden Grenzwerte in die Kästchen „Wert:“ bzw. „Kleinster Wert“ und „Größter Wert“ eintragen.
- ▷ Bestätigen Sie mit „Weiter“.

Abb. 11.6. Dialogbox „Bericht: Auswertung für“

Optionen. Man kann die Ausgabe des Reports so gestalten, dass entweder vor jeder neuen Gruppe eine oder mehrere Leerzeilen erscheinen oder eine neue Seite beginnt. Mit der neuen Seite kann auch gleichzeitig die Seitenzahl zurückgesetzt werden (Voreinstellung eine Leerzeile). Auch die Zahl der Leerzeilen vor der Gruppenstatistik kann man beeinflussen (Voreinstellung 0). Wollen Sie die Seitengestaltung in dieser Hinsicht beeinflussen, gehen Sie wie folgt vor:

- ▷ Klicken Sie in der Dialogbox „Bericht in Zeilen“ in der Gruppe „Break-Spalten“ auf die Schaltfläche „Optionen...“. Die Dialogbox „Bericht: Break-Optionen für“ erscheint (⇨ Abb. 11.7).
- ▷ Wählen Sie den gewünschten Optionsschalter. Geben Sie gegebenenfalls eine Zahl für die Leerzeilen ein und bestätigen Sie mit „Weiter“.

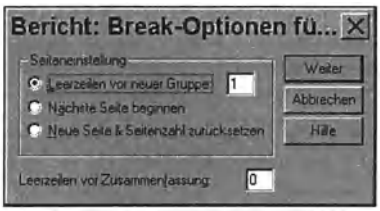


Abb. 11.7. Dialogbox „Bericht: Break-Optionen für“

Sortierfolge Break-Variable. Schließlich können Sie bestimmen, ob die Gruppen der Break-Variable(n) in aufsteigender oder absteigender Folge sortiert werden (Voreinstellung aufsteigend). Dazu markieren Sie in der Dialogbox „Bericht in

Zeilen" die jeweilige Break-Variable und klicken auf den entsprechenden Optionsschalter in der Gruppe „Sortierreihenfolge". Die gewählte Sortierfolge wird auch durch eine Klammerergänzung hinter dem Namen der Break-Variablen angezeigt. A steht für aufsteigend, D für absteigend. Bei numerischen Variablen bedeutet aufsteigend vom kleinsten zum größten Wert, bei Stringvariablen von A bis Z. Jede Break-Variable kann anders sortiert werden.

Gesamtstatistiken. Zusätzlich zu den zusammenfassenden Statistiken für die Gruppen der Break-Variablen, kann auch eine Gesamtstatistik für alle Fälle angefordert werden. Dafür ist in der Dialogbox „Bericht in Zeilen" die Schaltfläche „Auswertung..." in der Gruppe „Bericht" zuständig. Sie öffnet die Dialogbox „Bericht: Abschließende Auswertungszeilen". Diese ist im Aufbau identisch mit der Dialogbox „Bericht: Auswertung für" der Break-Spalten (⇨ Abb. 11.6). Wählen Sie darin die gewünschten Statistiken aus.

Optionen Bericht. Alle anderen Optionen der Gruppe „Bericht" dienen der Gestaltung des Reports.

▷ Das Anklicken von „Optionen..." in der Gruppe „Bericht" öffnet die in Abb. 11.8 dargestellte Dialogbox „Bericht: Optionen".

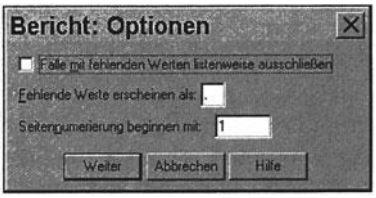


Abb. 11.8. Dialogbox „Bericht: Optionen"

❒ Markieren Sie *„Fälle mit fehlenden Werten listenweise ausschließen"*, dann werden Fälle mit fehlenden Werten aus der Falliste ganz ausgeschlossen. Per Voreinstellung werden diese dagegen angezeigt. (Bei zusammenfassenden Statistiken sind sie in jedem Falle ausgeschlossen).
❒ Unter *„Fehlende Werte erscheinen als:"* können Sie angeben, welches Zeichen zur Darstellung von fehlenden Werten verwendet wird. Dieses Zeichen steht dann sowohl für System-Missings als für nutzerdefinierte Missing-Werte (Voreinstellung: Punkt).
❒ Durch Eintrag eines Wertes in *„Seitennumerierung beginnen mit:"* bestimmen Sie die Seitennummer der ersten Seite des Reports (Voreinstellung: 1).

Berichtlayout. Durch Anklicken des Optionsschalters „Layout..." in der Gruppe „Bericht" öffnen Sie die Dialogbox „Bericht: Layout" (⇨ Abb. 11.9). Hier können Sie in der Gruppe „Seitenformat" bestimmen, in welcher Zeile die Ausgabe auf einer Seite beginnt und endet. Auch die erste und letzte Ausgabespalte wird hier festgelegt sowie die Ausrichtung des Reports „Linksbündig", „Mitte" oder „Rechtsbündig" innerhalb der Seitenränder. In der Gruppe „Titel und Fußzeilen der Seite" bestimmt man die Zahl der leeren Zeilen nach dem Titel und vor den Fuß-

zeilen. In der Gruppe „Break-Spalten" gestaltet man die Anzeige der Break-Variablen im Report. Wählt man „Alle Break-Variablen in der ersten Spalte", werden alle Break-Variablen in der ersten Spalte des Reports und nicht in getrennten Spalten angezeigt. Wählt man diese Option, kann man einen Wert in dem Eingabefeld „Bei jeder Break-Var. einrücken" festlegen, so dass die Ausprägung der Break-Variablen jeweils um diese Anzahl von Leerzeichen versetzt angezeigt werden. (Zu den Optionen der Gruppen „Spaltentitel" und „Beschriftung für Zeilen & Breaks der Datenspalte" ⇨ spaltenweise Berichte.)

Vorschau. Beim Anklicken des Kontrollkästchens „Vorschau" (⇨ Abb. 11.4) wird nicht der gesamte Report erstellt, sondern nur eine Musterseite ausgegeben. Das erspart Zeit, wenn man zunächst lediglich wünscht, das Layout zu überprüfen und gegebenenfalls zu verbessern.

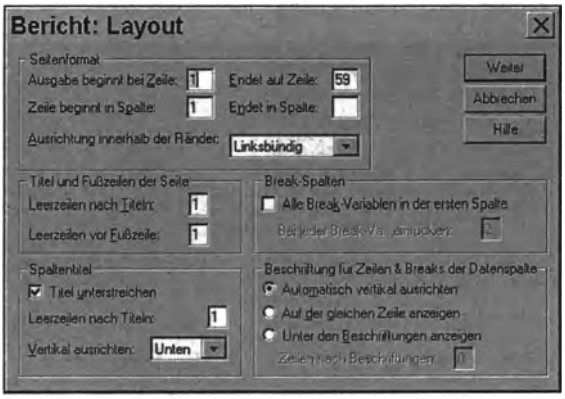


Abb. 11.9. Dialogbox „Bericht: Layout"

Die Abb. 11.9 zeigt die Voreinstellung des Layouts für den Gesamtbericht. Möchten Sie diese ändern, tragen Sie die gewünschten Werte in die entsprechenden Eingabekästchen ein. Die Ausrichtung der Absätze ändern Sie durch Anklicken des Pfeils auf der Seite des Auswahlkästchens. Eine Auswahlliste erscheint. Sie markieren die entsprechende Option. Bestätigen Sie die Eingaben mit „Weiter".

Titel. Schließlich können Sie den Report noch mit Titeln versehen. Diese können auch Datums- und Seitenangaben enthalten. Durch Anklicken des Optionsschalters „Titel..." in der Gruppe „Bericht" öffnet sich die in Abb. 11.10 dargestellte Dialogbox „Bericht: Titel".

Sie haben in dieser Box die Möglichkeit, jeweils bis zu zehn Kopf- und Fußzeilen zu definieren. Dabei kann Text entweder linksbündig oder rechtsbündig oder zentriert eingegeben werden. Jeweils für eine Zeile stehen deshalb drei Eingabekästchen zur Verfügung. In diese kann freier Text eingegeben werden. Es ist aber auch möglich, Variablennamen durch Markierung eines Variablennamens in der Quellvariablenliste oder Platzhalter durch Markieren einer Platzhalterbezeichnung in den Liste „Sondervariablen:" und Klicken auf ▶ zu

übertragen. Der Platzhalter „Date" steht für die Variable „Datum". Bei Ausgabe des Berichts wird das jeweilige Systemdatum eingesetzt. „Page" steht für die Variable „Seite". Es wird eine fortlaufende Seitennummer (beginnend mit der definierten) eingesetzt.

Abb. 11.10 zeigt die Eingabe für jeweils eine Kopf- und Fußzeile. In der Kopfzeile wird das Datum auf der linken Seite und eine Überschrift aus freiem Text zentriert ausgegeben. In der Fußzeile wird zentriert die Seitenzahl eingesetzt.

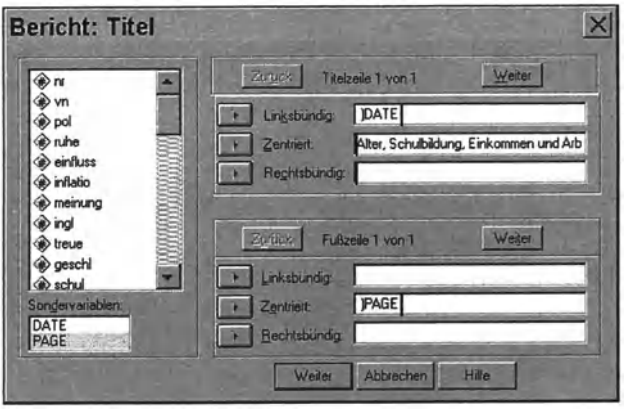


Abb. 11.10. Dialogbox „Bericht: Titel"

Tabelle 11.4. Mit dem Befehl „Bericht in Zeilen" erstellter zusammenfassender Bericht

05 Aug 98 Alter, Schulbildung, Einkommen und Arbeit

GESCHLECHT, BEFRAGTE<R>	Alter	höchster Schulabschluss	monatliches Nettoeinkommen	monatliche Arbeitsstunden
MAENNLICH				
Mittelwert	47	3	2506	172
Minimum	18	1	680	80
Maximum	83	5	7000	300
StdAbw	17	1	1197	32
WEIBLICH				
Mittelwert	48	3	1562	145
Minimum	18	1	129	68
Maximum	89	5	4200	240
StdAbw	19	1	775	45
Gesamtergebnis				
Mittelwert	48	3	2097	163
Minimum	18	1	129	68
Maximum	89	5	7000	300
StdAbw	18	1	1134	39

Sollen weitere Kopf- oder Fußzeilen definiert werden, schalten Sie jeweils durch Anklicken von „Weiter" in der Kopf- bzw. Fußzeilengruppe in einen neuen Eingabebereich und nehmen die gewünschten Eintragungen vor. Sie können auch durch Anklicken von „Zurück" in die vorhergehende Zeile schalten. Haben Sie sämtliche Kopf- und Fußzeilen auf diese Weise definiert, bestätigen Sie die gesamte Definition durch Anklicken von „Weiter" in der untersten Zeile der Dialogbox. Das Beispiel führt zu dem in Tabelle 11.4 dargestellten Report.

11.3.1.2 Auflistende Berichte

Man kann auch mit dem Befehl „Bericht in Zeilen" eine Auflistung von Fällen, ähnlich dem Befehl „Fälle zusammenfassen" erstellen. Die Vorgehensweise ist dieselbe wie beim zusammenfassenden Bericht. Jedoch werden für die Gruppen weder Statistiken noch Gesamtstatistiken aufgerufen. Der Hauptunterschied besteht darin, dass man in der Dialogbox „Bericht in Zeilen" das Auswahlkästchen „Fälle anzeigen" markiert. Außerdem hat jetzt auch die Option „Wertelabels" ihre Wirkung, wenn Sie in der Dialogbox „Bericht: Datenspaltenformat" gewählt wurde. Dies ist in unserem Beispiel für die Variable SCHUL der Fall.

Es stehen keine Optionen für den Umbruch zur Verfügung. Ein Fall wird immer nur in einer Zeile von Maximal 255 Zeichen Länge ausgegeben. Die Zeilenlänge kann durch Definition im Bericht-Layout weiter eingeschränkt sein. Reicht der Platz für die Ausgabe der Werte aller gewählten Variablen nicht aus, bricht das Programm mit einer Fehlermeldung ab.

Tabelle 11.5. Auszug aus einem auflistenden Report

05 Aug 98 Alter, Schulbildung, Einkommen und Arbeit

GESCHLECHT,	Alter	höchster Schulabschluss	monatliches Nettoeinkommen	monatliche Arbeitsstunden
MÄNNLICH	42	HAUPTSCHULABSCHLUSS	680	.
	34	MITTLERE REIFE	4800	180
	71	HAUPTSCHULABSCHLUSS	.	.
	80	HAUPTSCHULABSCHLUSS	2100	.
	.	.	.	.
	.	.	.	.

GESCHLECHT,	Alter	höchster Schulabschluss	monatliches Nettoeinkommen	monatliche Arbeitsstunden
WEIBLICH	37	MITTLERE REIFE	.	.
	25	MITTLERE REIFE	2800	160
	43	ABITUR	.	180
	47	MITTLERE REIFE	2900	160
	50	HAUPTSCHULABSCHLUSS	800	92
	.	.	.	.
	.	.	.	.

Tabelle 11.5 zeigt einen Ausschnitt aus dem auflistenden Report mit denselben Daten und Formatierungen, die für den zusammenfassenden Report verwendet wurden. Die Fälle sind in Zeilen aufgelistet, zunächst die Männer, dann die Frauen. Die Variablenwerte befinden sich in den Spalten.

11.3.1.3 Kombinierte Berichte

Eine Kombination von Auflistung und zusammenfassendem Bericht erhält man, wenn man sowohl das Auswahlkästchen „Fälle anzeigen" markiert als auch Statistiken für die Gruppen der Break-Variablen und/oder Gesamtstatistiken anfordert.

Tabelle 11.6 zeigt einen solchen Bericht für 10 ausgewählte Fälle mit den Variablen und Formatierungen unseres Beispiels. Zuerst werden jeweils die Fälle einer Gruppe aufgelistet, dann die Gruppenstatistiken ausgegeben. Am Ende des Reports finden sich die Gesamtstatistiken.

Tabelle 11.6. Auszug aus einem kombinierten Bericht

05 Aug 98 Alter, Schulbildung, Einkommen und Arbeitszeit

GESCHLECHT,	Alter	höchster Schulabschluss	monatliches Nettoeinkommen	monatliche Arbeitsstunden
MAENNLICH	35	HAUPTSCHULABSCHLUSS	4000	172
	48	ABITUR	3880	152
	69	MITTLERE REIFE	1500	.
	41	ABITUR	.	.
	83	HAUPTSCHULABSCHLUSS	.	.
Mittelwert	55	3	3127	162
Minimum	35	2	1500	152
Maximum	83	5	4000	172
StdAbw	20	2	1410	14
WEIBLICH	61	HAUPTSCHULABSCHLUSS	150	.
	89	HAUPTSCHULABSCHLUSS	1450	.
	26	MITTLERE REIFE	.	240
	53	HAUPTSCHULABSCHLUSS	.	240
	30	MITTLERE REIFE	1700	154
Mittelwert	52	2	1100	211
Minimum	26	2	150	154
Maximum	89	3	1700	240
StdAbw	26	1	832	50
Gesamtergebnis				
Mittelwert	54	3	2113	192
Minimum	26	2	150	152
Maximum	89	5	4000	240
StdAbw	22	1	1518	45

Für alle Berichtarten gilt einschränkend, dass die Länge der Zeile für die Ausgabe aller gewählten Variablen ausreichen muss. Sonst bricht das Programm mit einer Fehlermeldung ab.

Ergänzende Möglichkeiten bei Verwenden der Befehlssyntax. Beim Programmieren mit den Dialogboxen werden für alle Berichtsvariablen dieselben Statistiken definiert. Häufig ist das aber nicht sinnvoll. Besonders in auflistenden und kombinierten Berichten wird man auch Variablen aufnehmen wollen, die nicht das Messniveau besitzen, das eine Zusammenfassung mit den angebotenen Statistiken sinnvoll macht, denken wir an Schulbildung, Geschlecht u.ä.. In solchen Fällen wird man unterschiedliche Statistiken für die verschiedenen Berichtsvariablen anfordern. Das geht nur bei Verwendung der Befehlssyntax mit dem Unterbefehl SUMMARY.

Beispiel.

```
 /VARIABLES
vn (VALUES) (RIGHT) (OFFSET(0))
schul (VALUES) (RIGHT) (OFFSET(0))
alt (VALUES) (RIGHT) (OFFSET(0))
eink (VALUES) (RIGHT) (OFFSET(0))
/SUMMARY MEAN( alt ) MEAN( eink )
 'Mittelw.'
 /SUMMARY STDDEV( alt ) STDDEV( eink )
 'StdAbw.' .
```

Hier werden mit dem Unterbefehl VARIABLES vier Berichtsvariablen angefordert. Die beiden SUMMARY-Unterbefehle bilden aber nur für die Variablen ALT und EINK den Mittelwert bzw. die Standardabweichung und beschriften die Ausgabezeilen mit den Labels „Mittelw." bzw. „StdAbw".

11.3.2 Berichte in Spalten

Mit dem Befehl „Bericht in Spalten" können ebenfalls zusammenfassende Reports erstellt werden. Dagegen kann man keine Fälle auflisten. Der Unterschied zu den zeilenweisen Berichten liegt in der Art der Ausgabe der zusammenfassenden statistischen Maßzahlen. In zeilenweisen Reports werden die Berichtsvariablen in Spalten angeordnet, die verschiedenen Maßzahlen für eine Variable jedoch untereinander ausgegeben. Beim spaltenweisen Bericht ist dagegen für jede einzelne Maßzahl eine Spalte reserviert. Alle Ausgaben stehen nebeneinander in Spalten. Dadurch kann man oftmals die Ausgabe besser lesen. Allerdings verbraucht man wesentlich mehr Platz in einer Zeile, wenn mehrere Maßzahlen pro Variable angefordert werden. Deshalb ist die Zahl der gleichzeitig darstellbaren Variablen gegenüber dem Listenformat erheblich eingeschränkt. Im zeilenweisen Format werden für alle (!) Variablen gleichzeitig alle gewünschten statistischen Maßzahlen definiert. Das reduziert den Definitionsaufwand erheblich. (Will man die einzelnen Variablen unterschiedlich behandeln, geht das nur bei Anwendung der Befehlssyntax.) Im spaltenweisen Format dagegen muss jede Variablen-Maßzahlen-Kombination einzeln definiert werden. Das erfordert größeren Definitionsaufwand, hat auf

der anderen Seite den Vorteil, dass die verschiedenen Variablen unterschiedlich behandelt werden können. Schließlich lässt sich beim spaltenweisen Format mit einfachen Rechenoperationen aus zwei Variablenmaßzahlen eine neue bilden. Die weiteren Formatierungsmöglichkeiten sind bei beiden Formaten weitestgehend identisch. Deshalb werden sie nur in den Fällen näher behandelt, bei denen Abweichungen bestehen.

Beispiel. Es soll aus der Datei ALLBUS90.SAV ein zusammenfassender Report, gegliedert nach Geschlecht und Schulabschluss erstellt werden. Berichtsvariablen sind Alter (ALT), Einkommen (EINK) und monatliche Arbeitszeit (STDMON). Beim Alter und Einkommen interessiert nur das arithmetische Mittel, bei der monatlichen Arbeitszeit das arithmetische Mittel und die Standardabweichung. Eine weitere Variable Einkommen pro Arbeitsstunde (EINKSTD) soll errechnet und deren arithmetisches Mittel ebenfalls ausgegeben werden. Als Spaltenüberschrift benutzen wir für Alter, Geschlecht und Schulbildung die Voreinstellung. Bei Spaltenvariablen ergibt die Voreinstellung eine Überschrift, bestehend aus dem Variablen-Label und der Bezeichnung der angeforderten Maßzahl. Bei den Break-Variablen wird das Variablen-Label per Voreinstellung zur Beschriftung benutzt. Für die anderen Variablen sollen kurze Spaltenüberschriften mit einer Kurzbezeichnung für die Variable und die verwendete Kennzahl gebildet werden. Alle Spaltenüberschriften sollen zentriert ausgerichtet sein. Eine Seitenüberschrift, das Datum und die Seitenzahl werden wie oben definiert. Ansonsten sollen die Voreinstellungen beibehalten werden. (*Genereller Hinweis.* Break-Optionen für die Variable auf dem untersten Level, das ist die zuletzt angeführte Variable in der Liste der Break-Spalten [hier Schulbildung], werden ignoriert. Ebenso entfallen bei der unten verwendeten Formatierung, bei der alle Break-Variablen in einer Spalte ausgegeben werden, alle Spaltenüberschriften für Break-Variablen, außer für die erste.)

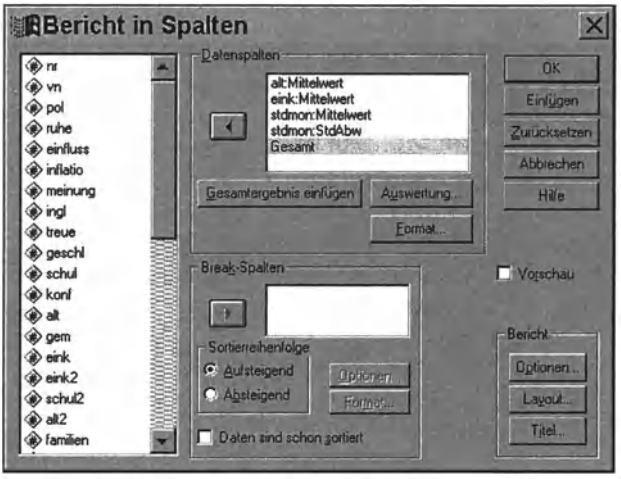


Abb. 11.11. Dialogbox „Bericht in Spalten“

Um einen spaltenweisen Report zu erzeugen, gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Berichte" und „Bericht in Spalten...". Die Dialogbox „Bericht in Spalten" öffnet sich (Abb. 11.11).
- ▷ Übertragen Sie die Berichtsvariablen in der gewünschten Reihenfolge in das Auswahlfeld „Datenspalten". Der Name der übertragenen Variablen erscheint dort jeweils mit dem voreingestellten Zusatz „Summe". Dies besagt, dass als Maßzahl die Summe der Werte ausgegeben werden soll. Wünschen Sie eine andere Maßzahl, müssen Sie das ändern. Wollen Sie für eine Variable mehrere Maßzahlen ermitteln, müssen Sie den Variablennamen für jede dieser Maßzahlen einmal übertragen und jedes Mal den Zusatz ändern.

Die gewünschte Maßzahl definieren Sie:

- ▷ Indem Sie den Variablennamen (mit Zusatz) markieren und die Schaltfläche „Auswertung..." anklicken. Es erscheint die Dialogbox „Bericht: Auswertung für" (⇨ Abb. 11.12) mit dem Namen der ausgewählten Variablen in der Überschrift.
- ▷ Klicken Sie dort den Optionsschalter neben der Bezeichnung der gewünschten Maßzahl an, und bestätigen Sie mit „Weiter".

Wiederholen Sie diesen Prozess für jede Berichtsvariable.

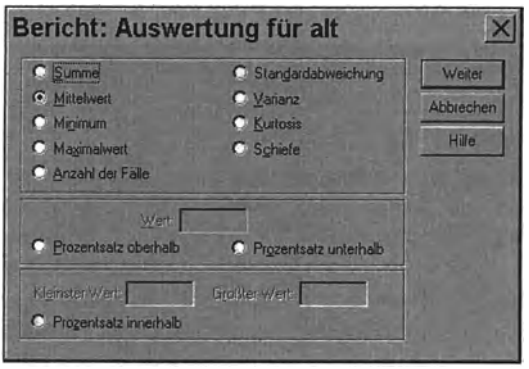


Abb. 11.12. Dialogbox „Bericht: Auswertung für"

Um die zusammenfassende Variable zu bilden, gehen Sie wie folgt vor:

- ▷ Klicken Sie auf die Schaltfläche „Gesamtergebnis einfügen". Der Variablenname „Gesamt" erscheint in der Liste „Datenspalten".
- ▷ Markieren Sie diesen Namen, und klicken Sie auf die Schaltfläche „Auswertung". Die Dialogbox „Bericht: Auswertungsspalte" erscheint (⇨ Abb. 11.13).
- ▷ Übertragen Sie die Namen der zusammenfassenden Variablen, aus denen die neue Variable gebildet werden soll, aus dem Feld „Datenspalten:" in das Feld „Zusammenfassungsspalte:". (Soll eine Division oder Subtraktion vorgenommen werden, dürfen es nur zwei Variablen sein, sonst können beliebig viele Variablen ausgewählt werden.)

▷ Klicken Sie auf den Pfeil neben dem Feld „Auswertungsfunktion:". Ein Auswahlfeld öffnet sich.
▷ Markieren Sie die gewünschte Funktion, und bestätigen Sie die Eingabe mit „Weiter".

Abb. 11.13. Dialogbox „Bericht: Auswertungsspalte"

Alle Funktionen bilden aus den zusammenfassenden Maßzahlen von zwei oder mehr Variablen (Spalten) ein Ergebnis. (*Beispiel:* Das arithmetische Mittel des durchschnittlichen Monatseinkommens einer Gruppe ist 2500, deren durchschnittliche monatliche Arbeitszeit 180. Daraus lässt sich mit der Funktion „1. Spalte / 2. Spalte" der Wert 13,89 für eine Totalvariable ermitteln. Hätte man als erste Variable das arithmetische Mittel, als zweite die Varianz des Einkommens ausgewählt, ergäbe sich aus deren Quotient ein Variabilitätskoeffizient usw..) Man kann mehrere unterschiedliche Totalvariablen bilden.

Verfügbare Funktionen (in Klammern der Namenszusatz) sind:

- *Summe der Spalten* (Summe). Summe der zusammenfassenden Werte der ausgewählten Variablen.
- *Mittelwert der Spalten* (Mittelwert). Deren arithmetisches Mittel.
- *Minimum der Spalten* (Minimum). Der kleinste Wert aller zusammenfassenden Werte der ausgewählten Variablen.
- *Maximum der Spalten* (Maximum). Deren größter Wert.
- *1. Spalte - 2. Spalte* (Differenz). Differenz zwischen einem ersten und zweiten zusammenfassenden Wert.
- *1. Spalte / 2. Spalte* (Verhältnis). Deren Quotient.
- *% 1. Spalte / 2. Spalte* (Prozentsatz). Prozentanteil des ersten Wertes am zweiten.
- *Produkt der Spalten* (Produkt). Produkt der zusammenfassenden Werte der ausgewählten Variablen.

Formatierung Spaltenvariablen. Durch Markieren des Variablennamens und Anklicken der Schaltfläche „Format" öffnet man die Dialogbox „Bericht: Datenspaltenformat für" (⇨ Abb. 11.14). Hier kann man auf dieselbe Weise wie beim zeilenweisen Report Überschriften und Layout der Spalten definieren.

Abb. 11.14. Dialogbox „Bericht: Datenspaltenformat für"

Sortierfolge. In der Gruppe „Break-Spalten" finden Sie mehrere Optionen zur Festlegung der Sortierfolge der Breakvariablen.

- Durch Markieren eines Variablennamens und Auswahl einer der Optionen „Aufsteigend" oder „Absteigend" in der Gruppe „Sortierfolge" bestimmen Sie, ob die Gruppen der jeweiligen Break-Variablen in aufsteigender oder absteigender Ordnung sortiert werden. Dies kann für die verschiedenen Break-Variablen unterschiedlich geschehen. Die Einstellung wird durch einen Klammerzusatz (A) oder (D) angezeigt.
- Sind die Daten bereits nach der oder den Break-Variablen sortiert, sollten Sie das Auswahlkästchen „Daten sind schon sortiert" markieren, um einen überflüssigen Sortierlauf zu vermeiden.

Optionen für die Break-Variablen. Durch Markieren einer Break-Variablen und Anklicken der Schaltfläche „Optionen" öffnet sich eine Dialogbox „Bericht: Break-Optionen für" (⇨ Abb. 11.15). Hier können sie parallel zu den Optionen beim zeilenweisen Format bestimmen, welcher Vorschub nach jeder Gruppe benutzt wird: Eine bestimmte Linienzahl, eine neue Seite usw..

Neu kommt die Möglichkeit hinzu festzulegen, ob auch Zwischenergebnisse angezeigt werden sollen, d.h. Zusammenfassungen für die Gruppen einzelner Break-Variablen. Das ist dann von Interesse, wenn mehr als eine Break-Variable benutzt wird. Bei Auswahl von „Zwischenergebnis anzeigen" wird eine Zusammenfassung für die einzelnen Gruppen ausgegeben. Nicht möglich ist das für die Break-Variable auf dem untersten Level. Wünschen Sie die Ausgabe von Zwischenergebnissen, dann gehen Sie wie folgt vor:

- Markieren Sie das Auswahlkästchen „Zwischenergebnis anzeigen". Es wird automatisch ein Label für die neue Spalte vorgeschlagen. (Voreinstellung: Zwischenergebnis und Variablennamen.)
- Sie können das durch Eintrag im Feld „Label" ändern.
- Außerdem können Sie durch Eintrag in das Feld „Leerzeilen vor Zwischenergebnis" bestimmen, wie viele Leerzeilen jeweils zwischen der vorhergehenden Anzeige und der Anzeige der Werte der Subtotals eingeschoben werden sollen (Voreinstellung 0).

Wiederholen Sie die Prozedur gegebenenfalls für andere Break-Variablen.

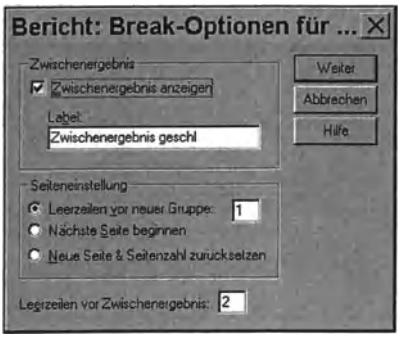


Abb. 11.15. Dialogbox „Bericht: Break-Optionen für"

Format für die Break-Variablen. Durch Markieren einer Break-Variablen und Anklicken von „Format" öffnet sich eine Dialogbox „Bericht: Break-Format für" (⇨ Abb. 11.16). Hier kann ein Spaltentitel, die Justierung des Titels und der Werte in den Spalten und die Spaltenbreite festgelegt werden. Außerdem bestimmt man hier, ob die Spalten mit den Werten der Subgruppen oder deren Labels beschriftet werden (Voreinstellung Labels). Die Dialogbox entspricht im Aufbau genau der entsprechenden Dialogbox beim zeilenweisen Report.

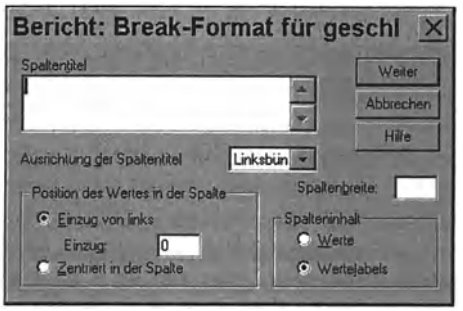


Abb. 11.16. Dialogbox „Bericht: Break-Format für"

Gestaltung des Gesamtlayouts. Alle Optionen in der Gruppe „Bericht" dienen der Gestaltung des Gesamtdokuments:

- *Optionen.* Öffnet die Dialogbox „Bericht: Optionen" (⇨ Abb. 11.17). Hier legen Sie fest, ob ein Gesamtergebnis („Grand total") ausgegeben wird. Ein- und Ausschaltung geschieht durch Anklicken des Kontrollkästchens „Gesamtergebnis anzeigen". Weiter können Sie für die entsprechende Spalte ein Label definieren (Voreinstellung: Gesamtergebnis). Sie können außerdem durch Anklicken des Kästchens „Fälle mit fehlenden Werten listenweise ausschließen" bestimmen, dass Fälle mit fehlenden Werten in irgendeiner Variablen ganz von der Berechnung ausgeschlossen werden sollen (Voreinstellung: Sie werden einbezogen). Schließlich bestimmen Sie im Kästchen „Fehlende Werte erscheinen

als:“ das Zeichen, das für fehlende Werte angezeigt wird (Voreinstellung Punkt) und durch Eingabe eines Wertes in „Seitennummerierung beginnt mit:“, mit welcher Nummer die Seitenzählung beginnt.

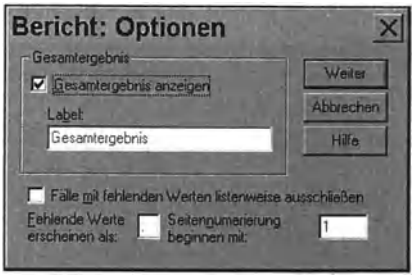


Abb. 11.17. Dialogbox „Bericht: Optionen“

❒ *Layout.* Öffnet die Dialogbox „Bericht: Layout für den Gesamtbericht“ (⇨ Abb. 11.18). Hier kann man das Seitenlayout für den Gesamtbericht beeinflussen. Wie beim zeilenweisen Report werden Beginn und Ende von Zeilen und Spalten auf einer Seite festgelegt, die Ausrichtung des Textes sowie der Abstand zwischen Text und Kopf- bzw. Fußzeilen. Die Dialogbox dient auch der Gestaltung der Spalten.

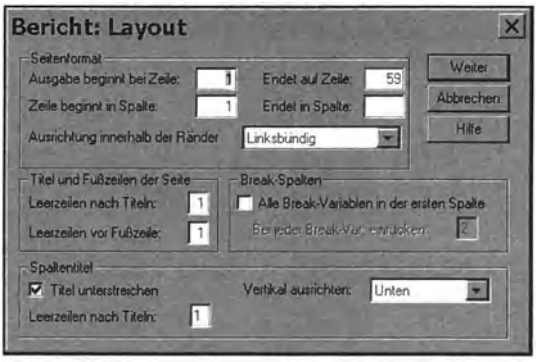


Abb. 11.18. Dialogbox „Bericht: Layout“

Neu hinzu kommt, dass in der Gruppe „Break-Spalten“ festgelegt werden kann, in welcher Spalte Break-Variablen ausgegeben werden. Voreingestellt ist, dass für jede neue Break-Variable eine neue Spalte angelegt wird. Diese Spalten stehen nebeneinander. Das kostet viel Platz in der Zeile. Sie können das ändern, indem Sie das Auswahlkästchen „Alle Break-Variablen in der ersten Spalte“ auswählen. Dann werden die Werte aller Break-Variablen in der ersten Spalte angezeigt. Durch Eingabe eines Wertes in das Feld „Bei jeder Break-Var. einrücken“ kann man aber bestimmen, um wie viele Leerstellen bei jeder neuen Break-Variablen

eingerückt wird. (Wird letzterer Modus verwendet, entfallen die Spaltenüberschriften für alle Break-Variablen, außer für die erste. Eine entsprechende Fehlermeldung wird ausgegeben.)

Ebenfalls neu ist, dass Sie in der Gruppe „Spaltentitel" die Unterstreichung der Spaltentitel gestalten können. Voreingestellt ist Unterstreichung. Durch Anklicken das Auswahlkästchens können Sie das ausschalten. Im Feld „Leerzeilen nach Titeln" bestimmen Sie durch Ihre Eingabe, wieviel Zeilen Abstand zwischen den Spaltenüberschriften und der ersten Datenzeile frei bleibt.

Im Auswahlfeld „Vertikal ausrichten:" können Sie schließlich bestimmen, wie die Spaltenüberschriften ausgerichtet sind. Das macht sich bemerkbar, wenn die Überschriften unterschiedlich viele Zeilen in Anspruch nehmen. Mit der Einstellung „Unten" (Voreinstellung) erreichen Sie, dass alle Überschriften auf derselben unteren Zeile enden. Dagegen beginnen sie bei der Einstellung „Oben" alle mit derselben Zeile, enden aber dann unterschiedlich. Die erste Einstellung ergibt gewöhnlich das bessere Bild.

Tabelle 11.7. Spaltenweiser Report für das Erläuterungsbeispiel

GESCHLECHT, BEFRAGTE<R>	ALTER Mittel	EINKOMMEN Mittel	stdmon Mittel	stdmon StdAbw	Stunden-lohn Mittel
MAENNLICH					
KEIN SCHULABSCHLUSS	40	2500	180	.	14
HAUPTSCHULABSCHLUSS	54	2207	170	31	13
MITTLERE REIFE	42	2895	175	41	17
FACHHOCHSCHULREIFE	45	2700	170	13	16
ABITUR	39	2670	171	19	16
NOCH SCHUELER	20	.	.	.	.
KEINE ANGABE	40	2800	167	18	17
ischenergebnis geschl	47	2506	172	32	15
WEIBLICH					
KEIN SCHULABSCHLUSS	62	1300	94	.	14
HAUPTSCHULABSCHLUSS	56	1329	133	58	10
MITTLERE REIFE	43	1898	155	32	12
FACHHOCHSCHULREIFE	44	1400	133	31	11
ABITUR	38	1900	150	48	13
NOCH SCHUELER	19	.	.	.	.
KEINE ANGABE	38	.	.	.	.
ischenergebnis geschl	48	1562	145	45	11
Gesamtergebnis	48	2097	163	39	13

Seitentitel. Durch Anklicken der Schaltfläche „Titel..." öffnet man die Dialogbox „Bericht: Titel". Sie entspricht vollkommen der Dialogbox „Bericht: Titel" beim zeilenweisen Report (⇨ Abb. 11.10). Hier kann man jeweils bis zu zehn Zeilen

Text in Titel- und Fußzeilen definieren. In jeder Zeile kann ein Teil des Textes linksbündig, rechtsbündig und zentriert ausgerichtet sein. Für jede Ausrichtung steht ein eigenes Eingabefeld zur Verfügung. Variablennamen können aus der Variablenliste übertragen werden. Für die Variablen „Datum" und „Seitenzahl" können Platzhalter aus der Gruppe „Sondervariablen:" übertragen werden.

Durch Anklicken des Kästchens „Vorschau" (⇨ Abb. 11.11) bewirken Sie, dass nicht der gesamte Report erstellt, sondern nur eine Musterseite angezeigt wird.

Tabelle 11.7 zeigt einen spaltenweisen Report für unser Beispiel. Wir sehen, dass die zusammenfassenden Werte für die einzelnen Variablen/Maßzahlen-Kombinationen in Spalten ausgegeben werden. Alle Spaltenvariablen, außer der ersten, sind mit durch den Nutzer definierten Überschriften versehen, die erste dagegen mit einer per Voreinstellung erzeugten Überschrift. Die Vorspalte enthält die Ausprägungen der Break-Variablen. Da alle in einer Spalte ausgegeben werden, ist sie nur mit dem Label der ersten Break-Variablen „Geschlecht" überschrieben. Die Gruppen sind, da wir Variablen-Labels angefordert haben, durch die Labels der Werte wie MÄNNLICH, KEIN SCHULABSCHLUSS usw. beschriftet. Für die Geschlechtsgruppen MÄNNLICH, WEIBLICH werden Zwischenergebnisse ausgegeben. Die Zeile ist mit „Zwischenergebnis geschl" beschriftet. Diese Beschriftung ist per Voreinstellung erzeugt. Schließlich werden in der letzten Zeile die zusammenfassenden Maßzahlen für die gesamte Population „Gesamtergebnis" angezeigt.

12 Analysieren von Mehrfachantworten

Im allgemeinen gilt die Regel, dass Messungen eindimensional sein und die verschiedenen Werte einer Variablen sich gegenseitig ausschließen sollen. Mitunter ist es aber sinnvoll, von dieser Regel abzuweichen. So kann es etwa bei einer Frage nach den Gründen für die Berufswahl zugelassen sein, dass sowohl „Interesse für den Berufsinhalt" als auch „Einfluss der Eltern" angegeben wird. Umgekehrt kann es notwendig sein, mehrere getrennte Messungen zu einer Dimension zusammenzufassen, etwa wenn man Zinssätze für den ersten, zweiten, dritten Kredit erfasst, man aber am durchschnittlichen Zinssatz interessiert ist, gleichgültig um den wievielten Kredit es sich handelt.

Solche Mehrfachmessungen auf derselben Dimension sind technisch schwer zu handhaben. In SPSS kann man je Variable nur einen Wert eintragen. Falls eine Mehrfachmessung vorliegt, muss sie für die Datenerfassung in mehrere Variable aufgeteilt werden, in denen jeweils nur ein Wert eingetragen wird. Dafür sind zwei verschiedene Verfahren geeignet:

- *Multiple Dichotomien-Methode.* Es wird für jeden Wert der Variablen eine eigene Variable gebildet. Auf dieser wird dann jeweils nur festgehalten, ob dieser Wert angegeben ist (gewöhnlich mit 1) oder nicht (gewöhnlich mit 0).
- *Multiple Kategorien-Methode.* Hier muss zunächst festgestellt werden, wie viele Nennungen maximal auftreten. Für jede Nennung wird dann eine eigene Variable gebildet. In der ersten dieser Variablen wird dann festgehalten, welcher Wert bei der ersten Nennung angegeben wurde, in der nächsten, welcher bei der zweiten usw.. Wenn weniger Nennungen maximal auftreten als die Ausgangsvariable Werte hat, kommt dieses Verfahren mit weniger neu gebildeten Variablen aus.

Mehrfachantworten müssen in SPSS also zunächst in Form mehrerer Elementarvariablen nach der multiple Dichotomien- oder multiple Kategorien-Methode abgespeichert werden. Zur Analyse können diese aber wieder in Form von multiple Dichotomie- oder multiple Kategorien-Sets zusammengefasst werden, für die man Häufigkeits- oder Kreuztabellen erstellen kann. Die Vorgehensweise wird zunächst an einem Beispiel nach der multiplen Kategorien-Methode dargestellt.

12.1 Definieren eines Mehrfachantworten-Sets (Multiple Kategorien-Set)

Beispiel. In einer Untersuchung eines der Autoren wurde bei überschuldeten Verbrauchern ermittelt, ob und bei welchen Banken sie für irgendeinen Kredit sittenwidrig hohe Zinsen bezahlt haben. Als sittenwidrig wurden von der Schuldnerberatung der Verbraucherzentrale gemäß der damaligen Rechtsprechung Kredite eingestuft, wenn die Zinsen den durchschnittlichen Marktpreis zum Zeitpunkt der Kreditvergabe um 100% und mehr überschritten. Der Marktpreis orientiert sich am Schwerpunktzinssatz der Deutschen Bundesbank. Manche der Verbraucher hatten für mehrere Kredite sittenwidrig hohe Zinsen bezahlt. Maximal waren es vier Kredite. Außerdem machte eine ganze Reihe von Banken solche rechtswidrigen Geschäfte. Es lag nahe, diese Daten nach der multiple Kategorien-Methode abzuspeichern. Dazu wurden in der Datei BANKEN.SAV vier numerische Variablen für den ersten bis vierten Kredit eingerichtet. In der ersten Variablen wurde abgespeichert, ob ein erster Kredit mit sittenwidrigen Zinsen vorlag. War dem nicht so, bekam der Fall den Kode 0, war das der Fall, die Kodenummer der Bank, eine Zahl zwischen 1 und 251. In der zweiten Variablen wurde nach demselben Verfahren abgespeichert, ob ein zweiter Kredit mit sittenwidrigen Konditionen vorlag und wenn ja, die Kodenummer der Bank usw. (die Namen der Banken wurden als Wertelabels eingegeben). Die Variablen, in denen diese Informationen abgespeichert sind, haben die Namen V043, V045, V047 und V049. Es soll jetzt eine „schwarze Liste" der Banken erstellt werden, die Kredite mit sittenwidrig hohen Zinsen vergaben. Ergänzend wird ermittelt, welchen Anteil an der Gesamtzahl der sittenwidrigen Kredite die einzelnen Banken haben. Dazu werden nur die Fälle ausgezählt, bei denen ein sittenwidriger Kredit vorliegt (gültige Fälle), also eine Kodenummer für eine Bank eingetragen ist. Ein Fall, bei dem gar kein sittenwidriger Kredit vorliegt, wird als ungültiger Fall behandelt.

Zunächst muss ein Mehrfachantworten-Set definiert werden. Gehen Sie dazu wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Mehrfachantworten ▷", „Sets definieren...". Es öffnet sich die Dialogbox „Mehrfachantworten-Sets" (⇨ Abb. 12.1).
- ▷ Wählen Sie aus der Variablenliste die Variablen aus, die zu einem Set zusammengefasst werden sollen.
- ▷ Klicken Sie den Optionsschalter „Kategorien" an, um festzulegen, dass ein nach der Methode „Multiple Kategorien" erstellter Datensatz verarbeitet werden soll.
- ▷ Geben Sie in die beiden Kästchen hinter „Bereich:" zunächst im ersten Kästchen den niedrigsten gültigen Wert ein (hier: 1), dann im zweiten Kästchen den höchsten gültigen Wert (hier: 251).
- ▷ Geben Sie im Feld „Name:" einen Namen für den so definierten Set ein.
- ▷ Geben Sie bei Bedarf im Feld „Label:" eine Etikette für den Set ein.
- ▷ Klicken Sie auf den Optionsschalter „Hinzufügen". In der Gruppe „Mehrfachantworten-Sets:" erscheint der Name der neuen Variablen (der definierte Namen

mit vorangestelltem $-Zeichen). Zugleich werden alle Definitionsfelder freigegeben. (Sie können im folgenden auf diese Weise weitere Sets definieren.) Der so definierte Set wird im folgenden innerhalb des Subprogramms „Mehrfachantworten" als Variablen verwendet.

▷ Beenden Sie die Definition mit „Schließen".

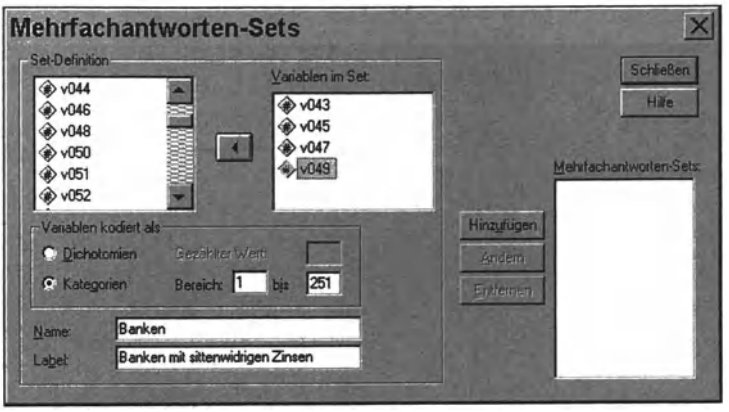


Abb. 12.1. Dialogbox „Mehrfachantworten-Sets"

Sie können später die definierten Sets löschen oder ändern. Dazu muss in der Gruppe „Mehrfachantworten-Sets:" der entsprechende Set-Name markiert werden. Durch Anklicken der Schaltfläche „Entfernen" wird der Set gelöscht. Umstellen zwischen multiplen Kategorien und Dichotomien-Sets ist durch Anwählen der entsprechenden Optionsschalter, Eingabe des zu zählenden Wertes bzw. Bereichs und Anklicken der Schaltfläche „Ändern" möglich.

12.2 Erstellen einer Häufigkeitstabelle für einen multiplen Kategorien-Set

Zum Erstellen einer Häufigkeitstabelle für einen Mehrfachantworten-Set gehen Sie wie folgt vor:

▷ Wählen Sie „Analysieren", „Mehrfachantworten ▷", „Häufigkeiten...". Es öffnet sich die Dialogbox „Mehrfachantworten Häufigkeiten" (⇨ Abb. 12.2).

▷ Wählen Sie aus der Liste „Mehrfachantworten-Sets:" den gewünschten Set aus.

In der Gruppe *„Fehlende Werte"* können Sie die Behandlung der fehlenden Werte bestimmen. Per Voreinstellung werden bei multiplen Kategorien-Sets nur solche Variablen ausgeschlossen, die bei allen Variablen einen fehlenden Wert aufweisen.

Wollen Sie alle Fälle ausschließen, bei denen irgendeine Variable einen fehlenden Wert aufweist, wählen Sie das Kontrollkästchen *„Für kategoriale Variablen Fälle listenweise ausschließen"*.

Abb. 12.2. Dialogbox „Mehrfachantworten Häufigkeiten"

In unserem Beispiel ist die Voreinstellung angemessen. Die meisten Verbraucher haben nur einen Kredit mit sittenwidrig hohen Zinsen. Würde man fehlende Werte listenweise ausschließen, würden nur die Fälle gezählt, die vier Kredite mit sittenwidrigen Zinsen haben. Das ist nicht der Sinn. Es sollen vielmehr alle Banken registriert werden, bei denen irgendein sittenwidriger Kredit vorliegt. Das Beispiel ergibt die Tabelle 12.1.

Die Tabelle ähnelt einer üblichen Häufigkeitstabelle, hat aber einige Besonderheiten. Zunächst werden nur gültige Werte verarbeitet. Wie man der Zeile unterhalb der Tabelle entnehmen kann, stehen 45 gültigen Fällen („valid cases"), bei denen also mindestens ein sittenwidriger Kredit vorlag, 87 nicht gültige Fälle („missing cases") gegenüber. In der Spalte „Count" sind die Häufigkeiten für die einzelnen Banken angegeben. Die Summe aller Antworten („Total responses") ist 59. Der Vergleich dieser gültigen Antworten mit den gültigen Fällen (45) verdeutlicht, dass in einer Reihe von Fällen mehrere sittenwidrige Kredite vorgelegen haben müssen.

In den letzten zwei Spalten sind zwei verschiedene Arten der Prozentuierung wiedergegeben. Die Spalte „Pct of Responses" gibt an, welchen Anteil der einzelne Wert an allen Antworten hat. Die Summe der Antworten ist 100 %. So sind bei der „Kundenbank" z.B. 26 von insgesamt 59 sittenwidrigen Krediten, das sind 44,1 %. Die Spalte „Pct of Cases" zeigt dagegen die Prozentuierung auf Basis der 45 gültigen Fälle. Diese sind gleich 100 % gesetzt. Da aber mehr Nennungen als Fälle auftreten, summiert sich hier der Gesamtprozentwert auf mehr als 100 % (im Beispiel sind es 131 %). Der Prozentwert für die „Kundenbank" beträgt so berechnet 57,8 %. Welche dieser Prozentuierungen angemessen ist, hängt von der Fragestellung ab. Interessiert in unserem Beispiel, welchen Anteil der sittenwidrigen Geschäfte an allen Banken die „Kundenbank" hat, ist die erste Prozentuierung angemessen, interessiert dagegen, wie viel Prozent der betroffenen Verbraucher von der

Kundenbank einen sittenwidrigen Kredit verkauft bekamen, ist es die zweite Prozentuierungsart.

Tabelle 12.1. Banken, die mindestens einen sittenwidrigen Kredit vergeben haben

Group $BANKEN Banken mit sittenwidrigen Zinsen

Category label*)	Code	Count	Pct of Responses	Pct of Cases
ABC Bank B.	1	1	1,7	2,2
ABC Privatbank K.	2	1	1,7	2,2
Absatzanstalt	5	1	1,7	2,2
Alemania	7	2	3,4	4,4
Allgemeine Bank	8	3	5,1	6,7
Allkredit D.	9	1	1,7	2,2
Badische K.	24	1	1,7	2,2
Bankhaus B. KG	25	2	3,4	4,4
Braunschweigische Bank	26	2	3,4	4,4
CTB Bank Th.	33	3	5,1	6,7
Gesellschaft für Finanzierung	38	1	1,7	2,2
Einkaufskredit K.	39	1	1,7	2,2
Hanseatic Bank	59	1	1,7	2,2
Hanseatische Kredit	65	3	5,1	6,7
Interbank	66	1	1,7	2,2
Kundenbank	71	26	44,1	57,8
N.-Bank	97	2	3,4	4,4
SKV Kredit	108	1	1,7	2,2
Süd-West Kredit	114	1	1,7	2,2
Teilzahlungs-Genossenschaft	116	3	5,1	6,7
Verba	132	1	1,7	2,2
WKV Bank N.	148	1	1,7	2,2
Total responses		59	100,0	131,1

87 missing cases; 45 valid cases

*) Die Namen wurden von den Autoren geändert

12.3 Erstellen einer Häufigkeitstabelle für einen multiplen Dichotomien-Set

Definieren eines Mehrfachantworten-Sets. *Beispiel.* In einer Untersuchung eines der Autoren wurde erfasst, ob die befragten Personen in ihrem Leben bereits einmal Rauschgift konsumiert hatten und wenn ja, welchen Stoff. Da einige Rauschgiftkonsumenten mehrere Mittel konsumiert haben, mussten Mehrfachangaben

verschlüsselt werden. Für die gebräuchlichsten Rauschgifte wurden eigene Elementarvariablen gebildet und in jeder dieser Variablen festgehalten, ob dieses Rauschgift benutzt wurde oder nicht. Dabei bedeutete 1 = „genannt", 2 = „nicht genannt", 9 = „nicht zutreffend oder keine Angabe". Die Daten sind in der Datei RAUSCH.SAV gespeichert, die zutreffenden Variablen haben die Namen V70 bis V76. Es soll jetzt eine zusammenfassende Häufigkeitstabelle für den Gebrauch dieser Rauschgifte erstellt werden. Mit Hilfe von „Mehrfachantworten" kann man eine zusammenfassende Variable bilden, bei der jede Elementarvariable einen Wert darstellt. Es wird ausgezählt, wie häufig eine gültige Nennung dieses Wertes auftritt. Zunächst muss ein Mehrfachantworten-Set definiert werden.

▷ Wählen Sie dazu „Analysieren", „Mehrfachantworten", „Sets definieren...". Es öffnet sich die Dialogbox „Mehrfachantworten-Sets" (⇨ Abb. 12.3).

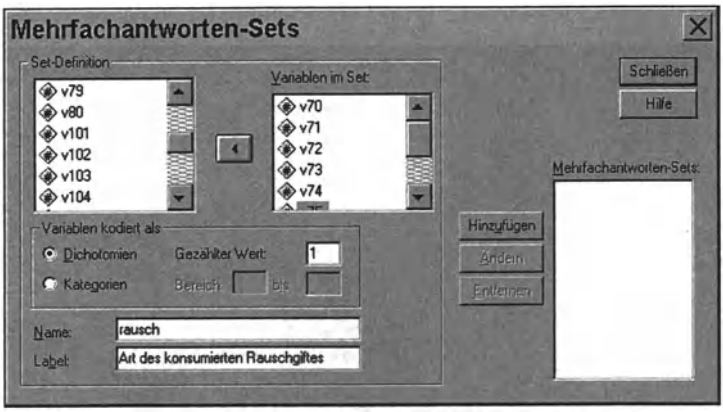


Abb. 12.3. Dialogbox „Mehrfachantworten-Sets"

▷ Wählen Sie aus der Variablenliste die Variablen aus, die zu einem Set zusammengefasst werden sollen.
▷ Klicken Sie den Optionsschalter „Dichotomien" an, um festzulegen, dass ein nach der Methode „multiple Dichotomien" erstellter Datensatz verarbeitet werden soll. Im Gegensatz zum „multiple Kategorien-Set" muss jetzt angegeben werden, welcher einzelne Wert der Elementarvariablen als gültiger Wert ausgezählt werden soll.
▷ Geben sie in dem Eingabefeld „Gezählter Wert:" den Variablenwert an, für den die Auszählung erfolgen soll. Im Beispiel ist das 1 = „genannt".
▷ Geben Sie im Feld „Name:" einen Namen für den so definierten Set ein (im Beispiel RAUSCH).
▷ Tragen Sie bei Bedarf im Feld „Label:" eine Etikette für den Set ein.
▷ Klicken Sie auf den Optionsschalter „Hinzufügen". In der Gruppe „Mehrfachantworten-Sets:" erscheint der Name der neuen Variablen. Zugleich werden alle

Definitionsfelder freigegeben. (Sie können auf diese Weise weitere Sets definieren.)

▷ Beenden Sie die Definition mit „Schließen".

Die so definierten Sets werden im folgenden innerhalb des Subprogramms „Mehrfachantworten" als Variablen verwendet.

Erstellen einer Häufigkeitstabelle. Zum Erstellen einer Häufigkeitstabelle für einen Mehrfachantworten-Set gehen Sie wie oben beschrieben vor.

Hinweis. Wollen Sie alle Fälle ausschließen, bei denen irgendeine Variable einen fehlenden Wert aufweist, wählen Sie in der in Abb. 12.2 dargestellten Dialogbox das Kontrollkästchen „Für dichotome Variablen Fälle listenweise ausschließen". Zu beachten ist dabei, dass es sich darum handelt, ob ein in den Elementarvariablen als fehlend deklarierter Wert auftritt, nicht darum, dass ein im Set als nicht zu zählend deklarierter Wert vorliegt. Setzt man diese Option nicht, werden alle Fälle ausgezählt, auch wenn in einer der dichotomisierten Variablen ein fehlender Wert vorliegt.

Haben Sie, wie in Abb. 12.3 dargestellt, einen Set $RAUSCH definiert und gespeichert und erstellen Sie für diesen eine Häufigkeitsauszählung, führt dies zu dem in Tabelle 12.2 enthaltenen Output.

Tabelle 12.2. Häufigkeit von Rauschgiftkonsum

```
Group $RAUSCH  Art des konsumierten Rauschgifts
     (Value tabulated = 1)
                                              Pct of      Pct of
Dichotomy label            Name      Count   Responses   Cases

Konsum Haschisch           V70          73      53,3       92,4
Konsum Kokain              V71          12       8,8       15,2
Konsum Opium               V72           7       5,1        8,9
Konsum Morphium            V73           5       3,6        6,3
Konsum Preludin            V74           3       2,2        3,8
Konsum Captagon            V75          16      11,7       20,3
Konsum Sonstiges           V76          21      15,3       26,6
                           -----      --------      --------
                  Total responses      137     100,0      173,4
181 missing cases;  79 valid cases
```

Der Aufbau der Tabelle entspricht dem der Häufigkeitstabelle, wie sie auch bei der multiple Kategorien-Methode ausgegeben wird. In unserem Beispiel gibt es 79 gültige Fälle („valid cases"), also haben 79 Personen mindestens einmal Rauschgift probiert. Aber es wurde 137 mal ein Rauschgift genannt („Total responses"). Also haben viele mehrere Rauschgifte versucht. Der Löwenanteil entfällt auf Haschisch. Es erhielt 53,3 % der Nennungen („Pct of Responses"). Genannt wurde es aber sogar von 92,4 % der Rauschgiftkonsumenten („Pct of Cases"). Auch hier erkennt

man deutlich die unterschiedliche Aussage der beiden Prozentuierungsarten auf Basis der Nennungen und auf Basis der Fälle.

12.4 Kreuztabellen für Mehrfachantworten-Sets

Beispiel. Es soll geprüft werden, ob sich die Konsummuster bei Rauschmittelkonsumenten zwischen Männern und Frauen unterscheiden (Datei RAUSCH.SAV). Dazu muss eine Kreuztabelle zwischen Geschlecht und Art der konsumierten Rauschmittel erstellt werden.

Mit der Prozedur „Kreuztabellen" können im Untermenü „Mehrfachantworten" Kreuztabellen zwischen einfachen Variablen, zwischen einfachen Variablen und Mehrfachantworten-Sets oder zwischen zwei Mehrfachantworten-Sets erstellt werden. Da eine dritte Variable als Kontrollvariable eingeführt werden kann, sind auch beliebige Mischungen möglich. (*Anmerkung:* Kreuztabellen zwischen einfachen Variablen wird man besser im Menü „Kreuztabellen" erstellen.)

Soll eine Kreuztabelle unter Einbeziehung eines Mehrfachantworten-Sets erstellt werden, muss dieser zunächst definiert sein. Das geschieht nach einer der beiden oben angegebenen Methoden. Um eine Kreuztabelle zu erstellen, gehen Sie wie folgt vor:

▷ Wählen Sie „Analysieren", „Mehrfachantworten ▷", „Kreuztabellen....". Es öffnet sich die Dialogbox „Mehrfachantworten: Kreuztabellen" (⇨ Abb. 12.4).

Abb. 12.4. Dialogbox „Mehrfachantworten: Kreuztabellen"

▷ Wählen Sie aus der „Variablenliste" oder im Fenster „Mehrfachantworten-Sets:" die Variable(n) oder den/die Mehrfachantworten-Set(s), welche in die Zeile(n) der Tabelle(n) kommen solle(n) (hier: $RAUSCH). Der Variablennamen oder Mehrfachantworten-Set Name erscheint im Feld „Zeile(n):". Sind darunter einfache Variablen, erscheint der Variablennamen mit einer Klammer,

in der zwei Fragezeichen stehen. Das bedeutet, dass für diese Variablen noch der Bereich definiert werden muss.

- ▷ In diesem Fall markieren Sie jeweils eine der betreffenden Variablen.
- ▷ Klicken Sie auf die Schaltfläche „Bereich definieren...". Es öffnet sich eine Dialogbox (⇨ Abb. 12.5), in der der höchste und der niedrigste gültige Wert der Zeilenvariablen angegeben wird.
- ▷ Tragen Sie den niedrigsten Wert in das Feld „Minimum:" und den höchsten in das Feld „Maximum:" ein und bestätigen Sie mit „Weiter".
- ▷ Wählen Sie aus der „Variablenliste" oder dem Feld „Mehrfachantworten-Sets:" die Variable(n) oder den/die Mehrfachantworten-Set(s) für das Feld „Spalte(n):" aus, die in die Tabellenspalte(n) kommen soll(en) (hier: V108). Für einfache Variablen wiederholen Sie die oben angegebenen Schritte zur Definition des gültigen Wertebereichs. (Wiederholen Sie die letzten Schritte gegebenenfalls für weitere Variablen.)
- ▷ Führen Sie gegebenenfalls dieselben Schritte zur Definition von Kontrollvariablen im Feld „Schicht(en)" durch.

Abb. 12.5. Dialogbox „Mehrfachantworten Kreuztabellen: Bereich definieren"

Die Abb. 12.4 und 12.5 zeigen die entsprechenden Schritte zur Vorbereitung einer Kreuztabelle zwischen Geschlecht (V108) und dem multiple Dichotomien-Set für Rauschgiftkonsum ($RAUSCH).

Optionen. Es müssen noch die Optionen für die Tabelle festgelegt werden.

- ▷ Klicken Sie auf die Schaltfläche „Optionen...". Es erscheint die Dialogbox „Mehrfachantworten Kreuztabellen: Optionen" (⇨ Abb. 12.6).
- ▷ Wählen Sie in der Gruppe *„Prozentwerte in den Zellen"* durch Anklicken der Kontrollkästchen eine oder mehrere Prozentuierungsarten aus.
  - ❒ *Zeile.* Es wird zeilenweise prozentuiert.
  - ❒ *Spalte.* Es wird spaltenweise prozentuiert.
  - ❒ *Gesamt.* Es wird auf die Gesamtzahl der Fälle prozentuiert.
- ▷ Wählen Sie in der Gruppe *„Prozentwerte bezogen auf"* aus, auf welcher Basis prozentuiert werden soll.
  - ❒ *Fälle.* Prozentuiert auf Basis der gültigen Fälle. Das ist die Voreinstellung.
  - ❒ *Antworten.* Prozentuiert auf der Basis aller gültigen Antworten.

- ▷ Wählen Sie gegebenenfalls in der Gruppe *„Fehlende Werte"* die Art, wie die fehlenden Werte von Mehrfachantworten-Sets verarbeitet werden soll. (Die Varianten sind oben genauer beschrieben.)
- ▷ Bestätigen Sie mit „Weiter" und „OK".

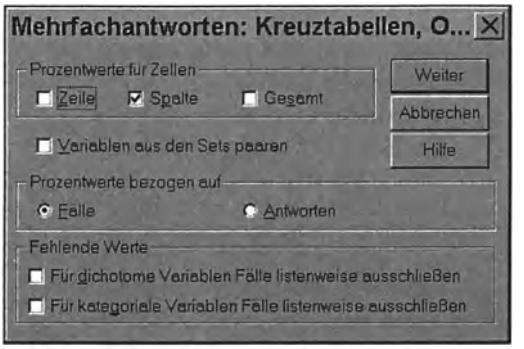


Abb. 12.6. Dialogbox „Mehrfachantworten Kreuztabellen: Optionen"

Die in den Abbildungen dargestellte Auswahl mit spaltenweiser Prozentuierung auf Basis der Fälle ergibt Tabelle 12.3.

Tabelle 12.3. Art des Rauschgiftkonsums nach Geschlecht

$RAUSCH (tabulating 1) Art des konsumierten Rauschgifts
by V108 Geschlecht

		V108		
	Count / Col pct	männlich	weiblich	Row Total
		1	2	
$RAUSCH				
	V70	43	30	73
Konsum Haschisch		91,5	93,8	92,4
	V71	11	1	12
Konsum Kokain		23,4	3,1	15,2
	V72	7	0	7
Konsum Opium		14,9	,0	8,9
	V73	3	2	5
Konsum Morphium		6,4	6,3	6,3

```
                V74        |      2  |      1  |      3
  Konsum Preludin          |    4,3  |    3,1  |    3,8
                           +---------+---------+
                V75        |     11  |      5  |     16
  Konsum Captagon          |   23,4  |   15,6  |   20,3
                           +---------+---------+
                V76        |     16  |      5  |     21
  Konsum Sonstiges         |   34,0  |   15,6  |   26,6
                           +---------+---------+
                  Column        47        32        79
                   Total      59,5      40,5     100,0

Percents and totals based on respondents

79 valid cases;  181 missing cases
```

Die Tabelle enthält zwei Spalten für die beiden als gültig deklarierten Ausprägungen der Spaltenvariablen „Geschlecht" (V108). In den Zeilen stehen alle Ausprägungen der Zeilenvariablen. In unserem Falle handelt es sich um den Mehrfachantworten-Set ($RAUSCH) mit den Ausprägungen für die Art des Rauschmittels. In den einzelnen Zellen steht oben die Absolutzahl der Nennungen („Count") für die jeweilige Kombination. Darunter steht der Prozentwert der spaltenweisen Prozentuierung („Col pct"). Aus der Tabelle geht deutlich hervor, dass (bei ansonsten ähnlichen Konsummustern) Männer häufiger bereits „Kokain", „Opium", „Captagon" und „Sonstige Rauschmittel" konsumiert haben.

Kreuzen mehrerer multipler Kategorien-Sets. Sollen zwei (oder mehr) multiple Kategorien-Sets miteinander gekreuzt werden, ist in der Dialogbox „Optionen" die Option „Variablen aus den Sets paaren" zu beachten.

Wenn zwei Mehrfachantworten-Sets miteinander gekreuzt werden, berechnet SPSS zunächst das Ergebnis auf der Basis der Elementarvariablen, d.h., es wird erst die erste Elementarvariable der ersten Gruppe mit der ersten Elementarvariablen der zweiten Gruppe gekreuzt, dann die erste Elementarvariable der ersten Gruppe mit der zweiten der zweiten usw.. Erst abschließend werden die ausgezählten Werte für die Zellen der Gesamttabelle zusammengefasst. Auf diese Weise kann es sein, dass manche Antworten mehrmals gezählt werden.

Bei multiple Kategorien-Sets kann man das durch Auswahl der Option „Variablen aus den Sets paaren" verhindern. Dann werden jeweils nur die erste Variable des ersten Sets mit der ersten des zweiten Sets, die zweite des ersten mit der zweiten des zweiten gekreuzt usw. und dann das Ergebnis in der Gesamttabelle zusammengefasst. Die Prozentuierung erfolgt bei Anwendung dieses Verfahrens auf jeden Fall auf Basis der Nennungen (Antworten) und nicht auf Basis der Fälle.

Weitere Möglichkeiten bei Verwenden der Befehlssyntax.

- Mit dem FORMAT-Unterkommando kann man das Ausgabeformat verändern. Die Labels können unterdrückt, die Tabelle mit doppelten Zeilenabstand gedruckt oder in kondensiertem dreispaltigem Format ausgegeben werden.
- Mit dem BY-Unterkommando können bis zu fünfdimensionale Tabellen erstellt werden.

Hinweis. Alle Prozentuierungen beziehen sich immer auf die gültigen Fälle. Es gibt bei der Kombination von Mehrfachantworten-Sets keine Möglichkeit, auf alle Fälle hin zu prozentuieren. Bisweilen ist das aber von Interesse. Dann bleibt nur die Umrechnung per Hand. Bei multiple Kategorien-Sets kann man dagegen eine Prozentuierung auf Basis der Gesamtzahl der Fälle mit einem Trick erreichen. In der ersten im Set enthaltenen Variablen muss eine Kategorie enthalten sein, die angibt, dass keiner der Werte zutrifft (gewöhnlich wird man diese bei allen Variablen mitführen). Diese Kategorie muss bei der ersten Variablen als gültig deklariert werden, bei den anderen (wenn vorhanden) als fehlender Wert. Die im Untermenü „Häufigkeiten" bzw. im Untermenü „Kreuztabellen" verfügbaren statistischen Maßzahlen, Tests und Grafiken stehen im Untermenü „Mehrfachantworten" nicht zur Verfügung. So wäre es für die Beispieluntersuchung auch nicht möglich, aus den entsprechenden Variablen für den 1., 2. usw. bis 4. Kredit einen durchschnittlichen Prozentsatz für die Zinsen dieser Kredite zu errechnen. Eine entsprechende Verarbeitung muss entweder per Hand geschehen oder nach Export der Ergebnisse von „Mehrfachantworten" in einem Tabellenkalkulationsprogramm.

12.5 Speichern eines Mehrfachantworten-Sets

Der so definierte Set kann nicht gespeichert werden. Er geht mit dem Ende der Sitzung verloren. Eine Wiederverwendung ist nur über die Syntax möglich. Diese können Sie aber nicht während der Definition eines Sets in der Dialogbox „Mehrfachantwortenset definieren" übertragen, sondern Sie müssen, während Sie eine Häufigkeitsauszählung oder Kreuztabellierung mit dem Set erstellen, den Befehl durch Anklicken der Schaltfläche „Einfügen" in das Syntaxfenster übertragen und anschließend die Syntax speichern. Bei einer späteren Sitzung starten Sie die Befehle in der so erstellten Syntaxdatei. Der definierte Set steht aber auch dann nicht in den Dialogboxen zur Verfügung. Sie können ihn nur innerhalb des Syntaxfensters verwenden (⇨ Kap. 2.6 und 4.2).

13 Mittelwertvergleiche und t-Tests

13.1 Überblick über die Menüs „Mittelwerte vergleichen“ und „Allgemein lineares Modell“

Die Kapitel 13 bis 15 bilden einen Komplex. Sie behandeln das Menü „Mittelwerte vergleichen“ mit seinen Untermenüs „Mittelwerte“, verschiedene „T-Test(s)“ und „Einfaktorielle ANOVA“ sowie das Menü „Allgemein lineares Modell“. In diesen Programmteilen geht es generell um Zusammenhänge zwischen zwei und mehr Variablen, wobei die abhängige Variable zumindest auf Intervallskalenniveau gemessen und per arithmetischem Mittel erfasst wird. Die unabhängigen Variablen dagegen sind kategoriale Variablen. Im Menü „Allgemein lineares Modell“ kann ergänzend eine oder mehrere mindestens auf Intervallskalenniveau gemessene Kovariate eingeführt werden.

- *Mittelwerte* (⇨ Kap. 13.2). Dieses Untermenü berechnet per Voreinstellung für jede Kategorie einer kategorialen Variable Mittelwerte und Standardabweichungen einer metrischen Variable (wahlweise zahlreiche weitere Statistiken). Ergänzt wird dieses durch die Option, eine Ein-Weg-Varianz-Analyse durchzuführen, samt der Berechnung von Eta²-Werten zur Erfassung des Anteils der erklärten Varianz. Außerdem stellt das Menü wahlweise einen Linearitätstest bereit, der es erlaubt zu prüfen, inwiefern ein Zusammenhang durch eine lineare Regression angemessen erfasst werden kann. (Letzteres ist nur bei Vorliegen einer metrischen unabhängigen Variablen – die allerdings in Klassen eingeteilt sein muss – sinnvoll.) Auf diese Optionen wird hier nicht eingegangen, weil „Einfaktorielle ANOVA“ in Kap. 14 diese ebenfalls abdeckt.
- *T-Tests* (⇨ Kap. 13.4). SPSS bietet drei Untermenüs für t-Tests. Zwei davon geben die Möglichkeit, die Signifikanz des Unterschieds von Mittelwerten zweier Gruppen zu überprüfen. Es kann sich dabei sowohl um zwei unabhängige als auch zwei abhängige (gepaarte) Stichproben handeln. Mit dem Ein-Stichproben-T-Test überprüft man den Unterschied zwischen einem Mittelwert und einem vorgegebenen Wert.
- *Einfaktorielle ANOVA* (⇨ Kap. 14). Dieses Menü zur Anwendung einer *Ein-Weg-Varianzanalyse* prüft – im Gegensatz zu den „t-Tests“ – die Signifikanz von Differenzen multipler Gruppen. Dabei wird der F-Test angewendet. Es ist aber nur möglich zu ermitteln, ob irgendeine Gruppe in signifikanter Weise vom Gesamtmittelwert abweicht. Dabei kann die Ein-Weg-Analyse – im Unterschied zur Mehr-Weg-Analyse – nur eine unabhängige Variable berücksichtigen. Aber „Einfaktorielle ANOVA“ ermöglicht auch eine Reihe von t-Tests

bzw. verwandten Tests zur Überprüfung der Signifikanz von Mittelwertdifferenzen zwischen allen bzw. beliebig vielen ausgewählten Gruppen. Diese werden als a priori bzw. post hoc Kontraste bezeichnet. Auch die Möglichkeit zur Erklärung der Variation in Form von Regressionsgleichungen ist eine Besonderheit von „Einfaktorielle ANOVA". Dabei können im Unterschied zum Linearitätstest in „Mittelwerte" auch nichtlineare Gleichungen in Form eines Polynoms bis zur 5. Ordnung verwendet werden.

- *Allgemein lineares Modell* (⇨ Kap. 15). Dieses Menü bietet die Möglichkeiten zur Mehr-Weg-Varianzanalyse. Es ist auch für Kovarianz- und Regressionsanalysen geeignet (darauf gehen wir im weiteren nicht ein). Für Post-hoc Gruppenvergleiche stehen dieselben Verfahren wie bei der einfaktoriellen ANOVA zur Verfügung. Die Möglichkeiten zur Kontrastanalyse sind etwas eingeschränkter (bei Heranziehung der Syntax allerdings umfassend). Weiterhin bestehen verschiedene Möglichkeiten, das Analysemodell zu beeinflussen. Verschiedene Optionen bieten Auswertungen für die Detailanalyse und die Überprüfung der Modellvoraussetzungen.

13.2 Das Menü „Mittelwerte"

Ähnlich wie das Menü „Kreuztabellen", dient „Mittelwerte" der Untersuchung von Zusammenhängen zwischen zwei und mehr Variablen. Die Befunde auf der abhängigen Variablen werden aber nicht durch die absolute oder relative Häufigkeit des Auftretens ihrer Ausprägungen ausgedrückt, sondern – in kürzerer Form – durch eine einzige Maßzahl, das arithmetische Mittel (andere Kennzahlen stehen wahlweise zur Verfügung). Die abhängige Variable muss, da zu ihrer Kennzeichnung gewöhnlich das arithmetische Mittel benutzt wird, zumindest auf dem Intervallskalenniveau gemessen sein. Für die unabhängige Variable genügt dagegen Nominalskalenniveau. Zur Prüfung einer Abhängigkeit wird berechnet, ob sich die Mittelwerte zwischen den verschiedenen Vergleichsgruppen (sie entsprechen den Kategorien oder Klassen der unabhängigen Variablen) unterscheiden oder nicht. Unterscheiden sie sich, spricht das dafür, dass die unabhängige Variable einen Einfluss auf die abhängige Variable besitzt, im anderen Falle muss man das Fehlen eines Zusammenhanges annehmen. Die Analyse kann, wie bei der Kreuztabellierung, durch Einführung von Kontrollvariablen verfeinert werden. Außerdem ist es möglich, einen Vergleich der Streuungen der abhängigen Variablen in den Untersuchungsgruppen (sowie wahlweise zahlreicher anderer deskriptiver Statistiken) durchzuführen.

13.2.1 Anwenden von „Mittelwerte"

Beispiel. Es soll geprüft werden, ob Männer mehr verdienen als Frauen (Datei: ALLBUS90.SAV). Zur Untersuchung dieser Fragestellung sei es ausreichend, den Durchschnittsverdienst zu betrachten, weitere Details seien nicht von Interesse (so werden mögliche weitere Einflussfaktoren wie geleistete Arbeitsstunden nicht be-

rücksichtigt). Dies ist mit dem Untermenü „Mittelwerte“ von „Mittelwerte vergleichen“ möglich.

Um eine Tabelle für den Mittelwertvergleich zu erstellen, gehen Sie wie folgt vor:

- ▷ Wählen Sie „Analysieren“, „Mittelwerte vergleichen ▷“, „Mittelwerte“. Es öffnet sich die Dialogbox „Mittelwerte“ (Abb. 13.1).
- ▷ Wählen Sie aus der Quellvariablenliste die abhängige Variable, und übertragen Sie diese in das Eingabefeld „Abhängige Variablen:“ (hier: EINK).
- ▷ Wählen Sie aus der Quellvariablenliste die unabhängige Variable, und übertragen Sie diese in das Eingabefeld „Unabhängige Variablen:“ (hier: GESCHL).
- ▷ Starten Sie den Befehl mit „OK“.

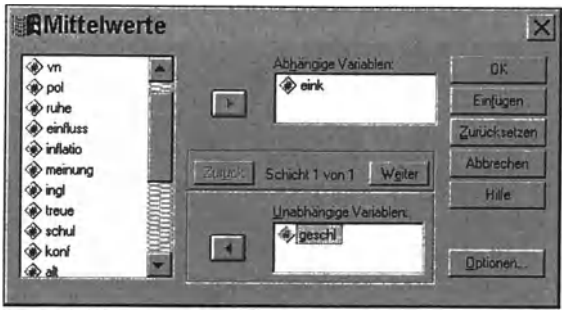


Abb. 13.1. Dialogbox „Mittelwerte“

Tabelle 13.1. Mittelwertvergleich für die Variable Einkommen nach Geschlecht

Bericht

EINK BEFR.: MONATLICHES NETTOEINKOMMEN

GESCHL GESCHLECHT,	Mittelwert	N	Standardabweichung
1 MAENNLICH	2506,30	81	1196,94
2 WEIBLICH	1561,77	62	774,57
Insgesamt	2096,78	143	1133,80

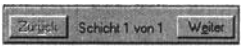
In der in Tabelle 13.1 dargestellten Ergebnisausgabe sind Geschlecht und Einkommen miteinander gekreuzt. Allerdings wird das Einkommen nur durch eine zusammenfassende Maßzahl, das arithmetische Mittel erfasst. Die beiden Ausprägungen der unabhängigen Variablen Geschlecht, MAENNLICH und WEIBLICH, bilden die Reihen dieser Tabelle. Für jede Gruppe ist in einer Spalte zunächst die hauptsächlich interessierende Maßzahl für die abhängige Variablen, das arithmetische Mittel („Mittelwert“) des Einkommens aufgeführt. In der Gesamtpopulation beträgt das Durchschnittseinkommen 2096,87 DM. Die Männer verdienen im Durchschnitt mit 2506,29 DM deutlich mehr, die Frauen mit 1561,77 DM im Monat deutlich weniger. Also hat das Geschlecht einen beträchtlichen Einfluss auf das Einkommen. Als weitere Informationen sind in den dahinter stehenden Spalten die Zahl der Fälle und die Standardabweichung aufgeführt. Die Zahl der Fälle interessiert vor allem, um die Basis der Befunde bewerten zu können. Die Stan-

dardabweichung kann ebenfalls einen interessanten Vergleich ermöglichen. So streuen in unserem Beispiel die Einkommen der Männer deutlich breiter um das arithmetische Mittel als die der Frauen. Offensichtlich handelt es sich bei den Männern um eine heterogenere Gruppe.

13.2.2 Einbeziehen einer Kontrollvariablen

Wie auch bei der Kreuztabellierung, interessiert, ob die Einbeziehung weiterer unabhängiger Variablen das Bild verändert. In unserem Beispiel wird wahrscheinlich das Einkommen auch vom Bildungsabschluss der Personen abhängen. Da Frauen bislang im Durchschnitt geringere Bildungsabschlüsse erreichen, wäre es deshalb z.B. denkbar, dass die Frauen gar nicht unmittelbar aufgrund ihres Geschlechts, sondern nur mittelbar wegen ihrer geringeren Bildungsabschlüsse niedrigere Einkommen erzielen. Auch andere Konstellationen sind denkbar. Näheren Aufschluss erbringt die Einführung von Kontrollvariablen. Das sind unabhängige Variablen, die auf einer nächsthöheren Ebene eingeführt werden. Hier wird SCHUL2 mit den Ausprägungen „Hauptschule", „Mittlere Reife" und „Abitur (einschl. Fachhochschulreife)" als Kontrollvariable verwendet.

- ▷ Zur Auswahl der unabhängigen und abhängigen gehen Sie zunächst wie in Kap. 13.2.1 beschrieben vor. Um eine Kontrollvariable einzuführen, müssen Sie dann die Ebene der unabhängigen Variablen weiterschalten.

  Dazu dient die Box 

- ▷ Klicken Sie auf die Schaltfläche „Weiter". Die Beschriftung ändert sich von „Schicht 1 von 1" in „Schicht 2 von 2" und das Eingabefeld „Unabhängige Variablen:" wird leer.
- ▷ Übertragen Sie aus der Variablenliste den Namen der Kontrollvariablen in das Eingabefeld „Unabhängige Variablen:" (hier: SCHUL2). Bestätigen Sie mit „OK".

Tabelle 13.2. Mittelwertvergleich der Einkommen nach Geschlecht und Schulbildung

Bericht

BEFR.: MONATLICHES NETTOEINKOMMEN

GESCHLECHT,	Schulbildung recodet	Mittelwert	N	Standardabweichung
MAENNLICH	Hauptschule	2214,63	40	1130,90
	Mittelschule	2895,24	21	1142,14
	Fachh/Abi	2675,00	19	1321,61
	Insgesamt	2502,63	80	1204,04
WEIBLICH	Hauptschule	1328,15	34	710,44
	Mittelschule	1897,92	12	633,39
	Fachh/Abi	1806,12	16	870,32
	Insgesamt	1561,77	62	774,57
Insgesamt	Hauptschule	1807,32	74	1053,22
	Mittelschule	2532,58	33	1091,13
	Fachh/Abi	2277,80	35	1204,87
	Insgesamt	2091,83	142	1136,26

Es ergibt sich der in Tabelle 13.2 enthaltene Output. Hier sind nun die Durchschnittseinkommen für die verschiedenen Kombinationen von Geschlecht und Schulbildung ausgewiesen. Man kann sehen, dass in jeder Schulbildungsgruppe die Frauen im Durchschnitt weniger verdienen. Also ist die Schulbildung nicht alleine der Grund für die niedrigeren Einkommen der Frauen. Allerdings hat auch die Schulbildung einen Einfluss auf das Einkommen, denn sowohl bei den Männern als auch den Frauen haben die Hauptschüler das geringste Durchschnittseinkommen, Personen mit Mittlerer Reife das höchste, und Personen mit Abitur/Fachhochschulreife liegen mit ihrem Durchschnittseinkommen dazwischen.

13.2.3 Weitere Optionen

Bei den bisherigen Beispielen wurde die Voreinstellung benutzt. Für die meisten Zwecke ergibt diese auch ein zweckmäßiges Ergebnis. Man kann allerdings mit dem Unterbefehl „Optionen...“ sowohl weitere Statistiken als auch eine Ein-Weg-Varianzanalyse sowie einen Test zur Prüfung auf eine linearen Zusammenhang zwischen den beiden Variablen anfordern.

▷ Klicken Sie auf die Schaltfläche „Optionen...“. Es erscheint die Dialogbox „Mittelwerte: Optionen“(⇨ Abb. 13.2). Sie enthält zwei Auswahlgruppen:

Abb. 13.2. Dialogbox „Mittelwerte: Optionen“

❒ *Zellenstatistik.* Im oberen Bereich des Dialogfeldes können verschiedene deskriptive statistische Kennzahlen gewählt werden, indem man sie aus der Liste „Statistik“ in das Auswahlfeld „Zellenstatistik“ überträgt. Voreingestellt sind „Mittelwert“ (arithmetisches Mittel), „Anzahl der Fälle“ und „Standardabweichung". Es die wichtigsten zahlreiche Lage- und Streuungsmaße zur Auswahl. Neben den gängigen Lagemaßen arithmetisches Mittel und Median (auch gruppiert) sind harmonische Mittel und geometrische Mittel verfügbar. Dazu kommen Schiefe und Kurtosis sowie deren Standardfehler wie auch der des

arithmetischen Mittels. Weiter Kennziffern sind Summe, höchster und niedrigster Wert. Letztlich ist es möglich, Prozente der Gesamtsumme und Prozente der Gesamtzahl ausgeben zu lassen, d.h. es wird für jede Gruppe angegeben, welchen Prozentanteil an der Gesamtsumme (z.B. des Einkommens) auf sie entfällt, bzw. welcher Anteil aller Fälle.

- ❒ *Statistik für die erste Schicht.* Es werden hier zwei statistische Analyseverfahren angeboten:
  - *ANOVA-Tabelle und Eta.* Es wird eine Ein-Weg-Varianzanalyse durchgeführt. Werden Kontrollvariablen verwendet, so werden sie bei der Varianzanalyse nicht berücksichtigt. Zusätzlich werden die statistischen Maßzahlen Eta und Eta^2 ausgegeben.
  - *Linearitätstest.* Dieser Test wird ebenfalls nur für die unabhängigen Variablen auf der ersten Ebene durchgeführt. Es wird immer das Ergebnis der Einweg-Varianzanalyse ausgegeben. Des weiteren Ergebnisse des Linearitätstests, Eta und Eta^2 sowie der Produkt-Moment-Korrelationskoeffizient R und das Bestimmtheitsmaß R^2 (nur bei unabhängigen Variablen mit mehr als zwei Ausprägungen).

Diese Analyseverfahren werden hier nicht behandelt, weil man sie auch mit dem (Unter-)Menü „Einfaktorielle ANOVA" durchführen kann (⇨ Kap. 14).

Weitere Möglichkeiten bei Verwenden der Befehlssyntax. Mit dem Unterkommando MISSING kann man die Behandlung der nutzerdefinierten fehlenden Werte beeinflussen. Der Befehl TABLE schließt sie für alle Variablen aus der Berechnung aus, der Befehl INCLUDE schließt sie in die Berechnung ein, und der Befehl DEPENDENT schließt die fehlenden Werte auf der abhängigen Variablen aus, nicht aber auf den unabhängigen Variablen.

13.3 Theoretische Grundlagen von Signifikanztests

Ein Anliegen der Forschung ist das empirische Prüfen von vermuteten Aussagen (Hypothesen) über Zusammenhänge zwischen Merkmalen in der Grundgesamtheit (Population). Gewöhnlich wird das Vorliegen eines solchen Zusammenhanges vermutet (= Hypothese H_1). Dieser Hypothese wird die Gegenhypothese H_0 gegenübergestellt, dass ein solcher Zusammenhang nicht existiere. Liegt eine Stichprobe vor, so könnte der empirisch zu beobachtende Zusammenhang (dieser zeigt sich in Unterschieden der Werte von Vergleichsgruppen oder im Unterschied der Werte einer empirischen und einer erwarteten Verteilung) der Variablen in der Stichprobe eventuell auch auf den Zufall zurückzuführen sein. Es wäre aber auch möglich, dass der Zusammenhang der Variablen in der Grundgesamtheit tatsächlich besteht. Um nun zu entscheiden, ob H_1 als statistisch gesichert angenommen werden kann oder H_0 vorläufig beibehalten werden sollte, wird ein Signifikanztest durchgeführt. Sozialwissenschaftliche Untersuchungen formulieren H_0 gewöhnlich als Punkthypothese (es besteht kein Unterschied), H_1 dagegen als Bereichshypothese (es besteht irgendeine Differenz). Die wahrscheinlichkeitstheoretischen Überlegungen gehen dann von der Annahme der Richtigkeit der Nullhy-

pothese aus, und die Wahrscheinlichkeitsverteilung wird auf dieser Basis ermittelt.[1] H_0 wird erst abgelehnt, wenn nur eine geringe Wahrscheinlichkeit von α oder eine kleinere (α = 5 % oder 1 %) dafür spricht, dass ein beobachteter Unterschied bei Geltung von H_0 durch die Zufallsauswahl zustande gekommen sein könnte. Die Hypothese H_1 wird bei dieser Art des Tests indirekt über Zurückweisen von H_0 angenommen. Deshalb wird auch H_1 üblicherweise als Alternativhypothese bezeichnet.

Die Hypothesen können sich auf sehr unterschiedliche Arten von Zusammenhängen bei unterschiedlichem Datenniveau beziehen. In den Kapiteln 14 und 15 wird die Varianzanalyse und der auf ihr basierende F-Test besprochen. Diese dienen dazu, mehrere Gruppen zugleich auf mindestens eine signifikante Differenz hin zu überprüfen. Kapitel 22 behandelt zahlreiche nichtparametrische Signifikanztests. Charakteristisch für sie ist zunächst das niedrige Messniveau der verwendeten Daten. Insbesondere aber beziehen sich die Hypothesen nicht auf einzelne Parameter, sondern auf ganze Verteilungen. So prüft der χ^2-Test, der schon in Kapitel 10.2 besprochen wurde, ob eine tatsächlich gefundene Verteilung signifikant von einer erwarteten Verteilung abweicht. Ist dies der Fall, wird die Hypothese H_1 angenommen, ansonsten H_0 beibehalten.

Anhand eines Beispiels für einen 1-Stichproben-t-Test (⇨ Kap. 13.4.1) soll das Testen von Hypothesen erläutert werden. Beispielsweise vermutet man, dass in der Bundesrepublik der durchschnittliche monatliche Nettoverdienst von männlichen Beschäftigten höher liegt als der durchschnittliche monatliche Nettoverdienst aller Beschäftigten, der für die Grundgesamtheit bekannt ist und DM 2100 beträgt. Zum Prüfen bzw. Testen dieser Hypothese stehen Verdienstdaten von Männern zur Verfügung, die aus einer Zufallsstichprobe aus der Grundgesamtheit stammen. Bei diesem Beispiel für einen statistischen Test lassen sich – wie generell bei jedem anderen statistischen Test auch – fünf Schritte im Vorgehen ausmachen (Bleymüller, Gehlert, Gülicher (2000), S. 102 ff.):

① *Aufstellung der Null- und Alternativ-Hypothese sowie Festlegung des Signifikanzniveaus.*

Die Nullhypothese (H_0) lautet: Der durchschnittliche Nettoverdienst von Männern beträgt DM 2100 (Punkthypothese). Erwartet man, dass die Männer durchschnittlich mehr verdienen, so lautet die Alternativhypothese (auch H_1-Hypothese genannt): Die durchschnittlichen Verdienste der Männer sind größer als DM 2100 (Bereichshypothese). Den durchschnittlichen Verdienst in der Grundgesamtheit (auch als Lage-Parameter bezeichnet) benennt man üblicherweise mit dem griechischen Buchstaben μ, zur Unterscheidung des Durchschnittswertes in der Stichprobe $\bar{x}$. Demgemäß lässt sich die Hypothesenaufstellung auch in folgender Kurzform formulieren:

---

[1] Werden dagegen (wie häufig in den Naturwissenschaften) genau spezifizierte Punkthypothesen gegeneinander getestet, können Wahrscheinlichkeitsverteilungen von beiden Hypothesen ausgehend konstruiert werden und die unten dargestellten Probleme der Überprüfung von H_0 lassen sich vermeiden, ⇨ Cohen, J. (1988). Das von SPSS vertriebene Programm Sample Power ist für solche Fragestellungen geeignet.

H_0: $\mu = 2100$ H_1: $\mu > 2100$

Wird die Alternativhypothese auf diese Weise formuliert, spricht man von einer gerichteten Hypothese oder einer *einseitigen* Fragestellung, da man sich bei der Alternativhypothese nur für nur eine Richtung der Unterscheidung von 2100 interessiert. Würde man die H_1-Hypothese als $\mu \neq 2100$ formulieren (weil man keine Vorstellung hat, ob der Verdienst höher oder niedriger sein könnte), so handelte es sich um eine ungerichtete Hypothese oder einen *zweiseitigen* Test.

Das Signifikanzniveau des Tests – meistens mit α bezeichnet – entspricht einer Wahrscheinlichkeit. Sie gibt an, wie hoch das Risiko ist, die Hypothese H_0 abzulehnen (weil die ausgewählten empirischen Daten der Zufallsstichprobe aufgrund eines relativ hohen durchschnittlichen Verdienstes dieses nahe legen), obwohl H_0 tatsächlich richtig ist. Die Möglichkeit, einen relativ hohen durchschnittlichen Verdienst in der Stichprobe zu erhalten, obwohl der Durchschnittsverdienst in der Grundgesamtheit tatsächlich DM 2100 beträgt, ist durch den Zufall bedingt: zufällig können bei der Stichprobenziehung hohe Verdienste bevorzugt in die Stichprobe geraten. Die Wahl eines Signifikanzniveaus in Höhe von z.B. 5 % bedeutet, dass man in 5 % der Fälle bereit ist, die richtige Hypothese H_0 zugunsten von H_1 zu verwerfen. Man bezeichnet das Signifikanzniveau α auch als Irrtumswahrscheinlichkeit, α-Fehler bzw. Fehler erster Art. Üblicherweise testet man in den Sozialwissenschaften mit Signifikanzniveaus von $\alpha = 0{,}05$ (= 5 %) bzw. $\alpha = 0{,}01$ (= 1 %).

② *Festlegung einer geeigneten Prüfgröße und Bestimmung der Testverteilung bei Gültigkeit der Null-Hypothese.*

Zur Erläuterung dieses zweiten Schritts muss man sich die Wirkung einer Zufallsauswahl von Stichproben verdeutlichen.

Dazu wollen wir einmal annehmen, dass 50000mal aus der Grundgesamtheit zufällige Stichproben gezogen und jeweils der Durchschnittsverdienst $\bar{x}$ berechnet wird. Wenn man nun eine Häufigkeitsverteilung für $\bar{x}$ bildet und grafisch darstellt, so kann man erwarten, dass sich eine glockenförmige Kurvenform ergibt, die sich über den Durchschnittsverdienst der Grundgesamtheit μ legt. Unter der Vorstellung wiederholter Stichprobenziehungen wird deutlich, dass $\bar{x}$ eine Zufallsvariable ist, die eine Häufigkeitsverteilung hat. Die Verteilung von $\bar{x}$ wird Stichprobenverteilung genannt. Aus der mathematischen Stichprobentheorie ist bekannt, dass die Verteilung von $\bar{x}$ sich mit wachsendem Stichprobenumfang n einer Normalverteilung mit dem Mittelwert μ und einer Standardabweichung von $\sigma_{\bar{x}} = \sigma_x / \sqrt{n}$ annähert (σ_x = Standardabweichung der Grundgesamtheit) (⇨ Kap. 8.4). In Abb. 13.3 rechts ist eine Stichprobenverteilung von $\bar{x}$ dargestellt. Unter der Annahme, dass H_0 richtig ist, überlagert die Verteilung den Mittelwert μ = 2100. Die Streuung der Verteilung wird mit wachsendem Stichprobenumfang n kleiner. Das Signifikanzniveau α ist am rechten Ende der Verteilung (wegen H_1:$\mu > 2100$) abgetrennt (schraffierte Teilfläche) und lässt erkennen: Wenn H_0 richtig ist, dann ist die Wahrscheinlichkeit in einer Stichprobe ein $\bar{x}$ zu erhalten, das in den schraffierten Bereich fällt, mit 5 % sehr klein. Aus diesem Grund wird die Hypothese H_0 verworfen und für die Ablehnung von H_0 entschieden. Das Ri-

siko, damit eine Fehlentscheidung zugunsten von H_1 zu treffen, also einen Fehler erster Art zu begehen (= Irrtumswahrscheinlichkeit), ist mit 5 % nur gering. Zur Durchführung des Tests wird als geeignete Prüfgröße aus Zweckmäßigkeitsgründen aber nicht $\bar{x}$, sondern die standardisierte Größe $z = \frac{(\bar{x} - \mu)}{\sigma_{\bar{x}}} = \frac{(\bar{x} - \mu)}{\sigma_x / \sqrt{n}}$ verwendet. Da aber in der Regel die Standardabweichung der Grundgesamtheit σ_x unbekannt ist, muss diese durch ihren aus der Stichprobe gewonnenen Schätzwert s ersetzt werden ($s = \sqrt{\frac{1}{n-1}\sum(x - \bar{x})^2}$). Als Prüfgröße ergibt sich dann $t = \frac{(\bar{x} - \mu)}{s / \sqrt{n}}$ bzw. in unserem Beispiel unter der Hypothese H_0: $t = \frac{(\bar{x} - 2100)}{s / \sqrt{n}}$.

Die Prüfgröße t folgt – ein Ergebnis der theoretischen Stichprobentheorie – näherungsweise einer t-Verteilung (auch Student-Verteilung genannt) mit n - 1 Freiheitsgraden (FG). Die Approximation ist umso besser, je größer der Umfang der Stichprobe ist. Man spricht daher von asymptotischen Tests. Sie wird Prüf- bzw. Testverteilung genannt.

Ist der Stichprobenumfang groß (Faustformel: $n > 30$), so kann die t-Testverteilung hinreichend genau durch die Standardnormalverteilung approximiert werden.

Bei anderen Testanwendungen ist die Prüfgröße eine andere und es wird daher auch die Prüfverteilung eine andere sein: z.B. die Standardnormalverteilung, F-Verteilung oder Chi-Quadratverteilung. Unter bestimmten Umständen ist die Approximation der Prüfgröße durch eine bekannte theoretische Wahrscheinlichkeitsverteilung zu ungenau. Dann sind exakte Tests angebracht (⇨ Kap. 29)

③ *Berechnung des Wertes der Prüfgröße.*

Die Berechnung der Prüfgröße ist in diesem Beispiel einfach. Man berechnet aus den Verdienstwerten der Stichprobe den Mittelwert $\bar{x}$ sowie den Schätzwert der Standardabweichung s und damit dann t gemäß obiger Formel. Hat man beispielsweise aus der Stichprobe mit einem Stichprobenumfang $n = 30$ einen Mittelwert von $\bar{x} = 2500$ und einer geschätzten Standardabweichung von $s = 850$ ermittelt, so erhält man $t = \frac{2500 - 2100}{850 / \sqrt{30}} = 2{,}58$.

④ *Bestimmung des Annahme- und Ablehnungsbereich.*

Die in Abbildung 13.3 dargestellte Testverteilung der Prüfgröße ist aus den genannten Gründen also eine t-Verteilung. Für unser Beispiel mit einer einseitigen Fragestellung ist die rechte Abbildung zutreffend. Das Signifikanzniveau $\alpha = 0{,}05$ (schraffierte Fläche) teilt die denkbar möglichen Werte der Prüfgröße t für die Hypothese H_0 in den Annahme- und den Ablehnungsbereich (auch kritischer Bereich genannt) auf.

Den Prüfwert (hier ein t-Wert), der die Bereiche trennt, nennt man auch den kritischen Wert (t_{krit}). Den kritischen Wert kann man aus tabellierten Prüfverteilungen für die Anzahl der Freiheitsgrade (FG) und dem Signifikanzniveau $\alpha = 0{,}05$ entnehmen. Für unser Beispiel eines einseitigen Tests ergibt sich für FG = $n - 1 = 29$ und $\alpha = 0{,}05$ $t_{krit} = 1{,}699$.

Wird aus einer Grundgesamtheit der Verdienste von Männern mit $\mu = 2100$ eine Stichprobe gezogen und die Prüfgröße t berechnet, so kann man in 5 % von Fällen erwarten, dass man eine derart hohe Prüfgröße erhält (bedingt durch die zufällige Auswahl), dass diese in den kritischen Bereich fällt. Bei einer zweiseitigen Fragestellung ist die linke Abbildung zutreffend. Sowohl sehr kleine Prüfgrößenwerte t (negative, da $\bar{x} - 2100$ negativ sein kann) als auch sehr hohe, können zur Ablehnung der Hypothese H_0 führen. Das Signifikanzniveau α verteilt sich je zur Hälfte auf beide Seiten der Prüfverteilung.

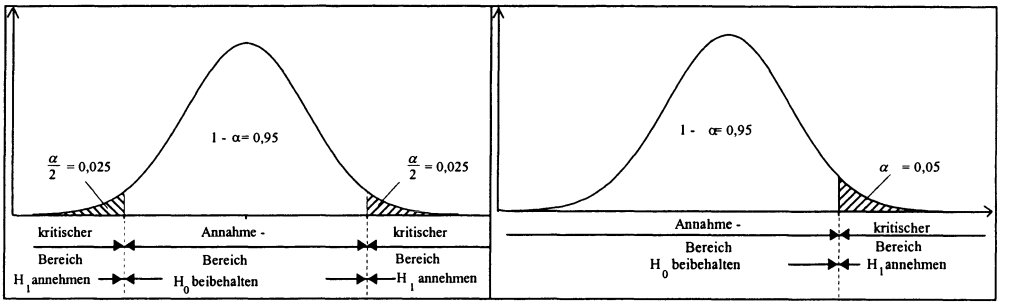


Abb. 13.3. Kritische Bereiche beim zweiseitigen und einseitigen Test

⑤ *Entscheidung für eine der Hypothesen.*

Für die in Schritt ③ berechnete Prüfgröße wird festgestellt, ob diese in den Annahmebereich oder kritischen Bereich fällt: ist sie also kleiner oder größer als der kritische Wert aus einer tabellierten Prüfverteilung. Fällt sie in den Annahmebereich, so entscheidet man sich für H_0 und fällt sie in den kritischen Bereich so für H_1. Für unser Anwendungsbeispiel kommt es wegen 2,58 > 1,699 zur Annahme von H_1.

Bei Verwendung von SPSS bleibt einem die Verwendung von tabellierten Prüfverteilungen erspart, weil SPSS nicht nur den Prüfwert t berechnet, sondern auch die zugehörige Wahrscheinlichkeit P angibt (in der Regel für den zweiseitigen Test), dass bei Geltung von H_0 der empirisch berechnete t-Prüfwert oder ein höherer zustande kommt. Führt man einen einseitigen Test durch, so wird der ausgewiesene zweiseitige P-Wert für den folgenden Vergleich halbiert. Ist $P > \alpha$, wird H_0 beibehalten. Ist umgekehrt $P < \alpha$, so entscheidet man sich für H_1.

Hinweis. Mit Hilfe der Verteilungsfunktion CDF.T(q,df) von SPSS im Menü „Transformieren" läßt sich für q = 2,58 und df = 29 berechnen, dass bei einer t-Verteilung mit 29 Freiheitsgraden die Wahrscheinlichkeit, ein t gleich oder größer 2,58 zu erhalten, gleich P = 0,01 beträgt. Wegen $P = 0{,}01 < \alpha = 0{,}05$ wird H_0 abgelehnt.

Man muss sich darüber im klaren sein, dass Signifikanztests lediglich eine Entscheidungshilfe bieten. Fehlentscheidungen werden durch sie nicht ausgeschlossen. Und zwar kann man sich sowohl fälschlicherweise für H_0 (Fehler erster Art) als auch fälschlicherweise für H_1 (Fehler zweiter Art) entscheiden. Einen

Überblick über die Fehlerrisiken gibt Tabelle 13.3. Im Signifikanztests werden die Risiken solcher Fehlentscheidung in Form einer Wahrscheinlichkeit kalkulierbar. Wird eine Punkt- gegenüber einer Bereichshypothese getestet, ist es wichtig zu sehen, dass diese beiden Risiken nicht gleichwertig behandelt werden. Die Wahrscheinlichkeit für einen Fehler erster Art steht unabhängig von anderen Faktoren mit der Wahl von α (meist 5% oder 1%) fest. Das Risiko eines Fehlers 2. Art (β) hängt nun von mehreren Faktoren ab: von α (je geringer α, desto größer β), von der Stichprobengröße n und von der Differenz der verglichenen Werte (Effektgröße). Lediglich α und die Stichprobengröße können wir frei bestimmen. Verringern wir aber mit α das Risiko eines Fehlers erster Art, erhöhen wir gleichzeitig das Risiko des Fehlers zweiter Art. Beide Entscheidungen sind also mit einem Fehlerrisiko behaftet. Im Sinne des konservativen Testmodells (⇨ Wolf (1980), Band 2, S. 89 ff.), das der Annahme einer neuen Hypothese von der Überwindung eines hohen Hindernissen abhängig macht, wirkt ein Punkt- gegen Bereichs-Test dann richtig, wenn H_1 die zu überprüfende Hypothese darstellt. Die Wahl von α sichert unabhängig von n mit hoher Wahrscheinlichkeit vor einem Fehler erster Art. Dagegen sichert sie nicht im Sinne dieses Modells, wenn H_0 die zu prüfende Hypothese darstellt, weil hier vorrangig ein Fehler zweiter Art zu vermeiden wäre und dies nicht von α, sondern vom nicht β abhängt. Soll auch das Fehlerrisiko β reduziert werden, ist das ausschließlich über die Vergrößerung der Stichprobe(n) möglich. Vorab ist dieses Fehlerrisiko nur ungenau kalkulierbar. Wegen der unterschiedlichen Art des Fehlerrisikos sprechen wir aber dann von der Annahme von H_1, wenn der Signifikanztest für H_1 spricht, dagegen bei der Entscheidung für H_0 von einem vorläufigen Beibehalten von H_0. (Sinnvollerweise wird ein Signifikanztest nur durchgeführt, wenn die Daten an sich für H_1 sprechen, z.B. die Mittelwerte von Gruppen differieren.)

Tabelle 13.3. Fehlermöglichkeiten bei der Anwendung von Signifikanztests

Entscheidung für	Objektiv richtig ist	
	H_0	H_1
H_0	richtig entschieden	Fehler 2. Art = β
H_1	Fehler 1. Art = α	richtig entschieden

Hinweise zu Problemen bei der Verwendung von Signifikanztests.

- Signifikanztests setzen voraus, dass Abweichungen gegenüber den wahren Werten als zufällig interpretiert werden können und nicht etwa auf die Wirkung systematischer Störvariablen zurückzuführen sind. Dies kann bei naturwissenschaftlichen Experimenten überwiegend vorausgesetzt werden, bei sozialwissenschaftlichen Untersuchungen aber häufig nicht.
- In sozialwissenschaftlichen Untersuchungen ist bei Verwendung sehr großer Stichproben praktisch jeder Unterschied signifikant. Deshalb ist bei Vorliegen großer Stichproben die Anwendung von Signifkanztests sinnlos. Dies liegt nicht daran, dass die Regeln der Wahrscheinlichkeitstheorie hier außer Kraft gesetzt wären. Vielmehr lie-

gen sehr schwache Beziehungen zwischen zwei Variablen bzw. schwache Wirkungen von Störvariablen praktisch immer vor. Bei solchen Untersuchungen ist nicht die Signifikanz, sondern die theoretische Bedeutsamkeit kleiner Differenzen das entscheidende Kriterium.

- Sehr oft wird in der Literatur auch die Gefahr von α-Fehlern (Fehler erster Art) beschworen. Werden in einer Untersuchung sehr viele Zusammenhänge getestet, so muss – auch wenn tatsächlich immer die Nullhypothese gilt – durch Zufallsfehler der eine oder andere Zusammenhang signifikant erscheinen. Dies spricht gegen das konzeptlose Erheben und Durchtesten von Daten. Allerdings trifft der Einwand nur dann zu, wenn es sich um eine Vielzahl unabhängig voneinander zufällig gemessener Zusammenhänge handelt. Zumeist aber bestehen zwischen den Messvariablen systematische Zusammenhänge, so dass man davon ausgehen muss, dass nicht bei jeder Variablen ein unabhängiger Zufallsfehler auftritt, sondern der einmal aufgetretene Zufallsfehler sämtliche Daten durchzieht. (*Beispiel:* Enthält eine Stichprobe per Zufall zu wenige Frauen, so wird sie deshalb auch evtl. zu wenige alte Personen enthalten, zu wenige mit höherer Schulbildung usw..)
- Bei kleinen Stichproben ist dagegen die Gefahr des β-Fehlers (Fehler zweiter Art) allgegenwärtig. Man kann zwar das Risiko eines Fehlers erster Art durch Festlegung des Signifikanzniveaus beliebig begrenzen, aber das Risiko des β-Fehlers steigt mit fallender Stichprobengröße (und geringerem Effekt) notgedrungen. Daran ändert auch die Verwendung exakter Tests nichts. Hier wird zwar die Wahrscheinlichkeit von Werten genau bestimmt, so dass der kritische Wert auch tatsächlich dem gewollten Signifikanzniveau entspricht. Damit ist das Risiko erster Art exakt unter Kontrolle, aber das Risiko für einen Fehler zweiter Art bleibt dasselbe. Die Konsequenz daraus ist: Ergibt eine kleine Stichprobe ein signifikantes Ergebnis für H_1, ist das Risiko eines Fehlers erster Art ebenso gering als hätten wir eine große Stichprobe untersucht. Müssen wir dagegen H_0 beibehalten, so kann man bei großen und mittleren Stichproben von einem geringen Fehlerrisiko zweiter Art ausgehen, bei kleinen dagegen ist dieses Fehlerrisiko sehr groß. Man sollte daher, wenn die deskriptiven Daten einer Untersuchung mit geringer Fallzahl für eine Hypothese sprechen, nicht voreilig die Hypothese verwerfen, wenn diese nicht signifikant abzusichern ist. Die Praxis, statistisch nicht signifikante Ergebnisse aus kleinen Stichproben nicht zu publizieren, lässt viele relevante Forschungsergebnisse verschwinden. Trotz der Einwände traditioneller wissenschaftstheoretischer Schulen, wird daher empfohlen, Daten kleinerer Studien zu demselben Gegenstand solange zu kumulieren, bis die Fallzahl einen hinreichend sicheren Schluss zwischen H_0 und H_1 zulässt.
- Dieses ganze Problem hängt damit zusammen, dass in den Sozialwissenschaften in der Regel eine Punkthypothese (H_0) gegen eine Bereichshypothese (H_1) getestet wird. Dadurch ist nur α, nicht aber β exakt bestimmbar. Würden Punkt- gegen Punkthypothesen getestet, könnten α und β im vorhinein festgelegt werden. Bei gegebenem Effekt kann dann auch die notwendige Mindestgröße der Stichprobe ermittelt werden, bei der eine Entscheidung mit vorgegebenem α und β möglich ist. Für solche Power-Analysen [⇨ Cohen, (1988)] bietet SPSS mit dem Programm Sample Power ein geeignetes Instrument.
- Ein besonderes Problem ergibt sich durch die Besonderheit des Punkt-gegen-Bereich-Signifkanztests, wenn die Nullhypothese die den Forscher eigentlich interessierende

Hypothese darstellt (z.B. das Vorliegen einer Normalverteilung). Will man ihm die Annahme einer falschen Nullhypothese ebenso erschweren wie dem Forscher, dessen Interesse H_1 gilt, muss man hier nicht α, sondern β niedrig ansetzen. Beta ist aber bei dieser Art von Hypothese nicht a priori bestimmbar. Bei großen und mittleren Stichproben ist das kein Problem, weil man davon ausgehen kann, dass das Fehlerrisiko Beta zwar unbekannt, aber gering ist. Dagegen stellt das bei kleinen Stichproben ein zentrales Problem dar. Man kann sagen: Ist die Stichprobe nur klein genug, kann man sicher sein, dass H_0 beibehalten werden muss. Der Forscher arbeitet also paradoxerweise mit einer Verkleinerung der Stichprobe zugunsten der Annahme seiner Hypothese. Das gilt insbesondere für Tests, die die Übereinstimmung einer Verteilung mit einer vorgegebenen Verteilungsform prüfen. Dabei handelt es sich nur um eine besondere Form der Nullhypothese. Auch hier behält man um so eher die Nullhypothese bei (und bejaht damit, dass die Verteilungsform den Bedingungen entspricht), je kleiner die Stichprobe ist. Diese Tests sind daher von zweifelhaftem Wert. Wichtig ist es, hier stattdessen geeignete Zusammenhangsmaße zu verwenden bzw. zu entwickeln.

13.4 T-Tests für Mittelwertdifferenzen

13.4.1 T-Test für eine Stichprobe

Das zur Erläuterung von Signifkanztests benutzte Beispiel soll jetzt in SPSS mit dem Datensatz ALLBUS90.SAV nachvollzogen werden. Die Hypothesen sind identisch. Die Stichprobengröße n = 81 unterscheidet sich, ebenso das für die Stichprobe der Männer ermittelte durchschnittliche Einkommen $\bar{x}$. Zur Vorbereitung wählen Sie zur Analyse nur die Männer aus (Befehlsfolge „Daten", „Fälle auswählen", „Falls Bedingung erfüllt ist", GESCHL=1).

- ▷ Wählen Sie „Analysieren", „Mittelwerte vergleichen", „T-Test bei einer Stichprobe...". Es erscheint die Dialogbox „T-Test bei einer Stichprobe" (⇨ Abb. 13.4).
- ▷ Übertragen Sie aus der Quellvariablenliste die Variable EINK in das Feld „Testvariable(n):".
- ▷ Tragen Sie in das Feld „Testwert" den gewünschten Wert ein (hier: 2100) und bestätigen Sie mit „OK".

Abb. 13.4. Dialogbox „T-Test bei einer Stichprobe"

In Tabelle 13.4 sieht man den Output. In der oberen Tabelle Spalte „Mittelwert" erkennt man, dass das mittlere Einkommen der 81 befragten Männer 2506,30 DM, also nicht 2100 beträgt. Die Differenz zum vorgegebenen Wert „Mittlere Differenz" (besser Mittelwertdifferenz) beträgt 406,30 DM. Die Frage ist, ob die Abweichung von 406,30 DM mit noch zu hoher Wahrscheinlichkeit zufallsbedingt sein könnte.

Tabelle 13.4. T-Test bei einer Stichprobe für die Differenz zwischen dem Mittwert des Einkommens der Männer und dem Testwert 2100

Statistik bei einer Stichprobe

	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
BEFR.: MONATLICHES NETTOEINKOMMEN	81	2506,30	1196,94	132,99

Test bei einer Sichprobe

	Testwert = 2100					
					95% Konfidenzintervall der Differenz	
	T	df	Sig. (2-seitig)	Mittlere Differenz	Untere	Obere
BEFR.: MONATLICHES NETTOEINKOMMEN	3,055	80	,003	406,30	141,63	670,96

Ein Maß für die Streuung der Werte in der Stichprobe ist die Standardabweichung 1196,94 DM. Aus der Standardabweichung in der Stichprobe und der Stichprobengröße kann man den Standardfehler des Mittelwertes von ± 132,99 für die Verteilung unendlich vieler Stichproben schätzen ($= 1196{,}934/\sqrt{81}$). Diesen nutzt man zur Konstruktion eines Konfidenzintervalls. SPSS gibt es für 95 %-Sicherheit (entspricht $\alpha = 0{,}05$) aus („95 % Konfidenzintervall der Differenz"). Die untere Grenze liegt bei 141,63, die obere bei 670,96. Schon daraus ersieht man, dass es unwahrscheinlich ist, dass eine Differenz von 406,30 zum H_0-Wert ($\mu = 2100$) durch Zufall zustande gekommen ist. Dieselbe Auskunft gibt der t-Test. Bei Geltung der Nullhypothese hat ein t von 3,055 (bzw. größer) bei df = 80 Freiheitsgraden eine Wahrscheinlichkeit von 0,003 [„Sig (2-seitig)"]. Das ist wesentlich weniger als der Grenzwert $\alpha = 0{,}05$. Die Hypothese H_1 wird also angenommen.

13.4.2 T-Test für zwei unabhängige Stichproben

Mit dem t-Test für Mittelwertdifferenzen werden die Unterschiede der Mittelwerte zweier Gruppen auf Signifikanz geprüft. Dabei ist zu unterscheiden, ob es sich bei den Vergleichsgruppen um unabhängige oder abhängige Stichproben handelt. Der übliche t-Test dient dem Vergleich zweier unabhängiger Stichproben. Mitunter werden aber auch abhängige Stichproben verglichen.

- *Unabhängige Stichproben.* Es sind solche, bei denen die Vergleichsgruppen aus unterschiedlichen Fällen bestehen, die unabhängig voneinander aus ihren Grundgesamtheiten gezogen wurden (z.B. Männer und Frauen).
- *Abhängige (gepaarte) Stichproben.* Es sind solche, bei denen die Vergleichsgruppen entweder aus denselben Untersuchungseinheiten bestehen, für die bestimmte Variablen mehrfach gemessen wurden (z.B. zu verschiedenen Zeitpunkten, vor und nach der Einführung eines experimentellen Treatments) oder bei denen die Untersuchungseinheiten der Vergleichsgruppen nicht unabhängig ausgewählt wurden. Letzteres könnte etwa vorliegen, wenn bestimmte Variablen für Ehemann und Ehefrau verglichen werden oder wenn die Vergleichsgruppen nach dem Matching-Verfahren gebildet wurden. Bei diesem Verfahren wird für jeden Fall einer Testgruppe nach verschiedenen relevanten Kriterien ein möglichst ähnlicher Fall für die Vergleichsgruppe(n) ausgewählt. Dadurch werden die Einflüsse von Störvariablen konstant gehalten.

Wir behandeln zunächst den t-Test für unabhängige Stichproben. Dabei macht es weiter einen Unterschied, ob die Varianzen der beiden Gruppen gleich sind oder sich unterscheiden. Man unterscheidet daher:

- Klassischer t-Test für unabhängige Gruppen mit *gleicher Varianz.*
- T-Test für unabhängige Gruppen mit *ungleicher Varianz.*

Test auf Gleichheit der Varianzen. Ist es unklar, ob die Varianzen der beiden Grundgesamtheiten als gleich angesehen werden können, sollte man zunächst einen Test auf Gleichheit der Varianzen durchführen. SPSS bietet den *Levene-Test* an (⇨ Kap. 9.3.1). Man sollte den „t-Test bei ungleicher Varianz" benutzen, wenn die Varianz ungleich ist, weil sonst falsche Ergebnisse herauskommen können. Andererseits führt die Anwendung dieses Tests bei gleicher Varianz zu einem etwas zu hohen Signifikanzniveau. Deshalb sollte, wenn es möglich ist, der t-Test für gleiche Varianz angewendet werden.

Die t-Tests setzen folgendes voraus:

- Die abhängigen Variable ist mindestens auf Intervallskalenniveau gemessen.
- Normalverteilung der abhängigen Variablen in der Grundgesamtheit.
- In der klassischen Version verlangt er Homogenität der Varianz, d.h. nahezu gleiche Varianz in den Vergleichsgruppen.
- Zufällige Auswahl der Fälle bzw. beim Vergleich abhängiger Stichproben, der Paare.

13.4.2.1 Die Prüfgröße bei ungleicher Varianz

Da der t-Test für unabhängige Gruppen mit ungleicher Varianz den allgemeineren Fall behandelt, erklären wir zuerst ihn. Dabei kann man an die Gleichung $t = \frac{(\bar{x} - \mu)}{s/\sqrt{n}}$ in Kap. 13.3 anknüpfen. Im Unterschied geht es nun nicht um den (im Zähler stehenden) Unterschied eines Stichprobenergebnisses $\bar{x}$ zum H_0-Wert μ, sondern um den Unterschied einer Stichprobendifferenz $\bar{x}_1 - \bar{x}_2$ zum H_0-Wert

$\mu_1 - \mu_2 = 0$. Auch die (im Nenner stehende) Standardabweichung der Stichprobenverteilung ist natürlich verschieden. Für die Stichprobenverteilung von $\bar{x}_1 - \bar{x}_2$ gilt, dass sie eine normalverteilte Zufallsvariable ist mit der Standardabweichung (= Standardfehler):

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{13.1}$$

Analog zu den Ausführungen in Kap. 13.3 sind s_1^2 und s_2^2 als Schätzwerte für die Varianzen der Grundgesamtheiten nach der Formel $\frac{\sum(x-\bar{x})^2}{n-1}$ zu berechnen (nicht mit n, sondern $n-1$ im Nenner).

Die Prüfgröße t ist unter der Hypothese H_0 (die Differenzen der Mittelwerte der beiden Grundgesamtheiten unterscheiden sich nicht, d.h. $\mu_1 - \mu_2 = 0$) die Differenz zwischen den beiden Samplemittelwerten, ausgedrückt in Einheiten des Standardfehlers:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{13.2}$$

Die Wahrscheinlichkeitsverteilung der Prüfgröße t entspricht einer t-Verteilung (auch Student Verteilung genannt). Aus ihr lässt sich die Wahrscheinlichkeit für einen empirisch ermittelten t-Wert bei den jeweils für die Stichprobengröße geltenden Freiheitsgraden ablesen. Für hinreichend große Stichproben (Faustformel $n \geq 30$) lässt sich die t-Verteilung durch die Normalverteilung approximieren.

Die Zahl der Freiheitsgrade (degrees of freedom df) ergibt sich aus der Formel:

$$df = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2)\right]^2}{\left[(s_1^2/n_1)^2/(n_1-1)\right] + \left[(s_2^2/n_2)^2/(n_2-1)\right]} \tag{13.3}$$

Es ergibt sich dabei gewöhnlich keine ganze Zahl, aber man kann näherungsweise die nächste ganze Zahl verwenden.

Die t-Tabelle enthält üblicherweise Angaben für bis zu 30 Freiheitsgrade. Bei höheren Stichprobengrößen n kann approximativ mit z-Werten der Standardnormalverteilung gearbeitet werden. So beträgt für hinreichend große Stichproben bei dem Wert 1,96 die Irrtumswahrscheinlichkeit 5 %.

13.4.2.2 Die Prüfgröße bei gleicher Varianz

Die obige Interpretation der Prüfgröße t – Differenz zwischen den beiden Samplemittelwerten, ausgedrückt in Einheiten des Standardfehlers – gilt auch hier. Aber der Standardfehler der Stichprobenverteilung wird in diesem Falle anders berechnet. Oben wird er auf Basis der – gegebenenfalls unterschiedlichen – beobachteten Varianzen der beiden verglichenen Stichproben geschätzt. Die Formel ist deshalb auf den Fall anwendbar, dass die Stichproben aus zwei Grundgesamthei-

ten mit unterschiedlicher Varianz stammen. Allerdings wird dadurch die Berechnung der Freiheitsgrade recht kompliziert.

Der hier besprochene klassische t-Test geht dagegen von gleichen Varianzen in den beiden Populationen aus. Wie alle Signifikanztests, geht auch der t-Test vom Ansatz her von der Nullhypothese aus. Diese unterstellt, dass die beiden Stichproben aus einer und derselben Grundgesamtheit mit demselben arithmetischen Mittel μ und derselben Varianz σ^2 stammen. Die empirisch beobachteten Unterschiede zwischen den arithmetischen Mitteln und den Varianzen der beiden Stichproben werden als durch die Zufallsauswahl entstanden unterstellt. Deshalb geht das klassische Modell des t-Tests auch davon aus, dass beide Vergleichsgruppen die gleiche Varianz haben. Entsprechend wird die Standardabweichung der Stichprobenverteilung (= Standardfehler)n icht auf Basis zweier unterschiedlicher Varianzen, sondern gleicher Varianzen geschätzt. Als Schätzwert für die wahre gemeinsame Varianz der beiden Stichproben wird daher das gewogene arithmetische Mittel beider Varianzen ermittelt (man spricht auch von gepoolter Varianz, deshalb Index P). Es ergibt sich für die geschätzte (gepoolte) Varianz:

$$s_p^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{(n_1 - 1) + (n_2 - 1)} \tag{13.4}$$

Anstelle der beiden Stichprobenvarianzen s_1^2 und s_2^2 wird dieser Schätzwert in die Gleichung 13.2 eingesetzt. Die Prüfgröße t errechnet sich demnach:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{13.5}$$

Die Zahl der Freiheitsgrade ist $df = n_1 + n_2 - 2$. Dieser Test wird auch als gepoolter t-Test bezeichnet.

13.4.2.3 Anwendungsbeispiel

Es soll untersucht werden, ob sich das Durchschnittseinkommen von Männern und Frauen unterscheidet (Datei: ALLBUS90.SAV). Dass dies in unserer Stichprobe der Fall ist, haben wir schon bei der Anwendung von „Mittelwerte" gesehen. Jetzt soll aber zusätzlich mit Hilfe des t-Tests geprüft werden, ob dieser Unterschied auf zufällige Auswahlschwankungen zurückzuführen sein könnte oder mit hinreichender Sicherheit ein realer Unterschied vorliegt.

Es handelt sich hier um zwei unabhängige Stichproben, nämlich um verschiedene Untersuchungsgruppen: Männer und Frauen. Der t-Test für unabhängige Stichproben kommt daher als Signifikanztest in Frage. Um einen t-Test durchzuführen, gehen Sie wie folgt vor:

- ▷ Wählen Sie falls nicht bereits erfolgt wieder alle Fälle zur Analyse aus.
- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Mittelwerte vergleichen ▷", „T-Test bei unabhängigen Stichproben...". Es erscheint die Dialogbox „T-Test bei unabhängigen Stichproben" (⇨ Abb. 13.5).

- ▷ Wählen Sie aus der Variablenliste zunächst die abhängige Variable (hier: EINK), und übertragen Sie diese in das Eingabefeld „Testvariable(n):".
- ▷ Wählen Sie aus der Variablenliste die unabhängige Variable (hier: GESCHL), und übertragen Sie diese in das Eingabefeld „Gruppenvariable:".
- ▷ Markieren Sie „geschl(? ?)", und klicken Sie auf die Schaltfläche „Gruppen def. ...". Die Dialogbox „Gruppen definieren" öffnet sich (⇨ Abb. 13.6).

Abb. 13.5. Dialogbox „T-Test bei unabhängigen Stichproben"

- ▷ Klicken Sie den Optionsschalter „Angegebene Werte verwenden" an, und geben Sie in die Eingabefeld „Gruppe 1:" und „Gruppe 2:" die Variablenwerte der beiden Gruppen an, die verglichen werden sollen (hier für GESCHL die Werte 1 und 2).
  (Hinweis. Liegt eine ordinalskalierte oder metrische Variable als unabhängige Variable vor, kann anstelle von diskreten Werten ein Teilungspunkt festgelegt werden. Dadurch werden zwei Gruppen, eine mit hohen und eine mit niedrigen Werten, gebildet, die verglichen werden sollen. In diesem Falle klicken Sie „Trennwert" an und geben den Teilungspunkt in das Eingabekästchen ein).
- ▷ Bestätigen Sie mit „Weiter".

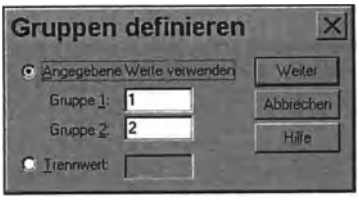


Abb. 13.6. Dialogbox „Gruppen definieren"

Optionen. Sollten Sie die Voreinstellung für das Signifikanzniveaus sowie der Behandlung der fehlenden Werte verändern wollen:

- ▷ Klicken Sie die Schaltfläche „Optionen..." an. Die Dialogbox „T-Test bei unabhängigen Stichproben: Optionen" erscheint (⇨ Abb. 13.7).

- *Konfidenzintervall.* Durch Eingabe eines beliebigen anderen Wertes in das Eingabefeld „Konfidenzintervall:" können Sie das Signifikanzniveau ändern. Üblich ist neben den voreingestellten 95 % (entspricht 5 % Fehlerrisiko) das Sicherheitsniveau 99 % (entspricht 1 % Fehlerrisiko).
- *Fehlende Werte.* Falls Sie mehrere abhängige Variablen definiert haben, können Sie in dieser Gruppe durch Anklicken von „Fallausschluss Test für Test" (Voreinstellung) dafür sorgen, dass nur Fälle ausgeschlossen werden, bei denen in den gerade analysierten abhängigen und unabhängigen Variablen ein fehlender Wert auftritt. „Listenweiser Fallausschluss" dagegen sorgt dafür, dass alle Fälle, in denen in irgendeiner dieser Variablen ein fehlender Wert auftritt, aus der Analyse ausgeschlossen werden.

▷ Bestätigen Sie mit „Weiter" (alle Eintragungen sind jetzt in der Dialogbox „T-Test bei unabhängigen Stichproben" vorgenommen) und „OK".

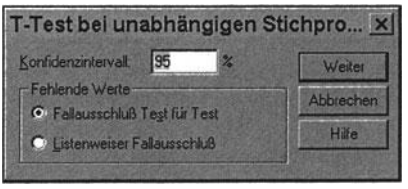


Abb. 13.7. Dialogbox „T-Test bei unabhängigen Stichproben: Optionen"

Die vorgeschlagenen Eingaben ergeben für das Beispiel die Tabelle 13.5. Es wird zunächst eine ähnliche Tabelle wie bei „Mittelwerte" ausgegeben. Wir sehen daraus, dass Angaben von 81 Männern und 62 Frauen vorliegen. Das Durchschnittseinkommen der Männer ist mit 2506,30 DM deutlich höher als das der Frauen mit 1561,77. Interessant sind die Angaben für die Standardabweichung in der vorletzten Spalte. Diese ist bei den Männern mit 1196,94 erheblich größer als bei den Frauen mit 774,57.

Das spricht dafür, dass wir es nicht mit Grundgesamtheiten mit gleicher Streuung zu tun haben. Das bestätigt auch *Levene-Test*, dessen Ergebnisse am Anfang der unteren Tabelle stehen. Dieser Test wird standardmäßig mitgeliefert. Es ist ein F-Test, der auf dem Vergleich der Varianzen beider Stichproben beruht. Der F-Wert beträgt laut Output 10,165.

Ein F dieser Größenordnung ist bei Geltung von H_0 – einer gleichen Varianz in den Gruppen – äußerst unwahrscheinlich. Die Wahrscheinlichkeit beträgt 0,2 Prozent („Signifikanz = 0,002"). Also stammen diese beiden Streuungen mit an Sicherheit grenzender Wahrscheinlichkeit nicht aus Grundgesamtheiten mit gleicher Varianz.

Deshalb müssen wir hier von den beiden ausgedruckten t-Test-Varianten die in der untersten Reihe („Varianzen sind nicht gleich") angegebene Variante für Stichproben mit ungleicher Varianz verwenden.

Hier ist der t-Wert mit 5,710 angegeben, die Zahl der Freiheitsgrade mit 137,045 und die Wahrscheinlichkeit dafür, dass ein solches Ergebnis bei Geltung

von H_0 – Differenz der Mittelwerte gleich Null – zustande kommen könnte, für einen zweiseitigen Test [„Sig (2-seitig)"]. Diese Wahrscheinlichkeit ist so gering, dass der Wert 0,000 angegeben ist. Also ist die Differenz der Einkommen zwischen Männern und Frauen mit an Sicherheit grenzender Wahrscheinlichkeit real und kein Produkt zufälliger Verzerrungen durch die Stichprobenauswahl.

Tabelle 13.5. T-Test für die Einkommensdifferenzen nach Geschlecht

Gruppenstatistiken

GESCHLECHT, BEFRAGTE<R>	BEFR.: MONATLICHES NETTOEINKOMMEN			
	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
MAENNLICH	81	2506,30	1196,94	132,99
WEIBLICH	62	1561,77	774,57	98,37

Test bei unabhängigen Stichproben

	BEFR.: MONATLICHES NETTOEINKOMMEN								
	Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
								95% Konfidenzintervall der Differenz	
	F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	Untere	Obere
Varianzen sind gleich	10,165	,002	5,405	141	,000	944,52	174,75	599,06	1289,99
Varianzen sind nicht gleich			5,710	137,504	,000	944,52	165,42	617,42	1271,62

Bei diesem Beispiel würde sich auch ein einseitiger Test rechtfertigen, da man ausschließen kann, dass Frauen im Durchschnitt mehr verdienen als Männer. Die Wahrscheinlichkeit könnte dann durch zwei geteilt werden. Da sie aber in diesem Fall ohnehin nahe Null ist, erübrigt sich das.

13.4.3 T-Test für zwei abhängige (gepaarte) Stichproben

Bestehen die abhängigen Vergleichsgruppen beispielsweise aus denselben Fällen, für die eine Variable mehrfach gemessen wurde, können zufällige Schwankungen bei der Stichprobenziehung keine Unterschiede zwischen den Vergleichsgruppen hervorrufen. Als zufällige Schwankungen sind lediglich noch zufällige Messfehler relevant. Deshalb werden bei abhängigen Samples auch nicht die Mittelwerte von Vergleichsgruppen als Zufallsvariablen behandelt, sondern die Differenzen der Messwerte von Vergleichspaaren. Die Zufallsvariable $D = x_1 - x_2$ wird aus der Differenz der beiden Werte für jedes Messpaar gebildet. D ist unter der Hypothese H_0 normal verteilt mit einem Mittelwert 0. T überprüft dann die Nullhypothese, dass die mittlere Differenz $\overline{D}$ zwischen den zwei Vergleichsmessungen in der Population gleich 0 ist. Die Prüfgröße t ist dann:

$$t = \frac{\overline{D}}{\frac{s_D}{\sqrt{n}}} \tag{13.6}$$

Wobei n die Zahl der Paare, s_D die Standardabweichung der Differenzen der paarweisen Vergleiche und $\overline{D}$ der Durchschnitt der Differenzen der Vergleichspaare ist. Die Prüfgröße ist t-verteilt mit $n-1$ Freiheitsgraden.

Beispiel. Es soll (Datei ABM.SAV) das Einkommen von Teilnehmern an einer Arbeitsbeschaffungsmaßnahme vor (VAR225) und nach (VAR310) der Maßnahme sowie vor und während der Maßnahme (VAR233) verglichen werden. Es handelt sich hier um zwei abhängige Stichproben, denn es wird das Einkommen derselben Personen zu jeweils zwei verschiedenen Zeitpunkten verglichen.

Um einen t-Test für zwei abhängige Stichproben (T-Test für gepaarte Stichproben) durchzuführen, gehen Sie wie folgt vor:

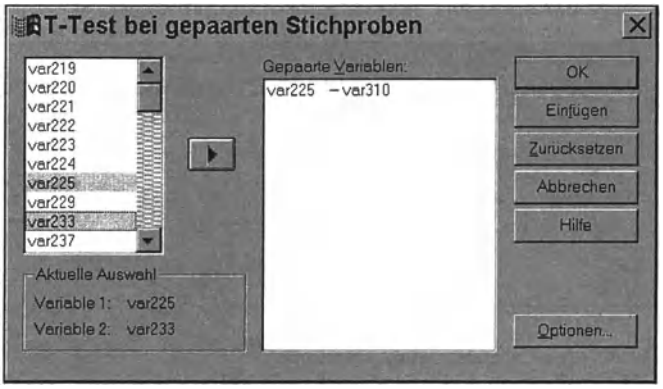


Abb. 13.8. Dialogbox „T-Test bei gepaarten Stichproben"

- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Mittelwerte vergleichen ▷", „T-Test bei gepaarten Stichproben...". Es erscheint die Dialogbox „T-Test bei gepaarten Stichproben" (⇨ Abb. 13.8).
- ▷ Übertragen Sie durch Anklicken aus der Quellvariablenliste die erste der beiden zu vergleichenden Variablen in das Feld „Aktuelle Auswahl" (hier: VAR225). Sie steht danach hinter „Variable 1:"
- ▷ Wiederholen Sie das für die zweite Vergleichsvariable (hier: VAR310).
- ▷ Klicken Sie dann auf ▸. Die beiden Variablen werden in das Eingabefeld Eingabefeld „Gepaarte Variablen:" übertragen.

(Sie können dies für weitere Vergleiche wiederholen, so dass mehrere Paare im Auswahlfeld untereinander stehen. Abb. 13.8 zeigt z.B. die Dialogbox vor Übertragung des zweiten Paars für die Variablen VAR225 und VAR233.)

Optionen. Wenn Sie wollen, können Sie die voreingestellten Werte für das Konfidenzintervall und die Behandlung der fehlenden Werte ändern.

- ▷ Klicken Sie dafür auf die Schaltfläche „Optionen...". Es erscheint die Dialogbox „T-Test bei gepaarten Stichproben: Optionen". Diese ist mit Ausnahme der Überschrift mit der in Abb. 13.7 dargestellten Dialogbox identisch.
- ▷ Nehmen Sie die Einstellungen vor (es sind dieselben Einstellmöglichkeiten wie beim t-Test bei unabhängigen Stichproben), und bestätigen Sie diese mit „Weiter". Bestätigen Sie die Eingaben mit „OK".

Tabelle 13.6. T-Test für die Differenzen zwischen den Einkommen vor und nach einer Arbeitsbeschaffungsmaßnahme (ABM)

Statistik bei gepaarten Stichproben

Paaren							
BRUTTOEINKOMMEN VOR ABM				ERSTES BRUTTOEINK.NACH ABM			
Mittelwert	N	Standardabweichung	Standardfehler des Mittelwertes	Mittelwert	N	Standardabweichung	Standardfehler des Mittelwertes
2783,54	80	1284,75	143,64	2631,80	80	920,82	102,95

Korrelationen bei gepaarten Stichproben

		N	Korrelation	Signifikanz
Paaren	BRUTTOEINKOMMEN VOR ABM & ERSTES BRUTTOEINK.NACH ABM	80	,644	,000

Test bei gepaarten Stichproben

		Gepaarte Differenzen					T	df	Sig. (2-seitig)
		Mittelwert	Standardabweichung	Standardfehler des Mittelwertes	95% Konfidenzintervall der Differenz Untere	Obere			
Paaren	BRUTTOEINKOMMEN VOR ABM - ERSTES BRUTTOEINK.NACH AB	151,74	987,72	110,43	-68,07	371,54	1,374	79	,173

Tabelle 13.6 zeigt die Ausgabe für den Vergleich zwischen dem Einkommen vor (VAR225) und nach (VAR310) der Arbeitsbeschaffungsmaßnahme. Die obere Tabelle enthält zunächst einige beschreibende Angaben. 80 Paarvergleiche haben stattgefunden („N"). Das Durchschnittseinkommen („Mittelwert") vorher war 2783,54, nachher 2631,80 DM. Es scheint also etwas gesunken zu sein. Die Streuung, gemessen durch die Standardabweichung, bzw. der Standardfehler des Mittelwertes waren vorher etwas größer als nachher. Eine zweite Teiltabelle gibt die Korrelation der Einkommen zwischen den beiden Zeitpunkten an. Sie ist mit 0,644 recht hoch und, wie die dazugehörige Fehlerwahrscheinlichkeit („Signifi-

kanz") ausweist, auch hoch signifikant. In der unteren Tabelle „Test bei gepaarten Stichproben" stehen die Angaben zum t-Tests für abhängige Stichproben. Das arithmetische Mittel („Mittelwert") der Differenz zwischen den Einkommen vor und nach der Maßnahme beträgt 151,74 DM. (Obwohl der zufällig gleich ausfällt, ist dies nicht zu verwechseln mit der Differenz zwischen den Mittelwerten zu beiden Zeitpunkten, hier wird zunächst für jeden Fall die Differenz berechnet und aus diesen Differenzen der Mittelwert gebildet). Die Standardabweichung dieser Differenzen beträgt ± 987,72 DM und der Standardfehler ± 110,43 DM. Um für den Mittelwert ein 95 %-Konfidenzintervall zu berechnen, multipliziert man den Standardfehler mit dem entsprechenden Sicherheitsfaktor t. Aus einer t-Tabelle kann man diesen bei $df = 79$ und $\alpha = 0{,}05$ mit $\approx 1{,}99$ ermitteln. Schlägt man den so ermittelten Wert dem Mittelwert zu, ergibt sich die Obergrenze des Konfidenzintervalls, vom Mittelwert abgezogen die Untergrenze (⇨ Kap. 8.4). Diese Intervallgrenzen betragen −68,07 DM und 371,54. Dieses Intervall („95% Konfidenzintervall der Differenz") ist in der Tabelle schon berechnet angegeben. In diesem Bereich liegt mit 95prozentiger Sicherheit der wahre Wert. Er könnte also auch 0 sein. Diesem Ergebnis entspricht, dass bei Geltung von H_0 der t-Wert 1,374 bei den gegebenen 79 Freiheitsgraden beim zweiseitigen t-Test eine Wahrscheinlichkeit von 0,173 oder 17 % aufweist. Es ist also nicht mit hinreichender Sicherheit auszuschließen, dass die Differenz nur auf Zufallsschwankungen zurückzuführen ist und keine reale Differenz existiert. H_0 wird vorläufig beibehalten.

14 Einfaktorielle Varianzanalyse (ANOVA)

Während der t-Test geeignet ist, zwei Mittelwerte zu vergleichen und ihre evtl. Differenz auf Signifikanz zu prüfen, können mit der Varianzanalyse mehrere Mittelwerte zugleich untersucht werden. Die Varianzanalyse hat dabei zwei Zielsetzungen:

- Sie dient der Überprüfung der Signifikanz des Unterschiedes von Mittelwertdifferenzen. Sie zeigt dabei auf, ob mindestens ein Unterschied zwischen multiplen Vergleichsgruppen signifikant ausfällt. Darüber, um welchen oder welche es sich handelt, ermöglicht sie keine Aussage. Als Signifikanztest wird der F-Test verwendet.
- Sie dient zur Ermittlung des von einer oder mehreren unabhängigen Variablen erklärten Anteils der Gesamtvarianz.

Voraussetzungen für die Varianzanalyse sind:

- Eine auf Intervallskalenniveau oder höher gemessene abhängige Variable, auch als Kriteriumsvariable bezeichnet.
- Normalverteilung der Kriteriumsvariablen in der Grundgesamtheit.
- Mindestens eine unabhängige Variable, die eine Aufteilung in Gruppen ermöglicht. Diese Variable wird auch als Faktor bezeichnet. Es reicht dazu eine auf Nominalskalenniveau gemessene Variable. Auch metrische Variablen können Verwendung finden. Aber bei kontinuierlichen oder quasi kontinuierlichen Variablen müssen geeignete Klassen gebildet werden. Sie werden danach wie kategoriale Variablen verwendet.
- Die Vergleichsgruppen müssen unabhängige Zufallsstichproben sein.
- Die Vergleichsgruppen sollten in etwa gleiche Varianzen haben.

Die *einfaktorielle (Ein-Weg) Varianzanalyse* berücksichtigt lediglich einen Faktor. Die *multifaktorielle (Mehr-Weg) Varianzanalyse* dagegen n Faktoren.

SPSS bietet im Menü „Mittelwerte vergleichen" sowohl im Untermenü „Mittelwerte" (als Option) als auch im Untermenü „Einfaktorielle ANOVA" eine Ein-Weg-Varianzanalyse an. Auch das Menü „Univariat", das einzige Untermenü des Menüs "Allgemeines lineares Modell" im Basismodul, das für Mehr-Weg-Analysen gedacht ist, kann für Ein-Weg-Analysen verwendet werden. Allerdings ist „Einfaktorielle ANOVA" etwas einfacher aufgebaut und bietet etwas andere Features zur Prüfung der Signifikanz von Einzeldifferenzen zwischen Gruppen und zur Prüfung verschiedener Gleichungsformen zur Varianzerklärung, die in den anderen Prozeduren entweder nicht oder (Univariat) in etwas eingeschränkter Form zur Verfügung stehen.

In diesem Kapitel wird auf die Anwendung von „Einfaktorielle ANOVA“ eingegangen.

14.1 Theoretische Grundlagen

Varianzzerlegung. Die Grundgedanken der Varianzanalyse sollen zunächst an einem fiktiven Beispiel mit wenigen Fällen dargestellt werden, das später mit realen Zahlen ausgebaut wird. Es sei das Einkommen von 15 Personen untersucht. Die Daten sind so konstruiert, dass die 15 Personen ein mittleres Einkommen von $\bar{x}_T$ = 2.500 DM haben (Index T für total). Die Einkommenswerte für die einzelnen Personen streuen um diesen Mittelwert. Die Streuung wird von der Variablen Schulbildung – auch als Faktor bezeichnet – beeinflusst: Personen mit mittlerer Reife (Index M) erhalten das Durchschnittseinkommen, Abiturienten (Index A) erhalten dagegen einen Zuschlag von DM 500, Hauptschulabsolventen (Index H) einen Abschlag derselben Größe. Innerhalb der Schulbildungsgruppen schwanken aufgrund nicht näher bestimmter Ursachen die Einkommen und zwar so, dass eine der fünf Personen genau das mittlere Einkommen der Gruppe verdient, zwei verdienen 100 bzw. 200 DM mehr als der Durchschnitt, zwei 100 bzw. 200 DM weniger. Tabelle 14.1 enthält die Daten der so konstruierten Fälle, bereits eingeteilt in die Gruppen des Faktors Schulbildung. In der Tabelle werden mit $\bar{x}$ auch die Durchschnittseinkommen der Personen einer jeden Schulbildungsgruppe ausgewiesen.

Tabelle 14.1. Einkommen nach Schulabschluss (fiktive Daten)

Hauptschule	Mittlere Reife	Abitur
1.800	2.300	2.800
1.900	2.400	2.900
2.000	2.500	3.000
2.100	2.600	3.100
2.200	2.700	3.200
Σ = 10.000	Σ = 12.500	Σ = 15.000
$\bar{x}_H$ = 2.000	$\bar{x}_M$ = 2.500	$\bar{x}_A$ = 3.000

Wie wir aus der beschreibenden Statistik wissen, sind die Variation (die Summe der quadratischen Abweichungen oder Summe der Abweichungsquadrate, abgekürzt SAQ), die Varianz und die Standardabweichung geeignete Maßzahlen für die Beschreibung der Streuung der Variablenwerte in einer Population. Die Variation ist definiert als:

$$SAQ = \sum (x - \bar{x})^2 \qquad (14.1)$$

Aus Stichprobendaten schätzt man die unbekannte Varianz σ^2 und die unbekannte Standardabweichung σ der Grundgesamtheit nach den Formeln:

$$s^2 = \frac{\sum(x-\bar{x})^2}{df} \text{ und } s = \sqrt{\frac{\sum(x-\bar{x})^2}{df}} \tag{14.2}$$

Dabei ist df (degrees of freedom = Freiheitsgrade) gleich $n-1$.

In der Varianzanalyse zerlegt man die Gesamtvariation der Kriteriumsvariablen (= SAQ_{Total}), im Beispiel ist das die Variation der Einkommen aller Personen, in einen durch den Faktor (hier: Schulbildung) erklärten Teil und in einen nicht erklärten Teil. In einem varianzanalytischen Test wird dann ein Quotient aus zwei auf Basis der Zerlegung der Variation vorgenommenen unterschiedlichen Schätzungen der Gesamtvarianz gebildet und mit einem F-Test geprüft, ob der Faktor einen statistisch signifikanten Einfluss auf die Kriteriumsvariable (hier: Einkommenshöhe) hatte oder nicht.

Die Gesamtvariation SAQ_{Total} berechnet sich für die Daten der Tabelle 14.1 als $(1.800 - 2.500)^2 + (1.900 - 2.500)^2 + \cdots + (3.200 - 2.500)^2 = 2.800.000$. Diese Gesamtvariation der Einkommen SAQ_{Total} stammt aus zwei Quellen und wird entsprechend zerlegt. Einmal ist sie durch den Faktor Schulbildung verursacht: diese Variation ist die zwischen den Gruppen (= $SAQ_{zwischen}$), denn die Abiturienten bekommen ja mehr als der Durchschnitt, die Hauptschüler weniger. Dazu kommt aber eine weitere Streuung. In jeder der Schulbildungsgruppen besteht eine Einkommensstreuung, für die jedoch keine Ursache angegeben wurde: diese Variation ist die innerhalb der Gruppen (= $SAQ_{innerhalb}$). Dementsprechend wird die Gesamtvariation SAQ_{Total} in diese zwei Teilvariationen zerlegt:

$$SAQ_{Total} = SAQ_{zwischen} + SAQ_{innerhalb} \tag{14.3}$$

Tabelle 14.2. Ausgangsdaten für die Varianzzerlegung

Hauptschule $\bar{x}_H = 2000$			**Mittlere Reife** $\bar{x}_M = 2500$			**Abitur** $\bar{x}_A = 3000$		
x	$(x-\bar{x})$	$(x-\bar{x})^2$	x	$(x-\bar{x})$	$(x-\bar{x})^2$	x	$(x-x)$	$(x-\bar{x})^2$
1.800	-200	40.000	2.300	-200	40.000	2.800	-200	40.000
1.900	-100	10.000	2.400	-100	10.000	2.900	-100	10.000
2.000	0	0	2.500	0	0	3.000	0	0
2.100	+100	10.000	2.600	+100	10.000	3.100	+100	10.000
2.200	+200	40.000	2.700	+200	40.000	3.200	+200	40.000
Σ		100.000			100.000			100.000

Variation und Varianz innerhalb der Gruppen. Der Tatsache, dass innerhalb der drei Schulbildungsgruppen nicht alle Personen, also z.B. nicht alle Abiturienten, das gleiche Einkommen haben, ist durch irgendwelche nicht näher erfassten Einflüsse bedingt. Diese berechnete Variation, ermittelt als *Variation innerhalb der Gruppen* ($SAQ_{innerhalb}$), wird im weiteren auch als *unerklärte* – unerklärt durch den Faktor Schulbildung – oder *Restvariation* bezeichnet.

In Tabelle 14.2. sind die Daten des Beispiels zur Berechnung von $SAQ_{innerhalb}$ aufbereitet. Getrennt für jede Gruppe wird im ersten Schritt die Summe der Abwei-

chungsquadrate SAQ auf der Basis des Mittelwerts der jeweiligen Gruppe berechnet. Es ergibt sich aus Tabelle 14.2:

$$SAQ_H = \sum (x - \bar{x}_H)^2 = 100.000 \tag{14.4}$$

$$SAQ_M = \sum (x - \bar{x}_M)^2 = 100.000 \tag{14.5}$$

$$SAQ_A = \sum (x - \bar{x}_A)^2 = 100.000 \tag{14.6}$$

Die Variation innerhalb der Gruppen ergibt sich im zweiten Schritt aus der Summation dieser Abweichungsquadratsummen:

$$SAQ_{innerhalb} = SAQ_H + SAQ_M + SAQ_A = 300.000 \tag{14.7}$$

Zur Berechnung der Varianz innerhalb der Gruppen wird die Variation innerhalb der Gruppen $SAQ_{innerhalb}$ durch die Anzahl der Freiheitsgrade geteilt. Die Anzahl der Freiheitsgrade (df) ergibt sich aus der Anzahl der Fälle n = 15 minus Anzahl der Gruppen k = 3, also n – k = 12:

$$s^2_{innerhalb} = \frac{SAQ_{innerhalb}}{df} = \frac{300.000}{12} = 25.000 \tag{14.8}$$

Variation und Varianz zwischen den Gruppen. Ermitteln wir jetzt die Variation zwischen den Gruppen, die auf den Faktor Schulbildung zurückzuführende Variation. Die Wirkung des Faktors besteht ja darin, dass nicht alle Gruppen den gleichen Mittelwert haben. Hauptschulabsolventen müssen ja einen Abschlag von DM 500 in Kauf nehmen, Abiturienten profitieren von einem Zuschlag von DM 500. Diese Streuung zwischen den k = 3 Gruppen berechnet sich dadurch, dass die quadrierte Abweichung jedes Gruppenmittelwertes $\bar{x}_i$ (hier i = 1 bis 3) vom Gesamtmittelwert $\bar{x}_T$ gebildet wird. Sodann wird jede dieser quadrierten Abweichungen mit der Zahl der Fälle n_i in der Gruppe gewichtet. $SAQ_{zwischen}$ ist die Variation der Einkommen zwischen den Gruppen. Die Varianz zwischen den Gruppen $s^2_{zwischen}$ ergibt sich durch Teilung der Abweichungsquadratsumme durch die Zahl der Freiheitsgrade df. Die Anzahl der Freiheitsgrade beträgt: Anzahl der Gruppen minus 1. In unserem Fall mit drei Gruppen: $k - 1 = 2$.

$$SAQ_{zwischen} = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x}_T)^2 \text{ und } s^2_{zwischen} = \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x}_T)^2}{df} \tag{14.9}$$

Wegen $\bar{x}_T = 2.500$ (Gesamtmittelwert) ergibt sich:

$$SAQ_{zwischen} = 5 \cdot (2.000 - 2.500)^2 + 5 \cdot (2.500 - 2.500)^2 + 5 \cdot (3.000 - 2.500)^2 = 2.500.000$$

$$\text{und } s^2_{zwischen} = \frac{2.500.000}{3-1} = 1.250.000.$$

Als Ergebnis der Varianzzerlegung ergibt sich gemäß Gleichung 14.3:

2.800.000 = 300.000 + 2.500.000

Verwenden wir nun die Varianzzerlegung zur Feststellung des durch einen Faktor *erklärten Anteils der Varianz*. Wir haben in unserem Falle einen einzigen Faktor, die Schulbildung. Der durch ihn erklärte Anteil der Varianz (genau genommen der Variation) drückt sich aus im Verhältnis der Summe der quadrierten Abweichungen zwischen den Gruppen zu der Summe der quadrierten Abweichungen insgesamt:

$$\text{eta}^2 = \frac{SAQ_{zwischen}}{SAQ_{Total}} = \frac{2.500.000}{2.800.000} = 0{,}89 \qquad (14.10)$$

Varianzanalytischer F-Test. Die Varianzzerlegung ist Ausgangspunkt für einen Signifikanztest. Wenn wir Zufallsstichproben vorliegen haben, müssen wir davon ausgehen, dass beobachtete Unterschiede von Mittelwerten zwischen den Gruppen eventuell auch per Zufall zustande gekommen sein könnten. Nach den Regeln der Signifikanztests ist so lange H_0 beizubehalten, als dies nicht als sehr unwahrscheinlich (Fehlerrisiko 5 % oder 1 %) angesehen werden kann.

Wir haben in unserer Untersuchung insofern mehrere Stichproben vorliegen, als jede Schulausbildungsgruppe als eine unabhängige Stichprobe interpretiert werden kann. Auf Basis dieser drei Stichproben kann man – ausgehend von der Varianzzerlegung – auf verschiedene Weise die Varianz der Grundgesamtheit σ^2_{Total} schätzen: mittels der Varianz innerhalb der Gruppen ($s^2_{innerhalb}$) und mittels der Varianz zwischen den Gruppen ($s^2_{zwischen}$). Beide Varianzen können als zwei verschiedene Schätzungen der wahren Varianz σ^2_{Total} in der Gesamtpopulation angesehen werden. Gilt jetzt die Nullhypothese, würden sich also alle Gruppen in ihren Einkommen nur durch Zufallsschwankungen voneinander unterscheiden, müssten beide Schätzungen zum gleichen Ergebnis führen. Dagegen führen sie zu unterschiedlichen Ergebnissen, wenn der Faktor Schulausbildung einen Einfluss auf das Einkommen hat und somit die Gruppen aus unterschiedlichen Grundgesamtheiten stammen. Dabei kann man davon ausgehen, dass die Varianz innerhalb der Gruppen einen ziemlich genauen Schätzwert der Varianz der Grundgesamtheit darstellt. Dagegen gilt das für die Varianz zwischen den Gruppen nur, wenn kein Einfluss des Faktors vorliegt und die Differenzen zwischen den Gruppen auf Zufallsschwankungen beruhen. Sind die beiden so geschätzten Varianzen also näherungsweise gleich, spricht das für die Nullhypothese: es gibt keinen Einfluss des Faktors auf das Einkommen. Ist die Varianz zwischen den Gruppen aber deutlich höher, muss zumindest in einer Gruppe eine deutliche Abweichung vom Zufallsprozess vorliegen. Der Quotient aus der Varianz zwischen den Gruppen und der Varianz in den Gruppen kann demnach als eine Testgröße dafür dienen, ob die Schwankungen zwischen den Gruppen zufälliger Natur sind oder nicht. Diese Größe wird als F bezeichnet:

$$F = \frac{s^2_{zwischen}}{s^2_{innerhalb}} = \frac{1.250.000}{25.000} = 50 \qquad (14.11)$$

Die Testgröße F hat eine F-Verteilung mit $df_1 = k - 1$ und $df_2 = n - k$ Freiheitsgraden. Aus der tabellierten F-Verteilung kann man unter Berücksichtigung der Freiheitsgrade für beide Varianzschätzungen die Wahrscheinlichkeit eines solchen

Wertes bei Geltung von H_0 – der Faktor Schulbildung hat keinen Einfluss – ermitteln. Ein Blick in eine F-Tabelle mit $df_1 = 2$ und $df_2 = 12$ ergibt bei einem Signifikanzniveau von 5 % ($\alpha = 0{,}05$) einen F-Wert = 3,34. Da der empirische F-Wert = 50 diesen kritischen bei weitem übersteigt, kann die Hypothese H_0 abgelehnt werden. Es liegt demnach ein signifikanter Effekt des Faktors vor (zu Hypothesentests ⇨ Kap. 13.3).

14.2 ANOVA in der praktischen Anwendung

Die Ein-Weg-Varianzanalyse soll nun für die gleichen Variablen der Datei ALLBUS90.SAV durchgeführt werden. Um die Ein-Weg-Varianzanalyse aufzurufen, gehen Sie wie folgt vor:

- ▷ Wählen Sie „Analysieren", „Mittelwerte vergleichen ▷", „Einfaktorielle ANOVA...". Es erscheint die in Abb. 14.1 abgebildete Dialogbox „Einfaktorielle ANOVA".
- ▷ Wählen Sie aus der Variablenliste die abhängige Variable, und übertragen Sie diese ins Feld „Abhängige Variablen" (hier: EINK).
- ▷ Übertragen Sie die unabhängige Variable in das Feld „Faktor" (hier: SCHUL2).
- ▷ Bestätigen Sie mit „OK".

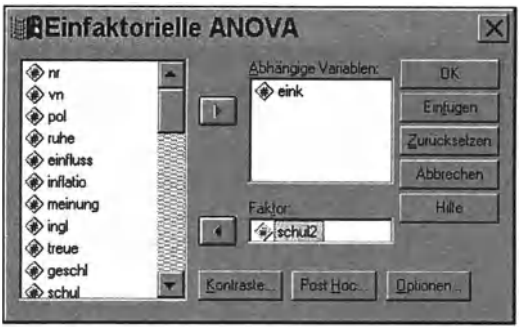


Abb. 14.1. Dialogbox „Einfaktorielle ANOVA"

Optionen. Wenn Sie mehr als die Standardergebnisausgabe erhalten wollen:

- ▷ Klicken Sie auf die Schaltfläche „Optionen...". Die Dialogbox „Einfaktorielle ANOVA: Optionen" erscheint (⇨ Abb. 14.2). Je nach Wunsch klicken Sie in der Auswahlgruppe „Statistik" bzw. „Diagramm der Mittelwerte" auf die Kontrollkästchen und wählen in der Gruppe „Fehlende Werte" die gewünschte Option aus. Bestätigen mit „Weiter" und „OK".

Folgende Auswahlmöglichkeiten bestehen:

- *Deskriptive Statistik.* Deskriptive Statistiken wie Mittelwerte, Standardabweichung, Standardfehler, die Konfidenzintervalle für die Mittelwerte sowie Minimum und Maximum werden für die Vergleichsgruppen ausgegeben.
- *Feste und zufällige Effekte.* Gibt Statistiken für ein Modell mit festen Effekten (Standardabweichung. Standardfehler und Konfidenzintervall) bzw. zufällige Effekte (Standardfehler, Konfidenzintervall, Varianz zwischen den Komponenten) aus.
- *Test auf Homogenität der Varianzen.* Damit wird der Levene-Test (in der klassischen Version) zur Prüfung von Homogenität (Gleichheit) der Varianzen aufgerufen, der bereits bei der Besprechung des t-Tests erläutert wurde (⇨ Kap. 9.3.1). Mit diesem können Sie prüfen, ob ungefähr gleiche Varianz in den Vergleichsgruppen gegeben ist, eine der Voraussetzungen der Varianzanalyse.
- *Brown-Forsythe.* Ein Test auf Gleichheit der Gruppenmittelwerte. Er hat dieselbe Funktion wie der F-Test, der in der Varianzanalyse als Standardtest fungiert. Dieser hat aber als Voraussetzung Gleichheit der Varianzen der Vergleichsgruppen. Der Brown-Forsythe-Test ist für den Fall entwickelt worden, dass diese Voraussetzung nicht zutrifft.
- *Welsh.* Dito.
- *Diagramm der Mittelwerte.* Erstellt ein Liniendiagramm mit den Mittelwerten der Vergleichgruppen als Punkten.
- *Fehlende Werte.* Durch Anklicken einer der Optionsschalter in dieser Auswahlgruppe bestimmen Sie, ob die fehlenden Werte fallweise Test für Test (Voreinstellung) oder listenweise (d.h. für die gesamte Analyse) ausgeschlossen werden sollen.

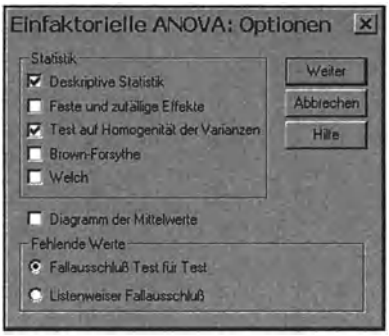


Abb. 14.2. Dialogbox „Einfaktorielle ANOVA: Optionen"

Die in Abb. 14.1 und 14.2 angezeigten Einstellungen führen zur Ergebnisausgabe in Tabelle 14.3 (durch Pivotierung leicht überarbeitet):

Zuerst sehen wir uns in der Mitte des Outputs das Ergebnis des Levene-Tests an. Falls die Voraussetzung homogener Varianzen verletzt sein sollte, könnten sich die weiteren Überlegungen erübrigen. Der Levene-Test ergibt, dass keine signifikanten Abweichungen der Varianzen in den Vergleichsgruppen vorliegen (wegen „Signi-

fikanz“ = 0,197 > Signifikanzniveau $\alpha = 0{,}05$). Demnach darf die Varianzanalyse angewendet werden.

Als nächstes betrachten wir in der Tabelle die eigentliche Varianzanalyse. Es wird die Zerlegung der summierten Abweichungsquadrate („Sum of Squares“) SAQ_{Total} („Gesamt“) gemäß Gleichung 14.3 in die zwischen den Gruppen $SAQ_{zwischen}$ („Zwischen den Gruppen“) und innerhalb der Gruppen $SAQ_{innerhalb}$ angegeben. Ebenso werden die Varianzen („Mittel der Quadrate“) zwischen ($s^2_{zwischen}$) und in den Gruppen ($s^2_{innerhalb}$) und die Freiheitsgrade („df“) ausgegeben. Als F-Wert ergibt sich 5,616. Man könnte diesen Wert nach Gleichung 14.11 auch selbst berechnen. Dieser Wert hat bei Freiheitsgraden $df_1 = 2$ und $df_2 = 139$ bei Geltung von H_0 eine Wahrscheinlichkeit von 0,005 oder ca. einem halben Prozent. Es liegt also ein signifikanter Einfluss der Schulbildung vor.

Tabelle 14.3. Ergebnisse einer einfaktoriellen Varianzanalyse für die Beziehung zwischen Einkommen und Schulbildung

Deskriptive Statistik

Abhängige Variable: EINK BEFR.: MONATLICHES NETTOEINKOMMEN

		Hauptschule	Mittelschule	Fachh/Abi	Gesamt
N		74	33	35	142
Mittelwert		1807,32	2532,58	2277,80	2091,83
Standardabweichung		1053,22	1091,13	1204,87	1136,26
Standardfehler		122,43	189,94	203,66	95,35
95%-Konfidenzintervall für den Mittelwert	Untergrenze	1563,31	2145,68	1863,91	1903,32
	Obergrenze	2051,33	2919,47	2691,69	2280,34
Minimum		129	850	800	129
Maximum		7000	5300	4800	7000

Test der Homogenität der Varianzen

BEFR.: MONATLICHES NETTOEINKOMMEN

Levene-Statistik	df1	df2	Signifikanz
1,643	2	139	,197

ANOVA

BEFR.: MONATLICHES NETTOEINKOMMEN

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	13610762,067	2	6805381,033	5,616	,005
Innerhalb der Gruppen	168433045,877	139	1211748,531		
Gesamt	182043807,944	141			

Allerdings zeigen die deskriptiven Statistiken der Ergebnisausgabe in der ersten Tabelle, dass das Einkommen nicht kontinuierlich mit der Schulbildung steigt (dasselbe zeigt das „Diagramm der Mittelwerte“, falls wir es anfordern). Das Einkommen der Personen mit Mittlerer Reife liegt im Durchschnitt höher als das der Hauptschulabsolventen. Das Einkommen der Abiturienten (einschl. Fachschulabsolventen) liegt aber etwas unter dem der Personen mit Mittlerer Reife. Die Be-

trachtung der 95%-Konfidenzintervalle für den Mittelwert macht deutlich, dass bei einem Sicherheitsniveau von 95 % sich nur die Konfidenzintervalle der Hauptschulabsolventen und der Personen mit Mittlerer Reife nicht überschneiden, also wahrscheinlich nur ein signifikanter Unterschied zwischen diesen beiden Gruppen existiert, die anderen Unterschiede hingegen nicht signifikant sind. Für die Einzelprüfung der Differenzen stehen allerdings die anschließend zu erörternden multiplen Vergleichstests zur Verfügung.

Einen eta²-Wert gibt „Einfaktorielle ANOVA" nicht aus. Dafür müssen wir entweder auf die Option „Mittelwerte" von „Mittelwerte vergleichen" oder auf „Univariat" im Menü „Allgemeines lineares Modell" zurückgreifen. Allerdings kann man eta² nach Gleichung 14.10 leicht selbst berechnen:

$$\text{eta}^2 = \frac{13.610.762{,}067}{182.043.807{,}944} = 0{,}0748$$

Obwohl zumindest eine signifikante Abweichung zwischen zwei Mittelwerten gefunden wurde, sehen wir, dass der Faktor Schulbildung nur 7,5 % der Varianz erklärt. Der Faktor hat also nur geringe Erklärungskraft.

14.3 Multiple Vergleiche (Schaltfläche „Post Hoc")

Mit dem F-Test kann lediglich geprüft werden, ob beim Vergleich der Mittelwerte mehrerer Gruppen die Differenz zwischen mindestens einem der Vergleichspaare signifikant ist. Nichts ergibt sich dagegen darüber, zwischen welchen Vergleichspaaren signifikante Unterschiede bestehen. Deshalb bietet „Einfaktorielle ANOVA" als Option zwei Typen von Tests an, die für alle Kombinationen von Vergleichspaaren die Mittelwertdifferenz auf Signifikanz prüfen.

- *Paarweise Mehrfachvergleiche.* Damit werden die Mittelwertdifferenzen aller möglichen Paare von Gruppen auf statistische Signifikanz überprüft. Die Ergebnisse sämtlicher Vergleiche erscheinen in einer Tabelle. Signifikante Differenzen werden durch ein Sternchen am entsprechenden Wert in der Spalte „Mittlere Differenz" gekennzeichnet.
- *Post-Hoc-Spannweitentests* (Bildung homogener Untergruppen). Untersucht umgekehrt die Vergleichsgruppen auf nicht signifikante Mittelwertdifferenzen. Jeweils zwei Gruppen, die sich nicht unterscheiden, werden als neue homogene Gruppe ausgewiesen. Die entsprechende Spalte enthält die Gruppenmittelwerte der beiden Gruppen und das Signifikanzniveau.

Einige der verfügbaren Test berechnen sowohl „paarweise Mehrfachvergleiche" als auch „homogene Gruppen". Beide Typen von Analysen beruhen auf der Signifikanzprüfung der Mittelwertdifferenz von Vergleichspaaren. Es handelt sich dabei um Abwandlungen des in Kap. 13.3 erläuterten t-Tests oder ähnlicher Tests. Diese modifizierten Tests berücksichtigen die durch den Vergleich mehrerer Gruppen veränderte Wahrscheinlichkeit, einen signifikanten Unterschied zu ermitteln.

Dies sei anhand des t-Tests erläutert. Werden lediglich die Mittelwerte zweier zufällig gezogener Stichproben (Gruppen) verglichen, entspricht die Wahrschein-

lichkeit, bei Geltung der Nullhypothese die empirisch festgestellte Differenz mit dem entsprechenden t-Wert zu erhalten, der in der t-Verteilung angegebenen Wahrscheinlichkeit. Natürlich können dabei auch einmal zufällig stark voneinander abweichende Mittelwerte gefunden werden. Aber die Wahrscheinlichkeit ist entsprechend der t-Verteilung einzustufen. Vergleicht man dagegen mehrere Stichproben (Gruppen) miteinander, werden mit gewisser Wahrscheinlichkeit auch einige stärker vom „wahren Wert" abweichende darunter sein. Sucht man daraus willkürlich die am stärksten voneinander differierenden heraus, besteht daher eine erhöhte Wahrscheinlichkeit, dass man zwei extreme Stichproben vergleicht und daher auch eine erhöhte Wahrscheinlichkeit, dass sich die Differenz nach den üblichen Testbedingungen als signifikant erweist. Die für die multiplen Vergleiche entwickelten Tests berücksichtigen dies dadurch, dass für ein gegebenes Signifikanzniveau von z.B. 5 % ($\alpha = 0{,}05$) beim multiplen Vergleich ein höherer Wert für die Testgröße verlangt wird als beim einfachen t-Test. Dieses kann anhand der Gleichung 13.5 in Kap. 13.3 näher erläutert werden.

Die Gleichung kann auch wie folgt geschrieben werden:

$$\bar{x}_1 - \bar{x}_2 \geq t_\alpha * s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (14.12)$$

$\bar{x}_1 - \bar{x}_2$ = Differenz der Mittelwerte von zwei Gruppen.

t_α = t - Wert, der dem Signifikanzniveau α entspricht.

s_P = Standardabweichung insgesamt, d.h. aller Fälle der beiden Gruppen.
Berechnet als gepoolte Standardabweichung (⇨ Gleichung 13.4 in Kap. 13.4.2.2).

n_1, n_2 = Stichprobengröße der beiden Vergleichsgruppen.

Die Gleichung kann man wie folgt interpretieren: damit eine Differenz $\bar{x}_1 - \bar{x}_2$ signifikant ist bei zweiseitiger Betrachtung und einem Signifikanzniveau von z.B. 5 % ($\alpha = 0{,}05$), muss die Differenz größer sein als die rechte Seite der Gleichung (es wird hier angenommen, dass jeweils die Gruppe mit dem größeren Mittelwert mit Gruppe 1 bezeichnet wird). In der multiplen Vergleichsanalyse wird nun bei gleichem Signifikanzniveau α davon ausgegangen, dass der Faktor t_α größer sein muss als beim t-Test (dieser größere Faktor wird in SPSS Range genannt). Insofern kann man auch sagen, dass zum Erreichen eines Signifikanzniveaus von α tatsächlich ein höheres Signifikanzniveau (d.h. ein kleineres α) erreicht werden muss. Bei der Ermittlung dieses höheren Signifikanzniveaus bzw. höheren t-Wertes (= Range in SPSS) gehen die verschiedenen Verfahrensansätze der multiplen Vergleiche unterschiedlich vor. Dabei spielt bei gegebenem zu erreichenden Signifikanzniveau von z.B. 5 % die Anzahl der Gruppen k eine Rolle. Eine größere Anzahl von Gruppen erhöht den Range-Wert. Bei manchen Verfahren wird der Range-Wert für alle Vergleichsgruppenpaare in gleicher Höhe angewendet, in anderen nicht. Ist letzteres der Fall, hängt die Höhe des Range-Wertes davon ab, wie weit das Vergleichsgruppenpaar in der Rangreihe aller Gruppen auseinander liegt. Je weiter die gepaarten Gruppen auseinander liegen, desto höher der Range-Wert.

Als Beispiel für „Paarweise Mehrfachvergleiche" wird der „Bonferroni-Test" vorgestellt. Der „Duncan-Test" dient zur Demonstration der Bildung „homogener Gruppen".

Um multiple Vergleiche aufzurufen, gehen Sie wie folgt vor:

- ▷ Gehen Sie zunächst so vor wie in Kap. 14.2 beschrieben. Die Eingaben entsprechen denen in Abb. 14.1 und 14.2.
- ▷ Klicken Sie in der Dialogbox „Einfaktorielle ANOVA" (⇨ Abb. 14.1) auf die Schaltfläche „Post Hoc...". Die Dialogbox „Einfaktorielle ANOVA: Post-Hoc-Mehrfachvergleiche" erscheint (⇨ Abb. 14.3). Sie können aus mehreren Testverfahren wählen.

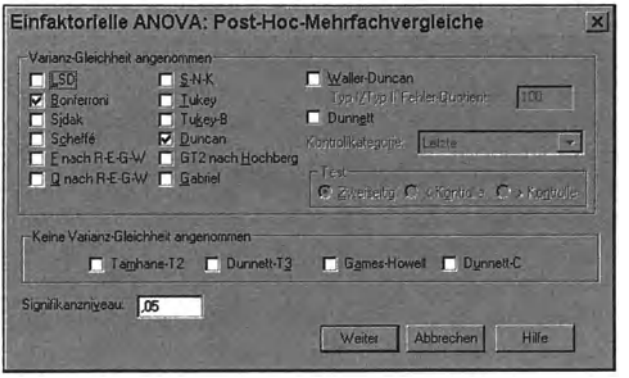


Abb. 14.3. Dialogbox „Einfaktorielle ANOVA: Post-Hoc-Mehrfachvergleiche"

Folgende Tests sind verfügbar:

① Tests für Mehrfachvergleiche, die Varianzgleichheit voraussetzen

- ❒ *LSD* (geringste signifikante Differenz). Entspricht einem t-Test zwischen allen Paaren von Gruppen, d.h. ohne den Range-Wert gegenüber dem t-Wert zu erhöhen. Da die Zahl der Gruppenvergleiche nicht berücksichtigt wird, steigt faktisch die Irrtumswahrscheinlichkeit mit der Zahl der Gruppen. Daher sollte dieser Test nicht oder allenfalls nach signifikantem F-Test verwendet werden.
- ❒ *Bonferroni* (modifizierter LSD). Es handelt sich um einen modifizierten LSD-Test. Die sich aus dem t-Tests ergebende Wahrscheinlichkeit α dafür, dass dies Ergebnis bei Geltung der Nullhypothese per Zufall zustande gekommen ist, wird mit der Zahl der Gruppen multipliziert. Also wird z.B. aus α=0,02 α=0,06. Er bringt bei ungleich großen Vergleichsgruppen ein exaktes Ergebnis.
- ❒ *Sidak.* Ähnlich Bonferroni, aber mit etwas geringerer Korrektur (engere Konfidenzintervalle).
- ❒ *Scheffé.* Er benutzt für alle Vergleichspaare einen einzigen Range-Wert. Er ist strenger als die anderen Tests. Die Werte sind auch für ungleich große Gruppen exakt. Bietet neben paarweisen Vergleichen auch homogene Subsets.

- *Tukey (HSD)* (ehrlich signifikante Differenz). Benutzt für alle Vergleichsgruppenpaare den gleichen Range-Wert, unabhängig davon, wie viele Mittelwerte verglichen werden. Der Range-Wert entspricht dem größten im Student-Newman-Keuls (SNK)-Test. Ergibt bei ungleichen Gruppengrößen nur einen Näherungswert.
- *GT2 Hochberg.* Ähnelt Tukey. Bietet neben paarweisen Vergleichen auch homogene Subsets.
- *Gabriel.* Ähnlich Hochberg. Ist genauer, wenn Zellengröße ungleich. Aber wird bei sehr ungleicher Zellengröße auch ungenau. Bietet neben paarweisen Vergleichen auch homogene Subsets.
- *Dunnett.* Ein besonderer Test. Er behandelt eine Gruppe als Kontrollgruppe und vergleicht alle Gruppen mit dieser Gruppe. Die Kontrollkategorie kann die erste oder die letzte – in der Reihenfolge der Eingabe – sein (Auswahl über: Auswahlliste „Kontrollkategorie"). Es ist der einzige Test, der auch einseitig durchgeführt werden kann. Die Auswahl zwischen zweiseitigem und (nach oben oder unten) einseitigem Test erfolgt über die Optionsschalter des Bereichs „Test".

② Spannweiten-Tests (Bildung homogener Untergruppen)

- *F nach R-E-G-W* (F -Test nach Ryan-Einot-Gabriel-Welsh). Bildet homogene Subsets nach einem mehrfachen Rückschrittverfahren, basierend auf dem F-Test, also nicht auf dem t-Test.
- *Q nach R-E-G-W* (Spannweitentest nach Ryan-Einot-Gabriel-Welsh). Bildet ebenfalls homogene Subsets nach einem mehrfachen Rückschrittverfahren, basierend auf der studentisierten Spannweite.
- *SNK* (Student-Newman-Keuls).Verwendet ein und denselben kritischen Wert über alle Tests. Er gibt nur einen näherungsweisen Wert, wenn gleiche Gruppengrößen gegeben sind.
- *Duncan (Duncans Test für multiple Mittelwertvergleiche).* Dieser Test verfährt ähnlich dem SNK, verwendet aber unterschiedliche Range-Werte für Gruppen in Abhängigkeit davon, wie weit die Gruppen auseinander liegen.
- *Tukey-B.* Verwendet als kritischen Wert den Durchschnitt aus dem von Tukey-HSD und SNK. Liegen ungleiche Gruppengrößen vor, ergibt sich nur ein Näherungswert.
- Homogene Untergruppen liefern außerdem noch *Tukey, GT2 nach Hochberg, Gabriel-Test* und *Scheffé-Test*, die auch Mehrfachvergleiche ausgeben.
- *Waller-Duncan.* Dieser Test nimmt wiederum eine Sonderstellung ein. Homogene Untergruppen werden auf Basis der t-Statistik unter Verwendung einer speziellen Bayesschen Methode gebildet. Als Besonderheit man kann einen „Type I/Type II Fehlerquotienten" einstellen (Voreinstellung = 100). Dadurch wird nicht mit einem fest vorgegebenen Signifikanzniveau α getestet, sondern auch der Fehler zweiter Art, d.h. die Fehlerwahrscheinlichkeit β kontrolliert. Bei gegebener Stichprobengröße ist das nicht absolut, sondern nur über das Verhältnis der beiden Fehlerwahrscheinlichkeiten möglich. Je niedriger der gewählte Wert des „Type I/Type II Fehlerquotienten", desto geringer die Wahrscheinlichkeit, einen Fehler II zu begehen. D.h.: bei einer solchen Vorgabe werden eher keine Zusammenfassungen vorgenommen.

③ Tests für Mehrfachvergleiche, die keine Varianzgleichheit voraussetzen

- ☐ *Tamhane-T2.* Paarweiser Vergleich auf Basis eines t-Tests. Bei Varianzgleichheit ergibt er dasselbe wie *Bonferroni.*
- ☐ *Dunnett-T3.* Paarweiser Vergleich auf Basis des studentisierten Maximalmoduls.
- ☐ *Games-Howell.* Paarweiser Vergleich. Ist geeignet, wenn die Varianzen ungleich sind.
- ☐ *Dunnett-C.* Paarweiser Vergleich auf Basis des studentisierten Bereichs. (Enthält im Vergleich zu den anderen Tests keine Spalte „Signifikanz" mit genauer Angabe der Wahrscheinlichkeit.)

▷ Durch Änderung des Wertes im Eingabefeld *„Signifikanzniveau"* können Sie selbst bestimmen, auf welchem Signifikanzniveau α die Mittelwert verglichen werden sollen. Bestätigen Sie Ihre Eingaben mit „Weiter" und starten Sie den Befehl mit „OK".

Bei den in der Abb. 14.3 angezeigten Einstellungen erscheint der Output von Tabelle 14.4 für die multiplen Vergleichsprozeduren:

Als erstes werden die Ergebnisse des Bonferroni-Tests ausgegeben, dann die des Duncan-Tests. Für den Bonferroni-Test soll dargestellt werden, wie der Wert sich aus Gleichung 14.12 ergibt.

Anstelle von t_α des einfachen t-Tests auf Differenz von zwei Mittelwerten wird – wie oben ausgeführt – ein höherer Wert RANGE eingesetzt. Bonferroni geht davon aus, dass für das angestrebte Signifikanzniveau von α ein höheres Signifikanzniveau von $\alpha'=\alpha/k$ erreicht werden muss. Dabei ist k die Zahl der Gruppen. In unserem Falle wäre bei einem angestrebten Signifikanzniveau von $\alpha = 0{,}05$ ein höheres Signifikanzniveau von $\alpha' = 0{,}5 : 3 = 0{,}017$ zu erreichen. RANGE gibt den entsprechenden Multiplikator für Gleichung 14.12 an, der benötigt wird, dieses höhere Signifikanzniveau zu erreichen.

Aus den Mittelwerten von k Gruppen lassen sich $\frac{k \cdot (k-1)}{2}$ Vergleichspaare bilden. Bei drei Gruppen sind es mithin drei Vergleichspaare.

Die Ergebnisse der Signifikanztests aller Paarvergleiche nach Bonferroni sehen wir in Tabelle 14.4. Die Informationen sind z.T. redundant, da Vergleiche zwischen zwei Gruppen in beiden Richtungen angegeben werden. Relevant ist zunächst die Spalte „Mittlere Differenz (I-J)". Hier können wir z.B. als erstes sehen, dass zwischen der Gruppe der Hauptschüler gegenüber den Mittelschülern eine Differenz im mittleren Einkommen von –725,25 DM besteht. Gleichzeitig signalisiert der *, dass diese Differenz auf dem gewählten Niveau (hier 0,05) signifikant ist. Die genaue Wahrscheinlichkeit für das Auftreten einer solchen Differenz bei Geltung von H_0 sieht man noch einmal in der Spalte „Signifikanz". Sie beträgt nur 0,006. Darüber hinaus werden der Standardfehler und die Ober- und Untergrenzen eines Konfidenzintervalles bei dem gewählten Signifikanzniveau für die Mittelwertdifferenz aufgeführt. Außer zwischen Hauptschülern und Mittelschülern existieren beim Einkommen keine weiteren signifikanten Mittelwertdifferenzen zwischen den Gruppen. Hätte man einfache t-Test für die Mittelwertdifferenzen der

Paare durchgeführt, wäre die jeweilige Wahrscheinlichkeit „Signifikanz" kleiner ausgefallen, nämlich nur ein Drittel so groß. Das ist einleuchtend, weil nach Bonferroni die Wahrscheinlichkeit eines einfachen t-Test mit der Zahl der Vergleichspaare zu multiplizieren ist. Im Vergleich von Hauptschüler und Mittelschülern beträgt der Wert des einfachen t-Test z.B. 0,002, nach Bonferroni 0,006. Sie können das nachprüfen, indem Sie einen LSD-Test durchführen und die Ergebnisse mit denen nach Bonferroni vergleichen. Bei diesem Test wäre dann auch eine weitere Differenz, nämlich die zwischen Hauptschülern und Fachhochschülern/Abiturienten, signifikant.

Die Ergebnisse des Duncan-Test zeigt Tabelle 14.5. Die Tabelle weist zwei homogene Subsets aus, die je zwei Gruppen zusammenfassen. Der erste Subset besteht aus „Hauptschülern" einerseits und „Fachhochschüler/Abiturienten" andererseits. Die Mittelwerte für das Einkommen dieser beiden Gruppen sind in Spalte 1 mit 1807,32 DM und 2277,80 DM angegeben. Zu einer homogenen Gruppe könnten diese beiden Gruppen zusammengefasst werden, weil sich ihre Mittelwerte auf den 5%-Niveau nicht signifikant unterscheiden. (Das kann man der Überschrift „Untergruppe für alpha = 0.05" entnehmen.) Auch der Wert 0,052 in der Zeile „Signifikanz" gibt dieselbe Auskunft. Da hier allerdings die genaue Wahrscheinlichkeit angegeben ist, erkennt man auch, dass die Differenz doch beinahe das Signifikanzniveau erreicht. Bei der zweiten Subset, bestehend aus „Fachhochschülern/Abiturienten" einerseits und „Mittelschülern" andererseits, liegen die Verhältnisse klarer. Die Differenz der Einkommensmittelwerte dieser beiden Gruppen liegt mit einer Wahrscheinlichkeit vom 0,292 weit entfernt von der kritischen Grenze α von 0,05.

Tabelle 14.4. Multiple Mittelwertvergleiche

Abhängige Variable: BEFR.: MONATLICHES NETTOEINKOMMEN

		Bonferroni				
(I) Schulbildung umkodiert	(J) Schulbildung umkodiert	Mittlere Differenz (I-J)	Standard-fehler	Signifikanz	95%-Konfidenzintervall Untergrenze	95%-Konfidenzintervall Obergrenze
Hauptschule	Mittelschule	-725,25*	230,423	,006	-1283,63	-166,87
	Fachh/Abi	-470,48	225,824	,117	-1017,71	76,76
Mittelschule	Hauptschule	725,25*	230,423	,006	166,87	1283,63
	Fachh/Abi	254,78	267,097	1,000	-392,48	902,03
Fachh/Abi	Hauptschule	470,48	225,824	,117	-76,76	1017,71
	Mittelschule	-254,78	267,097	1,000	-902,03	392,48

*. Die mittlere Differenz ist auf der Stufe .05 signifikant.

Am unteren Ende der Tabelle finden sich darüber hinaus in unserem Beispiel zwei Anmerkungen. Der Duncan Test setzt eigentlich gleich große Vergleichsgruppen voraus. Wenn diese Bedingung nicht gegeben ist, machen die Anmerkungen auf diese Tatsache aufmerksam. Bei Berechnung der Signifikanz wird dann als Gruppengröße automatisch das harmonische Mittel aus allen Gruppengrößen verwendet.

Die in der Zeile „Signifikanz" angegebenen Irrtumswahrscheinlichkeiten α sind dann nicht ganz exakt.

Die in der Zeile „Signifikanz" angegebenen Wahrscheinlichkeiten dafür, dass die Mittelwertdifferenz zwischen den beiden Gruppen bei Geltung von H_0 zustande gekommen ist, unterscheiden sich von den entsprechenden Angaben im Bonferroni-Test. Das liegt daran, dass Duncan, anders als Bonferroni, unterschiedliche Range-Werte benutzt, je nachdem, wie weit die verglichenen Gruppen in der nach Größe des Mittelwertes geordneten Reihe auseinander liegen. Nach Duncan ist der erforderliche Range-Wert um so größer, je mehr andere Gruppen mit ihrem Mittelwert zwischen denen der zwei verglichenen Gruppen liegen. Sind sie direkt benachbart, kommt Step 2 mit Range = 2,8 zum Zuge, liegt dazwischen eine andere Gruppe, ist es Step 3 mit Range = 2,95. Hätten wir mehr als drei Gruppen, kämen weitere Schritte hinzu. Step ist dabei ein Wert, der die Größe des Abstandes der verglichenen Gruppen innerhalb der geordneten Reihe der Gruppen repräsentiert. Diese Größe wird berechnet als Step = m + 2. Dabei ist m = Anzahl der in der geordneten Reihe zwischen den beiden verglichenen Gruppen liegenden Gruppen.

Bei nur drei Gruppen liegen die Vergleichsgruppen entweder unmittelbar nebeneinander: dann ist Step = 0 + 2 = 2 oder es liegt eine Gruppe dazwischen: dann ist Step = 1 + 2 = 3. Für den Duncan-Test liegen Tabellen vor, aus denen man in Abhängigkeit vom Signifikanzniveau α, der Distanz (= Step) und der Zahl der Freiheitsgrade $n - k$ den Range-Wert entnehmen kann. Dieser Tafel kann man für α = 0,05 und df = 139 die angegebenen Range-Werte von 2,80 (für Step = 2) bzw. 2,95 (für Step = 3) entnehmen.

Tabelle 14.5. Homogene Sets aus den Schulabschlussgruppen nach dem Duncan Test

BEFR.: MONATLICHES NETTOEINKOMMEN

			Untergruppe für Alpha = .05.	
	Schulbildung umkodiert	N	1	2
Duncan[a]	Hauptschule	74	1807,32	
	Fachh/Abi	35	2277,80	2277,80
	Mittelschule	33		2532,58
	Signifikanz		,052	,292

Die Mittelwerte für die in homogenen Untergruppen befindlichen Gruppen werden angezeigt.

a. Verwendet ein harmonisches Mittel für Stichprobengröße = 41,443.

Hinweis. Aufgrund der Eigenarten der Tests kann es vorkommen, dass beim multiplen Gruppenvergleich für einzelne Vergleichspaare signifikante Unterschiede anzeigt werden, obwohl der F-Test bei der Varianzanalyse keine signifikante Differenz entdeckt. Das kommt zwar selten vor, ist aber nicht ausgeschlossen. Außerdem kann es bei den Tests mit unterschiedlichen Range-Werten für die Vergleichsgruppenpaare in seltenen Fällen zu dem paradoxen Ergebnis kommen, dass eine geringere Mittelwertdifferenz zwischen zwei näher beieinander liegenden Gruppen als signifikant ausgewiesen wird, während die größere Mittelwertdifferenz weiter auseinander liegender Gruppen, zwischen denen die erste-

ren liegen, als nicht signifikant ausgewiesen wird. Wenn dieses auftritt, sollte die Signifikanz der geringeren Differenz ignoriert werden.

Weitere Möglichkeiten bei Verwenden der Befehlssyntax. Mit dem Befehl „Ranges" können zu den einzelnen Tests unterschiedliche Signifikanzniveaus eingegeben werden. Voreingestellt ist immer 0,05. LSD, MODLSD (= Bonferroni) und SCHEFFE können zwischen 0 und 0,5 variieren. (Beachten Sie, dass bei der Eingabe von Dezimalzahlen ein Dezimalpunkt verwendet wird.) DUNCAN hat nur die Wahl zwischen 0,01, 0,05 und 0,10. Alternativ dazu können (in Klammern und durch Komma oder Leerzeichen getrennt) bis zu $n-1$ beliebige Range-Werte eingegeben werden. Werden weniger als $n-1$ Range-Werte angegeben, wird für die verbleibenden höheren Ränge der letzte Wert verwendet. Dadurch können eigene Signifikanztests konstruiert werden.

14.4 Kontraste zwischen a priori definierten Gruppen (Schaltfläche „Kontraste")

Der Befehl „Kontraste..." in der Dialogbox „Einfaktorielle ANOVA" (Abb. 14.1) bietet zwei weitere Features an:

- Es können t-Tests für die Mittelwertdifferenz zweier a priori ausgewählter Gruppen durchgeführt werden. Dabei kann man durch eine Zusammenfassung von bestehenden Gruppen neue definieren.
- In einem Regressionsansatz kann die auf einen Faktor zurückgeführte Abweichungsquadratsumme $SAQ_{zwischen}$ in einen durch Terme eines Polynoms bis zur 5. Ordnung erklärten Anteil und einen Rest zerlegt werden.

T-Test der Mittelwertdifferenz zwischen a priori definierten Gruppen. Wir beschäftigen uns in diesem Abschnitt mit dem t-Test für a priori festgelegt Kontrastgruppen. Das zweite Feature wird in Kap. 14.5 erläutert.

Der Unterschied zu den oben behandelten post hoc Tests der Mittelwertdifferenz aller Gruppenpaarungen besteht darin, dass nur a priori festgelegte Paare auf signifikante Differenzen hin überprüft werden. Dadurch ist das Problem einer erhöhten Wahrscheinlichkeit für signifikante Differenzen nicht gegeben und der in Kap. 13.4 erläuterte t-Test könnte ohne Probleme Verwendung finden. Interessant ist das Feature nur deshalb, weil es für die Tests ohne Umkodieren möglich ist, mehrere Untergruppen zu einer neuen Gruppe zusammenzufassen.

Zur Bewältigung dieser Aufgaben werden Koeffizienten verwendet. Diese haben drei Funktionen:

- Sie bestimmen, welche Gruppen verglichen werden sollen.
- Gegebenenfalls geben Sie an, welche bestehenden Gruppen zu einer neuen zusammengefasst werden sollen.
- Sie sind ein Multiplikator für die Werte der durch sie bestimmten Vergleichsgruppen.

Die Verwendung von Koeffizienten kann am besten mit unserem Beispiel aus dem ALLBUS90.SAV verdeutlicht werden. In diesem sind drei Gruppen mit unter-

schiedlichem Schulabschluss enthalten: Gruppe 2 = Hauptschulabschluss, Gruppe 3 = Mittlere Reife, Gruppe 4 = Abitur. Die drei Gruppen sind in der angegebenen Reihenfolge geordnet. Will man jetzt zwei Gruppen daraus zum Vergleich auswählen, bekommen diese beiden einen Koeffizienten $\neq 0$ zugeordnet. Die Gruppe, die nicht in die Auswahl kommt, dagegen einen Koeffizienten = 0. Die Zahl der Koeffizienten muss der der Gruppen entsprechen. Die Koeffizienten der ausgewählten, verglichenen Gruppen müssen zusammen Null ergeben. Daraus ergibt sich, dass eine der beiden Gruppen einen negativen, die andere einen positiven Koeffizienten zugeordnet bekommt (z.B. −1 und +1). Sollen mehrere Ursprungsgruppen zu einer neuen zusammengefasst werden, bekommen sie den gleichen Koeffizienten (z.B. 0,5 und 0,5). Alle Koeffizienten müssen aber auch dann zu Null summieren. Daraus ergibt sich, dass beim Vergleich stärker zusammengefasster Gruppen mit weniger stark zusammengefassten, die Absolutwerte der Koeffizienten der zusammengefassten Gruppen entsprechend kleiner ausfallen müssen. Es ist günstig, wenn die Koeffizienten aller Teilgruppen einer zusammengefassten Gruppen sich jeweils auf +1 bzw. −1 summieren (z.B. -0,5 und -0,5). Dann sind nämlich alle Ergebnisausgaben unmittelbar interpretierbar. Ist das nicht der Fall, fallen die angegebenen Mittelwertdifferenzen und Standardfehler entsprechend dem gewählten Koeffizienten größer oder kleiner aus. Die t-Statistik dagegen ist korrekt, da der Koeffizient bei der Division der Mittelwertdifferenz durch den Standardfehler wieder weggekürzt wird. Die letztgenannte Empfehlung kann bei der Zusammenfassung einer ungeraden Zahl von Gruppen (etwa bei 3 oder 7) zu einer neuen Gruppe nicht zum Tragen kommen, weil die Koeffizienten als Dezimalzahlen mit einer Stelle hinter dem Komma eingegeben werden müssen und deshalb eine Aufsummierung auf 1 nicht möglich ist.

Am Beispiel für die Variable Schulbildung sollen fünf Vergleichspaare (= Kontraste) bestimmt werden:

① Kontrast zwischen Gruppe 2 (Hauptschule) und Gruppe 4 (Abitur) mit den Koeffizienten −1 und +1: −1 0 +1.
② Kontrast zwischen Gruppe 2 (Hauptschule) und Gruppe 4 (Abitur) mit den Koeffizienten −2 und +2: −2 0 +2.
③ Kontrast zwischen Gruppe 3 (Mittlere Reife) und Gruppe 4 (Abitur): 0 −1 +1.
④ Kontrast zwischen Gruppe 2 (Hauptschule) und Gruppe 3 (Mittlere Reife): −1 +1 0.
⑤ Kontrast zwischen Gruppe 2 (Hauptschule) und einer zusammengefassten Gruppe aus Gruppe 3 (Mittlere Reife) und 4 (Abitur): −1 + 0,5 + 0,5.

Es sind hier alle relevanten Fälle aufgeführt. Der zweite Fall dient dazu, den Unterschied zu demonstrieren, der auftritt, wenn die Koeffizienten einer Gruppe nicht +1 oder −1 betragen.

Zur Durchführung des a priori t-Tests gehen Sie wie folgt vor:

▷ Gehen Sie zunächst so vor wie in Kap. 14.2 beschrieben. Die Eingaben entsprechen denen in Abb. 14.1 und 14.2.
▷ Klicken Sie nun in der Dialogbox „Einfaktorielle ANOVA“ (⇨ Abb. 14.1) auf die Schaltfläche „Kontraste...“. Die Dialogbox „Einfaktorielle ANOVA: Kontraste“ (Abb. 14.4) erscheint.

- ▷ Geben Sie in das Eingabefeld „Koeffizienten:“ den ersten Koeffizienten (hier:–1) für den ersten gewünschten Vergleich bzw. Kontrast ein (hier: Fall ①).
- ▷ Klicken Sie auf die Schaltfläche „Hinzufügen“.
- ▷ Wiederholen Sie die beiden letzten Schritte so lange, bis alle Koeffizienten für den ersten Kontrast eingegeben sind (hier: zwei weitere Schritte mit der Eingabe von 0 und +1).
- ▷ Sollen weitere Kontraste definiert werden, klicken Sie auf die Schaltfläche „Weiter“ bei „Kontrast 1 von 1“. Die Beschriftung ändert sich in „Kontrast 2 von“ und die Eingabefelder stehen wieder bereit.
- ▷ Geben Sie dann, wie oben beschrieben, die Koeffizienten für den zweiten Kontrast ein.

Abb. 14.4. Dialogbox „Einfaktorielle ANOVA: Kontraste“

Der ganze Prozess kann für bis zu 10 Kontraste wiederholt werden. Die Anzeige im Informationsfeld „Koeffizientensumme:“ ermöglicht es Ihnen, gleich zu überprüfen, ob die definierten Kontraste auf Null summieren. Für Änderungen können Sie durch Anklicken von „Zurück“ auf früher definierte Kontraste zurückschalten. Die einzelnen Koeffizienten können durch Markieren und Anklicken von „Entfernen“ widerrufen werden, Änderungen können durch Markieren des zu ändernden Koeffizienten, das Neueintragen eines Wertes in das Feld „Koeffizienten:“ und Anklicken von „Ändern“ vorgenommen werden.

- ▷ Haben Sie die Definition der Kontraste beendet, bestätigen Sie mit „Weiter“.
- ▷ Starten Sie mit „OK“.

Für die geschilderten fünf Kontraste führt das zur Tabelle 14.6.

Zunächst ist in der Matrix der Kontrast-Koeffizienten noch einmal die Definition der Koeffizienten übersichtlich dargestellt.

Es folgen dann die Ergebnisse der eigentlichen Kontrastgruppenanalyse und zwar für beide Varianten des t-Tests, die mit gepoolter Schätzung der Varianz (Varianzen sind gleich) (⇨ Gleichung 13.4) und die mit separater Schätzung der Varianz (Varianzen sind nicht gleich). Wie wir schon oben gesehen haben, unterscheiden sich die Varianzen der einzelnen Stichproben nicht signifikant voneinander. Daher können wir hier den t-Test für gepoolte Varianzschätzung benutzen.

Betrachten wir die entsprechende Tabelle. In der Spalte „Kontraste“ ist die Mittelwertdifferenz $\bar{x}_1 - \bar{x}_2$ für die Vergleichsgruppen angegeben. Danach der Standardfehler. Diese beiden Angaben stimmen nur, wenn die Koeffizienten der Kontrastgruppen jeweils auf 1 bzw. −1 summieren. Das kann man bei dem Vergleich von Kontrast 1 und 2 sehen. In beiden Fällen werden dieselben Gruppen verglichen. Im ersten Fall betragen aber die Koeffizienten der Kontrastgruppen 1 bzw. −1, im zweiten 2 bzw. −2. Deshalb fallen Mittelwertdifferenz und Standardfehler im zweiten Falle doppelt so hoch aus. Ebenso kann man aber erkennen, dass beim t-Wert und den Freiheitsgraden kein Unterschied auftritt und das Ergebnis dasselbe ist. Mit Ausnahme des Kontrastes 3, bei dem die Gruppen 3 (Mittelschüler) und 4 (Abiturienten einschl. Fachschulabsolventen) verglichen werden, sind alle formulierten Kontraste gemäß „Signifikanz (2-seitig)“, der Wahrscheinlichkeit dafür, dass der t-Wert bei Geltung von H_0 aufgetreten ist, auf dem 5 %-Niveau signifikant (der genaue Wert ist 0,039 oder 3,9 % Irrtumswahrscheinlichkeit für die Kontraste 1 und 2 und 0,002 oder 0,2 % Irrtumswahrscheinlichkeit für die Kontraste 4 und 5). Die beiden letzten Kontraste wären auch auf dem 1 %-Niveau signifikant.

Tabelle 14. 6. T-Tests für durch apriori Kontraste gebildete Gruppen

Kontrast-Koeffizienten

	Schulbildung umkodiert		
Kontrast	Hauptschule	Mittelschule	Fachh/Abi
1	-1	0	1
2	-2	0	2
3	0	-1	1
4	-1	1	0
5	-1	,5	,5

Kontrast-Tests

		BEFR.: MONATLICHES NETTOEINKOMMEN				
	Kontrast	Kontrastwert	Standardfehler	T	df	Signifikanz (2-seitig)
Varianzen sind gleich	1	470,48	225,82	2,083	139	,039
	2	940,95	451,65	2,083	139	,039
	3	-254,78	267,10	-,954	139	,342
	4	725,25	230,42	3,147	139	,002
	5	597,86	184,96	3,232	139	,002
Varianzen sind nicht gleich	1	470,48	237,63	1,980	59,402	,052
	2	940,95	475,26	1,980	59,402	,052
	3	-254,78	278,49	-,915	65,898	,364
	4	725,25	225,98	3,209	59,605	,002
	5	597,86	185,42	3,224	134,571	,002

Wie man sieht, kann man eine ganze Reihe von Kontrasten bilden. Bei bis zu fünf Gruppen könnte man auf diese Weise genauso wie beim post hoc Vergleich alle Gruppenpaare vergleichen. Dann wäre aber die Voraussetzung für die Verwendung

des t-Test aufgehoben. Diese ist nur gegeben, wenn einzelne, zufällige Vergleiche vorgenommen werden. Werden mehrere Kontraste anstelle eines F-Tests überprüft, so soll die Erhöhung der Wahrscheinlichkeit für signifikante Ergebnisse dadurch vermieden werden, dass der Set der definierten Kontraste orthogonal ist. Das heißt die Kontraste sollen nicht redundant und statistisch voneinander unabhängig sein. Das wäre der Fall, wenn die Produkte der korrespondierenden Koeffizienten aller Paare von Kontrasten zu Null summieren:

Beispiel für vier Gruppen:

Kontrast 1:	1	−1	0	0
Kontrast 2:	0	0	1	−1
Kontrast 3:	0,5	0,5	−0,5	−0,5

Die Summe der Produkte zwischen Kontrast 1 und 2 ist: $1 * 0 + -1 * 0 + 0 * 1 + 0 * -1 = 0$. Dasselbe gilt für die beiden anderen Kombinationen.

14.5 Erklärung der Varianz durch Polynome

„Einfaktorielle ANOVA" bietet die Möglichkeit, die Abweichungsquadratsumme zwischen den Gruppen $SAQ_{zwischen}$ in einem regressionsanalytischen Verfahren durch lineare, quadratische usw. Terme eines Polynoms zu erklären bzw. vorherzusagen. Dabei wird $SAQ_{zwischen}$ in von der linearen, quadratischen usw. Komponente dieses Polynoms erklärten und einen nicht erklärten Anteil zerlegt. Die maximal mögliche Ordnung des Polynoms beträgt fünf:

$$y = a + b_1x + b_2x^2 + b_3x^3 + b_3x^4 + b_5x^5 \qquad (14.14)$$

Allerdings ist die höchste sinnvolle Ordnung eines Polynoms gleich der Zahl der Gruppen k minus 1: (=k−1). Unabhängig von der Benutzereingabe berechnet SPSS das Polynom nur bis zur Ordnung k−1.

Die Vorgehensweise entspricht einer Erweiterung der Linearitätstests von „Mittelwertvergleiche" (⇨ Kap. 13.2.3). Ein Anwendungsbeispiel und eine detailliertere Darstellung können Sie auf den zum Buch gehörigen Internetseiten finden (⇨ Anhang B).

15 Mehr-Weg-Varianzanalyse

Die Mehr-Weg-Varianzanalyse unterscheidet sich von der Ein-Weg-Varianzanalyse dadurch, dass nicht ein, sondern zwei und mehr Faktoren zur Erklärung der Kriteriumsvariablen verwendet werden. Dadurch ist zweierlei möglich:

- Der Beitrag jeder dieser Faktorvariablen zur Erklärung der Gesamtvariation kann für sich alleine genommen untersucht werden. Es kann aber auch die Wirkung ihrer spezifischen Kombinationen miteinander (Interaktion) mit geprüft werden. Den Beitrag der Hauptvariablen (ohne Berücksichtigung ihrer Interaktion) nennt man Haupteffekte (Main Effects). Effekte, die auf spezifische Kombinationen der Faktoren zurückzuführen sind, bezeichnet man als Interaktionseffekte (Interactions). Es gibt neben den Haupteffekten gegebenenfalls Interaktionen auf mehreren Ebenen. Die Zahl der Ebenen errechnet sich durch $m-1$. Dabei ist m die Zahl der einbezogenen Faktoren. So gibt es bei einer Zwei-Weg-Varianzanalyse mit den Faktoren A und B, neben den Haupteffekten A und B, nur eine Interaktionsebene (2-Weg-Interaktion) mit der Interaktion AB, bei einer Drei-Weg-Analyse mit den Faktoren A, B und C dagegen, neben den Haupteffekten A, B und C, die 2-Weg-Interaktionen AB, AC und BC sowie die 3-Weg Interaktion ABC. Wie man sieht, steigt die Zahl möglicher Interaktionen mit der Zahl der Faktoren überproportional stark an.
- Jeder dieser Beiträge kann mit Hilfe des F-Tests auf Signifikanz geprüft werden. Es gilt aber: Ist eine Interaktion signifikant, sind alle F-Test der Haupteffekte hinfällig, weil das Berechnungsmodell für die Haupteffekte dann nicht mehr zutrifft. Es muss also zuerst, nach der Prüfung des Gesamtmodells, immer die Signifikanz der Interaktionen geprüft werden. So wie man auf ein signifikantes Ergebnis trifft, sind alle weiteren Signifikanztests obsolet.

Man unterscheidet faktorielle Designs mit gleichen und ungleichen Zellhäufigkeiten. Dieser Unterschied hat Konsequenzen für die Berechnung der Effekte. Ist der Design orthogonal, d.h. sind alle Zellen mit der gleichen Zahl der Fälle besetzt, dann sind die Effekte alle wechselseitig voneinander unabhängig. Dann kann die klassische Berechnung der verschiedenen Statistiken der Varianzanalyse uneingeschränkt benutzt werden. Bis zu einem gewissen Grade gilt das auch, wenn die Zellenbesetzung proportional der Randverteilung ist. Dann sind zumindest die Haupteffekte voneinander unabhängig. Sind dagegen die Zellen ungleich besetzt, wird davon die Berechnung der verschiedenen Komponenten und die Interpretation der Resultate berührt. Die Effekte korrelieren miteinander, sind nicht statistisch unabhängig. Dadurch addieren z.B. die „Komponenten Abweichungsquadratsummen" (d.h. die Haupt- und Interaktionseffekte), wenn sie separat berechnet werden,

nicht auf die „Totale Abweichungsquadratsumme". Um das zu verhindern, wird nur ein Teil der Abweichungsquadratsummen separat berechnet. Andere Teile werden dagegen durch Differenzbildung zu den vorher berechneten gebildet. Man muss entsprechend gegebenenfalls eine Hierarchie der verschiedenen Effekte festlegen, um die Art der Berechnung der einzelnen Effekte zu bestimmen. Je nachdem, wie dies genau geschieht, können erheblich unterschiedliche Ergebnisse ermittelt werden. SPSS hält dafür drei verschiedene Verfahren bereit (⇨ Kap. 15.2).

Außerdem können sich Designs noch in mannigfaltigen anderen Eigenschaften unterscheiden. Wichtig ist z.B., ob sie nur „feste Faktoren" enthalten oder auch Zufallsfaktoren. Bei festen Faktoren sind alle relevanten Merkmale des Faktors durch die Untersuchungsanordnung vorgegeben. „Zufallsfaktoren" kommen dagegen durch hinsichtlich dieses Merkmals zufällige Zuweisung von Fällen zu den Untersuchungsgruppen zustande (z.B. Gruppenbildung nach dem Randomverfahren). Wir besprechen nur Modelle mit festen Faktoren. Weiter kann es wichtig sein, ob die Datenmatrix leere Zellen enthält oder nicht, ob die Werte der Faktoren selbst eine Zufallsauswahl darstellen etc. All dieses kann durch entsprechende Modellbildung mit der Syntax berücksichtigt werden, kann aber im Rahmen dieses Buches nicht behandelt werden. Schließlich ist das Menü nicht für Designs mit wiederholten Messungen vorgesehen. Dafür enthält das Modul „Advanced Statistik" ein eigenes Programm. Auch im Menü „Reliability" (Kap. 23.2.2) steht eine entsprechende Varianzanalyse zur Verfügung.

15.1 Faktorielle Designs mit gleicher Zellhäufigkeit

Beispiel. Zur Erläuterung eines Designs mit gleicher Zahl der Fälle in den Zellen sei das konstruierte Beispiel aus der Einweg-Varianzanalyse (⇨ Kap. 14.1) erweitert. Es war so konstruiert, dass die Kriteriumsvariable „Einkommen" (EINK) vom Faktor „Schulbildung" (SCHUL) beeinflusst war, und zwar führte höhere Schulbildung zu einem Aufschlag gegenüber dem Durchschnittseinkommen der Mittelschüler und geringere zu einem Abschlag. Dabei waren in jeder Gruppe (in der Varianzanalyse spricht man von *Faktorstufen*) fünf Fälle. Es sei jetzt die Zahl der Fälle verdoppelt, und es werde als weiterer Faktor „Geschlecht" (GESCHL) eingeführt. Je die Hälfte der Fälle jeder Schulbildungsgruppe sei männlichen und weiblichen Geschlechts. Daher sind in jeder Schulbildungsgruppe jetzt fünf Männer und fünf Frauen bzw. jede Kombination von Schulbildung und Geschlecht trifft für fünf Fälle zu. Das Beispiel wird so verändert, dass weibliches Geschlecht gegenüber dem Durchschnittswert einer Schulbildungskategorie zu einem Abschlag von 300 DM Einkommen führt, das männliche dagegen zu einem Zuschlag von 300 DM. Das gilt aber nicht für die Abiturienten. In dieser Schulbildungsgruppe haben Männer und Frauen dasselbe Einkommen. Durch die letzte Festlegung wird ein Interaktionseffekt (Wechselwirkung) produziert. Die Wirkung der Schulbildung ist jetzt nämlich nicht mehr unabhängig davon, welche Kategorie des Geschlechts vorliegt (bzw. des Geschlechts, welche Schulbildung), sondern es kommt auf die spezifische Kombination an. Die Daten des Beispiels (VARIANZ2.SAV) sind in

Tabelle 15.1 enthalten. Außerdem sind die wichtigsten für die Varianzanalyse benötigten Statistiken bereits berechnet: die Mittelwerte, Summierte Abweichungsquadrate (SAQ), Varianzen und Fallzahlen.

Tabelle 15.1. Einkommen nach Schulabschluss und Geschlecht (fiktive Daten)

Variable B: Schulabschluss	**Variable A: Geschlecht**		
	männlich	weiblich	gesamt
Hauptschule-	2.100 2.200 2.300 2.400 2.500	1.500 1.600 1.700 1.800 1.900	
	$\bar{x}_{mH}$=2.300 SAQ_{mH}=100.000 n_{mH}=5	$\bar{x}_{wH}$=1.700 SAQ_{wH}=100.000 n_{wH}=5	$\bar{x}_{H}$=2.000 n_{H}=10
Mittlere Reife	2.600 2.700 2.800 2.900 3.000	2.000 2.100 2.200 2.300 2.400	
	$\bar{x}_{mM}$=2.800 SAQ_{mM} = 100.000 n_{mM}=5	$\bar{x}_{wM}$=2.200 SAQ_{wM} = 100.000 n_{wM}=5	$\bar{x}_{M}$=2.500 n_{M}=10
Abitur	2.800 2.900 3.000 3.100 3.200	2.800 2.900 3.000 3.100 3.200	
	$\bar{x}_{mA}$=3.000 SAQ_{mA}=100.000 n_{mA}=5	$\bar{x}_{wA}$=3.000 SAQ_{wA}=100.000 n_{wA}=5	$\bar{x}_{A}$=3.000 n_{A}=10
Insgesamt	$\bar{x}_{m}$=2.700 n_{m}=15	$\bar{x}_{w}$=2.300 n_{w}=15	$\bar{x}_{T}$=2.500 SAQ_{T}=7.400.000 s^2_{T}=255.172,41 n_{T}=30

Die Berechnungen der Varianzanalyse erfolgen – mit Ausnahme der Interaktionen – genau wie bei der Ein-Weg-Analyse. Allerdings werden die Bezeichnungen etwas verändert. Die Summe der Abweichungsquadrate bzw. Varianzen innerhalb

der Gruppen werden als „Quadratsumme Fehler" und „Mittel der Quadrate Fehler" (SAQ_{Fehler} und s^2_{Fehler}) bezeichnet. Die entsprechenden Werte zwischen den Gruppen werden als SAQ_A und s^2_A, SAQ_B und s^2_B usw. bezeichnet, wobei A, B etc. für den Namen der Variablen steht.

Die Abweichungsquadratsummen insgesamt für alle Daten SAQ_T und die daraus errechnete Varianz s^2_T sind in der untersten Zeile der Tabelle enthalten.

Zur Berechnung der entsprechenden Angaben für jede der beiden Variablen führt man praktisch zwei Einweg-Varianz-Analysen durch. Man betrachtet die entsprechend vereinfachten Tabellen, deren Werte jeweils als Randverteilung der angegebenen Tabelle vorliegen. Die entsprechenden Ergebnisse sehen Sie in Tabelle 15. 2. Bei der Analyse können wir den „Konstanten Term" und „Gesamt", das den konstanten Term umfasst, vernachlässigen. (SPSS enthält seit der Version 8.0 für diese Art der Analyse das Menü „Univariat" als Untermenü von „Allgemeines lineares Modell". Es ist auch für Kovarianz- und Regressionsanalysen vorgesehen. Darauf kann hier nicht eingegangen werden. Teile des Outputs, die sich auf diese Analysetypen beziehen, bzw. entsprechende Optionen werden nicht besprochen.)[1]

Tabelle 15.2. Ausgabe einer Zwei-Weg-Varianzanalyse (gesättigtes Modell)

Tests der Zwischensubjekteffekte

Abhängige Variable: monatl. Nettoeinkommen

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Signifikanz
Korrigiertes Modell	6800000,000[a]	5	1360000,0	54,400	,000
Konstanter Term	187500000,000	1	1,87E+08	7500,000	,000
GESCHL	1200000,000	1	1200000,0	48,000	,000
SCHUL	5000000,000	2	2500000,0	100,000	,000
GESCHL * SCHUL	600000,000	2	300000,000	12,000	,000
Fehler	600000,000	24	25000,000		
Gesamt	194900000,000	30			
Korrigierte Gesamtvariation	7400000,000	29			

a. R-Quadrat = ,919 (korrigiertes R-Quadrat = ,902)

Für die Variable A (Geschlecht) können gemäß Gleichung 14.9 $SAQ_{zwischen}$ bzw. $s^2_{zwischen}$ aus den Angaben am unteren Rand der Tabelle errechnet werden:

$SAQ_A = 15 \cdot (2.700 - 2.500)^2 + 15 \cdot (2.300 - 2.500)^2 = 1.200.000$, df = 2-1 = 1 und $s^2_A = 1200.000 : 1 = 1.200.000$.

Die entsprechenden Werte für die Variable B (Schulabschluss) werden analog aus den Angaben in der rechten Randspalte berechnet:

[1] Dadurch wird das Menü „Einfach mehrfaktorielle ANOVA" ersetzt. Wer mit einer älteren Version arbeitet, kann dessen Beschreibung von den zum Buch gehörenden Internetseite downloaden (⇨ Anhang B). Bei neueren Versionen ist es per Syntax ebenfalls noch zugänglich.

$SAQ_B = 10 \cdot (2.000 - 2.500)^2 + 10 \cdot (2.500 - 2.500)^2 + 10 \cdot (3.000 - 2.500)^2 = 5.000.000$, $df = 3 - 1 = 2$ und $s^2_B = 5.000.000 : 2 = 2.500.000$.

Die Abweichungsquadratsumme der Haupteffekte A und B zusammen (die in der Ausgabe nicht angegeben ist) beträgt $SAQ_{Haupteffekte} = 1.200.000 + 5.000.000 = 6.200.000$, $df = 1 + 2 = 3$ und $s^2_{Haupteffekte} = 6.200.000 : 3 = 2.066.666{,}67$.

Die $Abweichungsquadratsumme_{Residuen}$ (Fehler) errechnet sich aus den Abweichungsquadratsummen der Zellen wie folgt:

$$SAQ_{Fehler} = 100.000 + 100.000 + 100.000 + 100.000 + 100.000 + 100.000 = 600.000$$

Das Besondere liegt jetzt in der Berechnung der entsprechenden Werte für die Interaktionen.

Wechselwirkung (Interaktion). Bevor wir auf die Berechnung eingehen, soll die Bedeutung von Wechselwirkungen anhand einer grafischen Darstellung verdeutlicht werden. Abb. 15.1 und 15.2 sind jeweils Darstellungen des Zusammenhanges zwischen der Kriteriumsvariablen „Einkommen" und den beiden Faktoren „Schulabschluss" und „Geschlecht". Dabei bilden die drei Schulabschlüsse „Hauptschulabschluss", „Mittlere Reife" und „Abitur" jeweils eine Zeile in der Tabelle 15.1 und sind in der Grafik auf der x-Achse abgetragen. Die Ausprägungen der Variablen Geschlecht, „weiblich" und „männlich", entsprechen den Spalten der Tabelle. In der Grafik ist das durch zwei unterschiedliche Einkommenskurven für Männer und Frauen repräsentiert. Das Ergebnis der jeweiligen Wertekombination von Schulabschluss und Geschlecht im Einkommen ergibt in einer Tabelle einen Zellenwert, in der Grafik einen Punkt auf einer dieser Kurven. Die durchschnittliche Einkommensgröße entspricht dem Abstand zwischen x-Achse und diesem Punkt. Die entsprechende Skala ist auf der y-Achse abgetragen.

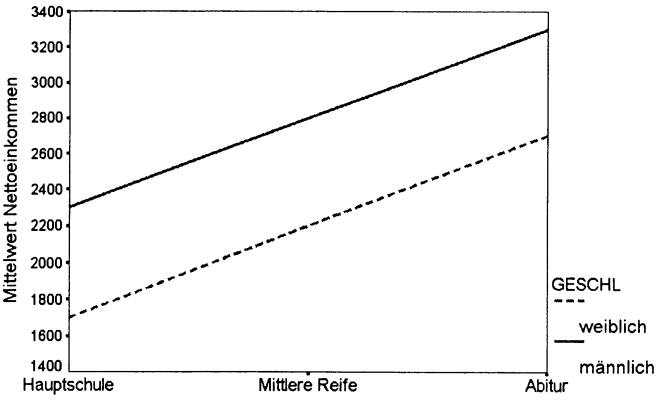


Abb. 15.1. Darstellung einer additiven linearen Wirkung von Schulabschluss und Geschlecht auf das Einkommen (Profilplot)

In Abb. 15.1 ist eine rein additive Wirkung der beiden Variablen „Schulabschluss" und „Geschlecht" dargestellt. Zudem sind die Beziehungen auch noch linear. Dass die Zeilenvariable „Schulbildung" einen Einfluss besitzt, zeigt sich darin, dass die Kurve nicht als Gerade parallel zur x-Achse verläuft. Dies wäre der Fall, wenn die Zeilenvariable keinen Einfluss hätte. Besitzt sie einen Einfluss, steigt oder fällt die Kurve. Sie kann auch in verschiedenen Abschnitten unterschiedlich verlaufen, aber nicht als Parallele zur x-Achse. Hat die Spaltenvariable (hier: Geschlecht) dagegen keinen Einfluss, müssen die Kurven, die für die verschiedenen Kategorien dieser Variablen stehen, zusammenfallen. Dies ist aber im Beispiel nicht der Fall. Die Kurve der Männer verläuft oberhalb derjenigen der Frauen. Das zeigt, dass die Variable Geschlecht einen Einfluss hat. Verlaufen die verschiedenen Kurven parallel – wie im Beispiel –, dann besteht ein additiver Zusammenhang. Linear sind die Beziehungen, da die Kurven als Geraden verlaufen. Das ist aber keine Bedingung für additive Beziehungen.

Abbildung 15.2 ist dagegen die Darstellung des oben beschriebenen Beispiels. Dort besteht – wie beschrieben – insofern eine Interaktion, als bei den „Hauptschulabsolventen" und den Personen mit „Mittlerer Reife" das Geschlecht einen Einfluss auf das Einkommen hat, bei den „Abiturienten" aber nicht. Das schlägt sich darin nieder, dass die beiden Kurven für Männer und Frauen am Anfang parallel verlaufen, am Ende aber nicht. Immer, wenn eine Interaktion vorliegt, verlaufen die Kurven zumindest in Teilbereichen nicht parallel. Sie können sich voneinander entfernen, sich nähern oder überschneiden.

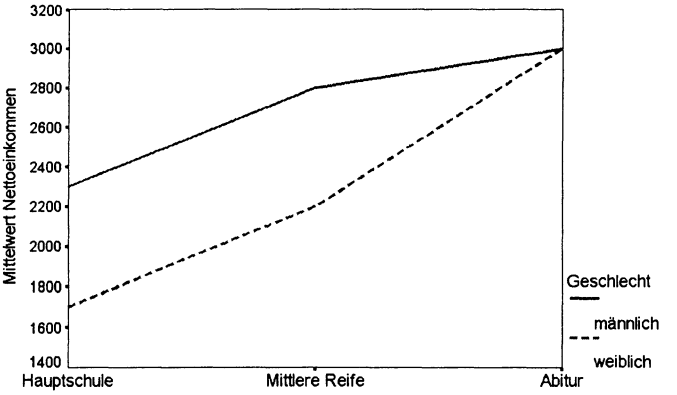


Abb. 15.2. Darstellung einer interaktiven Wirkung von Schulabschluss und Geschlecht auf das Einkommen (Profilplot)

Wir haben also drei Kennzeichen: Differenzen zwischen den auf der Abszisse abgetragenen Kategorien zeigen sich im „nicht-horizontalen" Verlauf der Kurve. Das zweite Kriterium ist „Abstand zwischen den Linien". Abstand ist ein Zeichen für die Differenz zwischen den Kategorien, die die Linien konstituieren. Das dritte Kriterium ist „Konstanz des Abstands" zwischen den Linien. Bleibt dieser kon-

stant, besteht keine Interaktion, verändert er sich, ist das ein Zeichen von Interaktion.

Kommen wir jetzt zur Berechnung von Interaktionseffekten. In unserem Beispiel kommt nur die Interaktion AB in Frage. Diese Berechnung geht von relativ komplizierten Überlegungen aus, die hier nur angedeutet werden können. Sie basiert zunächst auf einem Vergleich der tatsächlich beobachteten Abweichung der arithmetischen Mittelwerte der Zellen $\bar{x}_z$ (der Index z steht hier für Zelle, d.h. für alle Wertekombinationen der Variablen A und B) vom Gesamtmittelwert $\bar{x}_T$ mit der Abweichung, die erwartet würde, wenn keine Interaktion existierte. Dann müsste diese nämlich gleich der Summe der Abweichungen der dazugehörigen Reihen- und Spaltenmittelwerte vom Gesamtmittelwert sein: $(\bar{x}_r - \bar{x}_T) + (\bar{x}_s - \bar{x}_T)$.

Die Abweichung beider Werte voneinander ist dann:

$$d_{r*s} = (\bar{x}_z - \bar{x}_T) - [(\bar{x}_r - \bar{x}_T) + (\bar{x}_s - \bar{x}_T)] = \bar{x}_z - \bar{x}_r - \bar{x}_s + \bar{x}_T \qquad (15.1)$$

Um zur Varianz zu kommen, werden diese Abweichungsmaße quadriert, mit der Zahl der Fälle in den Zellen n_z gewichtet und summiert. Es ergibt sich:

$$\sum d^2_{r*s} = \Sigma\, n_z (\bar{x}_z - \bar{x}_r - \bar{x}_s + \bar{x}_T)^2. \qquad (15.2)$$

Das erste Glied in dieser Summe wird demnach berechnet:
$5 \cdot (2.300 - 2.000 - 2.700 + 2.500)^2 = 50.000$. Und insgesamt ergibt sich:

$$\sum d^2_{r*s} = SAQ_{AB} = 50.000 + 50.000 + 50.000 + 50.000 + 200.000 + 200.000 = 600.000.$$

Dies ist der Wert, den Sie in Tabelle 15.2. für die Interaktion GESCHL*SCHUL als $Quadratsumme_{Geschlcht*Schul}$ finden. Teilt man den Betrag durch die zugehörige Zahl der Freiheitsgrade (= 2), so erhält man die Varianz $s^2_{Geschlecht*Schul} = 300.000$.

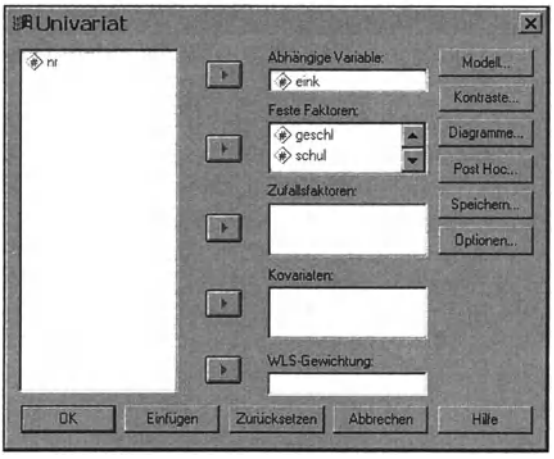


Abb. 15.3. Dialogbox „Univariat"

Das Menü bietet auch die Gelegenheit, in einer Dialogbox „Diagramme“ ein oder mehrere „Profildiagramm(e)“ (Profilplotts) anzufordern. Dies sind Liniendiagramme, welche den Zusammenhang zwischen höchstens zwei Faktoren und der abhängigen Variablen darstellen. Ein Faktor bildet in diesem Diagramm die x-Achse. Welcher das ist, bestimmt man durch Übertragen des Namens in das Feld „Horizontale Achse“ (am besten der Faktor mit den meisten Faktorstufen). Für den zweiten Faktor werden die Ausprägungen (Faktorstufen) als separate Linien dargestellt. Man überträgt seinen Namen in das Feld „Separate Linien:“ Für alle Stufen eines dritten Faktors können diese Zusammenhänge in gesonderten Diagrammen dargestellt werden. Das erreicht man, indem man den Namen dieses Faktors in das Feld „Separate Diagramme“ überträgt. Die Definition wird abgeschlossen mit „Hinzufügen“. Es können mehrere Diagramme nach einander definiert werden. Mit „Weiter“ schließen Sie die Gesamtdefinition ab.

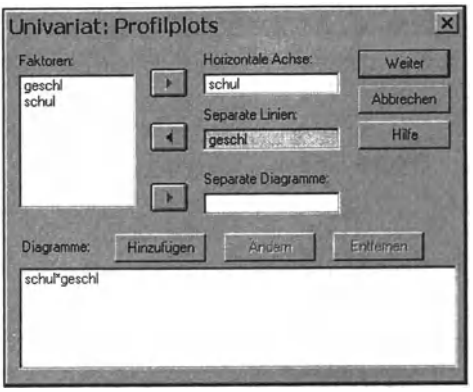


Abb. 15.4. Dialogbox „Univariat: Profilplotts“

Um den in Tabelle 15.2 angegebenen Output und das in Abb. 15.2 dargestellt Diagramm zu erhalten, gehen Sie wie folgt vor:

- ▷ Wählen Sie „Analysieren“ „Allgemeines lineares Modell“, „Univariat...“. Die Dialogbox „Univariat“ erscheint (⇨ Abb. 15.3).
- ▷ Wählen Sie die abhängige Variable (hier: EINK) aus der Variablenliste, und übertragen Sie diese in das Eingabefeld „Abhängige Variable:“.
- ▷ Wählen Sie die beiden Faktoren (hier: GESCHL und SCHUL) aus der Variablenliste, und übertragen Sie diese in das Eingabefeld „Feste Faktoren:“.
- ▷ Klicken Sie auf „Diagramme“. Die Dialogbox „Univariat: Profilplots“ erscheint.
- ▷ Übertragen Sie SCHUL in das Feld „Horizontale Achse:“, GESCHL in das Feld „Separate Linien:“, und klicken Sie auf „Hinzufügen“. Die Definition erscheint im Feld „Diagramme“.
- ▷ Starten Sie den Befehl mit „Weiter“ und „OK“.

15.2 Faktorielle Designs mit ungleicher Zellhäufigkeit

Dieselbe Analyse soll jetzt für die Daten der Datei ALLBUS90.SAV wiederholt werden. Hier sind aber die einzelnen Zellen, gemäß den Verhältnissen in der Realität, nicht gleich besetzt. Schulbildung der verschiedenen Kategorien ist unterschiedlich weit verbreitet. Aber auch Proportionalität zur Randverteilung ist nicht gegeben, denn Geschlecht und Schulbildung korrelieren miteinander. Es liegt demnach ein nicht-orthogonales Design vor. Dies führt zu unterschiedlichen Ergebnissen, je nach Wahl des Analyseverfahrens. Außerdem soll die Variable „Alter" (ALT) als Kovariate eingeführt werden.

Kovarianzanalyse. Die Einführung einer Kovariate heißt, dass zusätzlich zu den kategorialen Faktoren eine metrisch gemessene unabhängige Variable in die Analyse eingeführt wird. Dabei muss vorausgesetzt werden, dass zwischen Kovariate und Faktoren keine Korrelation besteht. Außerdem sollte eine lineare Beziehung zwischen Kovariate und der abhängigen Variablen in allen Gruppen bestehen[2].

Modellbildung. Da wir ein Design mit ungleichen Zellhäufigkeiten vorliegen haben, wäre evtl. an eine Veränderung des Modell zu denken. Das Modell kann in der Dialogbox „Univariat: Modell" auf zweierlei Art beeinflusst werden (⇨ Abb. 15.5).

- *Auswahl von Faktoren und Kovariaten*, die in das Modell eingehen. Zunächst ist durch Anwahl des Optionsschalters, ob ein gesättigtes oder ein angepasstes Modell verwendet werden soll.
  - *Gesättigtes Modell.* Alle in der Dialogbox „Univariat" ausgewählten Faktoren und Kovariate gehen in das Modell ein, aber nur Wechselwirkungen zwischen Faktoren. Diese aber vollständig.
  - *Anpassen.* Es kann ausgewählt werden, welche Faktoren bzw. Kovariate als Haupteffekte und welche ihrer Wechselwirkungen in das Modell aufgenommen werden sollen. Es können also weniger Terme aufgenommen werden, aber auch zusätzlich Wechselwirkungen zwischen Kovariaten bzw. Kovariaten und Faktoren. Um Haupteffekte auszuwählen, markiert man in der Liste „Faktoren und Kovariaten:" die gewünschte Variable, markiert in der Auswahlliste „Term(e) konstruieren" die Option „Haupteffekte" und überträgt die Variable in die Auswahlliste „Modell". Um Wechselwirkungen einer bestimmten Ebene auszuwählen, müssen *alle* in diese Wechselwirkung(en) eingehenden Variablen markiert werden. Dann wählt man in der Auswahlliste „Term(e) konstruieren" die Wechselwirkung der gewünschten Ordnung aus und überträgt sie in das Feld „Modell". Dass bei der Auswahl z.B. von Wechselwirkungen der 2ten Ordnung „**Alle** 2-Weg" aus der Liste zu wählen ist, ist etwas irreführend formuliert. Das bedeutet nur, dass man gleichzeitig zwischen mehr als zwei Variablen alle Zweiweg Interaktionen definieren kann. Das muss aber nicht sein. Man kann auch nur einzelne auswählen. (Bei

[2] In diesem Übungsbeispiel sind (wie wohl bei den meisten nicht experimentell gewonnen Daten) die Bedingungen nicht erfüllt. Inwieweit die Analyse dennoch durchgeführt werden kann, ist z.T. dem Fingerspitzengefühl des Forschers überlassen.

Verwendung vieler Faktoren schließt man gewöhnlich Interaktionen höherer Ordnung aus.) Es ist auch zu beachten, dass bei Verwendung eines hierarchischen Typs der Berechnung auch die Reihenfolge der Eingabe der Faktoren, Kovariaten und Interaktionen von Bedeutung ist.

❒ *Berechnung der Quadratsummen.* Die Berechnung der Summe der Abweichungen ist auf verschiedene Weise möglich. Das Programm bietet vier Berechnungsarten an. Sie unterscheiden sich in erster Linie dadurch, wie die Berechnung der Quadratsummen verschiedener Terme hinsichtlich der Wirkung anderer Terme angepasst (korrigiert) wird. Relevant sind vor allem Typ III und Typ I.

- *Typ I* (Hierarchisch). Jeder Term wird nur für die in der Liste vor ihm stehenden korrigiert. Dadurch wirkt sich die Reihenfolge der Auswahl der Terme auf das Ergebnis aus. Man kann z.B. steuern, ob die Berechnung der Faktorquadratsummen um die Wirkung der Kovariaten korrigiert werden soll oder nicht.
- *Typ III* (Voreinstellung). Hier wird die Berechnung der Quadratsumme eines Effektes um alle anderen Effekte bereinigt, die nicht im Effekt enthalten sind. Dieses Modell hat den Vorteil, dass es weitgehend gegenüber ungleichen Zellhäufigkeiten invariant ist. Deshalb sollte es für solche Designs in der Regel verwendet werden. Nicht geeignet ist dieser Typ allerdings, wenn leere Zellen auftreten.
- *Typ II* und *Typ IV*. Typ II ist ein Regressionsmodell. Es berechnet Haupteffekte um alle anderen Terme (außer Interaktionen) korrigiert. Typ IV ist speziell für Designs mit leeren Zellen entwickelt.

In unserem Beispiel werden wir zunächst zur Demonstration ein Modell anpassen (allerdings so, dass es dem gesättigten Modell entspricht). Wir rechnen mit dem voreingestellten Typ III die Quadratsummen.

Zur Durchführung der Analyse gehen Sie wie folgt vor:

▷ Wählen Sie zunächst die Befehlsfolge „Analysieren", „Allgemeines lineares Modell ▷", „Univariat..". Es öffnet sich die bekannte Dialogbox (⇨ Abb. 15.3).

▷ Geben Sie dann – wie oben beschrieben – die festen Faktoren (hier GESCHL und SCHUL2) ein.

▷ Wählen Sie die als Kovariate benutzte Variable aus der Variablenliste (hier: ALT), und übertragen Sie diese in das Eingabefeld „Kovariaten:".

▷ Klicken Sie auf die Schaltfläche „Modell...". Es öffnet sich die in Abb. 15.5 dargestellte Dialogbox.

▷ Klicken Sie auf den Optionsschalter „Anpassen". Übertragen Sie GESCHL und SCHUL2 und ALTER als Haupteffekte, indem Sie die drei Namen im Feld „Faktoren und Kovariaten:" markieren, in der Liste „Terme konstruieren:" die Option "Haupteffekte auswählen" und auf den Übertragungspfeil klicken. Markieren Sie dann nur die beiden Faktoren, wählen Sie in der Liste „Terme konstruieren:" die Option "Alle 2-Weg", und übertragen Sie diese Interaktion das in das Feld „Modell". Das Ergebnis sehen Sie in Abb. 15.5.

▷ Bestätigen Sie mit „Weiter".

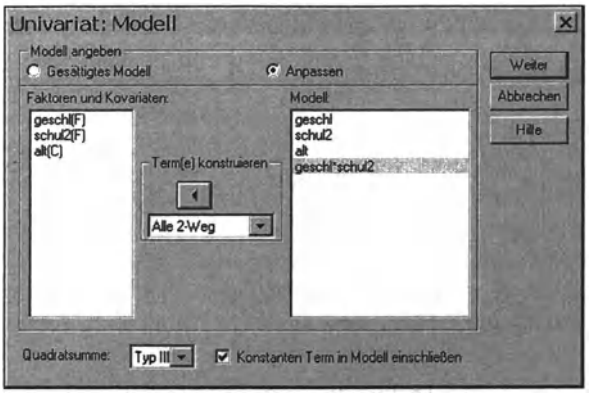


Abb. 15.5. Dialogbox „Univariat: Modell"

Außerdem wollen wir über die Dialogbox „Univariat: Optionen" zwei weitere Ausgaben anfordern.

▷ Klicken Sie auf „Optionen...". Die in Abb. 15.6. dargestellte Dialogbox erscheint.
▷ Wählen Sie „Schätzer der Effektgröße" und „Beobachtete Schärfe".
▷ Bestätigen Sie mit „Weiter", und schicken Sie den Befehl mit „OK" ab.

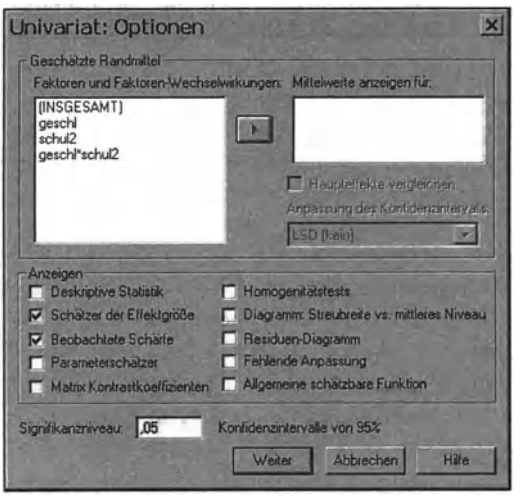


Abb. 15.6. Dialogbox „GLM - Allgemein mehrfaktoriell: Optionen"

Das Ergebnis sehen Sie in Tabelle 15.3. Die eigentliche Ausgabe der Varianzanalyse befindet sich darin in den ersten Spalten bis einschließlich der Spalte „Signifikanz". Die drei letzten Spalten sind Ausgaben der zusätzlich gewählten Optionen.

Die Ergebnisse zeigen zunächst in der Zeile „GESCHL*SCHUL2", dass keine signifikanten Interaktionen vorliegen (Sig. von $F > \alpha = 0,05$). Daher ist die Signifikanzprüfung der Haupteffekte sinnvoll. Von diesen hat Geschlecht eine signifikante Wirkung (Zeile: „GESCHL", Signifikanz 0,000 < 0,05). Die Wirkung der Schulbildung ist dagegen nicht signifikant (Zeile „SCHUL2", Signifikanz 0,10 > 0,05). Keine signifikante Wirkung hat die Kovariate Alter (Zeile: „ALT", Signifikanz 0,646 > 0,05).

Beobachtete Schärfe. Das Menü „Univariat" ist im Basismodul von SPSS das einzige, das dem Problem Rechnung trägt, dass bei statistischen Signifikanztests nicht nur Fehler erster Art, sondern auch Fehler zweiter Art auftreten können und von Interesse sind (⇨ Kapitel 13.3). Das Signifikanzniveau α bestimmt das Risiko, einen Fehler erster Art zu machen, also fälschlich die Nullhypothese abzulehnen. Dagegen hängt das Risiko β, einen Fehler zweiter Art zu begehen, nämlich fälschlich die Nullhypothese beizubehalten von α, der Größe des tatsächlichen Effekts und der Stichprobengröße n ab. Nun ist der Wissenschaftler nicht nur daran interessiert, einen Fehler erster Art zu vermeiden, sondern auch einen Fehler zweiter Art, nämlich tatsächlich vorhandene Effekte auch zu entdecken. Die Wahrscheinlichkeit, einen tatsächlich vorhandenen Effekt auch zu entdecken, nennt man „Schärfe"(Power) eines Tests. Sie beträgt $1 - \beta$. Wegen bestimmter statistischer Probleme (es müssen zwei Punkthypothesen gegeneinander getestet werden), benutzt man die Stärke in der Regel nur für die Kalkulation der in einer Untersuchung notwendigen Stichprobengröße (⇨ SPSS bietet dafür das Programml „Sample Power" im Programm). U.U. kann es aber auch nützlich sein, die „beobachtete Schärfe" zu beachten. Dann nimmt man einmal an, der beobachtete Effekt sei der tatsächliche und fragt sich: Mit welcher Wahrscheinlichkeit würde eine Stichprobe der gegebenen Größenordnung einen solchen Effekt auch entdecken, also nicht die Nullhypothese beibehalten. Das ist vor allem bei relativ kleinen Stichproben interessant. Da kann es nämlich vorkommen, dass der Test nicht die „Schärfe" besitzt, Effekte von einer inhaltlich relevanten Größenordnung zu entdecken. Stellt man dann fest, dass die Untersuchung einen Effekt von relevanter Größenordnung ausweist, dieser Effekt aber statistisch nicht signifikant ist, gleichzeitig der Test aber auch nur geringe Schärfe besitzt, ist es ungerechtfertigt, den Effekt einfach als unbedeutend aus dem Modell auszuschließen. Man sollte vielmehr durch Erhöhung der Fallzahl die Stärke des Tests erhöhen. In unserem Beispiel ist das evtl. für die Variable SCHUL2 zu überlegen. Sie weist keinen signifikanten Effekt auf (α=0,10). Gleichzeitig würde der Test einen Effekt der beobachteten Größe auch nur mit 78,8%iger Wahrscheinlichkeit („Beobachtete Schärfe" = 0,788) entdecken. Wenn dem Forscher diese „Schärfe" nicht ausreicht, muss er die Stichprobengröße erhöhen.

Messung der Effektgröße. Um die Erklärungskraft des Gesamtmodells und der einzelnen Faktoren, Kovariaten und Interaktionen abschätzen zu können, kann man auf die Eta-Statistik zurückgreifen. Sie wurde durch die Option „Schätzer der Effektgröße" angefordert und ist in der drittletzten Spalte von Tabelle 15.3 enthalten. Es handelt sich dabei um partielle Eta-Werte, d.h. der Zusammenhang wird um die

Wirkung der anderen Variablen bereinigt gemessen. Das Programm berechnet in diesem Falle die Werte aus der F-Statistik nach der Formel:

$$\text{Partial Eta}^2 = \frac{df_{Quelle} \cdot F_{Quelle}}{df_{Quelle} \cdot F_{Quelle} + df_{Fehler}} \tag{15.3}$$

wobei:
df_{Quelle} = Freiheitsgrade der untersuchten Einflussquelle
F_{Quelle} = F-Statistik der untersuchten Einflussquelle
df_{Fehler} = Freiheitsgrade der Variation innerhalb der Zellen

Für die Einflussquelle Geschlecht gilt etwa:

$$\text{Partial Eta}^2_{Geschl} = \frac{1 \cdot 24{,}59}{1 \cdot 24{,}59 + 135} = 0{,}154$$

Aus dem Vergleich der Partiellen Eta²-Werten für die verschiedenen Effekte ergibt sich, dass der Faktor „Geschlecht" eine stärkere Wirkung hat als der Faktor „Schulbildung". Er erklärt ca. 15% der Varianz, Schulbildung dagegen 6%. Die Wirkung von Alter und der Interaktion ist verschwindend gering. Das Gesamtmodell erklärt ca. 23% der Gesamtvarianz. Dieselbe Aussage gewinnen wir aus dem multiplen „R-Quadrat" am Fuß der Tabelle. Da dies die Erklärungskraft etwas überschätzt, wird auch noch ein korrigiertes R-Quadrat ausgegeben. Danach würde das Modell etwa 20% der Variation erklären.

Tabelle 15.3. Ergebnisse einer Mehrweg-Varianzanalyse für die Beziehung zwischen Einkommen, Schulabschluss und Geschlecht

Tests der Zwischensubjekteffekte

Abhängige Variable: BEFR.: MONATLICHES NETTOEINKOMMEN

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Signifikanz	Eta-Quadrat	Nichtzentralitäts-Parameter	Beobachtete Schärfe[a]
Korrigiertes Modell	42424611,068[b]	6	7070768,5	6,837	,000	,233	41,021	,999
Konstanter Term	63559012,059	1	63559012	61,456	,000	,313	61,456	1,000
GESCHL	25432596,662	1	25432597	24,591	,000	,154	24,591	,998
SCHUL2	9909843,044	2	4954921,5	4,791	,010	,066	9,582	,788
ALT	218591,240	1	218591,240	,211	,646	,002	,211	,074
GESCHL * SCHUL2	91478,764	2	45739,382	,044	,957	,001	,088	,057
Fehler	139619196,875	135	1034216,3					
Gesamt	803401284,000	142						
Korrigierte Gesamtvariation	182043807,944	141						

a. Unter Verwendung von Alpha = ,05 berechnet

b. R-Quadrat = ,233 (korrigiertes R-Quadrat = ,199)

Unterschiede bei der Verwendung verschiedener Typen der Berechnung der Variationen. Zur Erläuterung der Unterschiede der im Feld „Quadratsumme:" wählbaren Berechnungstypen sind die Ergebnisse der Berechnung mit Typen I bis III für dieselbe Analyse – ohne Kovariate – in Tabelle 15.4 nebeneinander gestellt (die für die Erläuterung irrelevanten Zeilen sind gelöscht). Wendet man die verschiedenen Berechnungsarten auf ein Design mit gleicher Zellhäufigkeit an (wie

VARIANZ2.SAV), unterscheiden sich die Ergebnisse der verschiedenen Berechnungstypen nicht, in unserem aktuellen Beispiel aber wohl.

Wie man sieht, unterscheiden sich die Ergebnisse allerdings bei der durch das Modell erklärten Variation („Korrigiertes Modell") und der entsprechenden F-Statistik nicht, ebensowenig beim unerklärten Rest („Fehler"). Dasselbe gilt auch für die 2-Weg-Wechselwirkung und die „Gesamtvariation". Diese werden bei allen Verfahren gleich berechnet, nämlich nicht hierarchisch, sondern um alle Effekte korrigiert. Unterschiede zeigen sich aber bei den Haupteffekten, also den Faktoren GESCHLECHT (Variable A) und SCHULBILDUNG (Variable B).

Tabelle 15.4. Ergebnisse verschiedener Berechnungstypen der Mehr-Weg-Varianzanalyse für die Beziehung zwischen Einkommen, Schulabschluss und Geschlecht

	Typ I		Typ II		Typ II	
	Quadrat-summe	F	Quadrat-summe	F	Quadrat-summe	F
Korrigiertes Modell	42206019	8,210	42206019	8,210	42206019	8,210
GESCHL	30919670	30,071	28512605	27,730	25215720	24,524
SCHUL2	11203697	5,448	11203697	5,448	10559924	5,153
GESCHL*SCHUL2	82652	,040	82652	,040	82652	,040
FEHLER	139837788		139837788		139837788	8,210
Korrigierte Gesamt-variation	182043807		182043807		182043807	

Die Ergebnisse von Typ I, II unterscheiden sich beim Faktor GESCHL. Das liegt daran, dass beim Typ I der Faktor GESCHL unkorrigiert berechnet wird, da ihm in der Liste kein Term vorausgeht. Bei Typ II wird eine Korrektur vorgenommen, allerdings nur hinsichtlich der Hauptfaktoren, bei Typ III hinsichtlich aller Terme. Typ I und II ergeben dagegen für SCHUL2 dasselbe Ergebnis. Die Berechnung ist in beiden Fällen um den zweiten Hauptfaktor korrigiert, bei Typ I, weil er in der Liste vorangeht. Typ III unterscheidet sich, weil auch noch um die Interaktion korrigiert wurde.

15.3 Mehrfachvergleiche zwischen Gruppen

Die Mehrweg-Varianzanalyse ermöglicht zunächst generelle Signifikanztests für die einzelnen Effekte. Ein signifikanter Wert besagt allerdings lediglich, dass wenigstens eine der Kategorien des Faktors vom Gesamtmittelwert signifikant abweicht. Um die genaueren Einflussbeziehungen zu klären, sind dagegen genauere Betrachtungen des Beziehungsgeflechtes nötig. Dazu bietet „Univariat" mehrere Hilfsmittel. Diese sind zweierlei Art:

- Ausgabe von Mittelwerten oder Mittelwertdifferenzen zwischen verschiedenen Gruppen.

- *Deskriptive Statistik.* Das ist möglich in der Dialogbox „Univariat: Optionen" über die Option „Deskriptive Statistik". Diese führt zu einer Tabelle mit den Mittelwerten, Standardabweichungen und Fallzahlen für jede Faktorstufenkombination.
- *Mittelwerte anzeigen für:.* Ist ein Auswahlfeld der Dialogbox „Univariat: Optionen", in dem man ebenfalls bestimmen kann, für welche Faktoren und Faktorkombinationen man Mittelwerte ausgegeben wünscht. Man überträgt sie dazu aus der Liste „Faktoren und Faktorwechselwirkungen". Anders als bei „Deskriptive Statistik" kann man auch die Mittelwerte für die Gruppen der einzelnen Faktoren sowie den Gesamtmittelwert anfordern. Zusätzlich werden hier „Standardfehler" sowie Ober- und Untergrenzen von „Konfidenzintervallen" (Voreinstellung: 95%-Sicherheit) für die Mittelwerte berechnet. Post hoc Tests können auch im Dialogfeld „Univariat: Optionen" angefordert werde (siehe Haupteffekte vergleichen).
- *Kontraste.* Werden in der Dialogbox „Univariat: Kontraste" Vergleichsgruppen definiert, erscheinen dieselben Angaben in etwas anderer Form im Output.

❒ *Signifikanztests für paarweise Mittelwertvergleiche.*

Sogenannte *„Post hoc-Tests"* können an zwei Stellen aufgerufen werden.

- *Haupteffekt vergleichen.* Drei Verfahren zum Post Hoc Gruppenvergleich (LSD, Bonferroni, Sidak) sind in der Dialogbox „Univariat: Optionen" verfügbar. Man muss dazu die Faktoren, für die der Signifikanztest durchgeführt werden soll, in das Fenster „Mittelwerte anzeigen für:" übertragen. Danach ist das Auswahlkästchen „Haupteffekte vergleichen" anzuklicken. Aus der Liste „Anpassung des Konfidenzintervalls" wählen Sie aus den drei verfügbaren Verfahren das gewünschte aus und bestätigen mit „Weiter".
- *Post hoc.* Hauptsächlich werden paarweise Mittelwertvergleiche aber in der Dialogbox „Univariat: Post-Hoc Mehrfachvergleiche für ..." aufgerufen, die sich beim Anklicken der Schaltfläche „Post-Hoc" im Dialogfenster „Univariat" öffnet. Werden Post Hoc-Tests durchgeführt, erscheinen neben dem eigentlichen Signifikanztest die Differenz der Mittelwerte zwischen den Vergleichsgruppen, deren Standardfehler und die Ober- und Untergrenze eines 95%-Konfidenzintervalls in der Ausgabe.

Multiple Vergleiche Post Hoc. Die Post Hoc Tests von „Univariat: Post-Hoc-Mehrfachvergleiche" sind vollkommen identisch mit den in Kapitel 14 ausführlich besprochenen Test des Menüs „Einfaktorielle ANOVA". Sie werden daher hier nicht besprochen (⇨ Kap 14.3). Der einzige Unterschied besteht darin, dass man mehrere Faktoren gleichzeitig auswählen kann. Es finden immer aber auch hier nur Vergleiche zwischen den Gruppen *eines* Faktors statt, also einfaktorielle Analysen. (Wie oben dargestellt, können die drei ersten Verfahren auch unter „Optionen" aufgerufen werden).

Kontraste zwischen a priori definierten Gruppen (Schaltfläche „Kontraste"). Auch Vergleiche von Gruppen über a priori definierte Kontraste entsprechen im Prinzip dem in Kap. 14 für die einfaktorielle ANOVA geschilderten Verfahren. Jedoch ist „Univariat" bei der Definition von Kontrasten über die Menüs nicht so

flexibel (mit der Syntax dagegen sind alle Möglichkeiten offen), sondern bietet einige häufig benutzte Kontraste zur Auswahl an. Diese sind:

- *Abweichung.* Vergleicht die Mittelwerte aller Faktorstufen (außer der Referenzkategorie) mit dem Gesamtmittelwert. Der Gesamtmittelwert ist allerdings das ungewogene arithmetische Mittel aller Faktorstufen (was bei ungleicher Besetzung der Zellen nicht dem wirklichen Gesamtmittelwert der Stichprobe entspricht).
- *Einfach.* Vergleicht die Mittelwerte aller Faktorstufen (außer der Referenzkategorie) mit dem Mittelwert der Referenzkategorie. Wenn der Design eine Kontrollgruppe enthält, ist diese als Referenzkategorie zu empfehlen.
- *Differenz.* Vergleicht den Mittelwert jeder Faktorstufe mit dem ungewogenen (!) arithmetischen Mittel der Mittelwerte *aller* vorherigen Faktorstufen. (Die erste Faktorstufe hat keine vorherige, daher werden f-1 Vergleiche durchgeführt, wobei f = Zahl der Faktorstufen ist.)
- *Helmert.* Umgekehrt. Vergleicht den Mittelwert jeder Faktorstufe mit dem ungewogenen (!) arithmetischen Mittel der Mittelwerte aller folgenden Faktorstufen.
- *Wiederholt.* Vergleicht den Mittelwert jeder Faktorstufe (außer der letzten) mit dem Mittelwert der folgenden Faktorstufe.
- *Polynomial.* Vergleicht den linearen, quadratischen etc. Effekt. Diese Kontraste werden verwendet, um polymoniale Trends zu schätzen.

Hinweis. Alle Vergleiche beziehen sich immer nur auf die Stufen eines Faktors, sind also einfaktoriell. U.U. ist die Reihenfolge der Stufen wichtig, weil bei einigen Verfahren mehrere Stufen zusammengefasst werden. Bei den Verfahren „Abweichung" und „Einfach" wird außerdem mit *Referenzkategorien* gearbeitet. Es kann entweder die „Erste" oder die „Letzte" (Voreinstellung) Faktorstufe als Referenzkategorie gewählt werden. Auch dafür ist die Anordnung der Faktorstufen wichtig.

Abb. 15.7. Dialogbox „Univariat: Kontraste" mit geöffneter Auswahlliste

Das Vorgehen sei für den Faktor SCHUL2 mit dem Verfahren „Einfach" und der Referenzkategorie „Letzte" demonstriert. Um diesen Kontrast zu definieren, gehen Sie wie folgt vor:

▷ Vollziehen zunächst alle bereits beschriebenen Schritte zur Anforderung der Varianzanalyse.

- ▷ Klicken Sie auf „Kontraste". Die Dialogbox „Univariat: Kontraste" öffnet sich (⇨ Abb. 15.7).
- ▷ Markieren Sie im Feld „Faktoren" den Faktor SCHUL2.
- ▷ Klicken Sie in der Gruppe „Kontrast ändern" auf den Pfeil neben dem Feld „Kontrast:". Wählen Sie aus der sich öffnenden Liste „Einfach".
- ▷ Markieren Sie den Optionsschalter „Letzte".
- ▷ Klicken Sie auf „Ändern". Die Bezeichnung in der Klammer hinter dem Faktornamen ändert sich in „Einfach".
- ▷ Bestätigen Sie mit „Weiter" und „OK".

Tabelle 15.5. Multipler Gruppenvergleich mittels apriori Kontrast

Kontrastergebnisse (K-Matrix)

Schulbildung umkodiert Einfacher Kontrast[a]			Abhängige Variable: BEFR.: MONATLICHES NETTOEINKOMMEN
Stufe 1 gegen Stufe 3	Kontrastschätzer		-469,176
	Hypothesenwert		0
	Differenz (Schätzung - Hypothesen)		-469,176
	Standardfehler		208,763
	Signifikanz		,026
	95% Konfidenzintervall für die Differenz	Untergrenze	-882,018
		Obergrenze	-56,335
Stufe 2 gegen Stufe 3	Kontrastschätzer		156,015
	Hypothesenwert		0
	Differenz (Schätzung - Hypothesen)		156,015
	Standardfehler		251,510
	Signifikanz		,536
	95% Konfidenzintervall für die Differenz	Untergrenze	-341,361
		Obergrenze	653,390

a. Referenzkategorie = 3

Testergebnisse

Abhängige Variable: BEFR.: MONATLICHES NETTOEINKOMMEN

Quelle	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Kontrast	10559924,321	2	5279962,2	5,135	,007
Fehler	139837788,116	136	1028219,0		

Das Hauptergebnis finden Sie in Tabelle 15.5. Unser Faktor SCHUL2 hat drei Faktorstufen „Hauptschüler" (Stufe 1), Mittelschüler" (Stufe 2) und „Abiturienten" (Stufe 3). Die Mittelwerte dieser Stufen sind uns bekannt. Sie betragen (für die gültigen Werte im Modell ohne Kovariate): Stufe 1 1771,39, Stufe 2 2396,58 und Stufe 3 2240,56. Beim Verfahren „Einfach" werden die Mittelwerte der Faktorstufen mit dem der Referenzkategorie verglichen. Wir haben „Letzte" ausgewählt, also „Abiturienten". Demnach werden die Mittelwerte der beiden anderen Katego-

rien mit dem Mittelwert dieser Stufe verglichen. Die Ergebnisse dieser Vergleiche stehen in der Zeile „Kontrastschätzer". Z.B. beträgt die Differenz zwischen Stufe 1 und Stufe 3: 1771,39 – 2240,56= –469,18. Es werden außerdem der Standardfehler und die obere und untere Grenze eines 95%-Konfidenzintervalls angegeben. Da 0 nicht in diesem Intervall liegt, kann es als gesichert angesehen werden, dass tatsächlich eine Differenz zwischen diesen Gruppen besteht. Dasselbe besagt der Wert 0,026 für Signifikanz (da die Irrtumswahrscheinlichkeit $\alpha < 0{,}05$). Wenn nicht mit Hilfe der Syntax anders definiert, wird immer davon ausgegangen, dass gegen die Nullhypothese getestet werden soll. Das ist hier auch der Fall. In der Tabelle schlägt sich das in „Hypothesenwert" 0 nieder.

Weiter gehört zur Ausgabe die Tabelle Testergebnisse. Aus dieser kann man entnehmen, ob sich das durch die Kontraste definierte Gesamtmodell signifikant von der Annahme fehlender Zusammenhänge unterscheidet. Das ist hier der Fall, was wir am Wert 0,007 in der Spalte „Signifikanz" erkennen (obwohl zwischen der Teilstufe 2 und 3, wie oben zu sehen, kein signifikanter Zusammenhang besteht).

Die anderen Berechnungsarten (außer Regression) sind in Tabelle 15.6 demonstriert.

Tabelle 15.6. Multipler Gruppenvergleich mittels apriori Kontrast nach verschiedenen Verfahren

		Abweichung		**Differenz**		**Helmert**		**Wiederholt**	
Stufe	Mittelwerte	Vergleichsstufen	Differenz	Vergleichsstufen	Differenz	Vergleichsstufen	Differenz	Vergleichsstufen	Differenz
I Hauptschule	1771,39	I vs Mittelw	-364,79	II vs I	625,19	I vs (II+III)	-547,18	I vs III	-625,19
II Mittelschule	2396,58	II vs Mittelw	260,40	III vs (I+II)	156,58	II vs III	156,01	II vs III	156,01
III Abitur	2240,56								
Gesamt	2136,17								

Weitere Optionen.

- *Signifikanzniveau.* Durch Veränderung des Wertes im Eingabekästchen „Signifikanzniveau" (Voreinstellung 0,05) verändert man das Signifikanzniveau sämtlicher abgerufener Signifkanztest, sofern sie nicht selbst den exakten α-Wert ausgeben (Schärfe, Bildung homogener Gruppen), und gleichzeitig das Sicherheitsniveau, das der Berechnung von Ober- und Untergrenzen von Konfidenzintervallen zugrunde gelegt wird.
- *Parameterschätzer.* Gibt die Parameter für die Terme einer Regressionsgleichung aus. (Ist bei der Anwendung für Regressionsanalysen relevant.)
- *Matrix Konstrastkoeffizienten.* Gibt mehrere Matrizen mit den in dem Modell verwendeten Kontrastkoeffizienten aus. Ist dann als Ausgangspunkt von Belang, wenn man eigene Modelle mit eigenen Kontrastkoeffizienten über die Syntax definieren will.

- *Allgemeine schätzbare Funktionen.* Gibt eine Kontrastmatrix für die verwendeten Terme aus.
- *Diagnostikfeatures.* Fast alle anderen Optionen dienen der Überprüfung der Voraussetzung homogener Varianz in den Vergleichsgruppen. *Homogenitätstest* führt den an andere Stelle bereits besprochenen „Levene-Test" durch. Die Diagramme „Streubreite vs. mittleres Niveau" (⇨ Kap. 9.3.1) und „Residuen-Diagramm" dienen demselben Zweck.
- *Unzureichende Anpassung.* Der Test sollte keinen signifikante Abweichung vom Modell ausweisen. (Probleme ⇨ Kap. 13.3.)

Tests auf unzureichende Anpassung

Abhängige Variable: BEFR.: MONATLICHES NETTOEINKOMMEN

Quelle	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
unzureichende Anpassung	34881223,381	46	758287,465	1,144	,340
Reiner Fehler	23856307,167	36	662675,199		

Speichern. Über die Schaltfläche „Speichern" gelangt man in eine Dialogbox, in der man festlegen kann, dass bestimmte Werte als neue Variablen gespeichert werden sollen. Gewählt werden können in der Gruppe „Vorhergesagte Werte" nicht standardisierte Werte, gewichtete Werte (falls Gewichtung vorgenommen wurde) und Standardfehler, in der Gruppe „Residuen" können nicht standardisierte, standardisierte (Residuen, geteilt durch den Standardirrtums) und studentisierte Residuen (Residuen geteilt durch ihren geschätzte Standardabweichung) angefordert werden. Studentisierte Residuen gelten als angemessener. „Ausgeschlossen" liefert für die Fälle das Residuum, das entstehen würde, wenn der betreffende Fall bei der Berechnung der Regressionsgleichung ausgeschlossen würde. In der Gruppe „Diagnose" steht die Cook-Distanz zur Verfügung, ein Maß, das angibt, wie stark sich die Residuen aller Fälle ändern würden, wenn der betrachtete Fall ausgeschlossen würde. Daneben der „Hebewert", ein Maß für den relativen Einfluss des speziellen Wertes auf die Anpassungsgüte des Modells. Bei kleineren Stichproben, in welchen Ausreißer das Ergebnis stark beeinflussen, sucht man mit Hilfe dieser Werte solche einflussreichen Werte und eliminiert sie gegebenenfalls. Ein Wert nahe Null signalisiert geringen Einfluss, je weiter der Wert von Null abweicht, desto kritischer ist der entsprechende Fall zu beurteilen. Schließlich kann die „Koeffizientenstatistik" in einer eigenen neuen (SPSS-Daten-)Datei gespeichert werden.

Gewichten. In der Dialogbox „Univariat" kann eine Gewichtung vorgenommen werden. Dazu muss vorher eine Gewichtungsvariable gebildet sein. Diese wird dann aus der Variablenliste in die das Feld „WLS – Gewichtung" übertragen. Durch die Gewichtung können Fälle von der Analyse ganz ausgeschlossen (Wert 0) werden oder mit geringerem oder größerem Gewicht in die Analyse eingehen (je nach relativer Größe des Wertes in der Gewichtungsvariablen).

16 Korrelation und Distanzen

Zur Messung der Stärke und Richtung des Zusammenhangs zwischen zwei Variablen werden *Korrelationskoeffizienten* berechnet. In Kap. 10.3 werden eine Reihe von Zusammenhangsmaßen bzw. Korrelationskoeffizienten erläutert, so dass hier zur Darstellung und Anwendung des Menüs „Korrelation" nur ergänzende Erörterungen erforderlich sind.

Das Menü „Korrelation" erlaubt es, bivariate und partielle Korrelationskoeffizienten zu berechnen. In den Anwendungsbeispielen zur bivariaten und partiellen Korrelation werden einige makroökonomische Datenreihen für die Bundesrepublik im Zeitraum 1960 bis 1990 genutzt (Datei MAKRO.SAV). Untersuchungsobjekte bzw. Fälle für Variablenwerte sind also die Jahre von 1960 bis 1990. Die Datenreihen von MAKRO.SAV sind mit dem Daten-Service erhältlich (⇨ Anhang B). Damit können Sie die Beispiele nachvollziehen sowie andere Optionen des Menüs ausprobieren

In Kap. 16.3 werden Distanz- und Ähnlichkeitsmaße behandelt.

16.1 Bivariate Korrelation

Theoretische Grundlagen. Das Messkonzept der (bivariaten) Korrelation lässt sich gut mit Hilfe eines *Streudiagramms* (englisch scatterplot) veranschaulichen, in dem die beiden Variablen x und y Achsen eines Koordinatensystems bilden. X/Y-Wertekombinationen von Untersuchungsobjekten (Fällen) bilden eine Punktwolke im Koordinatensystem (= Streudiagramm). Aus der Form der Punktwolke ergeben sich Rückschlüsse auf die Stärke und Richtung des Zusammenhangs der Variablen. In der folgenden Abb. 16.1 werden einige typische Formen dargestellt.

In a) von Abb. 16.1 wird ein positiver Zusammenhang der Variablen y und x sichtbar: Mit dem Anstieg von x wird auch y tendenziell y größer. Der Zusammenhang ist stark bis mittelstark: die Punkte um eine in die Punktwolke legbare Gerade zur Darstellung der Richtung des Zusammenhangs streuen nicht sehr stark um eine derartige Gerade. Ein berechneter Korrelationskoeffizient wird den positiven Zusammenhang durch ein positives Vorzeichen und die Stärke des Zusammenhangs durch die absolute Höhe des Koeffizienten ausweisen. Da die Punkte nicht sehr stark um eine in die Punktwolke legbare Gerade streuen, wird der Korrelationskoeffizient nahe dem (absoluten) maximalen Wert von eins liegen.

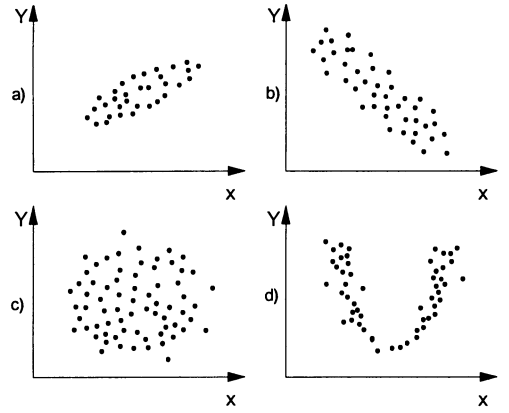


Abb. 16.1. Beispiele für Streudiagramme

In b) von Abb. 16.1 besteht ein negativer und mittelstarker Zusammenhang zwischen x und y. Ein berechneter Korrelationskoeffizient wird ein negatives Vorzeichen haben. Der Korrelationskoeffizient wird relativ groß sein, da die Werte eng um eine durch die Punktwolke legbare Gerade streuen.

In c) hat die Punktwolke keine Richtung. Die beiden Variablen stehen in keinem statistischen Zusammenhang. Der Korrelationskoeffizient hat einen Wert von Null bzw. nahe Null.

In d) wird ein enger nichtlinearer Zusammenhang zwischen y und x deutlich. Da Korrelationskoeffizienten die Stärke eines *linearen* Zusammenhangs messen, kann die Höhe des berechneten Korrelationskoeffizienten den tatsächlich bestehenden engen Zusammenhang aber nicht zum Ausdruck bringen. Eine Ermittlung eines Korrelationskoeffizienten für die (ursprünglichen) Werte von x und y verletzt eine Bedingung für die Anwendung: das Bestehen eines linearen Zusammenhangs. Dieses Beispiel zeigt wie wichtig es ist, zusammen mit der Berechnung von Korrelationskoeffizienten auch Streudiagramme zu erstellen (⇨ Kap. 24.11). Gelingt es, durch eine Transformation der Variablen den Zusammenhang der Variablen zu linearisieren (z.B. durch Logarithmierung der Variablen), so wird eine Anwendung eines Korrelationskoeffizienten auf die transformierten Variablenwerte sinnvoll.

Es muss davor gewarnt werden, aus der mit Hilfe eines Korrelationskoeffizienten gemessenen oder anhand eines Streudiagramms grafisch veranschaulichten statistischen Korrelation zwischen zwei Variablen auf das Bestehen eines Kausalzusammenhangs zu schließen. Zwei voneinander unabhängige Variable y und x können eine statistische Korrelation ausweisen, weil z.B. eine dritte Variable z sowohl auf y als auch x wirkt und sie sich verändert. Bei Vorliegen einer statistisch gemessenen Korrelation ohne Vorliegen eines Kausalzusammenhangs spricht man von *Scheinkorrelation*. Ein in der Literatur gern zitiertes Beispiel für eine Scheinkorrelation ist der gemessene positive und in mittlerer Größenordnung liegende Korrelationskoeffizient für den Zusammenhang zwischen der Anzahl der Geburten und der Zahl der gezählten Störche in einer Region.

Die Begründung für das Vorliegen eines Zusammenhangs zwischen Variablen sollte theoretisch bzw. durch Plausibilitätserklärung fundiert sein. Die Berechnung

eines Korrelationskoeffizienten kann lediglich die Stärke eines begründeten linearen Zusammenhangs messen.

Häufig beschränkt man sich bei einer Korrelationsanalyse von Daten nicht auf das Messen des Zusammenhangs der Variablen für den vorliegenden Datensatz im Sinne einer deskriptiven statistischen Untersuchung, sondern hat den Anspruch, allgemeinere Aussagen darüber zu treffen, ob ein Zusammenhang zwischen den Variablen besteht oder nicht. Dabei wird ein theoretischer Zusammenhang zwischen den Variablen im universelleren Sinne für eine tatsächlich existierende oder theoretisch gedachte Grundgesamtheit postuliert und der vorliegende Datensatz als eine Stichprobe aus der Grundgesamtheit interpretiert. Mit der Formulierung einer derartigen stichprobentheoretisch fundierten Korrelationsanalyse lassen sich Signifikanzprüfungen für die Höhe des Korrelationskoeffizienten vornehmen. Damit soll es ermöglicht werden, zwischen den Hypothesen des Bestehens und Nichtbestehens einer Korrelation zu diskriminieren.

Das Untermenü „Bivariat" von „Korrelation" erlaubt die Berechnung drei verschiedener Korrelationskoeffizienten (*Pearson*, *Kendall-Tau-b* und *Spearman*), die unter unterschiedlichen Anwendungsbedingungen gewählt werden können (⇨ Kap. 10.3).

Der Korrelationskoeffizient nach Pearson setzt eine metrische Skala beider Variablen voraus und misst Richtung und Stärke des linearen Zusammenhangs der Variablen. Der von SPSS berechnete Korrelationskoeffizient nach Pearson ist definiert (⇨ auch Kap. 10.3.3)

$$r_{x,y} = \frac{\frac{1}{n-1}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n-1}\sum(x-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum(y-\bar{y})^2}} \qquad (16.1)$$

Der Ausdruck $\frac{1}{n-1}\sum(x-\bar{x})(y-\bar{y})$ ist die geschätzte Kovarianz der Variablen x und y und misst die Stärke und Richtung des linearen Zusammenhangs zwischen den Variablen in Form eines nicht normierten Maßes. Die Ausdrücke

$$\sqrt{\frac{1}{1-n}\sum(y-\bar{y})^2} \text{ sowie } \sqrt{\frac{1}{n-1}\sum(y-\bar{y})^2}$$

sind die geschätzten Standardabweichungen s_x und s_y der Variablen. Der Korrelationskoeffizient r_{xy} ist also der Quotient aus Kovarianz und den Standardabweichungen der beiden Variablen. Er ist die mit den Standardabweichungen der beiden Variablen normierte Kovarianz. Die Normierung stellt sicher, dass der Korrelationskoeffizient im Falle eines vollkommenen (mathematischen) Zusammenhangs maximal den Wert eins annimmt.

Der Korrelationskoeffizient nach Pearson hat folgende Eigenschaften:

- Er hat je nach Richtung des Zusammenhangs ein positives oder negatives Vorzeichen.
- Er ist dimensionslos.
- Ein Vertauschen der Variablen berührt nicht den Messwert.

- Er kann absolut maximal 1 und minimal 0 werden.
- Er misst die Stärke eines linearen Zusammenhangs.

Soll statistisch getestet werden, ob ein linearer Zusammenhang zwischen den Variablen x und y für die Grundgesamtheit besteht, also die Hypothese geprüft werden, ob der unbekannte Korrelationskoeffizient der Grundgesamtheit - hier ρ genannt - sich signifikant von Null unterscheidet - so bedarf es spezieller Annahmen. Unter der Voraussetzung, dass die gemeinsame (bivariate) Verteilung der Variablen normalverteilt ist und die vorliegenden Daten aus dieser per Zufallsauswahl entnommen worden sind, hat die Prüfgröße (⇨ Kap 13.3)

$$t = r_{x,y}\sqrt{\frac{n-2}{1-r^2}} \tag{16.2}$$

für den Fall $\rho = 0$ eine Student's t-Verteilung mit $n-2$ Freiheitsgraden. Aus tabellierten t-Verteilungen lässt sich unter der Vorgabe einer Irrtumswahrscheinlichkeit von z.B. 5 % ($\alpha = 0{,}05$) für die Anzahl der Freiheitsgrade in Höhe von $n-2$ ein „kritischer" t-Wert ablesen. Ist der empirisch berechnete Wert (absolut) kleiner als der „kritische", so wird die Hypothese H_0, $\sigma = 0$ (kein Zusammenhang zwischen den Variablen), angenommen. Ist er (absolut) größer, so wird die Alternativhypothese H_1 (es besteht ein Zusammenhang) angenommen. Die Alternativhypothese wird dabei je nach Erwartung über die Richtung des Zusammenhangs unterschiedlich formuliert. Hat man keinerlei Erwartung über die Richtung des Zusammenhangs, so gilt H_1: $\rho \neq 0$. Es handelt sich dann um einen zweiseitigen Test. Erwartet man, dass die Variablen sich in gleicher Richtung verändern, so wird der positive Zusammenhang mit H_1: $\rho > 0$ formuliert. Bei Erwartung eines negativen Zusammenhangs gilt entsprechend H_1: $\rho < 0$. In diesen Fällen handelt es sich um einen einseitigen Test.

Anwendungsbeispiel. Im folgenden Beispiel soll untersucht werden, wie stark der private Konsum (CPR) mit der Höhe des verfügbaren Einkommens (YVERF) und dem Zinssatz (ZINS) korreliert. Erwartet wird, dass die Korrelation von CPR und YVERF positiv und sehr hoch und die von CPR und ZINS negativ und eher mittelmäßig stark ist.

Abb. 16.2 zeigt das SPSS-Daten-Editorfenster mit einem Ausschnitt aus der Datei MAKRO.SAV.

Zur Berechnung der Korrelationskoeffizienten gehen Sie wie folgt vor:

- ▷ Wählen Sie per Mausklick die Befehlsfolge „Analysieren", „Korrelation ▷„ „Bivariat...". Es öffnet sich die in Abb. 16.3 dargestellte Dialogbox.
- ▷ Übertragen Sie die zu korrelierenden Variablen aus der Quellvariablenliste in das Feld „Variablen:" (hier: CPR, YVERF und ZINS).
- ▷ Wählen Sie den gewünschten Korrelationskoeffizienten aus (hier: Pearson).
- ▷ Wählen Sie aus, ob Sie einen einseitigen oder zweiseitigen Signifikanztest durchführen wollen (hier: einseitig = one-tailed, da eine Erwartung über die Richtung der Zusammenhänge besteht).
- ▷ Wählen Sie, ob signifikante Korrelationen markiert werden sollen.

▷ Falls das Untermenü „Optionen...“ aktiviert werden soll, wird es angeklickt. Falls nicht, wird mit Klicken der Schaltfläche „OK“ die Berechnung gestartet.

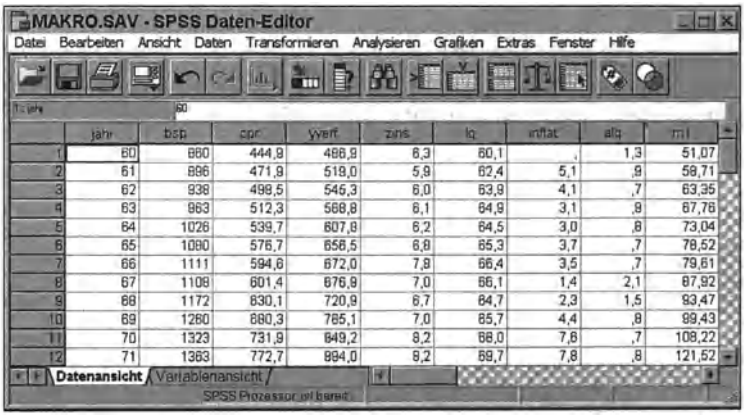


Abb. 16.2. SPSS-Daten-Editorfenster mit makroökonomischen Daten

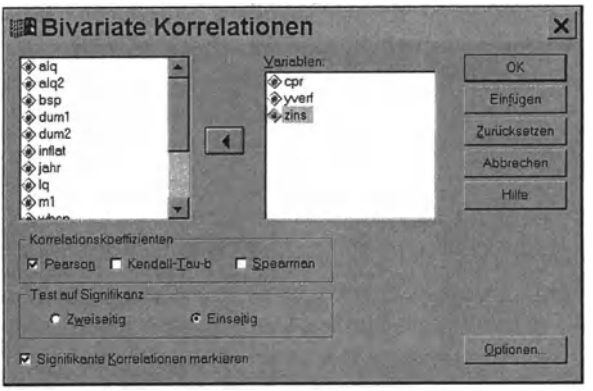


Abb. 16.3. Dialogbox „Bivariate Korrelationen“

Wahlmöglichkeiten.

① *Korrelationskoeffizienten.* Es kann der nach Pearson, Kendall-Tau-b sowie Spearman gewählt werden (⇨ Kap. 10.3).

② *Test auf Signifikanz.* Man kann sich entweder ein einseitiges oder ein zweiseitiges Signifikanzniveau angeben lassen.

③ *Signifikante Korrelationen markieren.* Kennzeichnung durch Sternchen.

④ *Optionen.* Anklicken der Schaltfläche „Optionen..." öffnet die in Abb. 16.4 dargestellte Dialogbox. Sie enthält zwei Auswahlgruppen:

- *Statistik.*
  - *Mittelwerte und Standardabweichungen.* Berechnung der arithmetischen Mittel sowie der Standardabweichungen.
  - *Kreuzproduktabweichungen und Kovarianzen.* Ausgabe der Kreuzprodukte der Abweichungen vom arithmetischen Mittelwert sowie der Kovarianzen.
- *Fehlende Werte.*
  - *Paarweiser Fallausschluss.* Fälle mit fehlenden Werten werden nur für die jeweiligen Variablenpaare, nicht aber für die gesamte Liste der zu korrelierenden Variablen, ausgeschlossen. Diese Option führt dazu, dass bei fehlenden Werten die Korrelationskoeffizienten einer Variablenliste auf der Basis unterschiedlicher Fälle berechnet werden und daher nur eingeschränkt vergleichbar sind.
  - *Listenweiser Fallausschluss.* Es werden die Fälle aller zu korrelierenden Variablen ausgeschlossen, sofern bei Variablen fehlende Werte auftreten. Es ist sorgfältig zu prüfen, ob eventuell ein systematischer Zusammenhang zwischen fehlenden Werten und Werten der Untersuchungsvariablen besteht. Nur wenn ein derartiger Zusammenhang nicht erkennbar ist, werden die ermittelten Korrelationskoeffizienten den Zusammenhang der Variablen unverzerrt widerspiegeln.

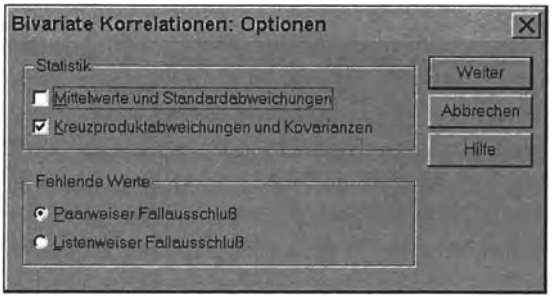


Abb. 16.4. Dialogbox „Bivariate Korrelation: Optionen"

Die in Abb. 16.3 und 16.4 gewählten Einstellungen führen zur in Tabelle 16.1 dargestellten Ausgabe. Die Ergebnisse werden in Matrixform dargestellt. Zunächst werden die Korrelationskoeffizienten aufgeführt. Die Korrelationskoeffizienten in der Diagonale von links oben nach rechts unten haben den Wert eins, da in diesen Fällen die Variablen mit sich selbst korreliert werden. Danach wird das einseitige Signifikanzniveau des Korrelationskoeffizienten aufgeführt. Dieses gibt an, mit welcher Irrtumswahrscheinlichkeit die H_0-Hypothese (es besteht kein Zusammenhang zwischen den Variablen) abgelehnt wird. Danach werden die Quadratsummen und Kreuzprodukte der Abweichungen vom Mittelwert, die Kovarianz und die

Fallanzahl N aufgeführt. N ist in diesem Beispiel trotz der gewählten Option „Paarweiser Ausschluss" 31, da für die gewählten Variablen keine Werte fehlen.

Bezüglich der Korrelation zwischen dem privaten Konsum (CPR) und dem verfügbaren Einkommen (YVERF) trifft mit $r_{cpr,yverf} = 0{,}9987$ die Erwartung eines sehr hohen Korrelationskoeffizienten mit positivem Vorzeichen zu. Das Signifikanzniveau P = 0,00 gibt an, dass mit dieser Irrtumswahrscheinlichkeit die H_0-Hypothese (es besteht kein Zusammenhang zwischen CPR und YVERF) abgelehnt und damit die H_1-Hypothese (es besteht ein positiver Zusammenhang) angenommen werden kann. Für CPR und ZINS wird mit dem Ergebnis $r_{cpr,zins} = 0{,}24$ die Hypothese eines erwarteten negativen Zusammenhangs nicht bestätigt. Das ermittelte Signifikanzniveau P = 0,097 liegt über der bei statistischen Tests üblichen Höhe von 5 % ($\alpha = 0{,}05$). Insofern wäre die H_0-Hypothese (es besteht kein Zusammenhang) beizubehalten. Bevor aber eine derartige Schlussfolgerung gezogen wird, sollte überlegt und geprüft werden, ob der korrelative Zusammenhang falsch gemessen wird, da der starke Einfluss von YVERF auf CPR den tatsächlichen Zusammenhang eventuell verdeckt. Die partielle Korrelation in Kap. 16.2 wird eine Klärung ermöglichen.

Tabelle 16.1. Ergebnisausgabe der bivariaten Korrelation

Korrelationen

		CPR	YVERF	ZINS
Korrelation nach Pearson	CPR	1,000	,999**	,240
	YVERF	,999**	1,000	,269
	ZINS	,240	,269	1,000
Signifikanz (1-seitig)	CPR	.	,000	,097
	YVERF	,000	.	,072
	ZINS	,097	,072	.
Quadratsummen und Kreuzprodukte	CPR	1568974	1830757	2191,967
	YVERF	1830757	2141581	2870,569
	ZINS	2191,967	2870,569	53,157
Kovarianz	CPR	52299,14	61025,25	73,066
	YVERF	61025,25	71386,04	95,686
	ZINS	73,066	95,686	1,772
N	CPR	31	31	31
	YVERF	31	31	31
	ZINS	31	31	31

**. Die Korrelation ist auf dem Niveau von 0,01 (1-seitig) signifikant.

Das Menü „Bivariate Korrelation" ermöglicht auch die Berechnung von Rangkorrelationskoeffizienten. Dabei stehen Kendall-Tau-b sowie der Korrelationskoeffizient nach Spearman zur Auswahl (⇨ Kap. 10.3). Rangkorrelationskoeffizienten werden berechnet, wenn entweder mindestens eine der beiden zu korrelierenden Variablen ordinalskaliert ist oder aber bei Vorliegen von metrischen Variablen ein statistischer Signifikanztest durchgeführt werden soll, aber die Voraussetzung einer bivariaten Normalverteilung nicht erfüllt ist.

16.2 Partielle Korrelation

Theoretische Grundlagen. Man kann im allgemeinen erwarten, dass der Zusammenhang zwischen zwei Variablen x und y durch den Einfluss weiterer Variablen beeinflusst wird. So kann z.B. der Einfluss einer dritten Variable z der Grund dafür sein, dass x und y statistisch korreliert sind, obwohl tatsächlich kein Zusammenhang zwischen ihnen besteht (Scheinkorrelation). Denkbar ist umgekehrt auch, dass ein tatsächlich bestehender Zusammenhang zwischen x und y durch den Einfluss von z statistisch verdeckt wird, so dass der Korrelationskoeffizient $r_{x,y}$ einen Wert nahe Null annimmt. Eine Scheinkorrelation oder verdeckte Korrelation kann mit Hilfe der partiellen Korrelation aufgedeckt werden. Eine partielle Korrelation entspricht dem Versuch, den korrelativen Zusammenhang zwischen x und y bei Konstanz der Variablen z zu messen. Damit wird eine Analogie zur Kreuztabellierung von Variablen unter Berücksichtigung von Kontrollvariablen deutlich. Im Unterschied zur Kreuztabellierung kann die Kontrolle nur statistisch unter der Voraussetzung linearer Beziehungen erfolgen: Es wird die Stärke des linearen Zusammenhangs zwischen x und y bei statistischer Eliminierung des linearen Effekts von z sowohl auf x als auch auf y gemessen.

Werden in zwei linearen Regressionsansätzen (⇨ Kap. 17)

$$y = a_1 + b_1 z + e_1 \tag{16.3}$$

$$x = a_2 + b_2 z + e_2 \tag{16.4}$$

sowohl die Variable y als auch x durch z erklärt, so sind

$$e_1 = y - (a_1 + b_1 z) \tag{16.5}$$

$$e_2 = x - (a_2 + b_2 z) \tag{16.6}$$

die Residualwerte, die jeweils die vom Einfluss der Variable z „bereinigten" Variablen x bzw. y darstellen. Der partielle Korrelationskoeffizient zwischen x und y wird häufig mit $r_{yx,z}$ bezeichnet. Er entspricht dem bivariaten Korrelationskoeffizienten nach Pearson zwischen den Variablen e_1 und e_2 ($r_{e_1e_2}$). In diesem Beispiel handelt es sich um eine partielle Korrelation erster Ordnung, da nur der Einfluss einer Variable z konstant gehalten (kontrolliert) wird. Ermittelbar sind in analoger Weise partielle Korrelationen höherer Ordnung, wenn zwei oder mehr Variable z_1, z_2 etc. in ihrer Wirkung auf x und y statistisch eliminiert werden, um die Stärke des Zusammenhangs zwischen y und x bei Kontrolle weiterer Variablen zu messen.

Auch partielle Korrelationskoeffizienten können auf statistische Signifikanz geprüft werden. Unter der Voraussetzung einer multivariaten Normalverteilung kann die H_0-Hypothese (partielle Korrelationskoeffizient der Grundgesamtheit $\rho = 0$) mit Hilfe folgender Prüfgröße

$$t = r\sqrt{\frac{n-\theta-2}{1-r^2}} \tag{16.7}$$

geprüft werden. Die Prüfgröße ist t-verteilt mit $n-\theta-2$ Freiheitsgraden. In der Gleichung ist r der partielle Korrelationskoeffizient, n die Anzahl der Fälle und θ die Ordnung des Korrelationskoeffizienten. Ist der empirische t-Wert gleich bzw. kleiner als ein (gemäß der Freiheitsgrade und einer vorgegebenen Irrtumswahrscheinlichkeit α) aus der tabellierten t-Verteilung entnehmbarer kritischer t-Wert, so wird die H_0-Hypothese (kein Zusammenhang zwischen den Variablen) angenommen. Für den Fall, dass der empirische t-Wert den kritischen übersteigt, wird die H_0-Hypothese abgelehnt und damit die H_1-Hypothese (es besteht ein Zusammenhang) angenommen.

Anwendungsbeispiel. Das im vorherigen Abschnitt 16.1 gewonnene Ergebnis einer positiven, aber nicht signifikanten Korrelation zwischen den Variablen CPR (privater Konsum) und ZINS (Zinssatz) entsprach nicht der Erwartung. Es soll nun mittels Berechnung eines partiellen Korrelationskoeffizienten zwischen CPR und ZINS bei Kontrolle des Einflusses der Variablen YVERF (verfügbares Einkommen) geprüft werden (= Korrelationskoeffizient erster Ordnung), ob das unplausible Ergebnis ein Resultat einer verdeckten Korrelation ist. Dazu gehen Sie wie folgt vor:

- ▷ Durch Mausklicken wird die Befehlsfolge „Analysieren" „Korrelation ▷„ „Partiell..." aufgerufen. Es öffnet sich die in Abb. 16.5 dargestellte Dialogbox. Übertragen sie aus der Quellvariablenliste die zu korrelierenden Variablen CPR und ZINS in das Feld „Variablen:".
- ▷ Übertragen Sie die Kontrollvariable YVERF in das Feld „Kontrollvariablen:".
- ▷ Wählen Sie, ob ein einseitiger oder zweiseitiger Signifikanztest vorgenommen werden soll.
- ▷ Wählen Sie, ob das Signifikanzniveau angezeigt werden soll.
- ▷ Falls weitere optionale Berechnungen durchgeführt werden sollen, muss die Schaltfläche „Optionen" angeklickt werden. Falls nicht, wird mit „OK" die Berechnung gestartet.

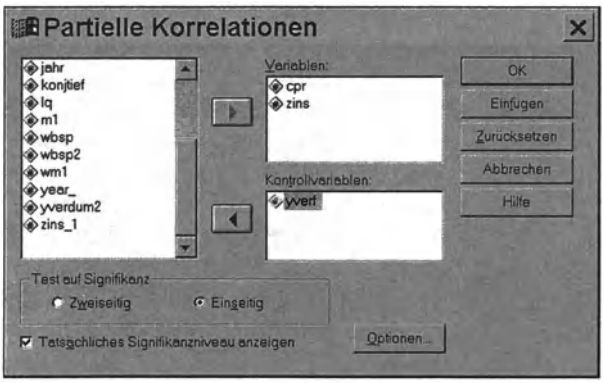


Abb. 16.5. Dialogbox „Partielle Korrelation"

Das in Tabelle 16.2 dargestellte Ergebnis ist Resultat der gewählten Einstellungen. Es ergibt sich mit $r = -0{,}5952$ das erwartete Ergebnis, dass CPR und ZINS mit mittlerer Stärke negativ korreliert ist, wenn die Haupteinfluss*s*variable YVERF kontrolliert wird. Unter dem Koeffizienten steht in Klammern die Anzahl der Freiheitsgrade (degrees of freedom, „D.F."). Sie beträgt 28 (wegen $D.F. = n - \theta - 2$), da $n = 31$ und $\theta = 1$. Mit $P = 0{,}000$ ist die Irrtumswahrscheinlichkeit so gering, dass die H_0-Hypothese (kein Zusammenhang zwischen den Variablen) zugunsten der H_1-Hypothese (negativer Zusammenhang) abgelehnt wird.

Wahlmöglichkeiten. Es bestehen die gleichen Optionen wie bei der bivariaten Korrelation (⇨ Kap. 16.1).

Tabelle 16.2. Ergebnisausgabe partieller Korrelation

```
P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N T S
Controlling for.. YVERF

                    CPR          ZINS
CPR              1,0000       -,5952
                (     0)     (    28)
                P= ,         P= ,000

ZINS             -,5952       1,0000
                (    28)     (     0)
                P= ,000      P= ,

(Coefficient / (D.F.) / 1-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

16.3 Distanz- und Ähnlichkeitsmaße

Messkonzepte für Distanz und Ähnlichkeit. Personen oder Objekte (Fälle) werden als ähnlich bezeichnet, wenn sie in mehreren Eigenschaften weitgehend übereinstimmen. Interessiert man sich z.B. für die Ähnlichkeit von Personen als Käufer von Produkten, so würde man Personen mit ähnlicher Einkommenshöhe, mit ähnlichem Bildungsstand, ähnlichem Alter und eventuell weiteren Merkmalen als ähnliche Käufer einordnen. Ähnliche Autos hinsichtlich ihrer Fahreigenschaften stimmen weitgehend überein in der Größe, der Motorleistung, den Beschleunigungswerten, dem Kurvenverhalten etc. Die multivariate Statistik stellt für unterschiedliche Messniveaus der Variablen (⇨ Kap. 8.3.1) etliche Maße bereit, um die Ähnlichkeit bzw. Unähnlichkeit von Personen bzw. Objekten zu messen.

Maße für die Unähnlichkeit von Objekten werden *Distanzen* genannt. Dabei gilt, dass eine hohe Distanz von zwei verglichenen Objekten eine starke Unähnlichkeit und eine niedrige Distanz eine hohe Ähnlichkeit der Objekte zum Ausdruck bringt. Für *Ähnlichkeitsmaße* gilt umgekehrt, dass hohe Messwerte eine starke Ähnlichkeit der Objekte ausweisen. Alle Distanz- und Ähnlichkeitsmaße beruhen auf einem

Vergleich von jeweils zwei Personen bzw. Objekten unter Berücksichtigung von mehreren Merkmalsvariablen.

Es gibt aber auch Ähnlichkeitsmaße für Variablen. Als zwei ähnliche Variable werden Variable definiert, die stark zusammenhängen. Daher handelt es sich bei den Ähnlichkeitsmaßen um Korrelationskoeffizienten bzw. um andere Zusammenhangsmaße.

Je nach Art der Daten bietet SPSS für eine Messung der Distanz (bzw. der Ähnlichkeit) eine Reihe von unterschiedlichen Maßen an. Für intervallskalierte Variablen und für binäre Variablen (Variablen mit nur zwei Merkmalswerten: 0 = eine Eigenschaft ist nicht vorhanden, 1 = eine Eigenschaft ist vorhanden) gibt es jeweils eine Reihe von Distanz- und Ähnlichkeitsmaßen. Liegen die Daten in Form von Häufigkeiten von Fällen vor, so kann aus zwei Distanzmaßen ausgewählt werden (vergl. die Übersicht in Tabelle 16.3).

Distanz- und Ähnlichkeitsmaße werden als Eingabedaten für die Clusteranalyse verwendet (⇨ Kap. 19).

Für jede der in Tabelle 16.3 aufgeführten fünf Gruppen von Maßen sollen nun exemplarisch mittels kleiner Beispiele die Definitionskonzepte der Maße erläutert werden. Auf der Basis dieser Erläuterungen kann man sich bei Bedarf sehr leicht Kenntnisse über alle anderen Maße beschaffen, wenn man über das SPSS-Hilfesystem das Syntaxhandbuch („Syntaxguide") von SPSS aufruft. Es öffnet sich dann automatisch Acrobat-Reader zum Lesen des Handbuchs. Man wähle dort die Seiten 667 bis 684 [mit der Maus auf das Bildrollfeld (⇨ Abb. 2.1) zeigen; linke Maustaste festhalten und Bildrollfeld ziehen bis die gewünschte Seite angezeigt wird]. Dort sind alle Formeln zur Definition der verschiedenen Maße aufgeführt.

Tabelle 16.3. Übersicht über Distanz- und Ähnlichkeitsmaße für unterschiedliche Daten

Maß	**Distanzmaße**			**Ähnlichkeitsmaße**	
Art der Daten:	Intervallskala	Häufigkeiten	binär	Intervallskala	binär
Anzahl der Maße:	6	2	7	2	20

Distanzmaße für intervallskalierte Variablen. Tabelle 16.4 enthält für einige Hamburger Stadtteile (= Objekte) vier metrische (intervallskalierte) Variable, die als Indikatoren für die soziale Struktur (Anteil der Arbeiter in %, Mietausgaben je Person in DM) einerseits sowie der urbanen Verdichtung (Bevölkerungsdichte, Anteil der Gebäude mit bis zu zwei Wohnungen in %) andererseits dienen (Datei ALTONA.SAV). Zur Berechnung der Distanz (Unähnlichkeit) von Stadtteilen hinsichtlich der sozialen Struktur und urbanen Verdichtung in einem multivariaten Messansatz wird häufig die *Euklidische Distanz* gewählt. Bezeichnet man (wie im Syntax Guide, S. 671) die Variablen eines Ortsteils mit x_i und die Variablen des Vergleichsortteils mit y_i, wobei der Index i die Variable angibt, so berechnet sich die Euklidische Distanz EUCLID(x,y) zwischen den Ortsteilen Flottbek (x) und Othmarschen (y) auf der Basis der vier Merkmale (i = 1,2,3,4) wie folgt:

$$EUCLID(x,y) = \sqrt{\sum_i (x_i - y_i)^2} = \tag{16.8}$$

$$\sqrt{(9{,}4-7{,}3)^2 + (385{,}9-471{,}7)^2 + (53{,}2-25{,}6)^2 + (78{,}1-77{,}7)^2} = 90{,}16$$

Berechnet man auf diese Weise die Distanz zwischen den Ortsteilen Flottbek und Ottensen 1, so ergibt sich EUCLID(x,y) = 205,72. Damit wird ausgewiesen, dass die Ortsteile Flottbek und Othmarschen sich hinsichtlich der vier Variablen weniger unterscheiden als die Ortsteile Flottbek und Ottensen 1. Aus Gleichung 16.8 kann man erkennen, dass kleine Unterschiede in den Messwerten der Variablen für die Vergleichsobjekte zu einer kleinen und hohe Unterschiede zu einer großen Distanz führen.

Auch alle anderen wählbaren Maße beruhen auf Differenzen in den Werten der Variablen für die jeweils zwei verglichenen Objekte. Das Maß Block (City-Block- bzw. Manhatten-Distanz) z.B. entspricht der Summe der absoluten Differenzen der Variablenwerte der Vergleichsobjekte.

Anhand des Beispiels wird deutlich, dass die Distanzmaße vom Skalenniveau der gemessenen Variablen abhängen. Ob z.B. die Arbeiterquote in % oder in Dezimalwerten gemessen wird, hat für die Einflußstärke (das Gewicht) der Variable bei der Distanzberechnung erhebliche Bedeutung. Da die Variablen Miete je Person sowie Bevölkerungsdichte auf der Zahlenskala ein höheres Niveau haben als die Arbeiterquote, gehen diese Variablen mit einem höheren Gewicht in die Berechnung des Distanzmaßes ein. Da dieses aber in der Regel unerwünscht ist, sollten vor der Distanzberechnung die Messwerte der Variablen transformiert werden. Dadurch erhält man für die Variablen vergleichbare Messskalen. Zur Transformation bietet SPSS mehrere Möglichkeiten an (⇨ Wahlmöglichkeiten).

Eine häufig gewählte Transformation ist die von Z-Werten der Variablen (⇨ Kap. 8.5). Werden für die Variablen Arbeiteranteil (ARBEIT), Miete je Person (MJEP), Bevölkerung je ha (BJEHA) und Anteil der Gebäude mit nicht mehr als zwei Wohnungen (G2W) der Ortsteile des Hamburger Bezirks Altona (Datei ALTONA.SAV) die Z-Werte berechnet, so ergeben sich für ausgewählte Ortsteile die in Tabelle 16.5 aufgeführten Werte.

Die Euklidische Distanz EUCLID(x,y) zwischen den Ortsteilen Flottbek (x) und Othmarschen (y) auf der Basis der Z-Werte der vier Variablen i (i = 1,2,3,4) ergibt

$$EUCLID(x,y) = \sqrt{\sum_i (x_i - y_i)^2} = \tag{16.8a}$$

$$\sqrt{(-1{,}45412 - -1{,}58592)^2 + (1{,}1584 - 2{,}18224)^2 + (-0{,}53367 - -0{,}93670)^2 + (1{,}05165 - 1{,}03968)^2} = 1{,}108.$$

Für die Euklidische Distanz zwischen Flottbek und Ottensen 1 ergibt sich 4,090.

Tabelle 16.4. Variable für Hamburger Ortsteile (ALTONA.SAV)

Stadtteil	Arbeiter	Miete/Person	BVG-Dichte	Bis 2 Wohng.
Flottbek	9,4	385,9	53,2	78,1
Othmarschen	7,3	471,7	25,6	77,7
Lurup	41,7	220,8	54,9	76,3
Ottensen 1	51,6	227,6	159,6	13,6

Tabelle 16.5. Z-Werte der Variablen in Tabelle 16.4

Stadtteil	ZARBEIT	ZMJEP	ZBJEHA	ZG2W
Flottbek	-1,45412	1,15840	-0,53367	1,05165
Othmarschen	-1,58592	2,18224	-0,93670	1,03968
Lurup	0,57306	-0,81171	-0,50884	0,99776
Ottensen 1	1,19439	-0,73056	1,02005	-0,87960

Distanzmaße für Häufigkeiten. In Tabelle 16.6 ist für Städte A, B und C die Anzahl von drei verkauften Produkten einer Firma je 10 Tsd. Einwohner aufgeführt [fiktives Beispiel, für die Städte A und B bzw. für die drei Produkte werden in eckigen Klammern auch die Zeilen- und Spaltensummen der Häufigkeiten aufgeführt (Zeilen- und Spaltensummen einer 2*3-Matrix) und in runden Klammern für jede Zelle der 2*3-Matrix erwartete Häufigkeiten gemäß des Distanzmaßes der Gleichung 16.9]. Auf der Basis dieser Daten soll die Ähnlichkeit und damit die Unähnlichkeit (Distanz) von Städten hinsichtlich des Absatzes der drei Produkte im Paarvergleich gemessen werden.

Tabelle 16.6. Absatzhäufigkeiten von Produkten in Städten

Stadt	Produkt 1	Produkt 2	Produkt 3	Summe
Stadt A	20 (16)	25 (32)	35 (32)	[$n_A = 80$]
Stadt B	10 (14)	35 (28)	25 (28)	[$n_B = 70$]
Summe	[$n_1 = 30$]	[$n_2 = 60$]	[$n_3 = 60$]	[$n = 150$]
Stadt C	20	32	28	

Die wählbaren Maße beruhen auf dem Chi-Quadrat-Maß zur Prüfung auf Unterschiedlichkeit von zwei Häufigkeitsverteilungen (⇨ Kap. 22.2.1). In der Gleichung 16.9 sind $E(x_i)$ und $E(y_i)$ erwartete Häufigkeiten unter der Annahme, dass die Häufigkeiten von zwei Objekten unabhängig voneinander sind. Die erwartete Häufigkeit einer Zelle i der 2*3-Matrix berechnet sich wie folgt: $\text{Zeilensumme}_i * \text{Spaltensumme}_i / \text{Gesamtsumme}$. Für die erwartete Häufigkeit z.B. der ersten Zelle der Matrix (Stadt A und Produkt 1) ergibt sich: $n_A * n_1 / n = 30 * 80 / 150 = 16$. Bezeichnet man (wie im Syntax Guide, S. 954 ff.) die Häufigkeiten der Variablen i einer Stadt mit x_i und die Häufigkeiten der Va-

riablen i der Vergleichsstadt mit y_i so berechnet sich Chi-Quadrat(x,y) für die Städte A (x) und B (y) auf der Basis der aufgeführten Variablen (i = 1,2,3) wie folgt:

$$\text{CHISQ}(x,y) = \sqrt{\sum_i \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum_i \frac{(y_i - E(y_i))^2}{E(y_i)}} \tag{16.9}$$

$$= \sqrt{\frac{(20-16)^2}{16} + \frac{(25-32)^2}{32} + \frac{(35-32)^2}{32} + \frac{(10-14)^2}{14} + \frac{(35-28)^2}{28} + \frac{(25-28)^2}{28}} = 2{,}455$$

Distanzmaße für binäre Variablen. In Tabelle 16.7 sind beispielhaft für drei Personen fünf Variablen aufgeführt. Die Variable FUSSBALL, TENNIS, SEGELN, AUTOR und SKI erfassen, ob ein Interesse für die Sportarten Fußball, Tennis, Segeln, Autorennen bzw. Skifahren besteht. Die Variablen sind binäre Variablen: der Merkmalswert 0 bedeutet, dass bei einer Person das Merkmal nicht und der Wert 1, dass das Merkmal vorhanden ist. Aus den Daten erschließt sich z.B., dass die Person A kein Interesse für Fußball und Autorennen wohl aber ein Interesse für Tennis, Segeln und Skifahren hat.

Tabelle 16.7. Binäre Merkmale von Personen

Person	Fußball	Tennis	Segeln	Autor	Ski
Person A	0	1	1	0	1
Person B	0	1	1	0	0
Person C	0	0	1	1	1

Vergleicht man die Werte der fünf Variablen für die Personen A und B, so lässt sich die in Abb. 16.8 dargestellte 2*2-Kontingenztabelle mit Häufigkeiten des Auftretens von Übereinstimmungen bzw. Nichtübereinstimmungen aufstellen. Bei zwei (allgemein: a) Variablen (TENNIS und SEGELN) besteht eine Übereinstimmung hinsichtlich eines Interesses an den Sportarten, bei zwei (allgemein: d) Variablen (FUSSBALL und AUTOR) besteht eine Übereinstimmung im Nichtinteresse an den Sportarten, bei keiner (allgemein: c) ein Interesse von Person B und nicht von Person A und bei einer (allgemein: b) Sportart (SKI) ist es umgekehrt.

Tabelle 16.8. Kontingenztabelle für Person A und B (gemäß Merkmalswerten in Tabelle 16.7)

Person A	Person B	
	Merkmalswert 1	Merkmalswert 0
Merkmalswert 1	a (= 2)	b (= 1)
Merkmalswert 0	c (= 0)	d (=2)

Alle Distanzmaße für binäre Variable beruhen auf den Häufigkeiten der in Abb. 16.8 dargestellten 2*2-Kontingenztabelle. Für die Euklidische Distanz ergibt sich

$$BEUCLID(x,y) = \sqrt{b+c} = \sqrt{1+0} = \sqrt{1} = 1 \qquad (16.10)$$

Auf die Aufführung der Formeln aller anderen Distanzmaße für binäre Variablen soll hier aus Platzgründen verzichtet werden (siehe Syntaxhandbuch im Hilfesystem, S. 673 ff.).

Ähnlichkeitsmaße für intervallskalierte Variablen. Hierbei handelt es sich um Korrelationsmaße. Im voreingestellten Fall wird der Korrelationskoeffizient gemäß Gleichung 16.1 berechnet.

Ähnlichkeitsmaße für binäre Variablen. Alle diese Maße stützen sich (wie auch die Distanzmaße) auf die Häufigkeiten einer in Tabelle 16.8 dargestellten 2*2-Kontingenztabelle. Das voreingestellte Maß nach Russel und Rao berechnet sich für die Daten in Tabelle 16.7 als

$$RR = \frac{a}{a+b+c+d} = \frac{2}{2+1+0+2} = \frac{2}{5} = 0{,}4 \qquad (16.11)$$

(Zu den weiteren Maßen siehe das Syntaxhandbuch im Hilfesystem, S. 673 ff.).

Anwendungsbeispiel. Das Menü „Distanzen" erlaubt es, verschiedene Maße für die Ähnlichkeit oder Unähnlichkeit von jeweils zwei verglichenen Personen bzw. Objekten (Fällen) zu berechnen.

Die mit dem Menü berechneten Ähnlichkeits- oder Distanzmaße können als Eingabedaten einer Clusteranalyse oder einer multidimensionalen Skalierung dienen.

Die Berechnung von Distanzen bzw. Ähnlichkeiten soll am Beispiel von Ortsteilen des Hamburger Stadtbezirks Altona erläutert werden (Datei ALTONA.SAV). Die Datei enthält vier Variable (ARBEIT, BJEHA, G2W und MJEP sowie die Z-Werte dieser Variablen; siehe Tabelle 16.4 und 16.5 mit den zugehörigen Erläuterungen). Da die Vorgehensweise unabhängig vom Typ des Distanz- bzw. des Ähnlichkeitsmaßes ist, können wir uns auf ein Beispiel beschränken (Euklidische Distanz zwischen Ortsteilen). Aus oben erörterten Gründen erfolgt die Berechnung auf der Basis von Z-Werten. Dazu gehen Sie nach Laden der Datei ALTONA.SAV wie folgt vor:

- ▷ Klicken Sie die Befehlsfolge „Analysieren" „Korrelation ▷„ „Distanzen..." zum Öffnen der in Abb. 16.6 dargestellten Dialogbox. Übertragen Sie aus der Quellvariablenliste die Variablen ZARBEIT, ZBJEHA, ZG2W und ZMJEP in das Feld „Variablen:".
- ▷ Übertragen Sie die Variable ORTN (= Ortsname) in das Feld „Fallbeschriftung". Beachten Sie, dass die Fallbeschriftungsvariable zum Ausweis der Fallnummer im Output eine String-Variable sein muss.
- ▷ Nun wählen Sie, ob Sie Unähnlichkeiten (= Distanzen) (bzw. Ähnlichkeiten) zwischen Fällen (hier: Fälle, ist voreingestellt) oder zwischen Variablen berechnen wollen.
- ▷ Danach wählen Sie durch Anklicken des entsprechenden Optionsschalters, ob Sie ein Unähnlichkeits- (= Distanzmaß, ist voreingestellt) oder ein Ähnlichkeitsmaß berechnen wollen. Durch Klicken auf „Maß" öffnet sich die in Abb. 16.7 dargestellte Dialogbox. Nun ist der zu verarbeitende Datentyp (Intervall,

Häufigkeiten oder Binär) auszuwählen (hier: Intervall, ist voreingestellt). Dann kann durch Öffnen einer Drop-Down-Liste (dazu ▼ anklicken) das gewünschte Maß gewählt werden (hier: Euklidische Distanz, ist voreingestellt).

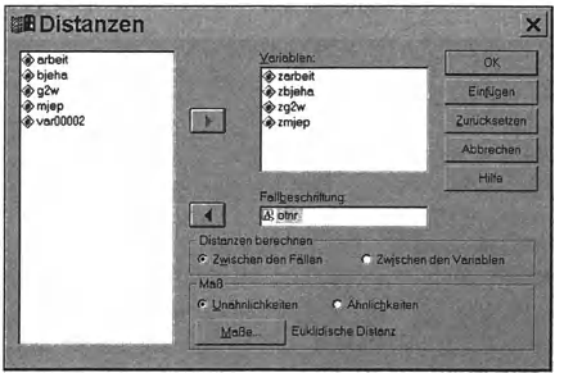


Abb. 16.6. Dialogbox „Distanzen"

Vor der Berechnung der Distanzen können die Variablen transformiert werden. Dafür kann aus mehreren Möglichkeiten gewählt werden.

Wahlmöglichkeiten.

① *Distanzen berechnen.* Es kann die Distanz bzw. die Ähnlichkeit zwischen Fällen oder zwischen Variablen berechnet werden.

② *Maß.* Es kann entweder ein Unähnlichkeits- (= Distanz-) oder ein Ähnlichkeitsmaß berechnet werden.

③ *Schaltfläche Maße.* Nach Anklicken der Schaltfläche „Maße..." öffnet die in Abb. 16.7 dargestellte Dialogbox. Sie enthält mehrere Auswahlgruppen:

- ❑ *Maß.*
  - *Intervall.* Wird gewählt, wenn man Maße für intervallskalierte Daten (⇨ Kap. 8.3.1) berechnen will. Aus einer Drop-Down-Liste wird das gewünschtes Maß gewählt.
  - *Häufigkeiten.* Wird gewählt, wenn die Daten im Dateneditor Häufigkeiten von Fällen sind. Ein gewünschtes Maß ist auszuwählen.
  - *Binär.* Wird gewählt, wenn man Maße für binäre Daten berechnen will. Binäre Daten haben nur zwei Werte. Ein gewünschtes Maß ist auszuwählen.
- ❑ *Werte transformieren.* Die für die Distanzmessung gewählten Variablen können vor der Distanzberechnung transformiert werden:
  - *Standardisieren.* Es gibt mehrere Möglichkeiten, die Variablen hinsichtlich ihres Werteniveaus zu vereinheitlichen:
    - *z-Werte.* Eine Transformation in z-Werte geschieht gemäß Gleichung 8.8 (⇨ Kap. 8.5).
    - *Bereich –1 bis 1.* Von jedem Wert einer Variable wird der größte Wert abgezogen und dann wird durch die Spannweite (kleinster minus größter Wert) dividiert.

- *Bereich 0 bis 1.* Von jedem Wert einer Variable wird der kleinste Wert abgezogen und dann wird durch die Spannweite (kleinster minus größter Wert) dividiert.
- *Maximale Größe von 1.* Jeder Wert einer Variable wird durch den größten Variablenwert dividiert.
- *Mittelwert 1.* Jeder Wert einer Variable wird durch den Mittelwert der Variable dividiert.
- *Standardabweichung 1.* Jeder Wert einer Variable wird durch die Standardabweichung der Variable dividiert.

● *Nach Variablen.* Die oben aufgeführten Transformationen werden für die Variablen durchgeführt.

● *Nach Fällen.* In diesem Fall werden die oben aufgeführten Transformationen für Fälle durchgeführt. Für die Transformation wird die Datenmatrix transponiert, d.h. um 90 Grad gedreht, so dass die Fälle zu Variablen und die Variablen zu Fällen werden.

❑ *Maße transformieren.* Hier kann man wählen, ob die berechneten Distanz- oder Ähnlichkeitsmaße transformiert werden sollen:

● *Absolutwerte.* Die Distanz- bzw. Ähnlichkeitsmaße werden ohne ihr Vorzeichen ausgegeben.

● *Vorzeichen ändern.* Ein berechnetes Maß mit einem negativen (positiven) Vorzeichen erhält in der Ergebnisausgabe ein positives (negatives) Vorzeichen.

● *Auf Bereich 0-1 skalieren.* Von jedem Distanzwert wird der kleinste Wert abgezogen und dann wird durch die Spannweite (kleinster minus größter Wert) dividiert.

Abb. 16.7. Dialogbox „Distanzen: Unähnlichkeitsmaße"

Tabelle 16.9 zeigt die Ergebnisausgabe einer Distanzberechnung gemäß den Einstellungen in den Dialogboxen der Abb. 16.6 und 16.7. Um Platz zu sparen, wird die Distanzmatrix auf die vier in Tabellen 16.4 und 16.5 aufgeführten Ortsteile beschränkt (Auswahl mit dem Menü Daten, Fälle). Die berechneten Euklidischen Distanzen in Höhe von 1,108 zwischen Flottbek und Othmarschen einerseits und die zwischen diesen Ortsteilen und Ottensen 1 in Höhe von 4,090 und 4,871 ande-

rerseits weisen deutlich aus, dass sich Flottbek und Othmarschen hinsichtlich der betrachteten Variablen stark ähnlich sind und Flottbek bzw. Othmarschen und Ottensen 1 sich stark unterscheiden.

Tabelle 16.9. Ergebnisausgabe: Euklidische Distanzen zwischen Ortsteilen

Näherungsmatrix

	Euklidisches Distanzmaß			
	10:Ottensen1	17:Flottbek	18:Othmarschen	19:Lurup
10:Ottensen1		4,090	4,871	2,501
17:Flottbek	4,090		1,108	2,827
18:Othmarschen	4,871	1,108		3,716
19:Lurup	2,501	2,827	3,716	

Dies ist eine Unähnlichkeitsmatrix

17 Lineare Regressionsanalyse

17.1 Theoretische Grundlagen

17.1.1 Regression als deskriptive Analyse

Lineare Abhängigkeit. Im Gegensatz zur Varianzanalyse und der Kreuztabellierung mit dem Chi-Quadrat-Unabhängigkeitstest befasst sich die *Regressionsanalyse* mit der Untersuchung und Quantifizierung von Abhängigkeiten zwischen metrisch skalierten Variablen (Variablen mit wohldefinierten Abständen zwischen Variablenwerten). Wesentliche Aufgabe ist dabei, eine lineare Funktion zu finden, die die Abhängigkeit einer Variablen - der *abhängigen Variablen* - von einer oder mehreren *unabhängigen Variablen* quantifiziert. Ist eine abhängige Variable y nur von einer unabhängigen Variablen x bestimmt, so wird die Beziehung in einer *Einfachregression* untersucht. Werden mehrere unabhängige Variablen, z.B. x_1, x_2 und x_3, zur Bestimmung einer abhängigen Variablen y herangezogen, so spricht man von einer *Mehrfach-* oder *multiplen Regression*. Die Regressionsanalyse kann in einfachster Form als beschreibendes, deskriptives Analysewerkzeug verwendet werden. In Abb. 17.1 wird in einem Streudiagramm (⇨ Kap. 20.12) die Abhängigkeit des makroökonomischen privaten Konsums (CPR) der Haushalte der Bundesrepublik vom verfügbaren Einkommen (YVERF) im Zeitraum 1960 bis 1990 dargestellt (Datensatz MAKRO.SAV, ⇨ Anhang B). Es ist ersichtlich, dass es sich bei dieser Abhängigkeit um eine sehr starke und lineare Beziehung handelt. Bezeichnet man den privaten Konsum (die abhängige Variable) mit y und das verfügbare Einkommen (die unabhängige Variable) mit x, so lässt sich für Messwerte i = 1,2,...,n der Variablen die Beziehung zwischen den Variablen durch die lineare Gleichung

$$\hat{y}_i = b_0 + b_1 x_i \qquad (17.1)$$

beschreiben. Dabei ist $\hat{y}_i$ (sprich y_i Dach) der durch die Gleichung für gegebene x_i vorhersagbare Wert für y_i und wird *Schätzwert* bzw. *Vorhersagewert* von y_i genannt. Dieser ist vom Beobachtungswert y_i zu unterscheiden. Nur für den Fall, dass ein Punkt des Streudiagramms auf der Regressionsgeraden liegt, haben $\hat{y}_i$ und y_i den gleichen Wert. Die Abweichung $e_i = (y_i - \hat{y}_i)$ wird *Residualwert* genannt. Die Koeffizienten b_0 und b_1 heißen *Regressionskoeffizienten*.

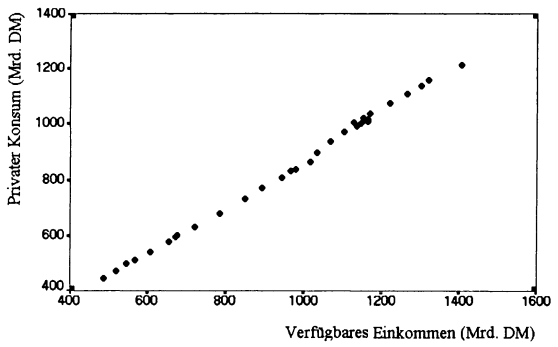


Abb. 17.1. Privater Konsum in Abhängigkeit vom verfügbaren Einkommen

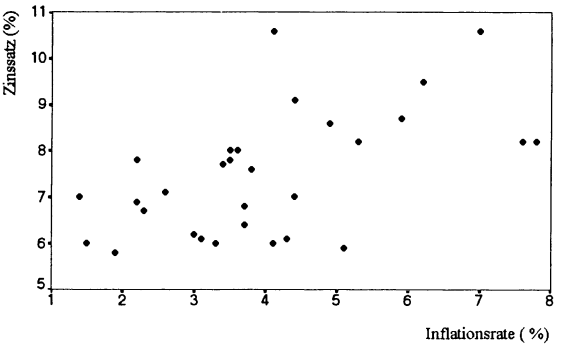


Abb. 17.2. Zinssatz in Abhängigkeit von der Inflationsrate, 1961-1990

Auch aus Abb. 17.2 wird eine lineare Beziehung zwischen zwei Variablen - der Zinssatz hängt von der Höhe der Inflationsrate ab - sichtbar. Im Vergleich zur Abb. 17.1 wird aber deutlich, dass die Beziehung zwischen den Variablen nicht besonders eng ist. Die Punkte streuen viel stärker um eine in die Punktwolke legbare Regressionsgerade. Daher geht es in der Regressionsanalyse nicht nur darum, die Koeffizienten b_0 und b_1 der obigen linearen Gleichung numerisch zu bestimmen, sondern auch darum, mit Hilfe eines statistischen Maßes zu messen, wie eng die Punkte des Streudiagramms sich um die durch die Gleichung beschriebene Gerade scharen. Je enger die Punkte an der Geraden liegen, um so besser ist die gewonnene lineare Gleichung geeignet, die beobachtete Abhängigkeit der Variablen zu beschreiben bzw. vorherzusagen. Das Maß zum Ausweis dieser Eigenschaft heißt *Bestimmtheitsmaß*.

Methode der kleinsten Quadrate. Die Berechnung der Regressionskoeffizienten basiert auf der *Methode der kleinsten Quadrate*, die im folgenden für den Fall einer Einfachregression ansatzweise erläutert werden soll. In Abb. 17.3 wird nur ein Punkt aus dem Streudiagramm der Abb. 17.2 dargestellt. Die senkrechte Abweichung zwischen dem Beobachtungswert y_i und dem mit Hilfe der Regressionsglei-

chung vorhergesagten Wert $\hat{y}_i$ ist der Residualwert (englisch *error*), hier mit e_i bezeichnet.

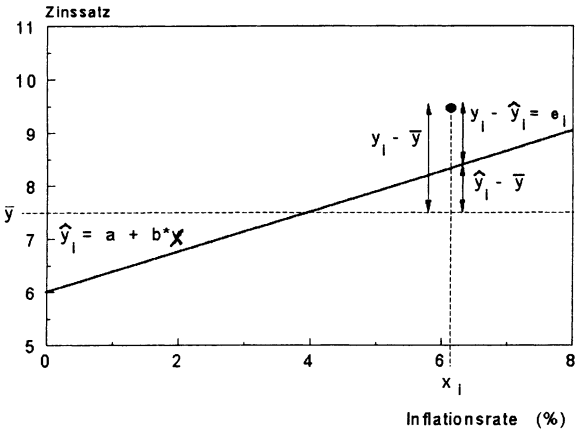


Abb. 17.3. Abweichungen der Beobachtungswerte vom mittleren Wert

Die Methode der kleinsten Quadrate bestimmt die Regressionskoeffizienten b_0 und b_1 derart, dass die Summe der quadrierten Residualwerte für alle Beobachtungen i ein Minimum annimmt:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \text{Minimum} \tag{17.2}$$

Ergebnis der Minimierung (mit Hilfe der partiellen Differentiation) sind zwei Bestimmungsgleichungen für die Koeffizienten $b_0$und b_1 (der Index i wird zur Vereinfachung im folgenden weggelassen):

$$b_0 = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2} \tag{17.3}$$

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \tag{17.4}$$

Die Summenbildung erfolgt jeweils über i = 1 bis n, wobei n die Zahl der beobachteten Wertepaare ist.

Definition des Bestimmtheitsmaßes. Die Abb. 17.3 dient auch zur Erläuterung des Bestimmtheitsmaßes. Da die Methode der kleinsten Quadrate die Eigenschaft impliziert, dass die Regressionsgerade durch den Punkt $(\bar{y},\bar{x})$, den Schnittpunkt der arithmetischen Mittel $\bar{y}$ und $\bar{x}$ verläuft, ist es zweckdienlich diesen als Ursprung eines neuen Koordinatensystems zu definieren. Vom neuen Koordinatensystem ausgehend, werden die Beobachtungswerte der beiden Variablen als Abweichung von ihren arithmetischen Mitteln gemessen. Die Abweichung $(y-\bar{y})$ wird in Abb. 17.3 durch die Regressionsgerade in $(y-\hat{y})$ und $(\hat{y}-\bar{y})$ zerlegt. Da mittels der Regressionsgleichung die Variation der abhängigen

Variable y statistisch durch die Variation der unabhängigen Variable x vorhergesagt bzw. statistisch „erklärt" werden soll, kann die Abweichung $(y-\bar{y})$ als (durch die Abweichung $x-\bar{x}$) zu erklärende Abweichung interpretiert werden. Diese teilt sich in die nicht erklärte $(y-\hat{y})$ (= Residualwert e) und die erklärte Abweichung $(\hat{y}-\bar{y})$ auf. Es gilt also für jedes beobachtete Wertepaar i:

zu erklärende Abweichung = nicht erklärte Abweichung + erklärte Abweichung

$$(y-\bar{y}) \quad = \quad (y-\hat{y}) \quad + \quad (\hat{y}-\bar{y}) \qquad (17.5)$$

Weitere durch die Methode der kleinsten Quadrate für lineare Regressionsgleichungen bedingte Eigenschaften sind der Grund dafür, dass nach einer Quadrierung der Gleichung und Summierung über alle Beobachtungswerte i auch folgende Gleichung gilt:

$$\sum(y-\bar{y})^2 = \sum(y-\hat{y})^2 + \sum(\hat{y}-\bar{y})^2 \qquad (17.6)$$

Damit erhält man eine Zerlegung der zu erklärenden Gesamtabweichungs-Quadratsumme in die nicht erklärte Abweichungs-Quadratsumme und die (durch die Regressionsgleichung) erklärte Abweichungs-Quadratsumme.

Das Bestimmtheitsmaß R^2 ist definiert als der Anteil der (durch die Variation der unabhängigen Variable) erklärten Variation an der gesamten Variation der abhängigen Variable:

$$R^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} \qquad (17.7)$$

Unter Verwendung von (17.6) gilt auch:

$$R^2 = \frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} \qquad (17.8)$$

Anhand dieser Gleichungen lassen sich die Grenzwerte für R^2 aufzeigen. R^2 wird maximal gleich 1, wenn $\sum(y-\hat{y})^2 = 0$ ist. Dieses ist gegeben, wenn für jedes Beobachtungspaar i $y = \hat{y}$ ist, d.h. dass alle Beobachtungspunkte des Streudiagramms auf der Regressionsgeraden liegen und damit alle Residualwerte gleich 0 sind. R^2 nimmt den kleinsten Wert 0 an, wenn $\sum(\hat{y}-\bar{y})^2 = 0$ bzw. gemäß (17.8) $\sum(y-\bar{y})^2 = \sum(y-\hat{y})^2$ ist. Diese Bedingung beinhaltet, dass die nicht erklärte Variation der gesamten zu erklärenden Variation entspricht, d.h. die Regressionsgleichung erklärt gar nichts. Damit ist als Ergebnis festzuhalten:

$$0 \le R^2 \le 1 \qquad (17.9)$$

Für den Fall nur einer erklärenden Variablen x gilt $R^2 = r_{yx}^2$. Im Falle mehrerer erklärender Variable gilt auch $R^2 = r_{y\hat{y}}^2$.

17.1.2 Regression als stochastisches Modell

Modellannahmen. In der Regel hat die lineare Regressionsanalyse ein anspruchsvolleres Ziel als die reine deskriptive Beschreibung von Zusammenhängen zwischen Variablen mittels einer linearen Gleichung. In der Regel interessiert man sich für den Zusammenhang zwischen der abhängigen und den unabhängigen Variablen im allgemeineren Sinne. Die per Regressionsanalyse untersuchten Daten werden als eine Zufallsstichprobe aus einer realen bzw. bei manchen Anwendungsfällen hypothetischen Grundgesamtheit aufgefasst. Die Grundlagen des *stichprobentheoretischen* bzw. *stochastischen Modells* der linearen Regressionsanalyse sollen nun etwas genauer betrachtet werden. Anschließend wird im nächsten Abschnitt anhand eines Anwendungsbeispiels aus der Praxis ausführlich auf Einzelheiten eingegangen.

Für die Grundgesamtheit wird postuliert, dass ein linearer Zusammenhang zwischen abhängiger und unabhängiger Variable besteht und dieser additiv von einer Zufallsvariable überlagert wird. So wird beispielsweise als Ergebnis theoretischer Analyse postuliert, dass der makroökonomische Konsum der Haushalte im wesentlichen linear vom verfügbaren Einkommen und vom Zinssatz abhängig ist. Daneben gibt es eine Vielzahl weiterer Einflussgrößen auf den Konsum, die aber jeweils nur geringfügig konsumerhöhend bzw. konsummindernd wirken und in der Summe ihrer Wirkung als zufällige Variable interpretiert werden können. Bezeichnet man den Konsum mit y_i, das verfügbare Einkommen mit $x_{1,i}$ und den Zinssatz mit $x_{2,i}$ sowie die Zufallsvariable mit ε_i, so lässt sich das theoretische Regressionsmodell für die Grundgesamtheit wie folgt formulieren:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i \qquad (17.10)$$

Die Variable y_i setzt sich somit aus einer systematischen Komponente $\hat{y}_i (= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i})$ und einer zufälligen (stochastischen) *Fehlervariable* ε_i zusammen.

Die abhängige Variable y wird durch die additive Überlagerung der systematischen Komponente mit der Zufallsvariable ε ebenfalls zu einer zufälligen Variablen, im Gegensatz zu den erklärenden Variablen x_1 und x_2, die als nicht-stochastische Größen interpretiert werden müssen.

Der Regressionskoeffizient β_1 gibt für die Grundgesamtheit an, um wieviel der Konsum steigt, wenn bei Konstanz des Zinssatzes das verfügbare Einkommen um eine Einheit steigt. Daher bezeichnet man ihn auch als partiellen Regressionskoeffizienten. Analog gibt β_2 an, um wieviel der Konsum sinkt bei Erhöhung des Zinssatzes um eine Einheit und Konstanz des verfügbaren Einkommens.

Damit die Methode der kleinsten Quadrate zu bestimmten gewünschten Schätzeigenschaften (beste lineare unverzerrte Schätzwerte, engl. BLUE) führt sowie Signifikanzprüfungen für die Regressionskoeffizienten durchgeführt werden können, werden für die Zufallsfehlervariable ε_i folgende Eigenschaften ihrer Verteilung vorausgesetzt:

❒ $E(\varepsilon_i) = 0$ für $i = 1,2,3,\cdots$ (17.11)

Der (bedingte) Erwartungswert (E), d. h. der Mittelwert der Verteilung von ε ist für jede Beobachtung der nicht-stochastischen Werte x_i gleich 0.

- $E(\varepsilon_i^2) = \sigma_\varepsilon^2$ = konstant für i = 1,2,3,... (17.12)

  Die Varianz der Verteilung der Zufallsvariable σ_ε^2 ist für jede Beobachtung der nicht-stochastischen Werte x_i konstant. Sie ist damit von der Höhe der erklärenden Variablen unabhängig. Ist diese Bedingung erfüllt, so besteht *Homoskedastizität* der Fehlervariable. Ist die Bedingung nicht erfüllt, so spricht man von *Heteroskedastizität.*

- $E(\varepsilon_i, \varepsilon_j) = 0$ für i = 1,2,... und j = 1,2,3, ... für $i \neq j$ (17.13)

  Die Kovarianz der Zufallsvariable ist für verschiedene Beobachtungen i und j gleich 0, d.h. die Verteilungen der Zufallsvariable für i und für j sind unabhängig voneinander. Ist die Bedingung nicht erfüllt, so besteht *Autokorrelation* der Fehlervariable ε: ε_i und ε_j korrelieren.

- ε_i ist für gegebene Beobachtungen i = 1,2,3... normalverteilt. Diese Voraussetzung ist nur dann erforderlich, wenn Signifikanzprüfungen der Regressionskoeffizienten durchgeführt werden sollen. (17.14)

In Abb. 17.4 a werden die Annahmen des klassischen linearen Regressionsmodells für den Fall einer erklärenden Variablen x veranschaulicht. Die Verteilung der Fehlervariable ε_i ist für alle Werte der unabhängigen Variable x_i unabhängig normalverteilt und hat den Mittelwert 0. In Abb. 17.4 b wird im Vergleich zur Abb. 17.4 a sichtbar, dass die Varianz der Fehlervariable ε_i mit zunehmenden Wert der erklärenden Variable x_i größer wird und damit Heteroskedastizität vorliegt.

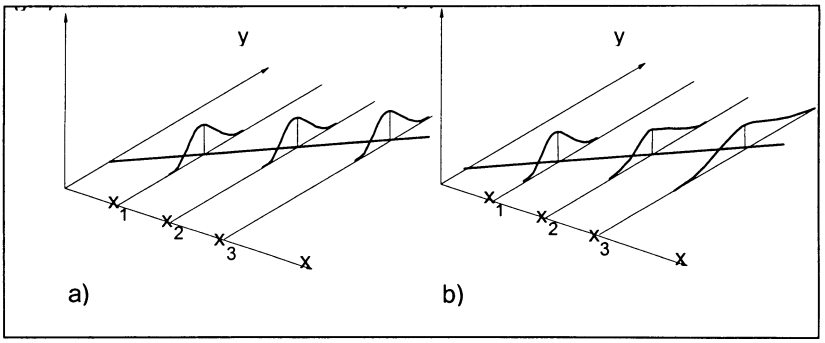


Abb. 17.4. Das lineare Regressionsmodell mit einer abhängigen Variable: a) Homoskedastizität, b) Heteroskedastizität.

Das stichprobentheoretisch fundierte Modell der linearen Regressionsanalyse geht davon aus, dass für gegebene feste Werte der unabhängigen Variable x_i der Wert der abhängigen Variable y_i zufällig ausgewählt wird. Bei den empirischen Beobachtungswerten der Variable y_i handelt es sich also um Realisationen einer Zufallsvariablen. Wird eine Stichprobe gezogen, so sind die für diese Stichprobe ermittelten Regressionskoeffizienten Schätzwerte für die unbekannten Regressi-

onskoeffizienten der Grundgesamtheit. Zur Unterscheidung werden sie mit b bezeichnet (⇨ Gleichung 17.1). Unter der Vorstellung wiederholter Stichprobenziehungen haben die Regressionskoeffizienten b eine normalverteilte Wahrscheinlichkeitsverteilung. Sie wird *Stichprobenverteilung der Regressionskoeffizienten* genannt. Im folgenden werden sowohl die Regressionskoeffizienten der Stichprobenverteilung als auch eine konkrete Realisierung dieser in einer bestimmten Stichprobe mit den gleichen Symbolen bezeichnet, da sich die Bedeutung aus dem Kontext ergibt. Für die Schätzwerte der Regressionskoeffizienten gilt folgendes (hier nur dargestellt für die Koeffizienten b_1 und b_2 in Gleichung 17.1):

❒ $E(b_k) = \beta$ für $k = 1,2$ (17.15)

Der Mittelwert [Erwartungswert (E)] der Stichprobenverteilung von b entspricht dem Regressionskoeffizienten der Grundgesamtheit. Es handelt sich um erwartungstreue, unverzerrte Schätzwerte.

❒ $$\sigma^2_{b_1} = \frac{1}{\sum(x_1 - \bar{x}_1)^2 (1 - R^2_{x1,x2})} \sigma^2_\varepsilon \qquad (17.16)$$

$$\sigma^2_{b_2} = \frac{1}{\sum(x_2 - \bar{x}_2)^2 (1 - R^2_{x2,x1})} \sigma^2_\varepsilon \qquad (17.17)$$

Die Varianz $\sigma^2_{b_i}$ der normalverteilten Stichprobenverteilung von b_i ist von der Variation der jeweiligen Erklärungsvariable x_j ($j = 1, 2$), der Varianz der Fehlervariable ε (σ^2_ε) sowie der Stärke des linearen Zusammenhangs zwischen den beiden erklärenden Variablen (gemessen in Form der Bestimmtheitsmaße $R^2_{x1,x2}$ bzw. $R^2_{x2,x1}$) abhängig. Aus den Gleichungen (17.16) sowie (17.17) kann man entnehmen, dass mit wachsendem Bestimmtheitsmaß - also wachsender Korrelation zwischen den erklärenden Variablen - die Standardabweichung des Regressionskoeffizienten zunimmt. Korrelation der erklärenden Variablen untereinander führt also zu unsicheren Schätzergebnissen. Die Höhe der Regressionskoeffizienten variiert dann stark von Stichprobe zu Stichprobe. Im Grenzfall eines Bestimmtheitsmaßes in Höhe von 1 wird die Standardabweichung unendlich groß. Die Koeffizienten können mathematisch nicht mehr bestimmt werden. Praktisch heißt das aber, dass die Variablen austauschbar sind und damit sowohl die eine als auch die andere alleine gleich gut zur Erklärung der unabhängigen Variable geeignet ist. Für den Fall nur einer bzw. mehr als zwei erklärenden Variablen sind die Gleichungen (17.16) bzw. (17.17) sinngemäß anzuwenden: Bei nur einer erklärenden Variable entfallen in den Formeln die Bestimmtheitsmaße $R^2_{x1,x2}$ bzw. $R^2_{x2,x1}$. Bei mehr als zwei erklärenden Variablen erfassen die Bestimmtheitsmaße den linearen Erklärungsanteil aller weiteren erklärenden Variablen. Der Sachverhalt einer hohen Korrelation zwischen den erklärenden Variablen wird mit *Multikollinearität* bezeichnet (⇨ Kap. 17.4.4).

❒ Da die Varianz der Fehlervariable σ^2_ε unbekannt ist, wird sie aus den vorliegenden Daten - interpretiert als Stichprobe aus der Grundgesamtheit - geschätzt. Ein unverzerrter Schätzwert für die Varianz ist

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum(y-\hat{y})^2}{n-m-1} = \frac{\sum e^2}{n-m-1} \tag{17.18}$$

Dabei ist $\sum e^2$ die Summe der quadrierten Residualwerte, n der Stichprobenumfang, d.h. die Anzahl der Beobachtungen i in den vorliegenden Daten und m die Anzahl der erklärenden Variablen. Die Differenz $n-m-1$ wird Anzahl der Freiheitsgrade (df) genannt, weil bei n Beobachtungen für die Variablen durch die Schätzung von m+1 Koeffizienten (einschließlich des konstanten Gliedes) $n-m-1$ Werte nicht vorherbestimmt sind. In unserem Beispiel zur Erklärung des Konsums durch das verfügbare Einkommen und den Zinssatz beträgt df = 28, da n = 31 und m = 2 ist. Wird $\hat{\sigma}_\varepsilon^2$ in Gleichung 17.16 bzw. 17.17 für σ_ε^2 eingesetzt, so erhält man Schätzwerte für die Varianzen der Regressionskoeffizienten: $\hat{\sigma}_{b_1}^2$ sowie $\hat{\sigma}_{b_2}^2$.

Die Wurzel aus $\hat{\sigma}_\varepsilon^2$ wird Standardfehler der Schätzung (standard error) oder auch Standardabweichung des Residualwertes genannt.

Testen von Regressionskoeffizienten. Die in der Praxis vorherrschende Anwendungsform des Testens von Regressionskoeffizienten bezieht sich auf die Frage, ob für die Grundgesamtheit der Variablen ein (linearer) Regressionszusammenhang angenommen werden darf oder nicht.

Ausgehend vom in Gleichung (17.10) formulierten Beispiel wäre zu prüfen, ob die Regressionskoeffizienten der Grundgesamtheit β_1 gleich 0 (kein linearer Zusammenhang) oder positiv (positiver linearer Zusammenhang) und β_2 gleich 0 (kein linearer Zusammenhang) oder negativ (negativer linearer Zusammenhang) sind. Die Hypothese, dass kein Zusammenhang besteht, wird als H_0-Hypothese und die Alternativhypothese als H_1-Hypothese bezeichnet (⇨ Kap. 13.3). In formaler Darstellung:

$$H_0: \beta_1 = 0 \qquad H_1: \beta_1 > 0 \tag{17.19}$$

$$H_0: \beta_2 = 0 \qquad H_1: \beta_2 < 0 \tag{17.20}$$

Besteht über das Vorzeichen des Regressionskoeffizienten keinerlei Erwartung, so lautet die Alternativhypothese $H_1: \beta \neq 0$. In diesem Fall spricht man von einem zweiseitigen Test im Vergleich zu obigen einseitigen Tests.

Ausgangspunkt des Testverfahrens ist die Stichprobenverteilung von b. Unter der Voraussetzung, dass die Annahmen (17.11) bis (17.14) zutreffen, unterliegt der standardisierte Stichproben-Regressionskoeffizient b (bei Verwendung des Schätzwertes der Standardabweichung)

$$t = \frac{b-\beta}{\hat{\sigma}_b} \tag{17.21}$$

einer t-Verteilung (auch *Student-Verteilung* genannt) mit $n-m-1$ Freiheitsgraden (df). Dabei ist n der Stichprobenumfang und m die Anzahl der erklärenden Variablen. Unter der Hypothese H_0 ($\beta = 0$) ist die Variable

$$t = \frac{b}{\hat{\sigma}_b} \tag{17.22}$$

die t-verteilte Prüfverteilung mit d.f. = $n-m-1$. Bei Vorgabe einer Irrtumswahrscheinlichkeit α (z.B. $\alpha = 0{,}05$) und der Anzahl der $df = n-m-1$ kann aus einer tabellierten t-Verteilung ein kritischer Wert für t (t_{krit}) entnommen werden, der den Annahmebereich und den Ablehnungsbereich für die Hypothese H_0 trennt (⇨ Kap. 13.3). Aus der vorliegenden Stichprobe ergibt sich mit dem Regressionskoeffizienten b_{emp} sowie der Standardabweichung $\hat{\sigma}_b$ ein empirischer Prüfverteilungswert

$$t_{emp} = \frac{b_{emp}}{\hat{\sigma}_b} \tag{17.23}$$

Je nachdem ob t_{emp} in den Ablehnungsbereich für H_0 ($t_{emp} > t_{krit}$) oder Annahmebereich für H_0 ($t_{emp} < t_{krit}$) fällt, wird entschieden, ob der Regressionskoeffizient mit der vorgegebenen Irrtumswahrscheinlichkeit α signifikant von 0 verschieden ist oder nicht.

Vorhersagewerte und ihre Standardabweichung. Für bestimmte Werte der beiden erklärenden Variablen (z.B. $x_{1,0}$ und $x_{2,0}$) lassen sich Vorhersagewerte aus der Schätzgleichung gemäß Gleichung 17.24 bestimmen.

$$\hat{y}_0 = b_0 + b_1 x_{1,0} + b_2 x_{2,0} \tag{17.24}$$

Bei dieser sogenannten Punktschätzung ist der Schätzwert sowohl für den durchschnittlichen Wert als auch für einen individuellen Wert von y bei $x_{1,0}$ und $x_{2,0}$ identisch (zur Erinnerung: für jeweils gegebene Werte der erklärenden Variablen hat y eine Verteilung mit dem Mittelwert $\hat{y}$). Anders sieht es aber bei einer Intervallschätzung analog der Schätzung von Konfidenzintervallen aus. Der Grund liegt darin, dass in diesen beiden Fällen die Varianzen bzw. Standardabweichungen des Schätzwertes verschieden sind. Aus Vereinfachungsgründen wird der Sachverhalt im folgenden für den Fall nur einer erklärenden Variablen x erläutert. Die Varianz des durchschnittlichen Schätzwertes $\hat{y}$ für x_0 ergibt sich gemäß folgender Gleichung:

$$\sigma_{\hat{y}}^2 = \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - x)^2}\right]\sigma_\varepsilon^2 \tag{17.25}$$

Aus (17.25) ist ersichtlich, dass bei gegebenem Stichprobenumfang n, gegebener Variation der Variablen x sowie gegebenem σ_ε^2 die Varianz $\sigma_{\hat{y}}^2$ mit zunehmender Abweichung des Wertes x_0 von $\bar{x}$ größer wird.

Der Schätzwert $\hat{\sigma}_{\hat{y}}^2$ ergibt sich durch Einsetzen von $\hat{\sigma}_\varepsilon^2$ gemäß Gleichung 17.18 in 17.25.

Die Varianz eines individuellen Wertes von y für x_0 ist größer, da die Varianz von y, die annahmegemäß die der Zufallsvariable ε entspricht, hinzukommt. Addiert man σ_ε^2 zu $\sigma_{\hat{y}}^2$ hinzu, so ergibt sich nach Ausklammern

$$\sigma_{\hat{y}}^2 = \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - x)^2}\right] \sigma_\varepsilon^2 \qquad (17.26)$$

Analog zur Berechnung von Konfidenzintervallen für die Parameter β der Grundgesamtheit, lassen sich Intervallschätzungen sowohl für den Mittelwert $\hat{y}$ als auch für individuelle Werte $\hat{y}_{ind}$ bestimmen (⇨ Gleichung 17.37). Dabei wird auch hier für σ_ε^2 der Schätzwert gemäß Gleichung 17.18 eingesetzt.

17.2 Praktische Anwendung

17.2.1 Berechnen einer Regressionsgleichung und Ergebnisinterpretation

Regressionsgleichung berechnen. Im folgenden sollen alle weiteren Erläuterungen zur Regressionsanalyse praxisorientiert am Beispiel der Erklärung des Konsums (CPR) durch andere makroökonomische Variablen vermittelt werden (Datei MAKRO.SAV, ⇨ Anhang B).

In einem ersten Schätzansatz soll CPR gemäß der Regressionsgleichung (17.10) durch YVERF (verfügbares Einkommen) und ZINS (Zinssatz) erklärt werden. Dabei sollen die Hypothesen über die Vorzeichen der Regressionskoeffizienten gemäß (17.19) und (17.20) geprüft werden. In einem zweiten Schätzansatz soll zusätzlich die Variable LQ (Lohnquote) in das Regressionsmodell eingeschlossen werden. Die Hypothese lautet, dass mit höherem Anteil der Löhne und Gehälter am Volkseinkommen der Konsum zunimmt, weil man erwarten darf, dass die durchschnittliche Konsumquote aus Löhnen und Gehältern höher ist als aus den Einkommen aus Unternehmertätigkeit und Vermögen.

Zur Durchführung der Regressionsanalysen gehen Sie wie folgt vor:

- ▷ Klicken Sie die Befehlsfolge "Analysieren", „Regression ▷ „„ „Linear..."
  Es öffnet sich die in Abb. 17.5 dargestellte Dialogbox „Lineare Regression".
- ▷ Wählen Sie aus der Quellvariablenliste die abhängige (zu erklärende) Variable CPR aus und übertragen Sie diese in das Eingabefeld „Abhängige Variable".
- ▷ Wählen Sie aus der Quellvariablenliste die erklärenden (unabhängigen) Variablen (YVERF und ZINS) aus und übertragen diese für „Block 1 von 1" (für die erste Schätzgleichung) in das Eingabefeld „Unabhängige Variable(n):". Falls keine weiteren Schätzgleichungen mit anderen erklärenden Variablen bzw. anderen Verfahren („Methode") berechnet werden sollen, kann man die Berechnung mit „OK" starten.
- ▷ Zur gleichzeitigen Berechnung der zweiten Schätzgleichung wählen Sie mit „Weiter" „Block 2 von 2" und übertragen aus der Variablenliste die gewünschte zusätzliche Erklärungsvariable LQ. Weitere Schätzansätze könnten mit weiteren

„Blöcken" angefordert werden. Mit „Zurück" kann man zu vorherigen „Blöcken" (Schätzansätzen) schalten und mit „Weiter" wieder zu nachfolgenden.

▷ Je nach Bedarf können Sie andere bzw. weitere optionale Einstellungen auswählen: Aus dem Auswahlfeld „Methode:" können andere Verfahren zum Einschluß der unabhängigen Variablen in die Regressionsgleichung gewählt werden (hier: für beide Blöcke „Einschluss"). Die Methode „Einschluss" ist die Standardeinstellung und bedeutet, dass alle gewählten unabhängigen Variablen in einem Schritt in die Regressionsgleichung eingeschlossen werden.

„Fallbeschriftungen" ermöglicht es, eine Variable zur Fallidentifizierung einzutragen. Für die mit der Schaltfläche „Diagramme" in Abb. 17.5 anforderbaren Streudiagrammen können im Diagramm-Editorfenster mit Hilfe des Symbolschalters [Fallbeschriftungen] einzelne Fälle identifiziert werden. Bei Verwendung einer Fallbeschriftungsvariable dient ihr Variablenwert zur Identifizierung eines Falles, ansonsten die Fallnummer (⇨ Kap. 21.4.2).

Mittels der Schaltflächen „WLS>>„ „Statistiken...", „Diagramme...", „Speichern...", „Optionen..." können per Dialogbox Untermenüs aufgerufen werden, die weitere ergänzende Berechnungen, Einstellungen etc. ermöglichen. Unten wird darauf ausführlich eingegangen. Mit „OK" wird die Berechnung gestartet.

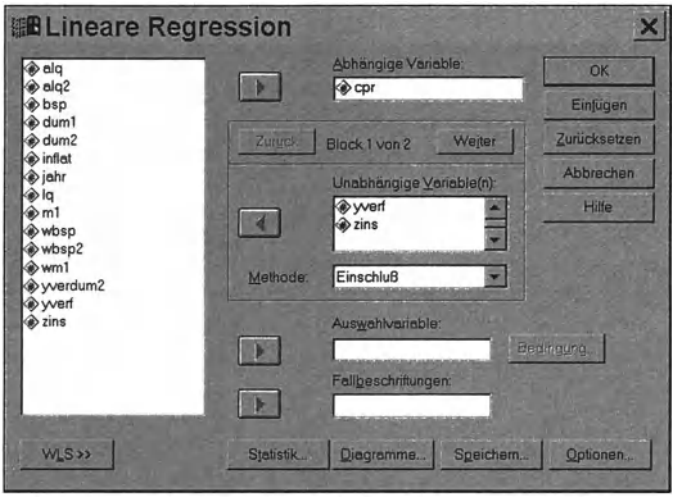


Abb. 17.5. Dialogbox „Lineare Regression"

Die folgenden Ergebnisausgaben beruhen auf den Einstellungen in Abb. 17.5 (ohne Aufruf von „WLS>>", „Statistiken...", „Diagramme...", „Speichern...", "Optionen...").

Regressionskoeffizienten. In Tabelle 17.1 werden die Regressionskoeffizienten der beiden Regressionsmodelle sowie Angaben zu Signifikanzprüfungen der Koeffizienten aufgeführt. Modell 1 enthält als erklärende Variablen YVERF und ZINS (siehe Block 1 in Abb. 17.5) und Modell 2 enthält zusätzlich die Variable LQ

(Block 2). In der Spalte „B" werden die (nicht standardisierten) Regressionskoeffizienten für die in der ersten Spalte genannten Variablen aufgeführt. Demnach lautet die Schätzgleichung für das erste Modell

$$\hat{CPR}_i = 51{,}767 + 0{,}862 * YVERF_i - 5{,}313 * ZINS_i \qquad (17.27)$$

Diese Schätzgleichung erlaubt es, bei vorgegebenen Werten für die beiden erklärenden Variablen, den Schätzwert der zu erklärenden Variablen (Vorhersagewert) zu berechnen. Die Vorzeichen der erklärenden Variablen entsprechen der Erwartung. In der Spalte „Standardfehler" werden die Schätzwerte für die Standardabweichungen der Regressionskoeffizienten [vergl. Gleichungen (17.16) und (17.17) in Verbindung mit (17.18)] aufgeführt. In der Spalte „T" sind die empirischen t-Werte gemäß Gleichung (17.23) als Quotient aus den Werten in Spalte „B" und Spalte „Standardfehler" aufgeführt. Geht man für den Signifikanztest der Regressionskoeffizienten (bei der hier einseitigen Fragestellung) von einer Irrtumswahrscheinlichkeit in Höhe von 5 % aus ($\alpha = 0{,}05$), so lässt sich für df (Freiheitsgrade) = $n - m - 1 = 28$ aus einer tabellierten t-Verteilung ein $t_{krit} = 1{,}7011$ entnehmen. Wegen $t_{emp} > t_{krit}$ (absolute Werte) sind bei einer Irrtumswahrscheinlichkeit von 5 % alle Regressionskoeffizienten signifikant von 0 verschieden. In der Spalte „Signifikanz" wird diese Information auf andere Weise von SPSS bereitgestellt, so dass sich das Entnehmen von t_{krit} aus Tabellen für die t-Verteilung erübrigt. „Signifikanz" ist die Wahrscheinlichkeit, bei Ablehnung von H_0 (keine Abhängigkeit), eine irrtümliche Entscheidung zu treffen. Da für alle Regressionskoeffizienten die „Signifikanz" kleiner als die vorgebene Irrtumswahrscheinlichkeit in Höhe von $\alpha = 0{,}05$ sind, ergibt sich auch so, dass die Koeffizienten signifikant sind.

Beta-Koeffizienten. In der Spalte „Beta" (Tabelle 17.1) werden die sogenannten Beta-Koeffizienten (bzw. standardisierte Koeffizienten) für die beiden Erklärungsvariablen aufgeführt. Beta-Koeffizienten sind die Regressionskoeffizienten, die sich ergeben würden, wenn vor der Anwendung der Regressionsanalyse alle Variablen standardisiert worden wären. Bezeichnet man mit $\bar{x}$ das arithmetische Mittel und mit s die Standardabweichung einer Variablen x, so wird

$$z = \frac{x - \bar{x}}{s} \qquad (17.28)$$

die standardisierte Variable genannt. Mit der Standardisierung werden die Abweichungen der Messwerte der Variablen von ihrem Mittelwert in Standardabweichungen ausgedrückt. Sie sind dann dimensionslos. Der Mittelwert einer standardisierten Variable beträgt 0 und die Standardabweichung 1. Im Unterschied zu den Regressionskoeffizienten sind die Beta-Koeffizienten deshalb von der Dimension der erklärenden Variablen unabhängig und daher miteinander vergleichbar. Es zeigt sich, dass der Beta-Koeffizient für das verfügbare Einkommen den für den Zinssatz bei weitem übersteigt. Damit wird sichtbar, dass das verfügbare Einkommen als bedeutsamste Variable den weitaus größten Erklärungsbeitrag liefert. Aus den Regressionskoeffizienten für die beiden Variablen ist dieses nicht erkennbar. Aufgrund der Größenverhältnisse der Regressionskoeffizienten könnte man eher

das Gegenteil vermuten. Allerdings darf bei dieser vergleichenden Beurteilung der relativen Bedeutung der Variablen zur statistischen Erklärung nicht übersehen werden, dass auch die Beta-Koeffizienten durch Multikollinearität nicht unabhängig voneinander und insofern in ihrer Aussagekraft eingeschränkt sind.

Um Beta-Koeffizienten zu berechnen, müssen die Variablen tatsächlich vor der Regressionsanalyse nicht standardisiert werden. Sie können für eine erklärende Variable wie folgt berechnet werden:

$$beta_k = b_k \frac{s_k}{s_y} \tag{17.29}$$

wobei b_k der Regressionskoeffizient, s_k die Standardabweichung der erklärenden Variable x_k und s_y die Standardabweichung der zu erklärenden Variable y bedeuten.

Tabelle 17.1. Ergebnisausgabe: Regressionskoeffizienten der multiplen Regression

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	Kollinearitätsstatistik	
		B	Standardfehler	Beta			Toleranz	VIF
1	(Konstante)	51,767	10,561		4,902	,000		
	YVERF	,862	,007	1,007	127,646	,000	,928	1,08
	ZINS	-5,313	1,355	-,031	-3,920	,001	,928	1,08
2	(Konstante)	-18,077	45,006		-,402	,691		
	YVERF	,844	,013	,986	64,897	,000	,237	4,22
	ZINS	-7,031	1,704	-,041	-4,127	,000	,557	1,80
	Lohnquote (%)	1,424	,893	,028	1,594	,123	,176	5,67

a. Abhängige Variable: CPR

Bestimmtheitsmaß. Die in Tabelle 17.2 dargestellte Ergebnisausgabe gehört zur Standardausgabe einer Regressionsschätzung (entspricht der Wahl von „Anpassungsgüte des Modells" in der Dialogbox „Statistiken"). In der Spalte „R-Quadrat" wird für beide Modelle das Bestimmtheitsmaß R^2 angegeben. Mit 0,998 ist der Wert nahezu 1, so dass fast die gesamte Variation von CPR durch die Variation von YVERF und ZINS erklärt wird. Man spricht von einem guten „Fit" der Gleichung. „R" ist die Wurzel aus R^2 und hat somit keinen weiteren Informationsgehalt. „Korrigiertes R-Quadrat" ist ein Bestimmtheitsmaß, das die Anzahl der erklärenden Variablen sowie die Anzahl der Beobachtungen berücksichtigt. Aus der Definitionsgleichung für R^2 in der Form (17.8) wird deutlich, dass mit zunehmender Anzahl der erklärenden Variablen bei gegebenem $\sum(y-\bar{y})^2$ der Ausdruck $\sum(y-\hat{y})^2$ kleiner und somit R^2 größer wird. Daher ist z.B. ein $R^2 = 0{,}90$ bei zwei erklärenden Variablen anders einzuschätzen als bei zehn. Des weiteren ist ein Wert

für R^2 basierend auf z.B. 100 Beobachtungen positiver zu sehen als bei 20. Das korrigierte Bestimmtheitsmaß versucht dieses zu berücksichtigen. Es wird von SPSS wie folgt berechnet (m = Anzahl der erklärenden Variablen, n = Zahl der Beobachtungsfälle):

$$R^2_{korr} = R^2 - \frac{m}{n-m-1}(1-R^2) = 0{,}998 \qquad (17.30)$$

Das korrigierte R^2 ist kleiner als R^2 (hier wegen nur drei Stellen nach dem Komma nicht sichtbar) und stellt für vergleichende Beurteilungen von Regressionsgleichungen mit unterschiedlicher Anzahl von Erklärungsvariablen bzw. Beobachtungswerten ein besseres Maß für die Güte der Vorhersagequalität der Regressionsgleichung dar.

„Standardfehler des Schätzers" ist der Schätzfehler der Regressionsgleichung und entspricht dem Schätzwert der Standardabweichung von ε_i gemäß Gleichung (17.18):

$$\sqrt{\frac{\sum(y-\hat{y})^2}{n-m-1}} = \sqrt{\frac{\sum e^2}{n-m-1}} = 9{,}518\,.$$

Er ist auch ein Maß für die Güte der Vorhersagequalität der Gleichung. Er ist im Unterschied zum korrigierten R^2 aber abhängig von der Maßeinheit der abhängigen Variablen.

Tabelle 17.2. Ergebnisausgabe: Bestimmtheitsmaß der multiplen Regression

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,999[a]	,998	,998	9,518
2	,999[b]	,999	,998	9,266

a. Einflußvariablen : (Konstante), ZINS, YVERF

b. Einflußvariablen : (Konstante), ZINS, YVERF, Lohnquote (%)

Varianzzerlegung und F-Test. In Tabelle 17.3 werden (hier nur für Modell 1) unter der Überschrift „ANOVA" Informationen zur varianzanalytischen Prüfung der Regressionserklärung mit Hilfe eines F-Testes bereitgestellt. Auch bei dieser Ergebnisausgabe handelt es sich um eine Standardausgabe einer Regressionsschätzung (entspricht der Wahl von „Anpassungsgüte des Modells" in der Dialogbox „Statistiken"). Es wird gemäß Gleichung (17.6) in der Spalte „Quadratsumme" die Zerlegung der Gesamt-Variation der zu erklärenden Variable $\sum(y-\bar{y})^2 = 1568974{,}2$ („Gesamt") in die durch die Regressionsgleichung erklärten $\sum(\hat{y}-\bar{y})^2 = 1566437{,}632$ („Regression") und nicht erklärten $\sum(y-\hat{y})^2 = 2536{,}568$ („Residuen") Variation angeführt. Durch Division der Werte der Spalte „Quadratsumme" durch die der Spalte „df" (= Anzahl der Freiheitsgrade) entstehen die Werte in der Spalte „Mittel der Quadrate", die durch-

schnittlichen quadrierten Abweichungen. Die Freiheitsgrade für das „Mittel der Quadrate" von „Regression" beträgt m = 2 und von „Residuen" n-m-1 = 28 (m= Anzahl der erklärenden Variablen und n = Anzahl der Fälle). Der Quotient aus der durchschnittlichen erklärten Variation (Varianz) und durchschnittlichen nicht erklärten Variation:

$$F_{emp} = \frac{\sum(\hat{y} - \bar{y})^2 / m}{\sum(y - \hat{y})^2 / (n - m - 1)} = 8645{,}589 \qquad (17.31)$$

folgt einer F-Verteilung mit $df_1 = m$ und $df_2 = n - m - 1$ Freiheitsgraden. Analog dem Signifikanztest für Regressionskoeffizienten wird bei Vorgabe einer Irrtumswahrscheinlichkeit α geprüft, ob das empirisch erhaltene Streuungsverhältnis (F_{emp}) gleich oder größer ist als das gemäß einer F-Verteilung zu erwartende kritische (F_{krit}). Aus einer tabellierten F-Verteilung kann man für $\alpha = 0{,}05$ und $df_1 = 2$ und $df_2 = 28$ entnehmen: $F_{krit} = 3{,}34$. Da $F_{emp} = 8645{,}589 > F_{krit} = 3{,}34$, wird die H_0-Hypothese - die Variablen x_1 und x_2 leisten keinen Erklärungsbeitrag (formal: $\beta_1 = 0$ und $\beta_2 = 0$) - abgelehnt mit einer Irrtumswahrscheinlichkeit von 5 %. „Signifikanz" = 0,00" in Tabelle 17.3 weist (ähnlich wie bei dem t-Test) den gleichen Sachverhalt aus, da das ausgewiesene Wahrscheinlichkeitsniveau kleiner ist als die gewünschte Irrtumswahrscheinlichkeit. Im Vergleich zum t-Test wird deutlich, dass der F-Test nur allgemein prüft, ob mehrere Erklärungsvariablen gemeinsam einen regressionsanalytischen Erklärungsbeitrag leisten, so dass sich das Testen einzelner Regressionskoeffizienten auf Signifikanz nicht erübrigt. Der F-Test kann auch interpretiert werden als Signifikanzprüfung, ob R^2 gleich 0 ist.

Tabelle 17.3. Ergebnisausgabe: Zerlegung der Varianz

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	1566437,632	2	783218,8	8645,589	,000[a]
	Residuen	2536,568	28	90,592		
	Gesamt	1568974,200	30			

a. Einflußvariablen : (Konstante), ZINS, YVERF

b. Abhängige Variable

Ergebnisvergleich von Modell 1 und Modell 2. Für die zweite Regressionsgleichung, die sich durch Hinzufügen der Variable LQ auszeichnet, können folgende typische Gesichtspunkte herausgestellt werden:

- Der Regressionskoeffizient hat - wie aus makroökonomischer Sicht erwartet - ein positives Vorzeichen. Aber der Regressionskoeffizient ist nicht signifikant von 0 verschieden bei einer Irrtumswahrscheinlichkeit von 5 % („Signifikanz" = 0,123 > 0,05). (⇨ Tabelle 17.1).
- Typischerweise verändern sich mit der zusätzlichen Variable die Regressionskoeffizienten („B") und auch die Standardabweichungen („Standardfehler") und damit die „T"-Werte bzw. „Signifikanz"-Werte der anderen Variablen (⇨ Ta-

belle 17.1). Dieses liegt daran, dass die Variable LQ mit den anderen erklärenden Variablen korreliert. Die Standardabweichungen („Standardfehler") der Regressionskoeffizienten werden größer. Dieses dürfte auch nicht überraschend sein, da es nur zu plausibel ist, dass mit zunehmender Korrelation der erklärenden Variablen die einzelne Wirkung einer Variable auf die abhängige Variable nicht mehr scharf isoliert werden kann und somit unsichere Schätzungen resultieren. Auch aus Gleichung (17.16) und (17.17) wird der Sachverhalt für den Fall von zwei erklärenden Variablen sichtbar. Nur für den in der Praxis meist unrealistischen Fall keiner Korrelation zwischen den erklärenden Variablen tritt dieser Effekt nicht auf. Das andere Extrem einer sehr starken Korrelation zwischen den erklärenden Variablen - als *Multikollinearität* bezeichnet - führt zu Problemen (⇨ Kap. 17.4.4).

❒ Das korrigierte R^2 wird größer (hier wegen nur drei Kommastellen nicht sichtbar) und der Standardfehler des Schätzers wird kleiner. Das Einbeziehen der Variable LQ führt insofern zu einem leicht verbesserten „Fit" der Gleichung (⇨ Tabelle 17.2).

Schaltfläche WLS >>. Hiermit kann eine gewichtete lineare Regressionsanalyse durchgeführt werden. Nach Klicken auf „WLS >>„ wird in der Dialogbox ein Eingabefeld „WLS-Gewichtung:" geöffnet, in das eine Gewichtungsvariable eingetragen werden kann.

17.2.2 Ergänzende Statistiken zum Regressionsmodell (Schaltfläche „Statistiken")

Durch Anklicken der Schaltfläche „Statistiken..." in der Dialogbox „Lineare Regression" (⇨ Abb. 17.5) wird die in Abb. 17.6 dargestellte Dialogbox geöffnet. Man kann nun zusätzliche statistische Informationen zu der Regressionsgleichung anfordern. Zum Teil dienen diese Informationen dazu, die Modellannahmen der linearen Regression zu überprüfen.

Voreingestellt sind „Schätzer" und „Anpassungsgüte des Modells" mit denen standardmäßig die Regressionskoeffizienten sowie Angaben zur Schätzgüte gemäß Tabelle 17.2 und 17.3 ausgegeben werden. Durch Anklicken weiterer Kontrollkästchen werden ergänzende Berechnungen ausgeführt. Nach Klicken von „Weiter" kommt man wieder auf die höhere Dialogboxebene zurück und kann die Berechnungen mit „OK" starten. Im folgenden werden alle Optionen anhand des ersten Modells unter Verwendung des Regressionsverfahrens „Einschluss" aufgezeigt.

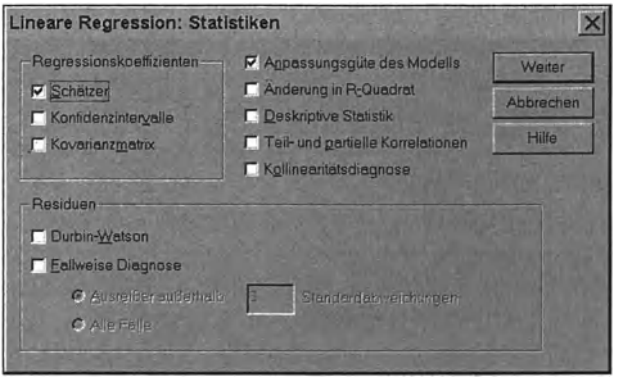


Abb. 17.6. Dialogbox „Lineare Regression: Statistiken"

Konfidenzintervalle. Unter der Vorgabe, dass das Regressionsmodell den Modellvoraussetzungen entspricht (vergl. Gleichungen 17.11 bis 17.14), können für die unbekannten Regressionskoeffizienten der Grundgesamtheit Konfidenzintervalle (auch Mutungs- oder Erwartungsbereiche genannt) bestimmt werden. Da die Schätzwerte der Regressionskoeffizienten t-verteilt sind mit $n-m-1$ Freiheitsgraden (vergl. die Ausführungen zu Gleichung 17.21), lässt sich ein mit einer Wahrscheinlichkeit von $1-\alpha$ bestimmtes Konfidenzintervall wie folgt ermitteln (zu Konfidenzbereichen ⇨ Kap. 8.4):

$$b \pm t_{\frac{\alpha}{2}} \hat{\sigma}_b \tag{17.32}$$

Werden Konfidenzintervalle angefordert, so werden 95 %-Konfidenzintervalle berechnet. Die Ergebnisse werden rechts an die Tabelle für die Ergebnisausgabe der Koeffizienten gehängt. In Tabelle 17.4 ist nur der angehängte Teil zu sehen.

Tabelle 17.4. Ergebnisausgabe: 95 %-Konfidenzintervall der Option „Statistiken"

Koeffizienten[a]

Modell		95%-Konfidenzintervall für B	
		Untergrenze	Obergrenze
1	(Konstante)	30,133	73,401
	YVERF	,848	,876
	ZINS	-8,089	-2,536

a. Abhängige Variable

Für die hier berechnete Regressionsgleichung ist df = n-m-1 =31-2-1 = 28. Für α = 0,05 (zweiseitige Betrachtung) ergibt sich bei df = 28 aus einer tabellierten t-Verteilung $t_{\frac{\alpha}{2}} = 2{,}0484$. Das in Tabelle 17.4 ausgewiesene Konfidenzintervall für die Variable ZINS errechnet sich dann gemäß Gleichung 17.32 wie folgt: Untergrenze = –5,313 – 2,0484*1,355 = – 8,089 und Obergrenze = –5,313 + 2,0484*1,355 = – 2,536 [die Werte für den Regressionskoeffizienten (b = -5,313) und den Standardfehler

dieser ($\hat{\sigma}_b$ =1,355) stehen in Tabelle 17.1]. Man kann also erwarten, dass (bei wiederholten Stichproben) mit einer Wahrscheinlichkeit von 95 % das unbekannte β der Grundgesamtheit in den berechneten Grenzen liegt. Zur Bestimmung eines Konfidenzbereichs mit einer anderen Wahrscheinlichkeit (z.B. 95 %) müsste man aus einer tabellierten t-Verteilung das entsprechende t auswählen und dann den Konfidenzbereich in analoger Weise berechnen.

Kovarianzmatrix. Im oberen Teil der Tabelle 17.5 stehen die Korrelationskoeffizienten der Regressionskoeffizienten. Im unteren Teil stehen die Varianzen bzw. Kovarianzen der Regressionskoeffizienten: z.B. ist 4,56E – 05 die wissenschaftliche Schreibweise für $4{,}56 * 10^{-5} = 0{,}0000456$ und diese Varianz ist das Quadrat der in Tabelle 17.1 aufgeführten Standardabweichung des Regressionskoeffizienten von YVERF (= 0,006753 aufgerundet zu 0,007).

Tabelle 17.5. Ergebnisausgabe: Kovarianzmatrix der Option „Statistiken"

Korrelation der Koeffizienten[a]

Modell			ZINS	YVERF
1	Korrelationen	ZINS	1,000	-,269
		YVERF	-,269	1,000
	Kovarianzen	ZINS	1,837	-2,5E-03
		YVERF	-2,5E-03	4,56E-05

a. Abhängige Variable

Änderung in R-Quadrat. Diese Option gibt für den hier betrachteten Fall der Anwendung der Methode „Einschluss" keinen Sinn (⇨ Kap. 17.2.6).

Deskriptive Statistik. Es werden die arithmetischen Mittel („Mittelwert"), die Standardabweichungen sowie die Korrelationskoeffizienten nach Pearson für alle Variablen in der Regressionsgleichung ausgegeben. Für die Korrelationskoeffizienten wird das Signifikanzniveau bei einseitiger Fragestellung ausgewiesen.

Teil- und partielle Korrelationen. Die Ergebnisausgabe wird rechts an die Tabelle zur Ausgabe der Regressionskoeffizienten gehängt. In Abb. 17.6 ist nur der angehängte Teil dargestellt. In der Spalte „Nullter Ordnung" stehen die bivariaten Korrelationskoeffizienten zwischen CPR und den Variablen YVERF sowie ZINS und in der Spalte „Partiell" die partiellen Korrelationskoeffizienten des gleichen Zusammenhangs bei Konstanthaltung der jeweils anderen erklärenden Variablen (⇨ Kap. 16).

Tabelle 17.6. Teil- und partielle Korrelationskoeffizienten der Option „Statistiken"

Koeffizienten[a]

Modell		Korrelationen		
		Nullter Ordnung	Partiell	Teil
1	YVERF	,999	,999	,970
	ZINS	,240	-,595	-,030

a. Abhängige Variable

Kollinearitätsdiagnose. Die von SPSS ausgegebenen statistischen Informationen zur Kollinearitätsdiagnose (⇨ Tabelle 17.7, die Maße „Toleranz" und „VIF" werden der Tabelle 17.1 angehängt) dienen zur Beurteilung der Stärke der Multikollinearität, d.h. der Abhängigkeit der erklärenden Variablen untereinander (⇨ Kap. 17.4.4). Toleranz ist ein Maß für die Stärke der Multikollinearität. Toleranz für z.B die Variable ZINS wird wie folgt berechnet: für die Regressionsgleichung $ZINS = b_1 + b_2 * YVERF$ wird R^2 berechnet. Als Toleranz für ZINS ergibt sich $1 - R^2$. Wären in der Regressionsgleichung zur Erklärung von CPR weitere erklärende Variablen enthalten, so wären diese ebenfalls in der Regressionsgleichung für ZINS als erklärende Variablen einzuschließen. Hat eine Variable eine kleine Toleranz, so ist sie fast eine Linearkombination der anderen erklärenden Variablen. Ist „Toleranz" kleiner 0,01, so wird eine Warnung ausgegeben und die Variable nicht in die Gleichung aufgenommen. Sehr kleine Toleranzen können zu Berechnungsproblemen führen. „VIF" (Variance Inflation Factor) ist der Kehrwert von „Toleranz" und hat daher keinen zusätzlichen Informationswert.

Aus den Eigenwerten der Korrelationsmatrix der Erklärungsvariablen leitet sich ein *Konditionsindex* ab. Als Faustregel gilt, dass bei einem *Konditionsindex* zwischen 10 und 30 moderate bis starke und über 30 sehr starke Multikollinearität vorliegt. Für unser Regressionsmodell kann man feststellen, dass keine sehr starke Multikollinearität vorliegt.

Tabelle 17.7. Ergebnisausgabe: „Kollinearitätsdiagnose" der Option „Statistiken"

Kollinearitätsdiagnose[a]

Modell	Dimension	Eigenwert	Konditionsindex	Varianzanteile		
				(Konstante)	YVERF	ZINS
1	1	2,941	1,000	,00	,01	,00
	2	4,402E-02	8,174	,08	,99	,12
	3	1,505E-02	13,981	,91	,01	,88

a. Abhängige Variable

Durbin-Watson. Die Ergebnisausgabe wird rechts an die Tabelle zur Ausgabe „Anpassungsgüte des Modells" angehängt. In Tabelle 17.8 wird nur der angehängte Teil mit der Durbin-Watson-Teststatistik aufgeführt.

Tabelle 17.8. Ergebnisausgabe von „Durbin Watson" der Option „Statistiken"

Modellzusammenfassung

Modell	Durbin-Watson-Statistik
1	,752

Diese Teststatistik erlaubt es zu prüfen, ob Autokorrelation der Residualwerte besteht oder nicht (vergl. Gleichung 17.13). Autokorrelation der Residualwerte spielt vorwiegend bei Regressionsanalysen von Zeitreihen eine Rolle. Man nennt sie dann auch serielle Korrelation. Auch bei räumlicher Nähe von Untersuchungseinheiten sollte auf Autokorrelation geprüft werden (spatial correlation). Bei Bestehen von Autokorrelation der Residualwerte sind zwar die Schätzwerte für die Regressionskoeffizienten unverzerrt, nicht aber deren Standardabweichungen. Konsequenz ist, dass die Signifikanztests fehlerbehaftet und somit nicht aussagekräftig sind. Autokorrelation der Residualwerte ist häufig eine Folge einer Fehlspezifikation der Regressionsgleichung. Zwei Gründe sind dafür zu unterscheiden:

- Die (lineare) Gleichungsform ist falsch.
- Es fehlt eine wichtige erklärende Variable in der Gleichung (⇨ Kap. 17.4.1).

Der Durbin-Watson-Test beschränkt die Prüfung auf eine Autokorrelation 1. Ordnung, d.h. der Residualwert ε_i ist positiv (oder negativ) vom Residualwert der vorherigen Beobachtung ε_{i-1} abhängig. Formal lässt sich das auch mittels einer linearen Gleichung so ausdrücken:

$$\varepsilon_i = \rho\,\varepsilon_{i-1} + \zeta_i \quad \text{mit} \quad |\rho| < 1 \tag{17.33}$$

wobei ρ eine Konstante und ζ_i eine zufällige Variable ist. Eine Prüfung auf Autokorrelation der Residualwerte mit dem Durbin-Watson-Test ist ein Test um zwischen den Hypothesen

$$H_0: \rho = 0 \qquad \text{und} \qquad H_1: \rho \neq 0 \tag{17.34}$$

zu diskriminieren. Die Durbin-Watson-Prüfgröße d ist wie folgt definiert:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{17.35}$$

Die Prüfgröße kann zwischen 0 und 4 schwanken. Besteht keine Korrelation aufeinanderfolgender Residualwerte ($\rho = 0$), so liegt die Prüfgröße nahe bei 2. Besteht eine positive Autokorrelation, so liegen e_i und e_{i-1} nahe beieinander mit der Konsequenz, dass d kleiner 2 wird. Besteht negative Korrelation, so folgen auf positiven e-Werten negative und umgekehrt. Konsequenz ist, dass d größer als 2 wird. Demnach besteht bei einer Prüfgröße d wesentlich kleiner 2 eine positive ($\rho > 0$) und bei d wesentlich größer 2 eine negative ($\rho < 0$) Autokorrelation. Durch Vergleich des empirisch erhaltenen d mit von Durbin und Watson vorgelegten tabellierten Werten kann für eine vorgegebene Irrtumswahrscheinlichkeit α auf Autokorrelation der Residualwerte getestet werden. Aus der von Durbin und Watson

vorgelegten Tabelle sind für die Anzahl der Beobachtungen n, die Anzahl der erklärenden Reihen m sowie der Irrtumswahrscheinlichkeit α jeweils eine kritische Untergrenze d_u sowie kritische Obergrenze d_o ablesbar. In Tabelle 17.9 sind fünf Entscheidungsbereiche in Abhängigkeit von d niedergelegt. Ist d kleiner als d_u oder größer als $4-d_u$, so besteht positive bzw. negative Autokorrelation. Im Indifferenzbereich kann keine sichere Entscheidung getroffen werden.

Tabelle 17.9. Bereiche der Durbin-Watson-Statistik d

H_0 ablehnen = positive Autokorrelation	Indifferenz-bereich	H_0 annehmen = keine Auto-korrelation-	Indifferenz-bereich	H_0 ablehnen = negative Autokorrelation
0 d_u		d_o 2	$4-d_o$	$4-d_u$ 4

Da Autokorrelation der Residualwerte eine schwerwiegende Verletzung der Modellvoraussetzungen ist, wird auch häufig d_o bzw. $4-d_u$ als kritischer Wert zur Abgrenzung des Annahme- oder Ablehnungsbereichs gewählt. Insofern wird der Indifferenzbereich gleichfalls als Ablehnungsbereich für H_0 gewählt.

Aus Tabelle 17.8 ergibt sich d = 0,752. Für n = 31, m = 2 und α = 0,05 ergibt sich aus der Durbin-Watson-Tabelle d_u = 1,30 und d_o = 1,57. Damit fällt die Prüfgröße in den Ablehnungsbereich für H_0: mit einer Irrtumswahrscheinlichkeit von 5 % wird die Hypothese H_0 (es besteht keine Autokorrelation der Residualwerte) verworfen. Es liegt demnach also eine positive Autokorrelation der Residualwerte vor. Mit diesem Ergebnis besteht Anlass, den Regressionsansatz hinsichtlich der Vollständigkeit der erklärenden Variablen sowie der Kurvenform zu überprüfen.

Eine positive Autokorrelation der Residualwerte kann auch grafisch verdeutlicht werden. Dazu wurden folgende Schritte unternommen: mittels der Option „Speichern" der Dialogbox „Lineare Regression" wurden die (unstandardisierten) Residualwerte RES_1 dem Datensatz hinzugefügt (⇨ Kap. 17.2.4). Dann wurde die um ein Jahr zeitverzögerte Residualgröße RES_1V gebildet [mit Hilfe des Menüs „Transformieren" und der Lag-Funktion von „Berechnen...", RES_1V = LAG(RES_1)]. Im letzten Schritt wurde mittels der Kommandofolge „Grafiken", „Streudiagramm.." und der Option „Einfach" sowie y-Achse: RES_1 und x-Achse: RES_1V ein Streudiagramm erzeugt, das optisch die positive Abhängigkeit der Residualwerte von zeitlich vorhergehenden verdeutlicht (⇨ Abb. 17.7). Berechnet man den bivariaten Korrelationskoeffizienten für RES_1 und RES_1V mittels der Befehlsfolge „Analysieren", „Korrelation ▷ Bivariat...", so ergibt sich ein Wert von 0,6115.

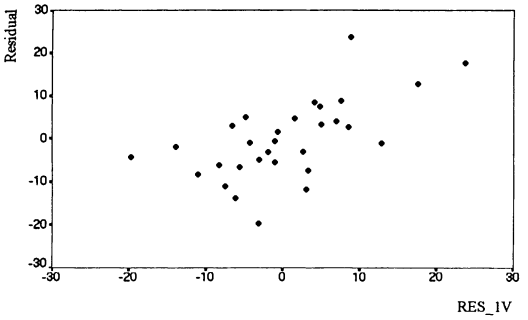


Abb. 17.7. Positive Autokorrelation der Residualwerte

Mit der Wahl der Durbin-Watson-Statistik werden in einer Tabelle (⇨ Tabelle 17.10) auch Ergebnisse für die Schätzwerte bzw. Vorhersagewerte $\hat{y}_i$ [vergl. Gleichung (17.27)] sowie für die Residualwerte e_i ausgegeben. Dabei wird zwischen nicht standardisierten und standardisierten Werten (d.h. in z-Werte transformierte Werte, ⇨ Gleichung 8.8) unterschieden. Es werden jeweils das Minimum, das Maximum, das arithmetische Mittel und die Standardabweichung aufgeführt.

Tabelle 17.10. Weitere Ergebnisausgabe von „Durbin-Watson" der Option „Statistiken"

Residuenstatistik

	Minimum	Maximum	Mittelwert	Standardabweichung	N
Nicht standardisierter vorhergesagter Wert	437,996	1222,980	837,792	228,505	31
Nicht standardisierte Residuen	-19,676	23,801	-1,5E-13	9,195	31
Standardisierter vorhergesagter Wert	-1,750	1,686	,000	1,000	31
Standardisierte Residuen	-2,067	2,501	,000	,966	31

Fallweise Diagnose. Je nach Wahl können die Residualwerte e_i entweder für alle Fälle oder nur für die Fälle mit Ausreißern in einer Tabelle aufgelistet werden. In beiden Fällen werden sie dann sowohl standardisiert (d.h. in z-Werte transformiert, ⇨ Gleichung 8.8) als auch nicht standardisiert ausgegeben. Außerdem wird die abhängige Variable und deren Vorhersagewert ausgegeben. Ausreißer liegen außerhalb eines Standardabweichungsbereichs um den Mittelwert von e_i ($\bar{e} = 0$) (voreingestellt ist der 3*Standardabweichungsbereich).

17.2.3 Ergänzende Grafiken zum Regressionsmodell (Schaltfläche „Diagramme“)

Durch Anklicken der Schaltfläche „Diagramme…“ in der Dialogbox „Lineare Regression“ (⇨ Abb. 17.5) können verschiedene Grafiken zu der Regressionsgleichung angefordert werden. Die Grafiken beziehen sich auf die Residual- und Vorhersagewerte in verschiedenen Varianten. Diese erlauben es, einige Modellvoraussetzungen bezüglich der Residualvariable zu überprüfen (⇨ Kap. 17.4). Das Regressionsgleichungsmodell kann nur dann als angemessen betrachtet werden, wenn die empirischen Residualwerte e_i ähnliche Eigenschaften haben wie die Residualwerte ε_i des Modells. Unter anderem werden auch Residual- und Vorhersagewerte unter Ausschluss einzelner Fälle bereitgestellt. Damit wird es möglich, den Einfluss von nicht recht in das Bild passenden Fällen („outliers“) für das Regressionsmodell zu bewerten.

Abb. 17.8 zeigt die (Unter-)Dialogbox „Grafiken“ mit einer Einstellung, die im folgenden erläutert wird.

Streudiagramm 1 von 1. In der Quellvariablenliste der Dialogbox stehen standardmäßig folgende Variablen, die zur Erstellung von Streudiagrammen (Scatterplots) genutzt werden können. Die mit einem * beginnenden Variablen sind temporär.

- DEPENDNT: abhängige Variable y.
- *ZPRED: Vorhersagewerte $\hat{y}_i$, in standardisierte Werte (z-Werte) transformiert.
- *ZRESID: Residualwerte e_i, in standardisierte Werte (z-Werte) transformiert.
- *DRESID: Residualwerte e_i bei Ausschluss (deleted) des jeweiligen Falles i bei Ermittlung der Regressionsgleichung.
- *ADJPRED: Vorhersagewerte $\hat{y}_i$ bei Ausschluss (deleted) des jeweiligen Falles i bei Ermittlung der Regressionsgleichung.
- *SRESID: Die Residualwerte e_i, dividiert durch den Schätzwert ihrer Standardabweichung, wobei diese je nach der Distanz zwischen den Werten der unabhängigen Variablen des Falles und dem Mittelwert der unabhängigen Variablen von Fall zu Fall variiert. Diese studentisierten Residuen geben Unterschiede in der wahren Fehlervarianz besser wieder als die standardisierten Residuen
- *SDRESID: Studentisiertes Residuum bei Ausschluss (deleted) des jeweiligen Falles i bei Ermittlung der Regressionsgleichung.

Abb. 17.8. Dialogbox „Lineare Regression: Grafiken“

Diese Variablen können in Streudiagrammen dargestellt werden, indem sie in die Felder für die y- bzw. x-Achse übertragen werden. Abb. 17.9 ist z.B. ein Ergebnis der Einstellungen in Abb. 17.8. Ein derartiges Streudiagramm kann Hinweise dafür geben, ob die Bedingung der Homoskedastizität erfüllt ist oder nicht (vergl. Gleichung 17.12). Aus dem Streudiagramm gewinnt man nicht den Eindruck, dass die Streuung der Residualwerte systematisch mit der Höhe der Vorhersagewerte variiert, so dass es gerechtfertigt erscheint, von Homoskedastizität auszugehen. Andererseits wirkt die Punktwolke aber auch nicht wie zufällig. Damit werden die im Zusammenhang mit einer Prüfung auf Autokorrelation aufgetretenen Zweifel hinsichtlich der Kurvenform oder der Vollständigkeit des Regressionsmodells bezüglich wichtiger Erklärungsvariable verstärkt. Das Modell ist nicht hinreichend spezifiziert. Es ist zu vermuten, dass eine oder mehrere wichtige Erklärungsvariable fehlen. Eine sinnvolle Ergänzung zu dem Streudiagrammen der Abb. 17.9 sind Streudiagramme, in denen die Residualwerte gegen die erklärenden Variablen geplottet werden. Dabei können auch erklärende Variable eingeschlossen sein, die bisher nicht im Erklärungsansatz enthalten waren. Derartige Streudiagramme und weitere lassen sich erzeugen, wenn die Residualwerte mit „Speichern" dem Datensatz hinzugefügt werden (⇨Kap. 17.2.4).

Im Rahmen des Untermenüs „Diagramme" können weitere Streudiagramme nach Klicken von „Weiter" angefordert werden.

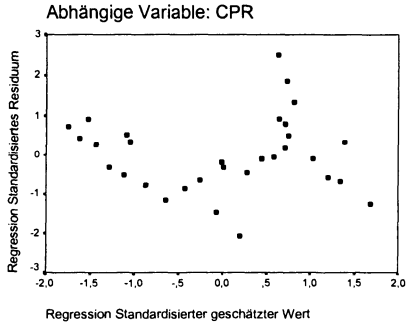


Abb. 17.9. Streudiagramm Residualwerte gegen Vorhersagewerte (jeweils standardisiert)

Diagramme der standardisierten Residuen. Mit Hilfe der Option „Diagramme der standardisierten Residuen" in Abb. 17.8 lassen sich weitere Untersuchungen der (standardisierten) Residualwerte vornehmen, insbesondere zur Prüfung der Frage, ob die Modellbedingungen erfüllt sind (⇨ Kap. 17.4).

① *Histogramm.* Abb. 17.10 bildet als Ergebnis der Option „Histogramm" die Häufigkeitsverteilung der Residualwerte ab. In die empirische Häufigkeitsverteilung ist die Normalverteilung mit den aus den empirischen Residualwerten bestimmten Parametern Mittelwert = 0 und Standardabweichung = 0,97 gelegt. Durch diese Darstellung kann geprüft werden, ob die Annahme einer Normalverteilung für die Residualvariable annähernd zutrifft. Unser Demonstrationsbeispiel hat allerdings nur 31 Fälle, so dass es schwerfällt, eine sichere Aussage bezüglich

Bestehens normalverteilter Residualwerte zu treffen. Da die Abweichungen der empirischen e_i-Werte von der Normalverteilung aber nicht sehr gravierend sind, kann man unterstellen, dass auch für die Zufallsvariable ε_i eine Normalverteilung als Voraussetzung zur Durchführung von Signifikanztests gegeben ist. Eine besser gesicherte Aussage lässt sich eventuell mit Hilfe von Tests durchführen (vergl. Kolmogorov-Smirnov-Test in Kap. 19.2.4 sowie die Tests mit „Explorative Datenanalyse" in Kap. 9.3.2).

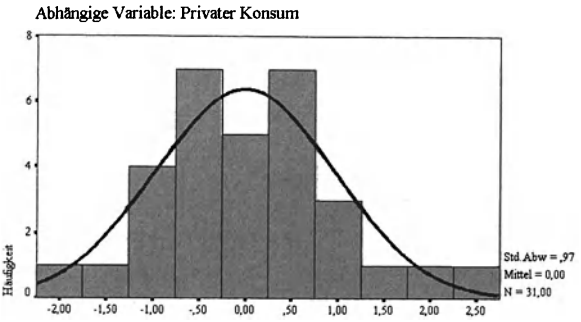


Abb. 17.10. Häufigkeitsverteilung der Residualwerte

② *Normalverteilungsdiagramm.* Abb. 17.11 hat die gleiche Aufgabenstellung wie Abb. 17.10: es soll festgestellt werden, ob die Residualwerte gravierend von der Normalverteilung abweichen. In dem Diagramm sind die bei Vorliegen einer Normalverteilung (theoretischen) und die empirischen kumulierten Häufigkeiten einander gegenübergestellt. Auch diese Darstellung bestätigt, dass die Abweichung von der Normalverteilung nicht gravierend ist (⇨ P-P-Diagramme in Kap. 20.13).

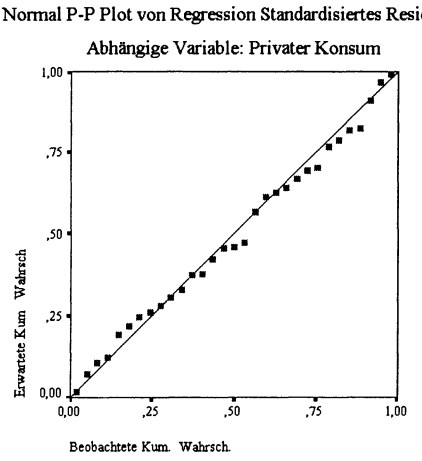


Abb. 17.11. P-P-Normalverteilungsdiagramm der standardisierten Residualwerte

Alle partiellen Diagramme erzeugen. Wird diese Option gewählt, so werden alle Streudiagramme erstellt, die partiellen Korrelationskoeffizienten entsprechen (⇨ Kap. 16.2). Diese Streudiagramme sind ein hilfreiches Mittel zur Prüfung der Frage, ob unter Berücksichtigung aller anderen erklärenden Variablen ein linearer Zusammenhang besteht (⇨ Kap. 17.4.1). Auch ist das Diagramm wertvoll, um zu sehen, ob eventuell „Ausreißer" einen starken Einfluss auf den partiellen Regressionskoeffizienten haben könnten.

In Abb. 17.12 ist als Beispiel CPR in Abhängigkeit von ZINS bei Eliminierung des linearen Effektes von YVERF sowohl aus CPR als auch aus ZINS dargestellt. Sichtbar wird eine negative Korrelation zwischen CPR und ZINS mittlerer Stärke, die ja auch im partiellen Korrelationskoeffizienten zwischen den Variablen in Höhe von −0,5952 zum Ausdruck kommt (⇨ Kap. 16.2). Der Zusammenhang ist durchaus linear.

Abb. 17.12 ließe sich auch (aber umständlicher) erzeugen, indem man folgende Schritte unternimmt: in einem ersten Regressionsansatz wird CPR mittels YVERF erklärt und die sich ergebenden Residualwerte RES_1 mit Hilfe der Option „Speichern" (⇨ 17.2.4) dem Datensatz hinzufügt. In einem zweiten Regressionsansatz wird ZINS mittels YVERF erklärt und die sich ergebenden Residualwerte RES_2 ebenfalls dem Datensatz hinzugefügt. Dann wird mit Hilfe der Befehlsfolge „Grafiken", „Streudiagramm..." mit den Optionen „Einfach" sowie y-Achse: RES_1 und x-Achse: RES_2 ein (partielles) Streudiagramm erzeugt, das dem in Abb. 17.12 entspricht.

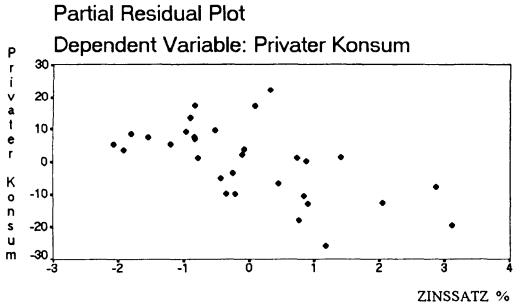


Abb. 17.12. Partielles Streudiagramm

17.2.4 Speichern von neuen Variablen des Regressionsmodells (Schaltfläche „Speichern")

Durch Anklicken der Schaltfläche „Speichern..." in der Dialogbox „Lineare Regression" (⇨ Abb. 17.5) wird die in Abb. 17.13 dargestellte Dialogbox geöffnet. Es lassen sich dann eine ganze Reihe im Zusammenhang mit einer Regressionsgleichung berechenbarer Variablen anfordern und zu den Variablen des Datensatzes hinzufügen. Der Sinn ist darin zu sehen, dass man anschließend die Variablen für umfassende Prüfungen hinsichtlich der Modellvoraussetzungen nutzen kann. Des weiteren dienen einige der Variablen dazu, zu prüfen, in welchem Maße „Ausreißer-Fälle" Einfluss auf die berechneten Ergebnisse haben. Fälle mit „unge-

wöhnlichen" Werten können identifiziert und ihr Einfluss auf Ergebnisse sichtbar gemacht werden.

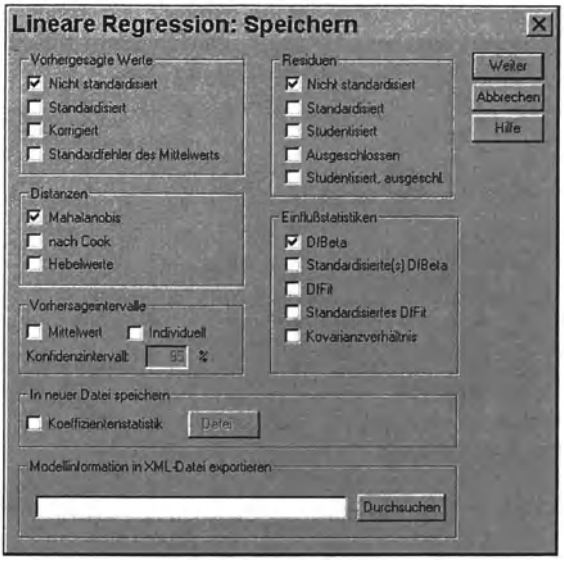


Abb. 17.13. Dialogbox „Lineare Regression: Speichern"

Die angeforderten und dem Datensatz hinzugefügten Variablen erhalten automatisch Variablennamen, die im folgenden erläutert werden. Sie enden jeweils mit einer an einen Unterstrich angehängten Ziffer. Die Ziffer gibt an, die wievielte Variable des Variablentyps dem Datensatz hinzugefügt worden ist. Beispielsweise bedeutet PRE_3, dass dem Datensatz inzwischen die dritte Variable PRE (die eines Vorhersagewertes) hinzugefügt worden ist. Sobald mindestens eine Variable zur Speicherung angefordert wird, wird eine mit „Residuenstatistik" überschriebenen Tabelle ausgegeben. In dieser Tabelle werden das Minimum, das Maximum, der Mittelwert, die Standardabweichung sowie die Anzahl der Fälle N für alle Variablen der Bereiche „Vorhergesagter Wert", „Residuen" und „Distanz" aufgeführt.

Folgende Variablen können dem Datensatz hinzugefügt werden (⇨ Abb. 17.13):

① *Vorhergesagte Werte*

- *Nicht standardisiert.* Nicht standardisierter vorhergesagter Wert $\hat{y}_i$ (PRE_: Predicted Value).
- *Standardisiert.* Standardisierter (in einen z-Wert transformierter) vorhergesagter Wert (ZPR_: Standardized Predicted value).
- *Korrigiert.* Korrigierter Vorhersagewert. Vorhersagewert bei Ausschluss des jeweiligen Falles i bei Ermittlung der Regressionsgleichung (ADJ_: Adjusted predicted Value).
- *Standardfehler des Mittelwerts.* Standardfehler des mittleren Vorhersagewerts $\hat{y}_i$ (SEP_: Standard error of predicted value, ⇨ Gleichung 17.25).

② *Distanzen.*

- *Mahalanobis.* Dieses Distanzmaß misst, wie stark ein Fall vom Durchschnitt der anderen Fälle hinsichtlich der erklärenden Variablen abweicht. Ein hoher Distanzwert für einen Fall signalisiert, dass dieser hinsichtlich der erklärenden Variablen ungewöhnlich ist und damit eventuell einen hohen Einfluss auf Ergebnisse haben könnte (MAH_: Mahalanobis' Distance).

  Mit Hilfe der Befehlsfolge „Analysieren", „Deskriptive Statistiken", ▷„ „Explorative Datenanalyse", Übertragen von MAH_ in das Eingabefeld „abhängige Variablen" und den Optionen „Statistik...", „Ausreißer" kann man sich z.B. die fünf größten sowie fünf kleinsten Werte des Distanzmaßes ausgeben lassen (⇨ Tabelle 17.11).

  Für unser Anwendungsbeispiel mit nur 31 Fällen bietet sich als Alternative eine Grafik zum Ausweis der Distanzmaße an. Mit dem Grafikbefehl „Linie···" und der Auswahlkombination „Einfach" sowie „Werte einzelner Fälle", der Definition „Linie entspricht:" = MAH_1 und „Kategorienbeschriftungen" „Variable:" = JAHR erhält man die folgende Abb. 17.14. Aus ihr wird deutlich, dass insbesondere die Jahre (= Fälle) 1974 und 1981 ungewöhnlich sind hinsichtlich der erklärenden Variablen.

Tabelle 17.11. Distanzmaß nach Mahalanobis: Fälle mit den fünf größten und fünf kleinsten Werten

Extremwerte

			Fallnummer	JAHR	Wert
MAH_1	Größte Werte	1	15	74	5,93362
		2	22	81	5,62722
		3	28	87	3,91993
		4	29	88	3,89916
		5	1	60	3,26889
	Kleinste Werte	1	17	76	,20597
		2	13	72	,36233
		3	10	69	,44637
		4	24	83	,47206
		5	12	71	,48334

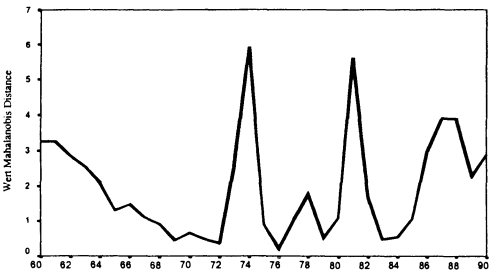


Abb. 17.14. Distanzmaß nach Mahalanobis für die Regressionsgleichung

- *nach Cook.* (COO_: Cooks's distance). Durch den Vergleich der Residualwerte („Nicht standardisiert" und „Ausgeschlossen") kann man ermessen, wie stark ein Fall auf Ergebnisse Einfluss nimmt. Nicht sehen kann man aber

daran, in welchem Ausmaß der Ausschluss eines Falles bei der Berechnung der Regressionsgleichung Wirkungen auf die Residualwerte aller anderen Fälle hat. Diese Information wird durch das Distanzmaß nach Cook vermittelt. Das Distanzmaß ist wie folgt definiert:

$$C_i = \frac{\sum_{j=1}^{n} (\hat{y}_j^{(i)} - \hat{y}_j)^2}{(m+1)\hat{\sigma}_\varepsilon^2} \qquad (17.36)$$

wobei $\hat{y}_j^{(i)}$ der Vorhersagewert für den Fall j ist, wenn die Regressionsgleichung bei Ausschluss des Falles i berechnet wurde, m ist die Anzahl der zu erklärenden Variablen und $\hat{\sigma}_\varepsilon^2$ ist der Schätzwert für die Varianz der Residualvariable (⇨ Gleichung 17.18). Das Maß C_i wird Null, wenn für alle Fälle j die $\hat{y}_j^{(i)}$ nicht von den $\hat{y}_j$ abweichen. Bestehen hohe Abweichungen, so wird C_i groß. Große Werte des Maßes weisen demnach Fälle aus, die hohen Einfluss auf die Ergebnisse haben. Werden die der Datei hinzugefügten Werte von C_i für alle Jahre (Fälle) in einer Liniengrafik dargestellt, so zeigt sich, dass in den Jahren 1975, 1981, 1983 und 1990 die Werte besonders hoch sind. Man kann davon ausgehen, dass diese Jahre den größten Einfluss auf die Regressionsergebnisse haben.

- *Hebelwerte (Leverage).* Ein Maß für den Einfluss, den eine Beobachtung i auf die Anpassung einer Regressionsfunktion besitzt. Der Wert für den Hebelwirkungseffekt ergibt sich aus der Mahalanobis-Distanz, dividiert durch $n-1$. (LEV_: leverage).

③ *Vorhersageintervalle.*

Analog zu den Konfidenzbereichen von Regressionskoeffizienten (⇨ Gleichung 17.32) können Konfidenzbereiche für die Vorhersagewerte bestimmt werden. Da sich die Varianzen der durchschnittlichen und individuellen Vorhersagewerte bei gegebenen Werten der erklärenden Variablen unterscheiden (⇨ Gleichungen 17.25 und 17.26), weichen auch die Berechnungen für Konfidenzintervalle voneinander ab.

- *Mittelwert.* Intervallschätzwerte (d.h. unterer und oberer Grenzwert) für das durchschnittliche $\hat{y}_i$. Ein mit einer Wahrscheinlichkeit $1-\alpha$ bestimmtes Konfidenzintervall für einen Fall i ergibt sich als

$$\hat{y}_i \pm t_{\alpha/2,n-m-1} * \hat{\sigma}_{\hat{y}} \qquad (17.37)$$

wobei $t_{\alpha/2,n-m-1}$ der t-Wert aus einer tabellierten t-Verteilung (für die zweiseitige Betrachtung) bei $n-m-1$ Freiheitsgraden und $\hat{\sigma}_{\hat{y}}$ der Schätzwert für die Standardabweichung des Schätzfehlers ist (Gleichung 17.25). Die Wahrscheinlichkeit, mit der ein Konfidenzintervall berechnet werden soll, kann durch Eingabe bestimmt werden. Voreingestellt ist 95 % ($= 1-\alpha$). Es kann ein anderer %-Wert eingegeben werden. Für $n-m-1 = 28$ entspricht dieses $t_{0,025,28} = 2{,}0484$ [(LMCI_: 95 % LCI for Variablennamen mean (L = lower,

der untere Wert), UMCI_: 95 %UCI for Variablennamen mean (U = upper, der obere Wert)].

- *Individuell.* Ein Konfidenzintervall in Analogie zur Gleichung 17.37 kann bestimmt werden. Im Unterschied muss aber der Schätzwert für die Standardabweichung gemäß Gleichung 17.26 gewählt werden. (LICI_: 95 % LCI for Variablennamen individual (L = lower, der untere Wert), UICI_: 95 %UCI for Variablennamen individual (U = upper, der obere Wert)].

④ *Residuen*

- *Nicht standardisiert.* Nicht standardisierte Residualwert e_i (RES_: residual).
- *Standardisiert.* Standardisierte (in z-Werte transformierte) Residualwerte, d.h. hier Residualwerte dividiert durch ihre Standardabweichung (ZRE: standardized residual).
- *Studentisiert.* Die Residualwerte e_i, dividiert durch den Schätzwert ihrer Standardabweichung, wobei diese je nach der Distanz zwischen den Werten der unabhängigen Variablen des Falles und dem Mittelwert der unabhängigen Variablen von Fall zu Fall variiert (SRE: studentized residual)
- *Ausgeschlossen.* Residualwert bei Ausschluss des jeweiligen Falles i bei Ermittlung der Regressionsgleichung (DRE_: deleted residual).
- *Studentisiert, ausgeschl.* Studentisierte Residuen bei Ausschluss des jeweiligen Falles i bei Ermittlung der Regressionsgleichung (SDR: studentized deleted residual).

Werden die nicht standardisierten Residuen und diejenigen, die sich bei Ausschluss des jeweiligen Falles (= Jahres) bei Berechnung der Regressionsgleichung ergeben, einander gegenübergestellt, so zeigt sich, dass die Unterschiede sehr gering sind.

⑤ *Einflussstatistiken* (Maße zur Identifizierung einfluss*reicher* Fälle).

- *DfBeta.* Differenz in den Regressionskoeffizienten bei Ausschluss des jeweiligen Falles i bei Ermittlung der Regressionsgleichung. Für jede erklärende Variable sowie das konstante Glied der Gleichung wird ein Variablennamen bereitgestellt (DFB0_: Dfbeta für das konstante Glied, DFB1_: Dfbeta für die erste erklärende Variable, DFB2_: Dfbeta für die zweite erklärende Variable usw.).
- S*tandardisierte(s) DfBeta.* Die oben beschriebenen DfBeta-Werte werden standardisiert zur Verfügung gestellt, d. h. in z-Werte transformiert (SDB0_: studentisierte Dfbeta für das konstante Glied, SDB1_: studentisierte Dfbeta für die erste erklärende Variable, SDB2_: studentisierte Dfbeta für die zweite erklärende Variable usw.).
- *DfFit.* Differenz von R^2 bei Ausschluss des jeweiligen Falles i bei Ermittlung der Regressionsgleichung (DFF_: dffit). Es zeigt sich, dass der Schwankungsbereich der Differenz zwischen ± 2 Prozentpunkte liegt, wobei die Jahre 1975, 1981 und 1990 R^2 relativ stark beeinflussen.
- S*tandardisiertes DfFit:* Die oben beschriebenen DfFit-Werte werden standardisiert bereitgestellt (SDF_: sdffit).
- *Kovarianzverhältnis (Covariance ratio).* Dieses Maß kann hier nicht näher erläutert werden. Es wird auch bei Ausschluss des jeweiligen Falles i berech-

net. Wenn der Quotient dicht bei 1 liegt, beeinflusst der weggelassene Fall die Varianz-Kovarianz-Matrix nur unwesentlich (COV_: Covratio).

Die Option „In neuer Datei speichern" (⇨ Abb. 17.13) ermöglicht es, die Regressionskoeffizienten und weitere statistische Informationen zu der geschätzten Regressionsgleichung in einer Datei zu speichern. Mit der Option „Modellinformation in XML-Datei exportieren", werden Modellinformationen in eine anzugebende Datei exportiert. Diese Datei kann von SmartScore und von neuen Versionen von WhatIf? verwendet werden.

17.2.5 Optionen für die Berechnung einer Regressionsgleichung (Schaltfläche „Optionen")

Die in Abb. 17.15 dargestellte Dialogbox „Lineare Regression: Optionen..." erscheint, wenn man in der Dialogbox „Lineare Regression" (⇨ Abb. 17.5) auf „Optionen" klickt In ihr lassen sich verschiedene Modalitäten für die Berechnung der Regressionsgleichung wählen:

- *Kriterien für schrittweise Methode.* Die Auswahlmöglichkeiten beziehen sich auf die anderen Verfahren zum Einschluss von unabhängigen Variablen in die Regressionsgleichung (also nicht für „Methode: Einschluss"). Daher werden diese unten im Zusammenhang mit den anderen Verfahren erläutert (⇨ Kap. 17.2.6).
- *Konstante in Gleichung einschließen.* Die Berechnung der Gleichung einschließlich des konstanten Gliedes ist die übliche und daher voreingestellte Variante. Nur in seltenen Ausnahmefällen macht die Restriktion, das konstante Glied gleich Null zu setzen, einen Sinn.
- *Fehlende Werte.* Die Option „Listenweiser Fallausschluss" ist die Voreinstellung und bedeutet, dass die Berechnungen nur auf Fälle basieren, die für alle Variablen des Regressionsmodells gültige Werte haben. Bei Wahl der Option „Paarweiser Fallausschluss" werden die als Basis aller Berechnungen dienenden Korrelationskoeffizienten für gültige Werte von jeweiligen Variablenpaaren kalkuliert. Bei der Option „Durch Mittelwert ersetzen" werden fehlende Werte von Variablen durch das arithmetische Mittel dieser substituiert (zu Ausreißer und fehlenden Werten ⇨ Kap. 17.4.5).

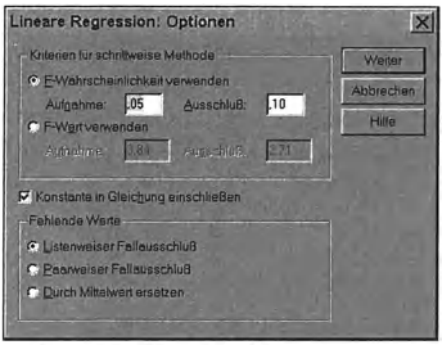


Abb. 17.15. Dialogbox „Lineare Regression: Optionen"

17.2.6 Verschiedene Verfahren zum Einschluss von erklärenden Variablen in die Regressionsgleichung („Methode")

Variablen können auf unterschiedliche Weise in die Regressionsgleichung eingeschlossen werden. Möglich sind die folgenden Verfahren, die im Auswahlfeld „Methode" der Dialogbox „Lineare Regression" wählbar sind (⇨ Abb. 17.5):

- *Einschluss.* Alle erklärenden Variablen werden in einem Schritt in die Gleichung einbezogen.
- *Schrittweise.* Die erklärenden Variablen werden schrittweise in die Gleichung aufgenommen. Die Reihenfolge richtet sich nach einem bestimmten Aufnahmekriterium, dessen Schwellenwerte man in der in Abb. 17.15 dargestellten Dialogbox festlegen kann. Werden schrittweise weitere Variablen aufgenommen, so wird nach jedem Schritt geprüft, ob die bislang in der Gleichung enthaltenen Variablen aufgrund eines Ausschlusskriteriums wieder ausgeschlossen werden sollen.
- *Ausschluss.* Diese Methode kann nur nach Einsatz eines anderen Verfahrens in einem ersten Block zum Zuge kommen. Zunächst werden alle erklärenden Variablen eingeschlossen. Mit „Ausschluss" werden die erklärenden Variablen, die ein Ausschlusskriterium erfüllen, wieder ausgeschlossen.
- *Rückwärts.* Zunächst werden alle Variablen eingeschlossen. In Folgeschritten werden Variablen, die ein bestimmtes Ausschlusskriterium erfüllen, ausgeschlossen.
- *Vorwärts.* Die erklärenden Variablen werden wie bei „Schrittweise" Schritt für Schritt einbezogen. Der Unterschied liegt aber darin, dass in Folgeschritten nicht geprüft wird, ob eine Variable wieder ausgeschlossen werden soll.

Im folgenden werden die Grundlagen der Verfahren am Beispiel von „Schrittweise" erläutert. Dazu wird ein Regressionsansatz gewählt, der CPR durch YVERF, ZINS und LQ erklären soll. In der in Abb. 17.5 dargestellten Dialogbox werden die erklärenden Variablen YVERF und ZINS um LQ ergänzt und die „Methode" „Schrittweise" gewählt. Ergebnistabellen werden im folgenden nur insoweit besprochen als es zum Verständnis der Methode nötig ist.

Grundlage für die Aufnahme- bzw. den Ausschluss einer Variable ist ein F-Test in Anlehnung an die Ausführungen im Zusammenhang mit Gleichung 17.31. Dieser sogenannte partielle F-Test prüft, ob durch die Aufnahme einer zusätzlichen erklärenden Variable das Bestimmtheitsmaß R^2 signifikant erhöht wird. Dieses entspricht der Prüfung, ob die zusätzliche Variable einen signifikant von Null verschiedenen Regressionskoeffizienten hat. Analog wird getestet, ob durch den Ausschluss einer Variable R^2 signifikant sinkt. Dieser Test kann auch angewendet werden für den Fall, dass in einem Schritt mehrere zusätzliche Variablen in die Regressionsgleichung aufgenommen (oder ausgeschlossen) werden sollen.

Die Prüfgröße ist

$$F = \frac{R^2_{Diff} / k}{(1 - R^2) / (n - m - 1)} \tag{17.38}$$

wobei R^2_{Diff} die Veränderung (Differenz) von R^2 bei Aufnahme (oder Ausschluss) einer (oder mehrerer) zusätzlichen erklärenden Variable, n der Stichprobenumfang, m die Anzahl der erklärenden Variablen und k die Anzahl der zusätzlich aufgenommenen (bzw. ausgeschlossenen) erklärenden Variablen ist. Unter der Nullhypothese (keine Veränderung von R^2) ist die Prüfgröße F-verteilt mit $df_1 = k$ und $df_2 = n - m - 1$ Freiheitsgraden. Durch Vergleich des aus Gleichung 17.38 erhaltenen empirischen F mit dem bei Vorgabe einer Irrtumswahrscheinlichkeit α und der Anzahl der Freiheitsgrade entnehmbaren F-Wert aus einer F-Tabelle, kann die H_0-Hypothese angenommen oder abgelehnt werden. Bei einer Irrtumswahrscheinlichkeit von 5 % ($\alpha = 0{,}05$) und $df_1 = k = 1$ und $df_2 = n - m - 1 = 31 - 3 - 1 = 27$, ergibt sich ein kritischer Wert $F_{krit} = 4{,}22$. Ist der empirische F-Wert nach Gleichung 17.38 kleiner als F_{krit}, so wird die Hypothese H_0 (keine signifikante Erhöhung von R^2 durch die zusätzliche Variable) angenommen, sonst abgelehnt. Alternativ kann auch die Wahrscheinlichkeit für den empirischen erhaltenen F-Wert mit der vorzugebenden Irrtumswahrscheinlichkeit verglichen werden. Die Vergleichskriterien für die Aufnahme und für den Ausschluss von erklärenden Variablen in die Regressionsgleichung können in der (Unter-) Dialogbox „Optionen" (⇨ Kap. 17.2.5) festgelegt werden.

Dieser F-Test zur Prüfung einer signifikanten Differenz von R^2 entspricht einem t-Test zur Prüfung der Signifikanz des Regressionskoeffizienten einer zusätzlichen Variable, da $t^2 = F$ ist.

In der Ausgabetabelle mit der Überschrift „Ausgeschlossene Variablen" (Tabelle 17.12) ist das Ergebnis der Regressionsgleichung hinsichtlich der nicht in die Regressionsgleichung aufgenommenen Variablen zu sehen. Im Anwendungsbeispiel wird im ersten Schritt die Variable YVERF eingeschlossen und ZINS sowie LQ ausgeschlossen (Modell 1) und dann im nächsten Schritt zusätzlich die Variable ZINS eingeschlossen (Modell 2).

Die Variable LQ wird (wie aus Tabelle 17.12 hervorgeht) nicht in das Modell eingeschlossen, weil das Einschlusskriterium (hier: Wahrscheinlichkeit des F-Wertes für die Aufnahme <= 0,05, ⇨ Abb. 17.15) für die Aufnahme nicht erreicht wird. Für die nicht in die Gleichung einbezogene Variable LQ (ausgeschlossene Variable) wird t = 1,594 ausgewiesen. Demnach ist $F = t^2 = 2{,}54$. Dieser F-Wert ist kleiner als $F_{krit} = 4{,}22$ und fällt insofern in den Annahmebereich für H_0. Daher wird LQ nicht in die Gleichung aufgenommen. Man kann dieses Ergebnis auch anhand des angegebenen Wertes für „Signifikanz" ablesen. Der Wert von „Signifikanz" beträgt im Modell 2 0,123. Da dieser Wert die Irrtumswahrscheinlichkeit von 5 % ($\alpha = 0{,}05$) übersteigt, wird die Variable LQ als nicht signifikant erkannt und deshalb nicht in das Modell eingeschlossen.

Tabelle 17.12. Ausschnitt aus der Ergebnisausgabe für schrittweise Regression

Ausgeschlossene Variablen[c]

Modell		Beta In	T	Signifikanz	Partielle Korrelation	Kollinearitätsstatistik: Toleranz
1	ZINS	-,031[a]	-3,920	,001	-,595	,928
	LQ	-,018[a]	-1,046	,304	-,194	,294
2	LQ	,028[b]	1,594	,123	,293	,176

a. Einflußvariablen im Modell: (Konstante), YVERF

b. Einflußvariablen im Modell: (Konstante), YVERF, ZINS

c. Abhängige Variable: CPR

Die Kriterien zur Aufnahme und zum Ausschluss einer Variable in die Gleichung können alternativ festgelegt werden (⇨ Abb. 17.15):

- *F-Wahrscheinlichkeit verwenden.* Eine Variable wird in die Gleichung aufgenommen, wenn die Wahrscheinlichkeit ihres F-Wertes kleiner ist als der in Abb. 17.19 eingetragene Aufnahmewert. Sie wird ausgeschlossen, wenn ihr F-Wahrscheinlichkeitswert größer ist als der in der Abbildung eingetragene Ausschlusswert. Der eingetragene Aufnahmewert muss kleiner als der Ausschlusswert sein.
- *F-Wert verwenden.* Eine Variable wird aufgenommen, wenn ihr F-Wert größer ist als der in Abb. 17.15 eingetragenen Aufnahme-F-Wert und ausgeschlossen, wenn er kleiner ist als der eingetragene F-Ausschlusswert.

17.3 Verwenden von Dummy-Variablen

In einer Regressionsanalyse kann man zusätzlich zu metrischen auch nominalskalierte (kategoriale) Variablen zur Erklärung einer metrischen Variable verwenden. Damit wird es möglich, auch qualitative Merkmale in das Modell einzubringen. Im Rahmen des Anwendungsbeispiels soll die Hypothese geprüft werden, ob die durch das OPEC-Kartell in den 70er Jahren verursachten schockartigen Preiserhöhungen für Rohöl den privaten Konsum der Haushalte beeinflusst haben. Es erscheint nicht unplausibel, dass durch die außerordentliche Situation, die über die zukünftige wirtschaftliche Entwicklung verunsicherten Verbraucher mit erhöhtem Sparen und damit kleinerem Konsum bei gegebener Höhe des verfügbaren Einkommens und Zinses reagiert haben. Durch Einführung einer Hilfsvariable - im angelsächsischen Dummy-Variable genannt - in das Regressionsmodell soll diese Hypothese getestet werden. Gleichzeitig kann auch geprüft werden, ob sich das bisherige Erklärungsmodell, das Schwächen hinsichtlich der Erfüllung von Modellvoraussetzungen zeigt, verbessert. Die Jahre der beiden „Ölkrisen" waren 1973-75 sowie 1978-79. Die Dummy-Variable DUM1, die die Hypothese prüfen soll, erhält in 1973-75 und 1978-79 den Wert 1 und in allen anderen Jahren den Wert 0. Die Gleichung des Modells lautet nun:

$$CPR_i = \beta_0 + \beta_1\, YVER_i + \beta_2\, ZINS_i + \beta_3\, DUM1_i + \varepsilon_i \qquad (17.39)$$

Die beiden zu prüfenden Alternativ-Hypothesen sind: H_0: $\beta_3 = 0$ und H_1: $\beta_3 < 0$. Wenn also die Verbraucher auf die „Ölschocks" in ihrem Konsumverhalten reagiert haben, so sollte in den Jahren der „Ölkrise" der Konsum um β_3 kleiner sein als man es im Vergleich zu den anderen Jahren aufgrund der Höhe von YVERF und ZINS erwarten kann. Das Ergebnis der Regressionsanalyse bezüglich der Regressionskoeffizienten ist in Abb. 17.13 zu sehen.

Der geschätzte Regressionskoeffizient $b_3 = -6{,}513$ bedeutet, dass in den Jahren der „Ölkrise" der private Konsum durchschnittlich um ca. 6,5 Mrd. DM kleiner gewesen ist als aufgrund der Höhe des verfügbaren Einkommens und des Zinses zu erwarten war. Es zeigt sich damit, dass das Vorzeichen für die Variable DUM erwartungsgemäß negativ ist. Aber der Regressionskoeffizient ist statistisch nicht gesichert („Signifikanz" = 0,196 > α = 0,05). Die Hypothese H_1 hat keine empirische Stützung erfahren. Weitere statistische Resultate werden hier nicht referiert. Es ist aber festzuhalten, dass auch weitere statistische Prüfungen des Modells zeigen, dass die Variable DUM1 nicht geeignet ist, das Regressionsmodell zu verbessern. Damit bleibt die Spezifizierung des Modells weiterhin unbefriedigend.

Tabelle 17.13. Ergebnisausgabe: Regression mit einer Dummy-Variablen

Koeffizienten[a]

	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
	B	Standardfehler	Beta	T	Signifikanz
(Konstante)	47,751	10,851		4,400	,000
YVERF	,862	,007	1,008	129,250	,000
ZINS	-4,695	1,416	-,027	-3,315	,003
DUM1	-6,513	4,907	-,011	-1,327	,196

a. Abhängige Variable

Eine Dummy-Variable kann auch verwendet werden, um einen „Strukturbruch" zu erfassen, der sich in der Veränderung eines Regressionskoeffizienten im Schätzungszeitraum ausdrückt. Das folgende Beispiel, das substanzwissenschaftlich fiktiven Charakter hat, soll diese Möglichkeit demonstrieren. Angenommen wird, dass es gute Gründe dafür gibt, dass ab den 80er Jahren die Konsumquote aus zusätzlichem verfügbaren Einkommen, also der Regressionskoeffizient für YVERF, angestiegen ist. Um diesen Bruch im Verhalten der Verbraucher zu erfassen, wird eine Dummy-Variable DUM2 eingeführt, die ab 1980 den Wert 1 und vorher den Wert 0 hat. Eine weitere Hilfsvariable, hier YVERDUM2 genannt, wird per „Transformieren" und „Berechnen" erzeugt. Sie ist definiert als YVERDUM2 = YVERF* DUM2. Der Regressionsansatz lautet nun:

$$CPR_i = \beta_0 + \beta_1\ YVER_i + \beta_2\ ZINS_i + \beta_3\ YVERDUM2_i + \varepsilon_i \qquad (17.40)$$

Aus der Gleichung ergibt sich, dass für die Jahre bis einschließlich 1979 der Koeffizient β_1 das Verbrauchsverhalten bezüglich des verfügbaren Einkommens erfasst (wegen DUM2 = 0 ist auch YVERDUM2 = 0). Das neue Verbrauchsverhalten bezüglich der Einkommensverwendung ab 1980 wird hingegen durch ($\beta_1 + \beta_2$) erfasst (wegen DUM2 = 1 ist YVERDUM2 = YVERF). In der folgenden Tabelle 17.14

wird ein Ausschnitt aus der Ergebnisausgabe für die Regressionsgleichung 17.40 aufgeführt:

Der Regressionskoeffizient der Hilfsvariable YVERDUM2 beträgt 0,0133 und ist auch statistisch gesichert. Tatsächlich ist aber die Erhöhung der Konsumquote so geringfügig, dass von einem Strukturbruch wohl keine Rede sein kann. Auch ist zu verzeichnen, dass das Modell sich durch die Einführung der Hilfsvariablen nicht wesentlich verbessert. Es ist auch möglich, explizit einen Test auf Vorliegen eines Strukturbruchs durchzuführen. Darauf kann hier aber nicht eingegangen werden.

Der Einsatz von Dummy-Variablen in der Variante DUM1 und DUM2 kann auch kombiniert werden. Immer sollte man sich aber sorgfältig versichern, ob die Verwendung einer zusätzlichen erklärenden Variablen wirklich sinnvoll ist. Ziel sollte sein, ein Modell mit möglichst wenigen erklärenden Variablen zu bilden, da dann sowohl eine Interpretation als auch Prognose mit dem Modell einfacher ist.

Tabelle 17.14. Ausschnitt aus der Ergebnisausgabe für eine Regression mit einer Dummy-Variablen

Koeffizienten

	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
	B	Standardfehler	Beta	T	Signifikanz
(Konstante)	61,871	9,495		6,516	,000
YVERF	,838	,009	,979	92,257	,000
ZINS	-4,387	1,189	-,026	-3,690	,001
YVERDUM2	1,33E-02	,004	,035	3,383	,002

17.4 Prüfen auf Verletzung von Modellbedingungen

Überprüfungen der Modellannahmen des Regressionsmodells basieren auf der Analyse der empirischen Residualwerte e_i. Basis für diese Vorgehensweise ist der Sachverhalt, dass für ein angemessenes Regressionsmodell die empirischen Residualwerte e_i ähnliche Eigenschaften haben sollen wie ε_i in der Grundgesamtheit (⇨ Gleichungen 17.11 bis 17.14). Bei den Überprüfungen bedient man sich sowohl der grafischen Analyse als auch statistischer Testverfahren. Im folgenden soll auf einige wichtige Aspekte eingegangen werden.

17.4.1 Autokorrelation der Residualwerte und Verletzung der Linearitätsbedingung

Autokorrelation der Residualwerte spielt vorwiegend bei Regressionsanalysen von Zeitreihen eine Rolle (⇨ Durbin-Watson, Kap. 17.2.2). Besteht Autokorrelation, so liegt eine sehr ernst zu nehmende Verletzung einer Modellvoraussetzung vor.

Autokorrelation der Residualwerte ist häufig eine Folge einer Fehlspezifikation der Regressionsgleichung. Dabei sind zwei Gründe zu unterscheiden:

① *Es wird eine falsche Gleichungsform angenommen.*
In der folgenden Abb. 17.16 soll gezeigt werden, dass eine falsche Gleichungsform Autokorrelation als Artefakt generiert.

Aus der Teilabbildung a) wird sichtbar, dass ein offensichtlich nichtlinearer Zusammenhang zwischen y und einer erklärenden Variablen x, der fälschlicherweise mittels einer linearen Gleichung erfasst werden soll, ein Muster der Residualwerte erzeugt, das nicht zufällig ist. Die Residualwerte e sind positiv autokorreliert: ein z.B. hoher positiver Residualwert für den Fall i führt für den Fall i + 1 ebenso zu einem hohen positiven Residualwert.

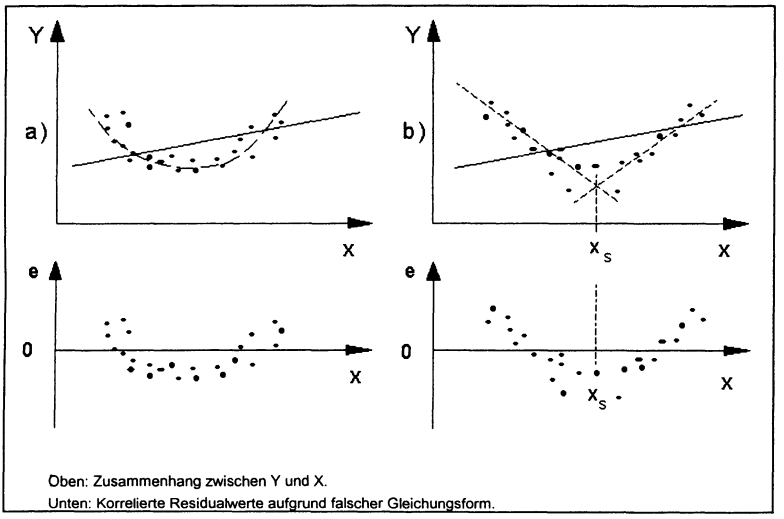


Abb. 17.16. Beispiele für entstehende Autokorrelation bei falscher Gleichungsform.

Zur Vermeidung dieses Problems bietet es sich an, die Variablen zu transformieren. In diesem Beispiel ist es sinnvoll, beide Variablen zu logarithmieren. In logarithmischer Darstellung wird der Zusammenhang linear, so dass für logarithmierte Werte eine lineare Regressionsanalyse vorgenommen werden kann und die Autokorrelation verschwindet. Für andere nichtlineare Zusammenhänge zwischen den Variablen müssen andere Transformationsformen gewählt werden.

Manchmal ist die Art des Zusammenhangs zwischen Variablen auch aus theoretischen Herleitungen bekannt. Dann bietet es sich an, auf dieser Basis eine Linearisierung durch Transformation der Variablen zu gewinnen. In Teilabbildung b) der Abb. 17.16 ist ebenfalls eine falsche Gleichungsform die Ursache für methodisch erzeugte Autokorrelation: Ein linearer Zusammenhang zwischen y und x ist zwar vorhanden, aber an einer Stelle von x ändert sich die Steigung des Zusammenhangs. Man spricht von einem „Strukturbruch" (⇨ Kap. 17.3).

Hier ist es hilfreich, den Zusammenhang der Variablen für Teilbereiche linear zu erfassen. Dabei ist es möglich, mittels einer Hilfsvariable (Dummy-Variable) beide lineare Teilstücke in einem Regressionsansatz zu schätzen (⇨ Kap. 17.3).

Aus der Abb. 17.16 wird deutlich, dass mit Hilfe von Streudiagrammen für die Residualwerte Fehlspezifikationen infolge einer falschen Gleichungsform aufgedeckt werden können. Hat man mehrere erklärende Variablen, so kann man zunächst einmal in einem Streudiagramm die Residualwerte e_i mit den Vorhersagewerten $\hat{y}_i$ auf der x-Achse darstellen. Ergänzt werden kann eine derartige Darstellung durch Streudiagramme mit jeweils den einzelnen erklärenden Variablen auf der x-Achse des Diagramms. Mit SPSS lässt sich dieses technisch ohne Mühe realisieren, indem bei der Berechnung der Regressionsgleichung zunächst die Residualwerte mittels der Option „Speichern" dem Datensatz hinzugefügt werden und dann per „Grafiken", „Streudiagramm" die Grafik erstellt wird.

② *Es fehlt mindestens eine wichtige erklärende Variable in der Gleichung.*
Auch fehlende erklärende Variable können Ursache für methodisch produzierte Autokorrelation sein. Um derartiges aufzudecken, macht es Sinn, die Residualwerte eines Regressionsansatzes mit Variablen, die vielleicht aus Signifikanzgründen bislang nicht in die Gleichung aufgenommen worden sind, auf der x-Achse in Streudiagrammen darzustellen. Falls es systematische Beziehungen zwischen den Residualwerten und einer bislang nicht aufgenommenen Variablen gibt, sollte man diese aufnehmen, um zu sehen, ob dadurch die Autokorrelation der Residualwerte verschwindet.

Zur Frage, ob Autokorrelation in den Residualwerten vorliegt, ist ein Test nach Durbin und Watson üblich (⇨ Durbin-Watson-Test in Kap. 17.2.2).

17.4.2 Homo- bzw. Heteroskedastizität

In Abb. 17.17 sind vier Muster des Verlaufs der Residualwerte e_i in Beziehung zu einer erklärenden Variable x in einem Streudiagramm dargestellt. In Teilabbildung a) wird ersichtlich, dass die Streuung der Residualwerte mit wachsendem Wert der erklärenden Variablen in etwa konstant bleibt. Dieses ist ein Indikator dafür, dass die Modellvoraussetzung der Homoskedastizität erfüllt ist (⇨ Gleichung 17.12). Im Vergleich zeigen die Teilabbildungen b), c) und d), dass die Residualwerte sich mit wachsendem Wert von x systematisch verändern. Man kann dann davon ausgehen, dass Heteroskedastizität der Residualwerte vorliegt.

Im Fall des starken Verdachts für das Vorliegen von Heteroskedastizität kann man versuchen, durch Transformation von Variablen diesen Mangel zu tilgen. Dabei kann man sich folgender Leitlinien bedienen:

- Ist σ_ε^2 proportional zu $\mu_{y/x}$ (dem Mittelwert von y bei gegebenem x), so sollte die Transformation $\sqrt{y}$ probiert werden (nur für positive Werte von x möglich).
- Ist σ_ε proportional zu $\mu_{y/x}$, so sollte eine Logarithmierung von y versucht werden.
- Ist σ_ε proportional zu $(\mu_{y/x})^2$, so ist die Transformation 1/y angebracht.

- Wenn y eine Quote oder eine Rate ist, so wird die Transformation in arc sin (Inverse einer Sinusfunktion) empfohlen.

Mit Hilfe von „Grafiken" und „Streudiagramm" lassen sich leicht Streudiagramme zur Prüfung der per „Speichern" dem Datensatz hinzugefügten Variablen auf Homoskedastizität herstellen.

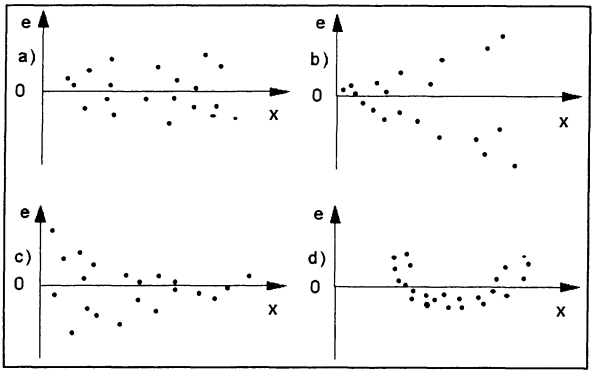


Abb. 17.17. Beispiele für Beziehungen zwischen Residualwerten und einer erklärenden Variable

17.4.3 Normalverteilung der Residualwerte

Ist die Modellbedingung der Normalverteilung verletzt, so können die statistischen Signifikanzprüfungen nicht mehr vorgenommen werden. Daher sollte man bei Verletzung der Normalverteilungsbedingung nach Möglichkeiten suchen, diese zu beheben. Auch hier kann eine Variablentransformation helfen. Bei schiefer Verteilung der Residualwerte kann man folgende Leitlinien zur Transformation der Variablen zu Rate ziehen:

- Bei positiver Schiefe ist häufig eine logarithmische Transformation der y-Variablen hilfreich.
- Bei negativer Schiefe wird eine quadratische Transformation empfohlen.

Die Prozedur Regression bietet per Option „Grafiken" die Möglichkeit zur grafischen Darstellung der Residualwerte im Vergleich zur Normalverteilung. Im Menü „Explorative Datenanalyse" können Tests auf Normalverteilung der Residualwerte vorgenommen werden (⇨ Kap. 9.3.2).

17.4.4 Multikollinearität

Multikollinearität, also eine Korrelation der erklärenden Variablen, kann verschiedene Grade annehmen (⇨ Kollinearitätsdiagnose in Kap. 17.2.2). Sind zwei erklärende Variablen vollständig (mathematisch) miteinander verbunden, so lassen sich die Regressionskoeffizienten nicht mehr mathematisch bestimmen. Dieser Fall ist andererseits aber kein Problem, da sowohl die eine als auch die andere Variable

gleich gut als Erklärungsvariable geeignet ist. Problematischer wird es, wenn - was in der Praxis auch viel häufiger vorkommt - zwar kein mathematisch vollständiger Zusammenhang zwischen den Variablen besteht, aber ein sehr hoher. Folge ist, dass die Regressionskoeffizienten von Stichprobe zu Stichprobe stark fluktuieren. Schon kleine Veränderungen in den Daten (z.B. Löschen von Fällen) können die Regressionskoeffizienten gravierend verändern. Auch sind die Standardfehler der Regressionskoeffizienten hoch. Des weiteren sind die Betakoeffizienten (⇨ Kap. 17.2.1) nicht mehr aussagekräftig. In solchen Fällen ist zu überlegen, ob aus den sehr hoch korrelierenden erklärenden Variablen nicht eine zusammenfassende Indexvariable konstruiert werden kann, die im Regressionsansatz Verwendung findet. Entfernen einer Variablen ist keine Lösung, da dieses zu verzerrten Regressionskoeffizienten für die anderen Variablen führt.

17.4.5 Ausreißer und fehlende Werte

Ausreißer. Fälle mit ungewöhnlichen Werten für erklärende Variablen können einen starken Einfluss auf die Ergebnisse der Regressionsanalyse nehmen. In Streudiagrammen zur Darstellung des Zusammenhangs zwischen der abhängigen und einer erklärenden Variable erscheinen solche Fälle als „Ausreißer", die dem generellen Muster des sichtbaren Zusammenhangs nicht entsprechen. SPSS bietet eine Fülle von Hilfen an, den Einfluss und die Bedeutung von „Ausreißern" zu beurteilen (⇨ Kap. 17.2.3 und 17.2.4).

Fehlende Werte. Bei fehlenden Werten von Variablen in Datensätzen sollte man mit Vorsicht walten. Zunächst sollte man prüfen, ob das Muster der fehlenden Werte zufällig ist oder ob es einen Zusammenhang zu der Variable mit fehlenden Werten oder anderen Variablen des Erklärungsmodells gibt. Bei Nichtzufälligkeit sollten Regressionsergebnisse unter Vorbehalt interpretiert werden. Im schlimmsten Fall sind die Daten für eine Analyse sogar unbrauchbar. Konzentrieren sich die Fälle mit fehlenden Werten auf wenige Variablen, so muss man sich überlegen, ob man nicht besser auf diese Variablen verzichtet. Bei Wahl der Option „Fallweiser Ausschluss" besteht die Gefahr, dass zu viele Fälle ausgeschlossen werden, so dass zu wenig übrig bleiben. Bei Wahl der Option „Paarweiser Ausschluss" besteht andererseits die Gefahr, dass aufgrund jeweils anderer Fälle und verschiedener Fallzahlen Inkonsistenzen entstehen.

18 Modelle zur Kurvenanpassung

18.1 Modelltypen und Kurvenformen

Bei der Statistik-Prozedur „Kurvenanpassung" geht es um die Frage der Vorhersage einer Variable y durch eine andere Variable x. Dabei sind zwei grundlegend verschiedene Modelltypen zu unterscheiden:

- *Regressionsmodell.* Die Entwicklung einer Variable y wird durch eine Erklärungsvariable x vorhergesagt. In Ergänzung der linearen Regressionsanalyse steht hier die Frage der Auswahl einer besten Kurvenform zur Vorhersage von y im Mittelpunkt der Analyse.
- *Trendmodell.* Die Entwicklung einer Variable y wird lediglich im Zeitablauf analysiert und durch die Zeitvariable x vorhergesagt. Auch hier geht es um die Frage, welche Kurvenform zur Vorhersage am besten geeignet ist.

Tabelle 18.1. Gleichungen der Modelle zur Kurvenanpassung

Modell	Gleichung	Gleichung linearisiert
Linear	$y = b_0 + b_1 x$	
Logarithmisch	$y = b_0 + b_1 \ln(x)$	
Invers	$y = b_0 + b_1 / x$	
Quadratisch	$y = b_0 + b_1 x + b_2 x^2$	
Kubisch	$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$	
Zusammengesetzt	$y = b_0 (b_1)^x$	$\ln(y) = \ln(b_0) + \ln(b_1) x$
Exponent	$y = b_0 x^{b_1}$	$\ln(y) = \ln(b_0) + b_1 \ln(x)$
S	$y = e^{(b_0 + b_1 / x)}$	$\ln(y) = b_0 + b_1 / x$
Wachstum	$y = e^{(b_0 + b_1 x)}$	$\ln(y) = b_0 + b_1 x$
Exponentiell	$y = b_0 e^{b_1 x}$	$\ln(y) = \ln(b_0) + b_1 x$
Logistisch	$y = 1 / \left[1/c + b_0 (b_1)^x\right]$	$\ln(1/y - 1/c) = \ln(b_0) + \ln(b_1) x$

b_0, b_1, b_2, b_3 = zu schätzende Koeffizienten

x = unabhängige Variable oder die Zeit mit x = 0,1,2,...

ln = natürlicher Logarithmus (zur Basis e ≈ 2,7183)

c = oberer Grenzwert des logistischen Modells

Für beide Modelltypen kann aus elf Kurvenformen ausgewählt werden. Die Gleichungen der Kurvenformen sind in Tabelle 18.1 aufgeführt. Sofern eine Gleichung nicht direkt geschätzt werden kann (weil sie nichtlinear ist), wird in der rechten Spalte die (lineare) Schätzungsform aufgeführt. Die Schätzmethode zur Bestimmung der Koeffizienten b_0 bis b_3 ist in allen Fällen die Methode der kleinsten Quadrate (⇨ Kap. 17.1.1). Für das logistische Modell kann zur Schätzung der Koeffizienten ein oberer Grenzwert c für die Variable y vorgegeben werden. Dieser muss größer als der maximale Wert von y sein. Verzichtet man auf die Vorgabe, so wird $1/c = 0$, d.h. c = unendlich gesetzt.

18.2 Modelle schätzen

Zur anwendungsorientierten Erläuterung sollen für die Entwicklung der Arbeitslosenquote in der Bundesrepublik von 1960-90 beispielhaft zwei Trendkurven ausgewählt und angepasst werden (Datensatz MAKRO.SAV). Nach Laden des Datensatzes geht man wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Regression", „Kurvenanpassung...". Es öffnet sich die in Abb. 18.1 dargestellte Dialogbox „Kurvenanpassung".
- ▷ Aus der Quellvariablenliste wird die Variable ALQ (Arbeitslosenquote) durch Markieren und Klicken auf den Pfeilschalter in das Feld „Abhängige Variable(n)" übertragen.
- ▷ Da die Arbeitslosenquote nicht durch eine Erklärungsvariable im Sinne eines Regressionsmodells, sondern durch die Zeit in einem Trendmodell vorhergesagt werden soll, wird als „Unabhängige Variable" „Zeit" gewählt. Bei Wahl von „Variable" müsste man eine erklärende Variable eines Regressionsmodells in das Variablenfeld übertragen.
- ▷ Aus den verfügbaren Modellen werden nun „Kubisch" und „Logistisch" ausgewählt. Für das logistische Modell wird „Obergrenze" auf 10 festgelegt.

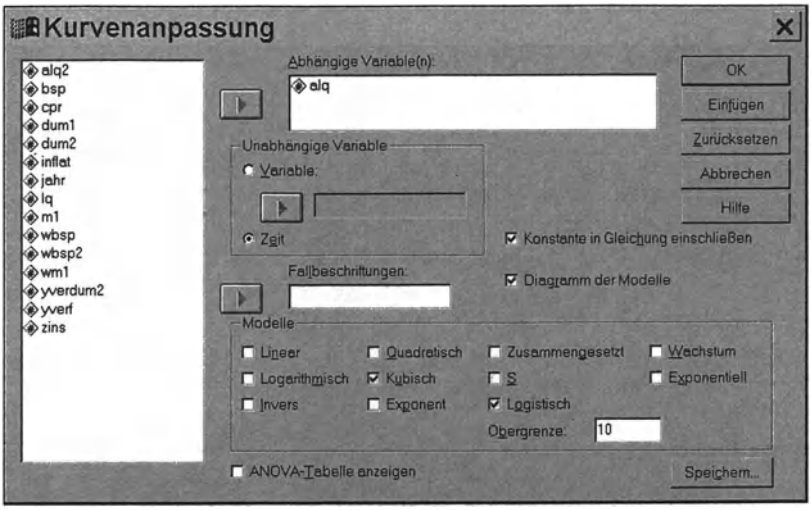


Abb. 18.1. Dialogbox „Kurvenanpassung"

In Tabelle 18.2 werden die statistischen Ergebnisse der Kurvenanpassung dokumentiert. Es bedeutet jeweils:

- *Mth.* Es werden die Methode „CUB" (Cubic) und „LGS" (Logistic) zur Kurvenanpassung verwendet.
- *Rsq.* Entspricht dem Bestimmtheitsmaß R^2 (⇨ Kap. 17.1.1). Das kubische Modell hat mit $R^2 = 0{,}904$ einen höheren Anteil der erklärten Varianz als das logistische. Der Grund ist darin zu sehen, dass durch die Schätzung von vier Koeffizienten gegenüber von zwei eine bessere Anpassung erreicht wird. Zwischen dem Vorteil einer besseren Anpassung und dem Nachteil eines komplexeren Modells infolge der größeren Anzahl von zu schätzenden Koeffizienten ist im Einzelfall abzuwägen. Bei der Wahl einer Anpassungskurve, die für Prognosen verwendet werden soll, sollte man sich nicht allein auf R^2 stützen, sondern sich auch davon leiten lassen, welche Kurve aus theoretischen Erwägungen zu bevorzugen ist. Ebenfalls ist es bei der Entscheidung für ein Modell hilfreich, die Residualwerte - die Abweichungen der beobachteten Werte von den geschätzten Werten von y - zu untersuchen.
- *d.f.* Anzahl der Freiheitsgrade (degrees of freedom). Die Anzahl der Freiheitsgrade im kubischen Modell ist um zwei kleiner, da zwei Koeffizienten mehr zu schätzen sind.
- *F.* F-Wert für den F-Test. (⇨ Kap. 17.2.1).
- *Sigf.* Signifikanzniveau für den F-Test (⇨ Kap. 17.2.1).
- *Upper bound.* Im logistischen Modell vorgegebener Wert für die Obergrenze von y.
- b_j, $j = 0,1,2,3$. Die geschätzten Koeffizienten des Modells.

Die Vorhersagegleichungen der Modelle lauten:

Kubisch: $y = 2{,}5496 - 0{,}7298x + 0{,}0686x^2 - 0{,}0013x^3$

Logistisch: $y = 1/\left[1/10 + 2{,}8028(0{,}8460)^x\right]$

Tabelle 18.2. Zusammenfassende statistische Angaben zur Modellanpassung

Independent: Time

Dependent	Mth	Rsq	d.f.	F	Sigf	Upper bound	b0	b1	b2	b3
ALQ	CUB	,904	27	85,22	,000		2,5496	-,7298	,0686	-,0013
ALQ	LGS	,814	29	127,30	,000	10,000	2,8028	,8460		

Wahlmöglichkeiten:

① *Konstante in Gleichung einschließen.* Es wird ein konstantes Glied in der Gleichung geschätzt. Diese Voreinstellung kann durch Mausklick deaktiviert werden.

② *ANOVA-Tabelle anzeigen.* Für jedes Modell wird eine zusammenfassende Tabelle zur varianzanalytischen Prüfung des Zusammenhangs der beiden Variablen ausgegeben. Sie entspricht der aus der Regressionsanalyse bekannten Tabelle (⇨ Tabelle 17.1 in Kap. 17.2.1).

③ *Diagramm der Modelle.* In einer Grafik werden die Werte der y-Variable gegen die Werte der x-Variablen geplottet.

④ *Speichern von vorhergesagten und Residualwerten.* Für jedes der geschätzten Modelle können bis zu vier bei der Modellschätzung entstehende neue Variablen zur weiteren Verarbeitung gespeichert werden. Zur Speicherung wird auf die Schaltfläche „Speichern" geklickt. Es öffnet sich dann die in Abb. 18.2 dargestellte Dialogbox „Kurvenanpassung: Speichern".

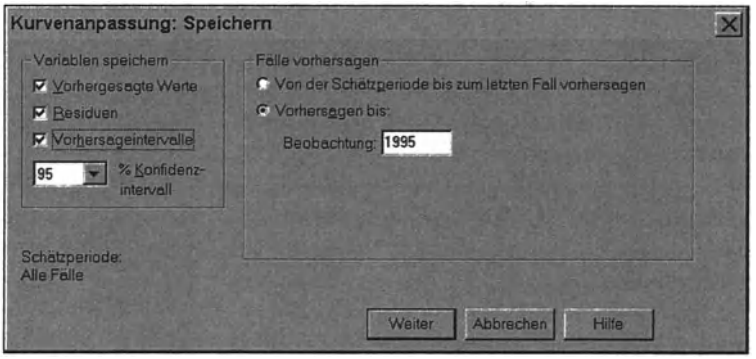


Abb. 18.2. Dialogbox „Kurvenanpassung: Speichern"

In zwei Gruppen stehen folgende Auswahloptionen bereit:

- *Variablen speichern.* Es können folgende Variable gespeichert werden:
  - *Vorhergesagte Werte.* Die Vorhersagewerte (Schätzwerte) des Modells.
  - *Residuen.* Abweichungen zwischen tatsächlichen und Vorhersagewerten.
  - *Vorhersageintervalle.* Es kann zwischen dem 95- (voreingestellt), 90- und 99-%-Konfidenzintervall für die vorhergesagten Werte gewählt werden.
- *Fälle vorhersagen.* Zur Vorhersage von Werten kann man zwischen folgenden Optionen wählen:
  - *Von der Schätzperiode bis zum letzten Fall vorhersagen.* Die Vorhersagewerte werden für die Fälle berechnet, die für die Schätzung der Gleichung zugrundegelegt worden sind. Mit der Befehlsfolge „Daten", „Fälle auswählen..." kann vorher aus den verfügbaren Fällen eine Auswahl für die Schätzung erfolgen.
  - *Vorhersagen bis: Beobachtung:.* Diese Option steht nur für das Trendmodell zur Verfügung. Mit ihr kann der Vorhersagezeitraum über das Ende der Zeitreihe hinaus verlängert werden. Für das Beispiel zur Vorhersage der Arbeitslosenquote wurde diese Option gewählt und in das Eingabefeld „Beobachtung" 1995 eingegeben. Diese Jahresangabe ist möglich, da für den Datensatz MAKRO.SAV mit der Befehlsfolge „Daten", „Datum definieren" die Datenreihen als Jahres-Zeitreihen mit 1960 als erstem Wert definiert worden sind. Für undatierte Daten hätte man in das Eingabefeld „Beobachtung" 35 eingeben müssen.

Dem Datensatz werden acht Datenreihen hinzugefügt. In Tabelle 18.3 wird die diesbezügliche Meldung im Ausgabefenster dokumentiert. FIT_1 und FIT_2 sind die Vorhersagewerte, ERR_1 und ERR_2 die Residualabweichungen, LCL_1 und LCL_2 die unteren (lower confidence limit), UCL_1 und UCL_2 die oberen (upper confidence limit) Konfidenzgrenzen für das kubische und logistische Modell.

Dem Datensatz sind für die Vorhersage- und die Konfidenzbereichswerte für 1991 bis 1995 fünf Fälle hinzugefügt worden („5 new cases have been added").

Tabelle 18.3. Speichern von vorhergesagten Residual- und Konfidenzbereichswerten

```
Name        Label

FIT_1       Fit for ALQ from CURVEFIT, MOD_1 CUBIC
ERR_1       Error for ALQ from CURVEFIT, MOD_1 CUBIC
LCL_1       95% LCL for ALQ from CURVEFIT, MOD_1 CUBIC
UCL_1       95% UCL for ALQ from CURVEFIT, MOD_1 CUBIC
FIT_2       Fit for ALQ from CURVEFIT, MOD_1 LGSTIC
ERR_2       Error for ALQ from CURVEFIT, MOD_1 LGSTIC
LCL_2       95% LCL for ALQ from CURVEFIT, MOD_1 LGSTIC
UCL_2       95% UCL for ALQ from CURVEFIT, MOD_1 LGSTIC

5 new cases have been added.
```

Die dem Datensatz hinzugefügten Variablen können weiterverarbeitet werden. So können z.B. wie in Abb. 18.3 die tatsächlichen und die mit den beiden Modellen vorhergesagten Werte in einer Grafik dargestellt werden (als Mehrfachliniendiagramm).

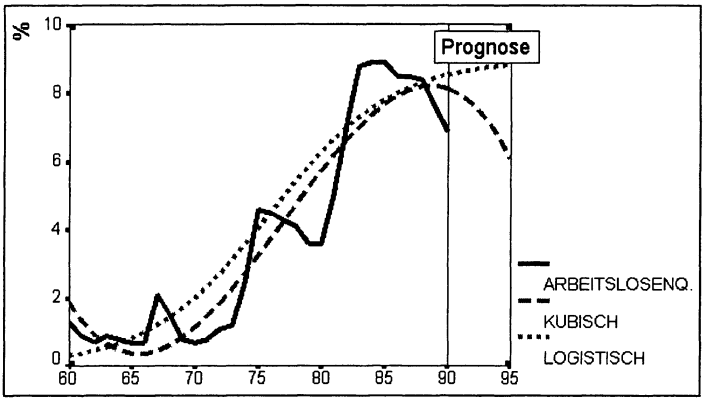


Abb. 18.3. Beobachtete Arbeitslosenquote und geschätzte Trendwerte mit Prognose ab 1990

19 Clusteranalyse

19.1 Theoretische Grundlagen

Einführung. Die Clusteranalyse ist eine multivariate statistische Methode mit der Zielsetzung, Objekte bzw. Personen (Fälle) - für die mehrere Merkmale (Variablen) vorliegen - derart in Gruppen (Cluster) zu ordnen, dass in einem Cluster hinsichtlich der Variablen möglichst gleichartige bzw. ähnliche Objekte zusammengefasst werden. Dem Anspruch, Cluster mit möglichst homogenen Objekten zu bilden steht gegenüber, dass die gebildeten Cluster sich möglichst stark voneinander unterscheiden. Bevor die eigentliche Clusteranalyse angewendet wird, muss die Ähnlichkeit der Objekte mit Hilfe von Distanz- oder Ähnlichkeitsmaßen berechnet werden. Ergebnis dieses Berechnungsschrittes ist eine Matrix, in der für alle Objektpaare das gewählte Distanz- oder Ähnlichkeitsmaß steht (⇨ Kap. 16.3). Die Clusteranalyse hat anschließend die Aufgabe, auf der Basis dieser Matrix die Gruppierung in Cluster vorzunehmen.

Die Clusteranalyse findet in vielen Bereichen Anwendung. So werden z.B. in der Marktforschung Städte (oder andere regionale Gebiete) in möglichst homogene Gruppen zusammengefasst zum Testen von unterschiedlichen Marketingstrategien für vergleichbare Städte. Oder es werden Personen auf der Basis erhobener Merkmalsvariablen über Einkommen, Bildung, Interessen und Einstellungen zu Käuferschichten geclustert. In der Mediaforschung werden Personen, deren Sendungsvorlieben, Sehgewohnheiten und weitere Merkmale erhoben wurden, zu Zuschauertypen (z.B. „Informationsorientierte", „Kulturorientierte", „TV-Abstinenzler" etc.) zusammengefasst.

Neben dieser primären Form der Anwendung einer Clusteranalyse (auch *objektorientierte* genannt) kann die Clusteranalyse auch für eine *variablenorientierte* Clusterung eingesetzt werden. Dann besteht die Aufgabe darin, mehrere Variablen in Variablengruppen einzuordnen. In einer Variablengruppe sollen jeweils ähnliche Variablen (korrelierte Variablen) zusammengefasst werden.

Zur Clusterung werden von SPSS zwei grundlegende Verfahren angeboten: die *Clusterzentrenanalyse* und die *hierarchische Clusteranalyse*.

Hierarchische Clusteranalyse. Hierbei handelt es sich um eine Gruppe von Verfahren. Sie eignen sich nicht für eine hohe Fallanzahl, da sie hohe Anforderungen an Speicherplatz und Rechenzeit voraussetzen. Allen diesen Verfahren ist gemeinsam, dass die Clusterbildung in nacheinander folgenden Schritten abläuft: im ersten Schritt bildet jedes Objekt ein Cluster, im zweiten Schritt werden zwei Objekte zu einem Cluster vereinigt, im dritten Schritt wird entweder diesem ersten Cluster

ein Objekt hinzugefügt oder es werden zwei weitere Objekte zu einem neuen Cluster vereinigt. In den weiteren Schritten werden entweder Objekte zu schon gebildeten Clustern hinzugefügt, so dass ein neues Cluster entsteht oder es werden zwei in vorherigen Stufen gebildete Cluster zu einem neuen Cluster vereinigt bis schließlich im letzten Schritt alle Objekte in einem einzigen Cluster enthalten sind. Auf diese Weise entstehen Stufen (Hierarchien) der Clusterbildung (agglomerative Clusterung).

Auf jeder Stufe werden das Objektpaar (bzw. das Objekt und das Cluster bzw. das Clusterpaar) zu einem neuen Cluster vereinigt, das die kleinste Distanz (bzw. die größte Ähnlichkeit) hat. Daher müssen ausgehend von der Matrix der Distanz- oder Ähnlichkeitsmaße aller Objekte (⇨ Kap. 16.3) auf allen Stufen Distanzen (bzw. Ähnlichkeiten) zwischen Objekten und Clustern (bzw. zwischen allen Clusterpaaren) berechnet werden zum Aufbau einer jeweils neuen Distanz- (Ähnlichkeits-)-Matrix. Die hierarchischen Clustermethoden unterscheiden sich darin, wie die Distanz (Ähnlichkeit) von Clustern (bzw. Objekten) berechnet wird:

- *Linkage zwischen den Gruppen* (average linkage between groups). Die Distanz zwischen zwei Clustern (bzw. einem Objekt und einem Cluster) berechnet sich als ungewichtetes arithmetische Mittel der Distanzen zwischen allen Objektpaaren der beiden Cluster. Es werden dabei nur die Objektpaare berücksichtigt, bei denen ein Objekt aus dem einen und das andere aus dem anderen Cluster kommt.
- *Linkage innerhalb der Gruppen* (average linkage within groups). Bei dieser Methode wird die Distanz zwischen Clustern (bzw. einem Objekt und einem Cluster) ebenfalls als arithmetische Mittel der Distanzen von Objektpaaren berechnet. Im Unterschied zu oben werden aber alle Objektpaare (auch die innerhalb der beiden Cluster) einbezogen.
- *Nächstgelegener Nachbar* (nearest neighbor bzw. single linkage). Als Distanz zwischen zwei Clustern (bzw. einem Objekt und einem Cluster) wird die Distanz zwischen zwei Objekten der beiden Cluster gewählt, die am kleinsten ist.
- *Entferntester Nachbar* (complete linkage). Als Distanz zwischen zwei Clustern (bzw. einem Objekt und einem Cluster) wird die Distanz zwischen zwei Objekten der beiden Cluster gewählt, die am größten ist.
- *Zentroid-Clustering.* Die Distanz zwischen zwei Clustern (bzw. einem Objekt und einem Cluster) wird auf jeder Stufe als Distanz zwischen den Zentren der Clusterpaare berechnet. Das Zentrum (Zentroid) eines Clusters ist durch die arithmetischen Mittel der Variablen für die Objekte innerhalb eines Clusters gegeben. Das Zentrum eines Clusters kann man sich als ein fiktives Objekt des Clusters vorstellen, das zum Repräsentanten des Clusters wird. Zur Berechnung des Zentrums von zwei vereinigten Clustern wird das gewichtete arithmetische Mittel der Zentren der individuellen Cluster berechnet. Dabei wird mit der Größe der Cluster (Anzahl der Objekte in den Clustern) gewichtet.
- *Median-Clustering.* Hier handelt es sich um eine Variante des Zentroid-Clustering. Der Unterschied liegt darin, dass bei der Berechnung des Zentrums keine Gewichtung vorgenommen wird (ein einfaches arithmetisches Mittel der Zentren entspricht dem Median der Zentren).

- *Ward-Methode.* Im Unterschied zu den anderen Methoden werden bei jedem Schritt nicht die Clusterpaare mit der kleinsten Distanz (bzw. größten Ähnlichkeit) fusioniert. Es werden vielmehr Cluster (bzw. Cluster und Objekte) mit dem Ziel vereinigt, den Zuwachs für ein Maß der Heterogenität eines Clusters zu minimieren. Als Maß für die Heterogenität wird die Summe der quadrierten Euklidischen Distanzen (auch Fehlerquadratsumme genannt) der Objekte zum Zentrum des Clusters (Zentroid) gewählt. Auf jeder Stufe wird also das Clusterpaar fusioniert, das zum kleinsten Zuwachs der Fehlerquadratsumme im neuen Cluster führt.

Die Methoden Linkage zwischen den Gruppen, Linkage innerhalb der Gruppen, Nächstgelegener Nachbar sowie Entferntester Nachbar können sowohl für Distanz- als auch Ähnlichkeitsmaße verwendet werden. Zentroid-Clustering und Median-Clustering sind nur für die quadrierte Euklidische Distanz sinnvoll.

Clusterzentrenanalyse (K-Means). Dieses Clusterverfahren eignet sich nur für metrische Variablen. Es verwendet als Distanzmaß die Euklidische Distanz (⇨ Kap. 16.3). Im Unterschied zu den hierarchischen Methoden ist bei diesem Verfahren die Anzahl der zu bildenden Cluster vorzugeben. Das Verfahren hat dann die Aufgabe, eine optimale Zuordnung der Objekte zu den Clustern vorzunehmen. Dieses geschieht in iterativen Schritten. Ausgehend von einer Anfangslösung der Zuordnung der Objekte zu Clustern wird in nachfolgenden Schritten eine bessere Gruppierung (hinsichtlich der Zielsetzung, homogene Cluster zu erhalten) angestrebt. Dabei können Objekte (im Unterschied zu hierarchischen Verfahren) die schon einem Cluster zugeordnet sind, diesem Cluster wieder entnommen und einem anderen Cluster zugeordnet werden. Ausgehend von der Anfangslösung wird analog dem Zentroid-Verfahren das Zentrum der Cluster berechnet. Danach werden alle Objekte derart in die Cluster eingruppiert, dass sie die kleinste Euklidische Distanz zum Zentrum der Cluster haben. Nach dieser Umordnung der Objekte werden die Zentren der Cluster erneut berechnet und die Objekte erneut umgruppiert. Diese iterativen Schritte setzen sich fort bis eine optimale Clusterlösung gefunden wird. Mit diesem Verfahren wird (wie im Modellansatz von Ward) die Streuungsquadratsumme innerhalb der Cluster minimiert.

Der Vorteil dieser Clustermethode gegenüber der hierarchischen Clusterung besteht darin, dass sie nicht so viel Hauptspeicherplatz (RAM) benötigt und schneller ist und daher auch bei sehr großen Datensätzen angewendet werden kann. Der Grund dafür ist, dass keine Distanzen zwischen allen Paaren von Fällen berechnet werden müssen. Diesem Vorteil stehen aber Nachteile gegenüber: die Anzahl der Cluster muss vor Anwendung des Verfahrens bekannt sein; es ist im Vergleich zur den hierarchischen weniger flexibel, da bei diesen je nach Messniveau der Variablen mehrere Distanz- oder Ähnlichkeitsmaße zur Auswahl stehen.

Zweckmäßig ist es, mit einer hierarchischen Methode zunächst die Anzahl der Cluster zu bestimmen (⇨ Kap. 19.2.1) und dann mit der Clusterzentrenanalyse (⇨ Kap. 19.2.2) die Clusterung zu verbessern. Bei sehr großen Datensätzen bietet es sich an, die Anzahl der Cluster anhand einer Zufallsstichprobe des großen Datensatzes mit einer hierarchischen Methode zu bestimmen.

19.2 Praktische Anwendung

19.2.1 Anwendungsbeispiel zur hierarchischen Clusteranalyse

Es sollen Ortsteile in Hamburg geclustert werden. Dafür werden die schon in Kap. 16.3 zur Berechnung von Distanzen verwendeten Daten über Ortsteile im Hamburger Bezirk Altona genutzt (Datei ALTONA.SAV). Es handelt sich dabei um vier metrische (intervallskalierte) Variable, die als Indikatoren für die soziale Struktur und für die Verdichtung anzusehen sind: Anteil der Arbeiter in %, Mietausgaben je Person, Bevölkerungsdichte, Anteil der Gebäude mit bis zu zwei Wohnungen. Wegen des ungleichen Werteniveaus dieser Variablen werden zur Distanzmessung die in z-Werte transformierte Variable (ZARBEIT, ZMJEP, ZBJEHA, ZG2W) verwendet (⇨ Kap. 16.3).

Als Clusterverfahren soll das Zentroid-Verfahren mit der quadratischen Euklidischen Distanz als Distanzmaß eingesetzt werden. Als Ergebnis erweist sich, dass eine Clusterung der Ortsteile Altonas in drei Cluster ein gute Lösung darstellt. Da aber die Ergebnisausgaben dieses Beispiels mit insgesamt 26 Ortsteilen sehr groß sind, beschränken wir uns hier auf die Darstellung der letzten beiden Cluster (es handelt sich um die letzten 11 Fälle der Datei ALTONA.SAV; die in der Datei ALTONA1.SAV gespeichert sind). Nach Laden der Datei ALTONA1.SAV gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Analysieen", „Klassifizieren ▷„ „Hierarchische Cluster...". Es öffnet sich die in Abb. 19.1 dargestellte Dialogbox.
- ▷ Übertragen Sie die Variablen ZARBEIT, ZMJEP, ZBJEHA, ZG2W aus der Quellvariablenliste in das Feld „Variable(n)". Um die Clusterung auf Basis der z-Werte vorzunehmen, kann man prinzipiell hier auch die Originalvariable ARBEIT, MJEP, BJEHA und ZG2W übertragen und dann im Dialogfeld „Hierarchische Clusteranalyse: Methode" im Feld „Standardisieren" von „Werte transformieren" z-Werte anfordern. Hier verbietet sich diese Vorgehensweise, weil die Clusterung nur für 11 Fälle der Datei ALTONA.SAV dargestellt wird.
- ▷ Zur Fallbeschriftung in der Ergebnisausgabe übertragen Sie die Variable ORTN (Ortsteilname) in das Eingabefeld „Fallbeschriftung".
- ▷ Im Feld „Cluster" wählen Sie „Fälle", da die Ortsteile geclustert werden sollen.
- ▷ Klicken Sie nun auf die Schaltfläche „Methode". Es öffnet sich die in Abb. 19.2 dargestellte Dialogbox. Wählen Sie nun die gewünschte Cluster-Methode für den zu verarbeitenden Datentyp aus. In diesem Anwendungsfall werden metrische Variablen zugrundegelegt und als Clusterverfahren soll das Zentroid-Verfahren für quadratische Euklidische Distanzen eingesetzt werden. Da aus oben genannten Gründen die z-Werte der Variablen in das Feld „Variable(n)" der Abb. 19.1 übertragen wurden, wird folglich die Standardeinstellung „keine" für „Standardisieren" gewählt.

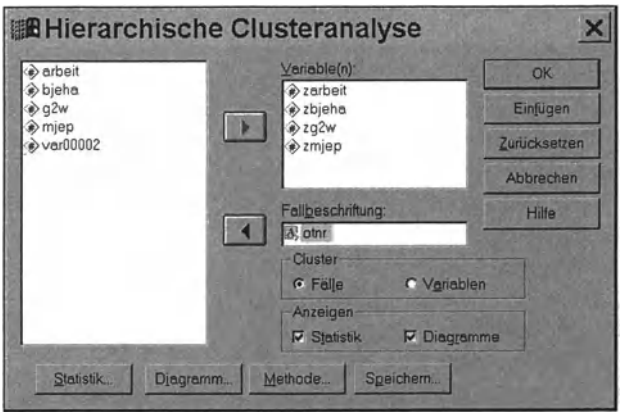


Abb. 19.1. Dialogbox „Hierarchische Clusteranalyse"

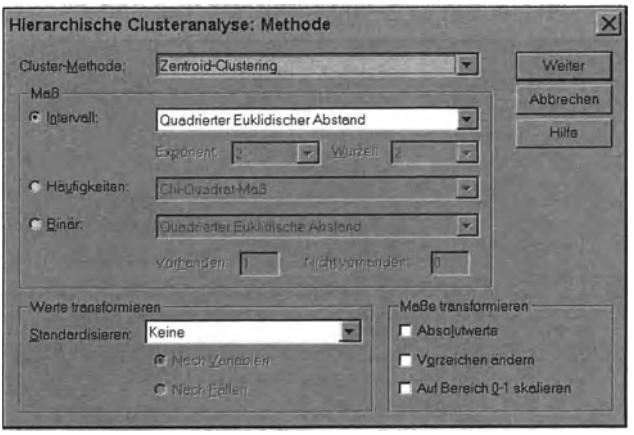


Abb. 19.2. Dialogbox „Hierarchische Clusteranalyse: Methode"

Die in Abb. 19.1 und 19.2 gewählten Einstellungen führen zu folgenden Ergebnissen. In Tabelle 19.1 ist die Ergebnisausgabe „Zuordnungsübersicht" zu sehen. In einzelnen Schritten wird in der Spalte „Zusammengeführte Cluster" aufgezeigt, welche Ortsteile bzw. Cluster (d.h. schon zusammengeführte Ortsteile) jeweils in einzelnen Schritten zu einem neuen Cluster zusammengeführt werden. Im ersten Schritt wird Fall 3 (Othmarschen) und Fall 8 (Blankenese 2) zu einem Cluster zusammengeführt. Dieses Cluster behält den Namen des Falles 3. Im Schritt 2 und 3 werden zum Cluster 3 (Othmarschen und Blankenese 2) die Fälle 7 (Blankenese 1) und 6 (Nienstedten) hinzugefügt. Im Schritt 4 werden die Fälle 9 (Iserbrook) und 10 (Sülldorf) zusammengeführt etc. Im vorletzten Schritt sind die 11 Ortsteile in zwei Cluster aufgeteilt (Othmarschen, Blankenese 2, Blankenese 1, Nienstedten, Flottbek und Rissen einerseits sowie Bahrenfeld 3, Osdorf, Iserbrook, Lurup, und

Sülldorf andererseits). Im zehnten Schritt werden dann diese beiden zu einem Cluster zusammengeführt, so dass alle Fälle ein Cluster bilden.

In der Spalte „Koeffizienten" wird die quadratische Euklidische Distanz aufgeführt. Diese entspricht in den Schritten, in denen zwei Ortsteile zusammengeführt werden (Schritt 1 und Schritt 4) der Distanz zwischen diesen Ortsteilen. Nach einer Fusion von Ortsteilen zu einem Cluster wird die Distanzmatrix neu berechnet. Bei Anwendung des Zentroid-Verfahrens geschieht dieses auf der Basis des Zentroids des Clusters (⇨ Kap. 16.3). Der Koeffizient steigt von Schritt zu Schritt zunächst kontinuierlich an und macht von Schritt 9 auf 10 einen großen Sprung. Diese Sprungstelle kann als Indikator für die sinnvollste Clusterlösung dienen. Danach ist es sinnvoll, die Clusterlösung im neunten Schritt zu wählen, die die elf Ortsteile in zwei (oben aufgeführten) Cluster ordnet.

In den Spalten „Erstes Vorkommen des Clusters" und „Nächster Schritt" wird dargelegt, in welchen Schritten es zur Fusion von Fällen und Clustern zu schon bestehenden Clustern kommt. So wird z.B. in Schritt 1 (in dem Fall 3 und 8 fusioniert werden) angeführt, dass in Schritt 2 diesem Cluster ein Fall bzw. Cluster (nämlich Fall 7) hinzugefügt wird. In Schritt 2 wird – wie oben ausgeführt – dem Cluster 3 (bestehend aus Fall 3 und 8) der Fall 7 hinzugefügt. Daher wird unter „Cluster 1" verbucht, dass das Cluster 3 in Schritt 1 entstanden ist. In „Nächster Schritt" wird Schritt 3 aufgeführt, weil dem Cluster 3 (nunmehr bestehend aus Fall 3, 8 und 7) in Schritt 3 der Fall 6 hinzugefügt wird.

Tabelle 19.1. Ergebnisausgabe „Zuordnungsübersicht"

Zuordnungsübersicht

Schritt	Zusammengeführte Cluster		Koeffizienten	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	8	3,668E-02	0	0	2
2	3	7	8,556E-02	1	0	3
3	3	6	,249	2	0	8
4	9	10	,269	0	0	6
5	1	5	,341	0	0	6
6	1	9	,448	5	4	9
7	2	11	,532	0	0	8
8	2	3	,900	7	3	10
9	1	4	1,175	6	0	10
10	1	2	6,421	9	8	0

Im vertikalen *Eiszapfendiagramm* (Tabelle 19.2) werden die Clusterlösungen der einzelnen Hierarchiestufen grafisch dargestellt. Im Fall nur eines Clusters sind natürlich alle 11 Ortsteile vereinigt. Bei zwei Clustern sind Nienstedten, Blankenese 1 und 2, Othmarschen, Rissen und Flottbek einerseits sowie Lurup, Sülldorf, Iserbrook, Osdorf und Bahrenfeld 3 andererseits in den Clustern vereinigt. Im Fall von drei Clustern bildet Lurup und im Fall von 4 Clustern Rissen und Flottbek ein weiteres Cluster.

Tabelle 19.2. Ergebnisausgabe „Vertikales Eiszapfendiagramm"

Vertikales Eiszapfendiagramm

Anzahl der Cluster	Fall																				
	6:Nienstedten		7:Blanken. 1		8:Blanken. 2		3:Othmarschen		11:Rissen		2:Flottbek		4:Lurup		10:Sülldorf		9:Iserbrook		5:Osdorf		1:Bahrenf. 3
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X		X		X	X	X	X	X	X	X
4	X	X	X	X	X	X	X		X	X	X		X		X	X	X	X	X	X	X
5	X	X	X	X	X	X	X		X		X		X		X	X	X	X	X	X	X
6	X	X	X	X	X	X	X		X		X		X		X	X	X		X	X	X
7	X	X	X	X	X	X	X		X		X		X		X	X	X		X		X
8	X	X	X	X	X	X	X		X		X		X		X		X		X		X
9	X		X	X	X	X	X		X		X		X		X		X		X		X
10	X		X		X	X	X		X		X		X		X		X		X		X

Wahlmöglichkeiten.

① *Statistik.* Nach Klicken auf die Schaltfläche „Statistik..." (Abb. 19.1) öffnet sich die in Abb. 19.3 dargestellte Dialogbox. Neben der Zuordnungsübersicht (⇨ Tabelle 19.1) kann eine Distanzmatrix angefordert werden (⇨ Kap. 16.3). In „Cluster-Zugehörigkeit" kann neben „Keine" aus folgenden Alternativen gewählt werden:

- *Einzelne Lösung.* Die Anzahl der Cluster ist im Eingabefeld einzugeben. Es wird dann für jede Clusterlösung die Zugehörigkeit der Objekte zu den Clustern ausgegeben.
- *Bereich von Lösungen.* In den Eingabefeldern ist anzugeben, für welche der Clusterlösungen die Clusterzugehörigkeit der Objekte ausgegeben werden soll.

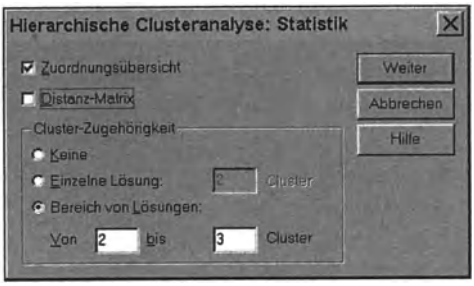


Abb. 19.3. Dialogbox „Hierarchische Clusteranalyse: Statistik"

② *Diagramm.* Nach Klicken der Schaltfläche „Diagramm..." wird die in Abb. 19.4 dargestellte Dialogbox geöffnet. In „Eiszapfen" kann „Alle Cluster" (Standardeinstellung), „Angegebener Clusterbereich" oder „keine" angefordert werden.

Für einen angefordeten Clusterbereich ist eine Start- und Stopeingabe sowie die Schrittweite anzugeben. Im in Abb. 19.4 gezeigten Fall wird ein Eiszapfendiagramm für alle Clusterlösungen von 1 bis 5 Cluster ausgegeben.

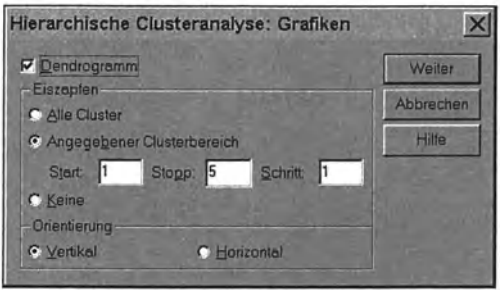


Abb. 19.4. Dialogbox „Hierarchische Clusteranalyse: Grafiken"

Es kann auch ein *Dendrogramm* angefordert werden. Das Dendrogramm wird in Tabelle 19.3 gezeigt. Aus dem Dendogramm kann man für die einzelnen Schritte der Clusterbildung sehen, welche Fälle bzw. Cluster zusammengeführt werden und welche Höhe die Distanz-Koeffizienten in den jeweiligen Clusterlösungen der Schritte haben. Die Koeffizienten werden dabei nicht gemäß Tabelle 19.1 in absoluter Größe grafisch abgebildet, sondern in einer Skala mit dem Wertebereich von 0 bis 25 transformiert. Auch im Dendrogramm kann man den großen Sprung (im 9. Schritt) in der Höhe des Koeffizienten erkennen und somit den Hinweis erhalten, dass eine 2er-Clusterlösung sinnvoll ist.

Tabelle 19.3. Ergebnisausgabe „Dendrogramm"

```
* H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * *

Dendrogram using Centroid Method

       Rescaled Distance Cluster Combine

    C A S E          0         5        10        15        20        25
 Label        Num  +---------+---------+---------+---------+---------+

 Othmarschen    3   -+
 Blanken. 2     8   -+
 Blanken. 1     7   -+------+
 Nienstedten    6   -+      +----------------------------------------+
 Flottbek       2   ---+---+                                         I
 Rissen        11   ---+                                             I
 Iserbrook      9   -+-+                                             I
 Sülldorf      10   -+ +------+                                       I
 Bahrenf. 3     1   ---+      +----------------------------------------+
 Osdorf         5   ---+      I
 Lurup          4   ---------+
```

③ *Methode.* Klicken auf die Schaltfläche „Methode...“ (Abb. 19.1) öffnet die in Abb. 19.2 dargestellte Dialogbox. Man kann aus der Drop-Down-Liste von „Cluster-Methode“ eine Methode auswählen (zu den Methoden ⇨ Kap. 19.1). Außerdem kann das Maß für die Distanzmessung gewählt und bestimmt werden, ob die Variablen und/oder das Distanzmaß transformiert werden soll. Da diese Möglichkeiten mit denen des Untermenüs „Distanzen“ von „Korrelation“ übereinstimmen, kann auf die Darstellung in Kapitel 16.3 verwiesen werden.

④ *Speichern.* Anklicken der Schaltfläche „Speichern...“ öffnet die in Abb. 19.5 dargestellte Dialogbox. Man kann hier auswählen, ob keine, für einen bestimmten Bereich von Clusterlösungen (z.B. 2 Cluster wie in Abb. 19.5) oder für eine bestimmte Clusterlösung die Clusterzugehörigkeit der Fälle gespeichert werden soll. Damit entsprechen diese Möglichkeiten denen im Untermenü „Statistik“ (⇨ Abb. 19.3). Der Unterschied besteht nur darin, dass hier die Clusterzugehörigkeit unter einem Variablennamen in der Arbeitsdatei gespeichert wird, während im Untermenü „Statistik“ die Ausgabe der Clusterzugehörigkeit im Ausgabefenster erfolgt.

Abb. 19.5. Dialogbox „Hierarchische Clusteranalyse: Neue Variablen speichern“

19.2.2 Anwendungsbeispiel zur Clusterzentrenanalyse

Die Clusterzentrenanalyse soll auf die in Kap. 19.2.1 dargestellte Clusterung von Ortsteilen im Hamburger Stadtbezirk Altona angewendet werden Aus der Anwendung der hierarchischen Clusterung bei Verwendung des Zentroid-Verfahrens hat sich ergeben, dass man die Ortsteile Altonas sinnvoll in drei Cluster ordnen kann (aus Gründen einer knappen und übersichtlichen Darstellung wurden in Kap. 19.2.1 aber nur zwei dieser Cluster dargelegt). Deshalb werden für die Clusterzentrenanalyse drei Cluster gewählt. Da die Clusterzentrenanalyse im Vergleich zur hierarchischen Clusterung das Ergebnis der Clusterbildung optimiert, kann auch überprüft werden, ob das in Kap. 19.2.1 erzielte Ergebnis sich verbessert. Zur Clusterung der Ortsteile gehen Sie nach Laden der Datei ALTONA.SAV wie folgt vor:

▷ Wählen Sie die Befehlsfolge „Analysieren“, „Klassifizieren ▷“, „Clusterzentrenanalyse...“. Es öffnet sich die in Abb. 19.6 dargestellte Dialogbox.

▷ Übertragen Sie die Variablen ZARBEIT, ZMJEP, ZBJEHA, ZG2W aus der Quellvariablenliste in das Feld „Variablen“. Es handelt sich dabei um die z-Werte der vier oben genannten Variablen. Diese wurden mit dem Menü „Deskriptive Statistiken“ (⇨ Kap. 8.5) erzeugt.

- ▷ Zur Fallbeschriftung in der Ergebnisausgabe übertragen Sie die Variable ORTN (Ortsteilname) in das Eingabefeld „Fallbeschriftung:".
- ▷ Im Feld „Anzahl der Cluster" ersetzen wir die voreingestellte „2" durch „3", um drei Cluster zu erhalten.

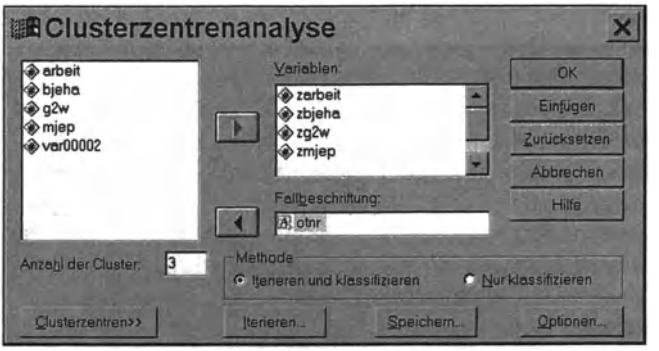


Abb. 19.6. Dialogbox „Clusterzentrenanalyse"

Die in Abb. 19.6 gewählten Einstellungen führen zu folgenden Ergebnisausgaben. In Tabelle 19.4 links werden die anfänglichen und rechts die Clusterzentren (Zentroide) der endgültigen Clusterlösung aufgeführt. Das Clusterzentrum eines Clusters wird durch die vier Durchschnittswerte der vier Variablen (hier: z-Werte der Variablen) aller im Cluster enthaltenen Fälle (Ortsteile) bestimmt. Als anfängliche Clusterlösung werden von SPSS einzelne Fälle gewählt. Daher handelt es sich z.B. bei dem Zentrum des ersten Clusters mit den Variablenwerten (0,57306, -0,50884, 0,99776, -81117) um den Ortsteil Lurup (⇨ Tabelle 16.4). In iterativen Schritten wird die endgültige Clusterlösung erreicht. Die sieben Ortsteile des ersten Clusters (Bahrenfeld 1 bis 3, Lurup, Osdorf, Iserbrook, Sülldorf) haben im Durchschnitt folgende Werte für die vier Variablen ZARBEIT bis ZMJEP: (0,00552, -0,82489, 0,69149,-0,46787). Diese Durchschnittswerte definieren das Zentrum dieses Clusters (⇨ Tabelle 19.4 rechts)

Im Iterationsprotokoll (⇨ Tabelle 19.5 links) wird die Änderung in den Clusterzentren aufgeführt. Eine weitere Tabelle zeigt die Anzahl der Fälle der Cluster (⇨ Tabelle 19.5 rechts).

Hat man in der in Abb. 19.6 dargestellten Dialogbox anstelle „Iterieren und Klassifizieren" „Nur Klassifizieren" gewählt, so erhält man als Ausgabeergebnisse nur die auf den rechten Seite der Tabellen 19. 4 und 19.5 gezeigten Ergebnisse.

Als Ergebnis der Clusterlösung zeigt sich, dass die Lösung der Clusterzentrenanalyse sich leicht von der der hierarchischen Clusterlösung unterscheidet. Die Ortsteile Bahrenfeld 1 und Bahrenfeld 2 sind in der Clusterzentrenanalyse zusammen mit Bahrenfeld 3 etc. zusammengefasst. In der hierarchischen Clusterlösung sind diese Ortsteile nicht alle im gleichen Cluster enthalten. Dieses zeigt, dass die hierarchische Clusterlösung nicht unbedingt zu einem optimalen Clusterergebnis führt.

Tabelle 19.4. Anfängliche (links) und endgültige (rechts) Clusterzentren

Anfängliche Clusterzentren

	Cluster		
	1	2	3
ZARBEIT	,57306	,70485	-1,58592
ZBJEHA	-,50884	1,86408	-,93670
ZG2W	,99776	-,95745	1,03968
ZMJEP	-,81171	-,67090	2,18224

Clusterzentren der endgültigen Lösung

	Cluster		
	1	2	3
ZARBEIT	,00552	,67734	-1,47400
ZBJEHA	-,82489	,86436	-,91042
ZG2W	,69149	-,88144	1,10305
ZMJEP	-,46787	-,52000	1,67251

Tabelle 19.5. Iterationsprotokoll (links) und Fälle je Cluster (rechts)

Iterationsprotokoll[a]

	Änderung in Clusterzentren		
Iteration	1	2	3
1	,796	1,014	,526
2	,000	,000	,000

a. Erzielte Konvergenz aufgrund keiner oder geringer Distanzänderung. Die maximale Distanz, um die ein Zentrum verändert wurde, ist ,000. Die aktuelle Iteration ist 2. Die minimale Distanz zwischen anfänglichen Zentren ist 3,081.

Anzahl der Fälle in jedem Cluster

Cluster	1	7,000
	2	13,000
	3	6,000
Gültig		26,000
Fehlend		,000

Wahlmöglichkeiten.

① *Clusterzentren>>*. Standardmäßig wählt SPSS als anfängliche Clusterlösung einzelne Fälle. Man kann aber den in iterativen Schritten sich vollziehenden Prozess des Auffindens der endgültigen optimalen Clusterlösung abkürzen, indem man in einer Datei Anfangswerte für Clusterzentren bereitstellt. Außerdem kann man die Clusterzentren der endgültigen Lösung in einer SPSS-Datei speichern. Nach Klicken auf die Schaltfläche „Clusterzentren>>„ verlängert sich die in Abb. 19.6 dargestellte Dialogbox um den in Abb. 19.7 dargestellten Bereich. Die gewünschte Option kann nun gewählt werden.

Abb. 19.7. Bereich „Clusterzentren" der Dialogbox „Clusterzentrenanalyse"

② *Iterieren*. Der Iterationsprozess des Auffindens einer optimalen endgültigen Lösung kann hier beeinflusst werden, indem man die Anzahl der Interationsschritte sowie das Konvergenzkriterium vorgibt. Das Konvergenzkriterium bestimmt, wann die Iteration abbricht. Nach Klicken der Schaltfläche „Iterieren..." wird die in Abb. 19.8 dargestellte Dialogbox geöffnet. Das Gewünschte kann eingetragen werden. Die Fälle werden der Reihe nach dem jeweils nächsten Clusterzentrum zugewiesen. Wenn „Gleitende Mittelwerte verwenden" gewählt wird, so wird das Zentrum nach jedem hinzugefügten Fall aktualisiert, ansonsten erst nachdem alle Fälle hinzugefügt wurden.

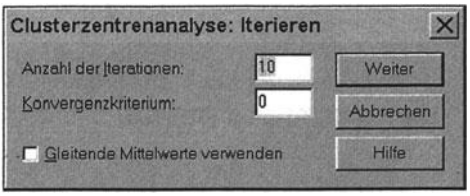


Abb. 19.8. Dialogbox „Clusterzentrenanalyse: Iterieren"

③ *Speichern..* Klicken auf die Schaltfläche „Speichern..." öffnet die in Abb. 19.9 dargestellte Dialogbox. Durch Anklicken von „Cluster-Zugehörigkeit" wird mit der Variable qcl_1 die endgültige Clusterzugehörigkeit der Fälle und mit qcl_2 die der Distanz der Fälle vom jeweiligen Clusterzentrum gespeichert.

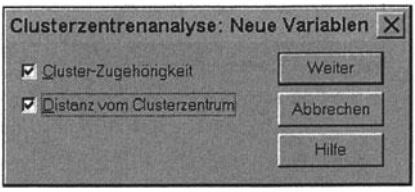


Abb. 19.9. Dialogbox „Clusterzentrenanalyse: Neue Variablen"

④ *Optionen.* Anklicken der Schaltfläche „Optionen..." öffnet die in Abb. 19.10 dargestellte Dialogbox. Man kann hier die Vorgehensweise bei Vorliegen von fehlenden Werten sowie zusätzliche statistische Informationen anfordern:

- *Anfängliche Clusterzentren.* Dieses ist die Standardeinstellung und erzeugt die in Abb. 19.4 links dargelegten Clusterzentren der Anfangslösung.
- *ANOVA-Tabelle.* Optional kann eine varianzanalytische Zerlegung der Varianz der einzelnen Variablen angefordert werden. Analog der Gleichung 14.3 wird die gesamte Variation einer Variablen[1] $\sum_{i=1}^{26}(x_i-\bar{x})^2=25$ (Zahlen für die Variable ZARBEIT) in die Variation innerhalb der Cluster $\sum_{i=1}^{7}(x_{i,1}-\bar{x}_1)^2+\sum_{i=1}^{13}(x_{i,2}-\bar{x}_2)^2+\sum_{i=1}^{6}(x_{i,3}-\bar{x}_3)^2=5{,}9996$ und zwischen den Clustern $\sum_{k=1}^{3}n_k(x_k-\bar{x})^2=19{,}004$ zerlegt. Unter Berücksichtigung der Anzahl der Freiheitsgrade für die Variation zwischen den Gruppen (Clusteranzahl minus 1 = 3 – 1 = 2) und für die Variation innerhalb der Cluster (Fallzahl minus Clusteranzahl = 26 – 3 = 23) ergibt sich gemäß Gleichung 14.11:

[1] Z-transformierte Variable haben eine Varianz = 1, d.h. $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=1$. Für n = 26 folgt $\sum_{i=1}^{26}(x_i-\bar{x})^2=25$.

$$F = \frac{s_{zwischen}}{s_{innerhalb}} = \frac{19{,}004/2}{5{,}9996/23} = \frac{9{,}5}{0{,}261} = 36{,}42$$

Tabelle 19.6. Ergebnisausgabe: Varianzzerlegung

ANOVA

	Cluster		Fehler			
	Mittel der Quadrate	df	Mittel der Quadrate	df	F	Sig.
ZARBEIT	9,500	2	,261	23	36,420	,000
ZBJEHA	9,724	2	,241	23	40,290	,000
ZG2W	10,374	2	,185	23	56,111	,000
ZMJEP	10,916	2	,138	23	79,227	,000

Die F-Tests sollten nur für beschreibende Zwecke verwendet werden, da die Cluster so gewählt wurden, daß die Differenzen zwischen Fällen in unterschiedlichen Clustern maximiert werden. Dabei werden die beobachteten Signifikanzniveaus nicht korrigiert und können daher nicht als Tests für die Hypothese der Gleichheit der Clustermittelwerte interpretiert werden.

- *Cluster-Informationen für jeden Fall.* Für jeden Fall wird die Clusterzugehörigkeit, die Distanz eines jeden Falles zum jeweiligen Clusterzentrum und eine Distanzmatrix der endgültigen Clusterlösung ausgegeben. Die Distanzen der Cluster sind Distanzen zwischen den Zentren der Cluster.

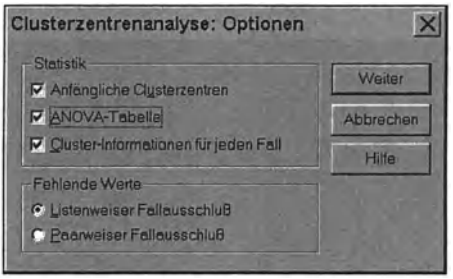


Abb. 19.10. Dialogbox „Clusterzentrenanalyse: Optionen"

19.2.3 Vorschalten einer Faktorenanalyse

Die für die Clusteranalyse verwendeten Variablen ARBEIT, MJEP, G2W und BJEHA sind korreliert. Dieses ist nicht unproblematisch, wenn Distanzen als Inputvariable für die Clusteranalyse berechnet werden. Eine Faktorenanalyse (⇨ Kap. 21) für alle 182 Ortsteile Hamburgs mit den vier Variablen hat gezeigt, dass sich hinter den Variablen zwei Dimensionen verbergen, die man als soziale Struktur und Verdichtung bezeichnen könnte. Daher sollte man bei Durchführung einer Clusteranalyse überlegen und prüfen, ob man der Clusteranalyse eine Faktorenanalyse vorschalten sollte. Zur Veranschaulichung ist die Faktorenanalyse zur Extraktion von zwei Faktoren mit der Hauptkomponentenmethode und anschlie-

ßender Varimax-Rotation auf den Datensatz der Datei ALTONA.SAV angewendet worden. Eine anschließende hierarchische Clusteranalyse der Faktorwerte nach der Zentroid-Methode mit der quadratischen Euklidischer Distanz ist zu einer Clusterlösung gekommen, die der Clusterzentrenanalyse angewendet auf die z-Werte der Ausgangsvariablen entspricht. Auch die Clusterzentrenanalyse für die Faktoren kommt zu diesem Ergebnis. Das Vorliegen von nur zwei Dimensionen erleichtert die Interpretation der Clusterlösung. Wegen der Zweidimensionalität lassen sich die Cluster grafisch anschaulich darstellen.

Abb. 19.11 ist ein Streudiagramm für die Faktorwerte der beiden Faktoren. Faktor 1 (score 1) lädt hoch auf die Variablen G2W (mit negativem Vorzeichen) und BJEHA und kann als Verdichtung, Faktor 2 (score 2) lädt hoch auf die Variablen ARBEIT (mit negativem Vorzeichen) und MJEP und kann als Sozialstruktur interpretiert werden. Im Streudiagramm sind die Faktorwerte der 28 Ortsteile von Altona abgebildet. Man sieht deutlich, dass sich drei Cluster voneinander abgrenzen. Im Cluster links oben sind die Ortsteile mit einer geringen Verdichtung und einer hohen sozialen Struktur (Flottbek, Othmarschen, Nienstedten, Blankenese 1 und 2, Rissen) zusammengefasst. Im Cluster links unten zeigen sich die Ortsteile mit einer niedrigeren sozialen Struktur und einer kleineren Verdichtung (Bahrenfeld 1 bis 3, Lurup, Osdorf, Iserbrook, Sülldorf). Im Cluster rechts unten sind die anderen Ortsteile zusammengefasst. Die Zuordnung der beiden Ortsteile Bahrenfeld 1 und Bahrenfeld 2 unterscheiden sich (wie oben dargelegt) in der Lösung der hierarchischen Cluster- und der Clusterzentrenanalyse. Daher werden sie in der Grafik angezeigt.

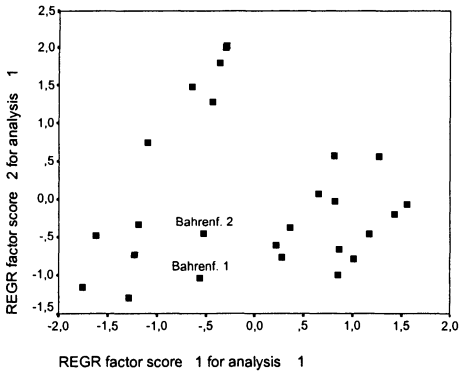


Abb. 19.11. Streudiagramm der Faktorwerte

20 Diskriminanzanalyse

20.1 Theoretische Grundlagen

Bei der multivariaten statistischen Methode der Diskriminanzanalyse geht es um die Vorhersage der Gruppenzugehörigkeit von Personen oder Objekten durch mehrere metrische Variablen. Es kann sich dabei um zwei oder auch mehrere Gruppen handeln, wobei die Gruppenzugehörigkeit bei Anwendung der Diskriminanzanalyse bekannt ist. Wenn es sich dann zeigt, dass die metrischen Variablen zur Vorhersage der Gruppenzugehörigkeit geeignet sind, kann das Ergebnis der Diskriminanzanalyse dazu verwendet werden, für weitere Objekte oder Personen, für die nur die metrischen Variablen aber nicht die Gruppenzugehörigkeit bekannt ist, eine Gruppenzuordnung vorzunehmen.

Mit einem Beispiel aus dem Bereich der Medizin soll das statistische Verfahren zunächst für den Zwei-Gruppenfall erläutert werden. Zur Diagnose von Lebererkrankungen wie der viralen Hepatitis dienen Leberfunktionstests. Dabei spielen Messergebnisse zu verschiedenen Enzymen eine besondere Rolle. Bei der Diagnose von Lebererkrankungen hat es sich gezeigt, dass es aber nicht möglich ist, anhand nur einer der Enzymvariablen klare Anhaltspunkte dafür zu gewinnen, ob ein Patient eine bestimmte Lebererkrankung hat (z. B. eine virale Hepatitis). Vielmehr ist man zu der Erkenntnis gekommen, dass sich aus dem Zusammentreffen von Werten mehrerer Enzymvariablen bessere Belege für eine bestimmte Diagnose ergeben. Zu der Frage, welche der Variablen dafür besonders bedeutsam sind und in welcher Wertekombination der verschiedenen Variablen, kann die Diskriminanzanalyse einen Beitrag leisten.

Der Grundgedanke der Diskriminanzanalyse soll zunächst durch eine grafische Darstellung erläutert werden. Dabei werden Daten aus der Datei LEBER.SAV verwendet[1]. Für 218 Fälle von Lebererkrankungen wird in der Variable GRUP1 die Lebererkrankung erfasst (0 = virale Hepatitis, 1 = andere Lebererkrankung). Mit den Variablen AST, ALT, OCT und GIDH werden Messwerte für vier Enzyme erfasst und mit LAST, LALT, LOCT und LGIDH die logarithmierten Messwerte. Da die für eine Diskriminanzanalyse verwendeten metrischen Variablen für jede Gruppe annähernd normalverteilt sein sollten, werden anstelle der Originalmesswerte logarithmierte Messwerte verwendet. Für die folgende grafische Darstellung

[1] Die Datei LEBER.SAV wurde uns freundlicherweise von Prof. Dr. Berg vom Universitätskrankenhaus Eppendorf in Hamburg zur Verfügung gestellt. Die Daten entstammen der Literatur (Plomteux, Multivariate Analysis of an Enzymic Profile for the Differential Diagnosis of Viral Hepatitis. In: Clinical Chemistry, Vol. 26, No. 13, 1980, S. 1897-1899).

beschränken wir uns auf zwei der vier metrischen Variablen: LAST und LALT. In Abb. 20.1 sind die 218 Krankheitsfälle in einem Streudiagramm mit den beiden Variablen LAST und LALT dargestellt. Durch eine unterschiedliche Markierung der Fälle im Streudiagramm werden die beiden Gruppen (virale Hepatitis und andere Lebererkrankungen) sichtbar. Es zeigt sich deutlich, dass die beiden Gruppen im Streudiagramm überlappende Punktwolken mit voneinander verschiedenen Zentren bilden. Die Aufgabe der Diskriminanzanalyse besteht darin, mit Hilfe der metrischen Variablen die beiden Gruppen möglichst gut zu trennen. Aus der Grafik wird ersichtlich, dass weder die Variable LAST noch die Variable LALT allein gut zur Trennung der Gruppen geeignet sind, weil sich die Punktwolken überlappen. Eine beispielhaft in das Streuungsdiagramm eingezeichnete Trennlinie für die Variable LALT mit einem (beispielhaft angenommenen) kritischen Trennwert $LALT_{krit}$ zeigt, dass eine derartige Trennung unbefriedigend ist, weil die Überlappung der Verteilungen beträchtlich ist. Werden die Punkte im Streudiagramm auf die LALT-Achse projiziert, so werden die Verteilungen der Variable LALT für die beiden Gruppen abgebildet.

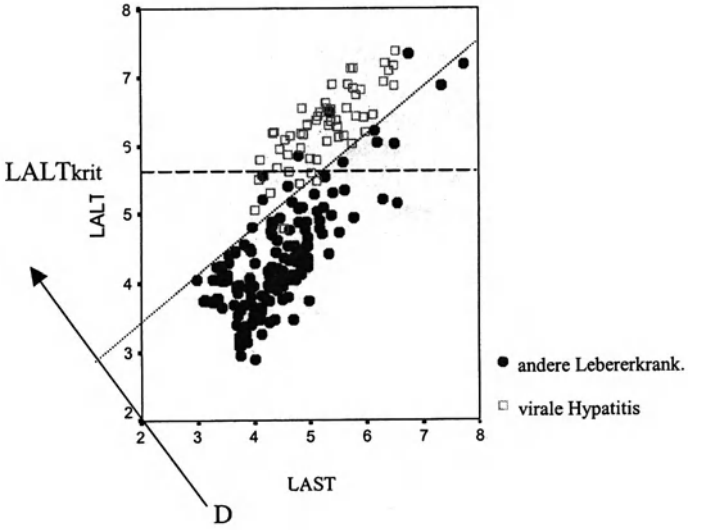


Abb. 20.1 Überlappende Punktwolken im Streudiagramm

In Abb. 20.2 werden diese Verteilungen idealisiert als Normalverteilungen mit gleicher Streuung dargestellt: Wegen der Überlappung beider Verteilungen kann ein zufriedenstellende Trennung beider Gruppen mit Hilfe eines Trennwertes von LALT nicht gelingen.

Die Trennung der beiden Gruppen gelingt wesentlich besser, wenn die von links unten nach rechts oben verlaufende Trennlinie in Abb. 20.1 gewählt wird. Mit einer derartigen Trennung wird für die Punktwolke aller 218 Fälle ein neues Koordinatensystem gewählt. Die im Winkel von neunzig Grad zur Trennlinie stehende D-Achse bildet die Grundachse des neuen Koordinatensystems. Die Messwerte auf

der D-Achse ergeben sich aus einer Linearkombination der Messwerte der Variablen LALT und LAST gemäß Gleichung 20.1. Diese Gleichung, die einer Regressionsgleichung ähnelt, nennt man eine Diskriminanzfunktion. Die Messwerte D heißen Diskriminanzwerte. Die Koeffizienten b_1 und b_2 sind die Gewichte der Linearkombination und werden Diskriminanzkoeffizienten genannt. Durch die Koeffizienten der Diskriminanzfunktion wird die Steigung der D-Achse bestimmt.

$$D = b_0 + b_1 LALT_1 + b_2 LAST \tag{20.1}$$

Die Koeffizienten der Gleichung - und hier liegt der Unterschied zu einer Regressionsgleichung - sollen derart bestimmt werden, dass die Werte von D möglichst gut die beiden im Datensatz enthaltenen Gruppen (Fälle mit viraler Hepatitis bzw. einer anderen Lebererkrankung) trennen. Projiziert man in Abb. 20.1 die Punkte des Streudiagramms auf die D-Achse, so wird klar, dass große Werte von D die Fälle einer viralen Hepatitis und kleine Werte die Fälle einer anderen Lebererkrankung ausweisen.

Durch eine Projektion der Punkte des Streudiagramms auf die D-Achse wird eine der Abb. 20.2 analoge Darstellung der Häufigkeitsverteilungen der beiden Gruppen mittels der Diskriminanzwerte D erstellt (hier ebenfalls idealisiert durch Normalverteilungen mit gleicher Streuung). Aus Abb. 20.3 kann man erkennen, dass auch diese beiden Verteilungen sich überlagern. Im Unterschied zu Abb. 20.2 ist die Überlagerung aber wesentlich reduziert. Dieses bedeutet, dass die Trennung der Gruppen mit Hilfe der Diskriminanzfunktion (einer Linearkombination der Ursprungsmesswerte) besser gelingt als mit den Ursprungswerten selber. Im Idealfall gelingt die Trennung ohne Überlappung der beiden Verteilungen. Noch besser als im dargelegten Fall von zwei Enzymvariablen (den unabhängigen Variablen) gelingt die Trennung der beiden Gruppen, wenn alle vier Enzymvariablen einbezogen werden. Bezeichnet man die vier logarithmierten Eynzymvariablen mit x_1 bis x_4, so lautet die lineare Diskriminanzfunktion:

$$D = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \tag{20.2}$$

Wenn die Koeffizienten der Diskriminanzfunktion bekannt sind, kann die Funktion zur Vorhersage der Gruppenzugehörigkeit (virale Hepatitis liegt vor oder nicht) für einen nicht im Datensatz enthaltenen Krankheitsfall benutzt werden. Dafür müssen die Werte der vier Variablen erhoben und dann in die Gleichung eingesetzt werden. Damit eine Zuordnung in eine der beiden Gruppen anhand der Höhe des für die Person berechneten Wertes von D möglich wird, muss ein kritischer Wert für D (ein Trennwert) bekannt sein oder - wie es bei SPSS der Fall ist - die Zuordnung auf andere Weise vorgenommen werden.

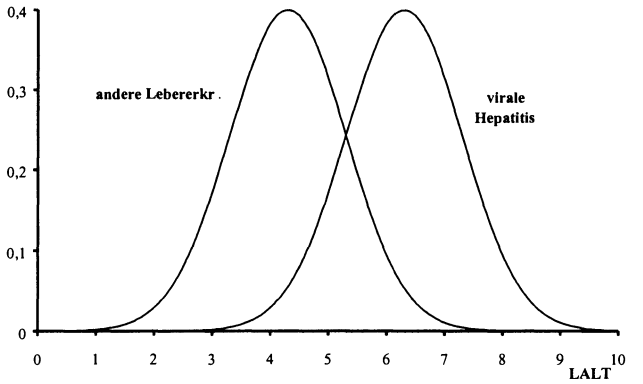


Abb. 20.2 Häufigkeitsverteilungen der beiden Gruppen auf der LALT-Achse

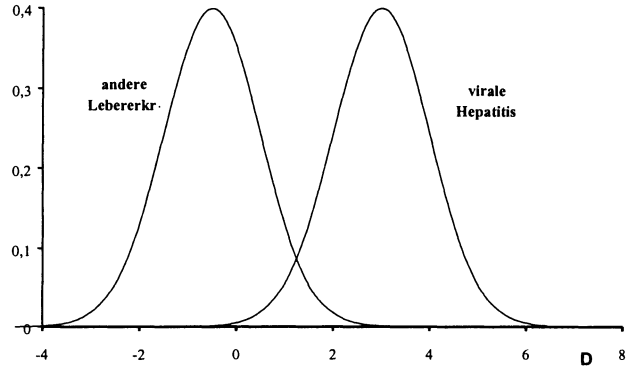


Abb. 20.3 Häufigkeitsverteilungen der beiden Gruppen auf der D-Achse

Aus Abb. 20.3 kann man intuitiv erfassen, unter welchen Bedingungen eine Trennung der beiden Gruppen mit Hilfe einer Diskriminanzfunktion besonders gut gelingt (d. h. die beiden Verteilungen sich möglichst wenig überlappen): die Mittelwerte der beiden Gruppen $\overline{D}_1$ bzw. $\overline{D}_2$ sollten möglichst weit auseinanderliegen und die Streuung der beiden Verteilungen sollten möglichst klein sein. Gemäß dieser beiden Zielsetzungen wird die Lage der D-Achse bestimmt (und damit die Diskriminanzkoeffizienten b). Das Optimierungskriterium zur Bestimmung der Diskriminanzkoeffizienten knüpft somit an das statistische Konzept der Varianzanalyse an (⇨ Kap. 14).

Die gesamte Streuung der Diskriminanzwerte D lässt sich aufteilen in die Streuung (gemessen als Summe der Abweichungsquadrate vom Mittelwert = SAQ) zwischen den beiden Gruppen und innerhalb der beiden Gruppen (mit Fallzahlen n_1 und n_2):

$$SAQ_{Total} = SAQ_{zwischen} + SAQ_{innerhalb} \qquad (20.3)$$

$$\sum_{i=1}^{n}(D_i-\overline{D})^2=\left[n_1(\overline{D}_1-\overline{D})^2+n_2(\overline{D}_2-\overline{D})^2\right]+\left[\sum_{i=1}^{n_1}(D_{1,i}-\overline{D}_1)^2+\sum_{i=1}^{n_2}(D_{2,i}-\overline{D}_2)^2\right]$$

$SAQ_{zwischen}$ erfasst die Streuung, die sich durch die Abweichungen der Gruppenmittelwerte $\overline{D}_1$ bzw. $\overline{D}_2$ vom gesamten Mittelwert $\overline{D}$ ergeben. Diese quadrierten Abweichungen werden (gewichtet mit den Fallzahlen der Gruppen n_1 bzw. n_2) summiert. $SAQ_{innerhalb}$ ist die Summe der Streuung der beiden Verteilungen. $SAQ_{zwischen}$ wird auch als die durch die Diskriminanzfunktion erklärte und $SAQ_{innerhalb}$ als die nicht erklärte Streuung bezeichnet. Die Diskriminanzkoeffizienten werden derart bestimmt, dass der Quotient aus den Streuungen gemäß Gleichung 20.4 maximiert wird.

$$\frac{SAQ_{zwischen}}{SAQ_{innerhalb}}=\text{Max!} \tag{20.4}$$

Diese Maximierungsaufgabe läuft auf die Bestimmung des Eigenwerts einer Matrix hinaus und soll hier nicht weiter betrachtet werden. Mit der Lösung der Maximierungsaufgabe sind die Koeffizienten b der unabhängigen Variablen in ihren Relationen zueinander bestimmt. Anschließend werden von SPSS zwei weitere Berechnungsschritte vorgenommen. Im ersten Schritt werden die Diskriminanzkoeffizienten derart normiert, dass die Varianz (Summe der Abweichungsquadrate dividiert durch die Anzahl der Freiheitsgrade df) innerhalb der Gruppen eins wird ($SAQ_{innerhalb}/df=1$ mit df = n–k, n = Fallzahl, k = Gruppenanzahl). Im zweiten Schritt wird die Konstante in der Diskriminanzfunktion b_0 derart bestimmt, dass der Mittelwert der Diskriminanzwerte gleich Null wird ($\overline{D}=0$).

Die Zuordnung der Fälle zu den Gruppen (d. h. die Vorhersage der Gruppenzugehörigkeit) mit Hilfe der Diskriminanzwerte D_i beruht bei SPSS auf einem wahrscheinlichkeitstheoretischen Theorem von Bayes. Die Wahrscheinlichkeit P (= A-posteriori-Wahrscheinlichkeit), dass ein Fall mit einem Diskriminanzwert $D_i = d$ (d sei ein konkreter Wert) zur Gruppe G gehört (im Zwei-Gruppenfall ist G = 1,2; im k-Gruppenfall ist G = 1, 2, ...k), wird berechnet durch

$$P(G/D_i=d)=\frac{P(D_i\geq d/G)P(G)}{\sum_{i=1}^{k}P(D_i\geq d/G)P(G)} \tag{20.5}$$

P(G) ist die Wahrscheinlichkeit dafür, dass ein Fall zur Gruppe G (G = 1, 2, ... k) gehört (= A-priori-Wahrscheinlichkeit. Bezogen auf das Beispiel für G = 1: Die Wahrscheinlichkeit, dass ein Leberkranker eine virale Hepatitis hat). $P(D_i \geq d/G)$ ist die bedingte Wahrscheinlichkeit des Auftretens eines Diskriminanzwertes $D_i \geq d$ bei bekannter Gruppenzugehörigkeit G. $P(D_i \geq d/G)$ wird wie folgt geschätzt: Es wird die quadrierte Distanz nach Mahalanobis (⇨ Kap. 16.3) eines Falles vom Zentrum (Zentroid) einer Gruppe G bestimmt und ihre Wahrscheinlichkeit mit Hilfe der Dichtefunktion der Normalverteilung berechnet (⇨ Backhaus u. a., S. 132 ff.).

Ein Fall wird der Gruppe G zugeordnet, für die die geschätzte Wahrscheinlichkeit $P(G/D_i = d)$ am größten ist.

20.2 Praktische Anwendung

Diskriminanzanalyse für zwei Gruppen. Für das in Kapitel 20.1 benutzte Beispiel soll nun unter Einschluss aller vier unabhängigen Enzymvariablen (LALT, LAST, LOCT und LGIDH) eine Diskriminanzanalyse durchgeführt werden. Nach Laden der Datei LEBER.SAV gehen Sie wie folgt vor:

- ▷ Wählen Sie per Mausklick die Befehlsfolge „Analysieren", „Klassifizieren ▷", „Diskriminanzanalyse". Es öffnet sich die in Abb. 20.4 dargestellte Dialogbox.
- ▷ Übertragen Sie die Variable GRUP1, die die Gruppenzuordnung der Fälle enthält (0 = virale Hepatitis, 1 = andere Lebererkrankung) in das Feld „Gruppenvariable".
- ▷ Klicken auf die Schaltfläche „Bereich definieren" öffnet die in Abb. 20.5 dargestellte Dialogbox zur Festlegung des Wertebereichs der Gruppenvariable. In die Eingabefelder „Minimum" und „Maximum" sind die Werte der Gruppenvariable zur Definition der Gruppen einzutragen (hier: 0 und 1). Anschließend klicken Sie die Schaltfläche „Weiter".
- ▷ Übertragen Sie die Variablen LALT, LAST, LOCT und LGIDH in das Eingabefeld „Unabhängige Variable(n)". Die Voreinstellung „Unabhängige Variablen zusammen aufnehmen" wird beibehalten.
- ▷ Mit Klicken der Schaltfläche „OK" wird die Berechnung gestartet.

Abb. 20.4. Dialogbox „Diskriminanzanalyse"

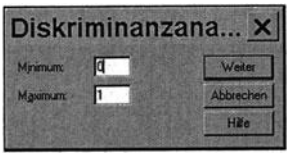


Abb. 20.5. Dialogbox "Diskriminanzanalyse: Bereich definieren"

Die in Abb. 20.4 und 20.5 gewählten Einstellungen führen zu folgenden Ergebnissen. In Tabelle 20.1 wird der Eigenwert der diskriminanzanalytischen Aufgabenstellung aufgeführt. Er entspricht dem maximalen Optimierungskriterium gemäß Gleichung 20.4 ($\frac{SAQ_{zwischen}}{SAQ_{innerhalb}} = 1{,}976$).[2] Der Eigenwert ist ein Maß für die Güte der Trennung der Gruppen. Ein hoher Wert spricht für eine gute Trennung. Da wir es mit einer Diskriminanzanalyse für zwei Gruppen zu tun haben, gibt es nur eine Diskriminanzfunktion, so dass diese Funktion die gesamte Varianz erfasst.

Mit dem kanonischen Korrelationskoeffizienten wird ein Maß aufgeführt, das die Stärke des Zusammenhangs zwischen den Diskriminanzwerten D_i und den Gruppen zum Ausdruck bringt. Er entspricht dem eta der Varianzanalyse (⇨ Kap. 14.1). Im hier dargestellten Zwei-Gruppenfall entspricht eta dem Pearson-Korrelationskoeffizienten zwischen der Diskriminanzvariable D_i und der Gruppenvariablen GRUP1 mit den Werten 0 und 1.

$$eta = \sqrt{\frac{SAQ_{zwischen}}{SAQ_{Total}}} = \sqrt{\frac{\text{erklärte Streuung}}{\text{gesamte Streuung}}} = 0{,}815 \qquad (20.6)$$

Tabelle 20.1. Ergebnisausgabe: Eigenwert der Diskriminanzanalyse

Eigenwerte

Funktion	Eigenwert	% der Varianz	Kumulierte %	Kanonische Korrelation
1	1,976[a]	100,0	100,0	,815

a. Die ersten 1 kanonischen Diskriminanzfunktionen werden in dieser Analyse verwendet.

In Tabelle 20.2 wird das Maß Wilks' Lambda (λ) zusammen mit einem Chi-Quadrat-Test aufgeführt. Wilks' Lambda ist das gebräuchlichste Maß für die Güte der Trennung der Gruppen mittels der Diskriminanzfunktion. Da

$$\lambda = \frac{SAQ_{innerhalb}}{SAQ_{Total}} = \frac{\text{nicht erklärte Streuung}}{\text{gesamte Streuung}} = 0{,}336 \qquad (20.7)$$

gilt, wird deutlich, dass ein kleiner Wert für eine gute Trennung der Gruppen spricht. Etwa 34 % der Streuung wird nicht durch die Gruppenunterschiede erklärt. Aus den Gleichungen 20.6 und 20.7 ergibt sich, dass Wilks' Lambda und eta^2 zueinander komplementär sind, da sie sich zu eins ergänzen ($\lambda + eta^2 = 1$).

Durch die Transformation

$$\chi^2 = -\left[n - \frac{m+k}{2} - 1\right] \ln(\lambda) = -(218 - \frac{4+2}{2} - 1) * \ln(0{,}336) = 233{,}4 \qquad (20.8)$$

[2] Speichert man die Diskriminanzwerte und rechnet eine Varianzanalyse (einfaktorielle ANOVA) mit Dis1_1 als abhängige Variable und GRUP1 als Faktor, so erhält man die Aufteilung von SAQ_{Total} in $SAQ_{innerhalb}$ und $SAQ_{zwischen}$.

wird Wilks' Lambda (λ) in eine annähernd chi-quadratverteilte Variable mit df = m(k–1) Freiheitsgraden überführt (n = Fallanzahl, m = Variablenanzahl, k = Gruppenanzahl). Mit einem Chi-Quadrat-Test kann geprüft werden, ob sich die Gruppen signifikant voneinander unterscheiden oder nicht. Bei einem Signifikanzniveau von 5 % (α = 0,05) und df = 4 ergibt sich aus einer tabellierten Chi-Quadrat-Verteilung ein kritischer Wert in Höhe von 9,49. Da der empirische Chi-Quadratwert mit 233,4 (⇨ Tabelle 20.2) diesen übersteigt, wird die H_0-Hypothese (die beiden Gruppen unterscheiden sich nicht) abgelehnt und die Alternativhypothese (die Gruppen unterscheiden sich) angenommen. Diese Schlussfolgerung ergibt sich auch daraus, dass der Wert von „Signifikanz" in Tabelle 20.2 kleiner ist als α = 0,05.

Tabelle 20.2. Ergebnisausgabe: Wilks' Lambda

Wilks' Lambda

Test der	Wilks-Lambda	Chi-Quadrat	df	Signifikanz
1	,336	233,400	4	,000

In Tabelle 20.3 werden standardisierte Diskriminanzkoeffizienten ausgegeben. Die Höhe der Diskriminanzkoeffizienten gemäß Gleichung 20.2 erlaubt es nicht, Aussagen darüber zu treffen, wie stark der relative Einfluss der unabhängigen Variablen zueinander ist. Dieses liegt daran, dass die Einflussstärke einer unabhängigen Variablen auf die Diskriminanzwerte auch durch die Streuung der unabhängigen Variablen beeinflusst wird. Analog den Beta-Koeffizienten in der Regressionsanalyse (⇨ Kap. 17.2.1) werden deshalb standardisierte Diskriminanzkoeffizienten gemäß folgender Gleichung berechnet:

$$b_{x_j}^{standardisiert} = b_{x_j} s_{x_j}^{innerhalb} \qquad (20.9)$$

Der standardisierte Koeffizient $b_{x_j}^{standardisiert}$ einer unabhängigen Variablen x_j ergibt sich durch Multiplikation des unstandardisierten Koeffizienten b_{x_j} mit der Standardabweichung der unabhängigen Variablen innerhalb der Gruppen $s_{x_j}^{innerhalb}$ ($s_{x_j}^{innerhalb} = \sqrt{SAQ_{innerhalb} / df}$; $(s_{x_j}^{innerhalb})^2$ steht in der Diagonale der Kovarianz-Matrix innerhalb der Gruppen, die in der in Abb. 20.6 dargestellten Dialogbox angefordert werden kann).

Die in Tabelle 20.3 aufgeführten standardisierten Diskriminanzkoeffizienten zeigen, dass die Variablen LALT und LAST den größten Einfluss auf die Diskriminanzwerte haben.[3] Da hohe Diskriminanzwerte eine virale Hepatitis und niedrige eine andere Lebererkrankung anzeigen (⇨ Abb. 20.1 und Abb. 20.3), wird aufgrund der Vorzeichen der Koeffizienten der Variablen LALT und LAST deutlich, dass hohe Werte von LALT und niedrige Werte von LAST mit dem Vorliegen ei-

[3] Analog der Interpretation von standardisierten Koeffizienten einer Regressionsgleichung (⇨ Kap.17.2.1) gilt auch hier, dass die relative Größe der standardisierten Koeffizienten wegen Multikollinearität nur Anhaltspunkte für die relative Bedeutung der Variablen geben.

ner viralen Hepatitis verbunden sind. Der Koeffizient von LOCT ist mit 0,066 so klein, dass zu fragen ist, ob man diese Variable überhaupt berücksichtigen sollte. Damit wird deutlich, dass eine Diskriminanzanalyse auch leistet, geeignete und weniger geeignete Variablen für die Gruppenvorhersage zu unterscheiden.

Tabelle 20.3. Ergebnisausgabe: standardisierte Diskriminanzkoeffizienten

Standardisierte kanonische Diskriminanzfunktionskoeffizienten

	Funktion
	1
LALT	1,411
LAST	-,554
LGLDH	-,362
LOCT	,066

Die Koeffizienten der Struktur-Matrix (⇨ Abb. 20.4) bieten ebenfalls Informationen über die (relative) Bedeutung der Variablen für die Diskriminanzfunktion. Das aus den standardisierten Diskriminanzkoeffizienten gewonnene Bild hinsichtlich ihrer Rolle in der Diskriminanzfunktion wird bestätigt.

Tabelle 20.4. Ergebnisausgabe: Strukturmatrix

Struktur-Matrix

	Funktion
	1
LALT	,850
LAST	,344
LOCT	,231
LGLDH	,067

Gemeinsame Korrelationen innerhalb der Gruppen zwischen Diskriminanzvariablen und standardisierten kanonischen Diskriminanzfunktionen. Variablen sind nach ihrer absoluten Korrelationsgröße innerhalb der Funktion geordnet.

Bei den in Abb. 20.5 aufgeführten Gruppen-Zentroiden handelt es sich um die durchschnittlichen Diskriminanzwerte der beiden Gruppen: $\overline{D}_1 = 2{,}352$ und $\overline{D}_2 = -0{,}833$.

Tabelle 20.5. Ergebnisausgabe: Gruppen-Zentroide

unktionen bei den Gruppen-Zentroiden

	Funktion
GRUP1	1
virale Hepatitis	2,352
andere Lebererkrankung	-,833

Nicht-standardisierte kanonische Diskriminanzfunktionen, die bezüglich des Gruppen-Mittelwertes bewertet werden

Wahlmöglichkeiten. Durch Klicken von Schaltflächen (⇨ Abb. 20.4) können weitere Ergebnisausgaben etc. angefordert werden:

① *Auswählen* >>. Es können mit einem Wert einer zu übertragenden Variable Fälle ausgewählt werden, für die die Analyse angewendet werden soll.

② *Statistik.* Klicken auf die Schaltfläche „Statistik" öffnet die in Abb. 20.6 dargestellte Dialogbox. Es können folgende Berechnungen angefordert werden:

❑ *Deskriptive Statistiken.*

- *Mittelwert.* Es werden Mittelwerte und Standardabweichungen der unabhängigen Variablen ausgegeben.
- *Univariate ANOVA.* Für jede der unabhängigen Variablen wird ein varianzanalytischer F-Test auf Gleichheit der Mittelwerte für die Gruppen durchgeführt (⇨ Kap. 14.1). Die Testgröße F ist gemäß Gleichung 14.11 der Quotient aus der Varianz (SAQ dividiert durch die Anzahl der Freiheitsgrade df) der unabhängigen Variablen zwischen und innerhalb der Gruppen. Bei einem Signifikanzniveau von $\alpha = 0{,}05$ besteht wegen 0,168 > 0,05 bei der Variable LGLDH keine signifikante Differenz der Mittelwerte der beiden Gruppen (⇨ Tabelle 20.6). Dieses bedeutet aber nicht unbedingt, dass diese Variable aus dem Diskriminanzalysemodell ausgeschlossen werden sollte. Eine Variable, die alleine keine diskriminierende Wirkung hat, kann simultan mit anderen Variablen sehr wohl dafür einen Beitrag leisten (siehe dazu die Überlegungen zu Abb. 20.1). Umgekehrt gilt natürlich für signifikante Variablen, dass sie nicht unbedingt geeignet sein müssen.

Tabelle 20.6. Ergebnisausgabe: Varianzanalytischer Test

Gleichheitstest der Gruppenmittelwerte

	Wilks-Lambda	F	df1	df2	Signifikanz
LALT	,412	308,444	1	216	,000
LAST	,810	50,610	1	216	,000
LGLDH	,991	1,911	1	216	,168
LOCT	,905	22,686	1	216	,000

- *Box-M.* Mit dem dazugehörigen F-Test (⇨ Tabelle 20.7). wird eine Voraussetzung der Anwendung der Diskriminanzanalyse geprüft: gleiche Kovarianz-Matrizen der Gruppen (d.h. gleiche Varianzen und Kovarianzen der Variablen im Gruppenvergleich). Da „Signifikanz" mit 0,000 < 0,05 ist, wird bei einem Signifikanzniveau von 5 % die Hypothese gleicher Kovarianz-Matrizen abgelehnt. Das Ergebnis des Box-M-Tests ist aber sehr von der Stichprobengröße (den Fallzahlen) abhängig. Auch ist der Test anfällig hinsichtlich der Abweichung der Variablen von der Normalverteilung. Um zu erreichen, dass die Variablen in den Gruppen annähernd normalverteilt sind, haben wir die Variablen logarithmiert.

  Wegen der angesprochenen Schwächen des Box-M-Tests sollte man nicht auf das Box-M-Testergebnis vertrauen. Zur Prüfung der Annahme gleicher Kovarianz-Matrizen der Gruppen wird empfohlen, die Kovari-

anz-Matrix auszugeben (⇨ Dialogbox „Statistik" in Abb. 20.6) und diese hinsichtlich der Höhe und der Vorzeichen der Kovarianzen im Gruppenvergleich zu prüfen. Die $x_i x_j$-Kovarianz der einen Gruppe sollte die der anderen Gruppe um nicht mehr als das 10-fache übersteigen und die Vorzeichen sollten sich nicht unterscheiden.

Tabelle 20.7. Ergebnisausgabe: Box-M-Test

Textergebnisse

Box-M		35,531
F	Näherungswert	3,453
	df1	10
	df2	52147,178
	Signifikanz	,000

Testet die Null-Hypothese der Kovarianz-Matrizen gleicher Grundgesamtheit.

- *Funktionskoeffizienten.* Es handelt sich hierbei um die Koeffizienten von Klassifizierungsfunktionen.
  - *Fisher.* Es werden die Koeffizienten der Klassifizierungsfunktion nach R. A. Fisher ausgegeben (⇨ Backhaus u. a., S. 125 ff.).
  - *Nicht standardisiert.* Die nicht standardisierten Diskriminanzkoeffizienten (⇨ Tabelle 20.8) sind Grundlage der Berechnung der Diskriminanzwerte für einzelne Fälle. Analog einer Regressionsgleichung errechnen sich die Diskriminanzwerte durch Einsetzen der Werte der unabhängigen Variablen in die Diskriminanzfunktion 20.2:

$$D = -5{,}070 + 1{,}934 * LALT - 0{,}719 * LAST - 0{,}522 * LGLDH + 0{,}067 * LOCT$$

Tabelle 20.8. Nicht-standardisierte Diskriminanzkoeffizienten

Kanonische Diskriminanzfunktion skoeffizienten

	Funktion
	1
LALT	1,934
LAST	-,719
LGLDH	-,522
LOCT	,067
(Konstant)	-5,070

Nicht-standardisierte Koeffizienten

- *Matrizen.* Es werden Korrelationskoeffizienten und Kovarianzen (jeweils für innerhalb der Gruppen, für einzelne Gruppen und für insgesamt) der Variablen berechnet und in Matrizenform dargestellt.

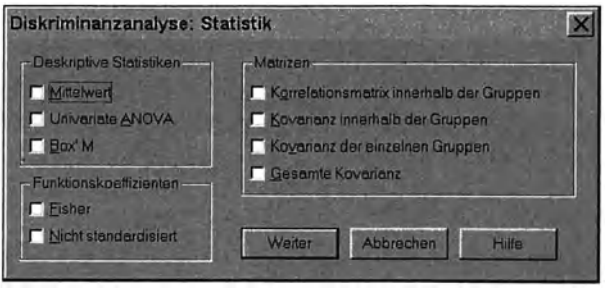


Abb. 20.6. Dialogbox "Diskriminanzanalyse: Statistik"

③ *Methode.* Ähnlich wie bei der Regressionsanalyse ist es auch in der Diskriminanzanalyse möglich, die unabhängigen Variablen schrittweise in die Berechnung einer Diskriminanzanalyse aufzunehmen. Dabei können sowohl Variablen aufgenommen als auch wieder ausgeschlossen werden. Wählt man in der Dialogbox „Diskriminanzanalyse" (⇨ Abb. 20.4) die Option „Schrittweise Methode verwenden", so wird die Schaltfläche „Methode" aktiv. Nach Klicken von „Methode" öffnet sich die in Abb. 20.7 dargestellte Dialogbox „Diskriminanzanalyse: Schrittweise Methode" mit folgenden Wahlmöglichkeiten:

❑ *Methode.* Man kann eine der nachfolgend aufgeführten statistischen Maßzahlen wählen, die Grundlage für die Aufnahme oder für den Ausschluss von Variablen werden sollen. Als Kriterium für die Aufnahme bzw. für den Ausschluss einer Variablen dient ein partieller F-Test. Die Prüfvariable für den F-Test ist dabei mit der jeweiligen statistischen Maßzahl verknüpft, wie hier nur am Beispiel von Wilks´ Lambda näher erläutert werden soll.

- *Wilks´ Lambda.* Bei jedem Schritt wird jeweils die Variable aufgenommen, die Wilks´ Lambda (λ) gemäß Gleichung 20.7 am meisten verkleinert. Die Prüfvariable des partiellen F-Tests zur Signifikanzprüfung für die Aufnahme einer zusätzlichen Variablen (bzw. den Ausschluss einer Variablen) berechnet sich gemäß Gleichung 20.10 (n = Fallanzahl, m = Variablenanzahl, k = Gruppenanzahl, λ_m = Wilks´ Lambda bei Einschluss von m unabhängigen Variablen, λ_{m+1} = bei Einschluss oder Ausschluss einer weiteren Variablen). In der Dialogbox kann man unter „Kriterien" die Grenzwerte von F für die Aufnahme und den Ausschluss festlegen. Voreingestellte Werte sind 3,84 und 2,71. Alternativ kann man anstelle von F-Werten Wahrscheinlichkeiten vorgeben. Voreingestellte Werte sind $\alpha = 0{,}05$ und $\alpha = 0{,}1$.

$$F = \left(\frac{n-k-m}{k-1}\right)\left(\frac{1-\lambda_{m+1}/\lambda_m}{\lambda_{m+1}/\lambda_m}\right) \qquad (20.10)$$

- *Nicht erklärte Varianz.* Bei jedem Schritt wird jeweils die Variable aufgenommen, die $SAQ_{innerhalb}$ (= nicht erklärte Streuung) am meisten verringert.

- *Mahalanobis-Abstand.* Dieses Distanzmaß misst, wie weit die Werte der unabhängigen Variablen eines Falles vom Mittelwert aller Fälle abweichen. Bei jedem Schritt wird jeweils die Variable aufgenommen, die den Abstand am meisten verkleinert.
- *Kleinster F-Quotient.* Es wird bei jedem Schritt ein F-Quotient maximiert, der aus der Mahalanobis-Distanz zwischen den Gruppen berechnet wird.
- *Rao V.* Es handelt sich um ein Maß für die Unterschiede zwischen Gruppenmittelwerten. Bei jedem Schritt wird die Variable aufgenommen, die zum größten Rao V führt. Der Mindestanstieg von V für eine aufzunehmende Variable kann festgelegt werden.

❑ *Kriterien.* Hier werden Grenzwerte für den partiellen F-Test festgelegt. Sie können entweder die Option „F-Wert verwenden" oder „F-Wahrscheinlichkeit verwenden" wählen. Die voreingestellten Werte können verändert werden. Mit einer Senkung des Aufnahmewertes von F (bzw. Erhöhung der Aufnahmewahrscheinlichkeit) werden mehr Variable aufgenommen und mit einer Senkung des Ausschlusswertes (bzw. Erhöhung der Ausschlusswahrscheinlichkeit) weniger Variablen ausgeschlossen.

❑ *Anzeigen.*

- *Zusammenfassung der Schritte.* Nach jedem Schritt werden Statistiken für alle (ein- und ausgeschlossenen) Variablen angezeigt.
- *F für paarweise Distanzen.* Es wird eine Matrix paarweiser F-Quotienten für jedes Gruppenpaar angezeigt. Dieses Maß steht in Verbindung zur Methode „Kleinster F-Quotient".

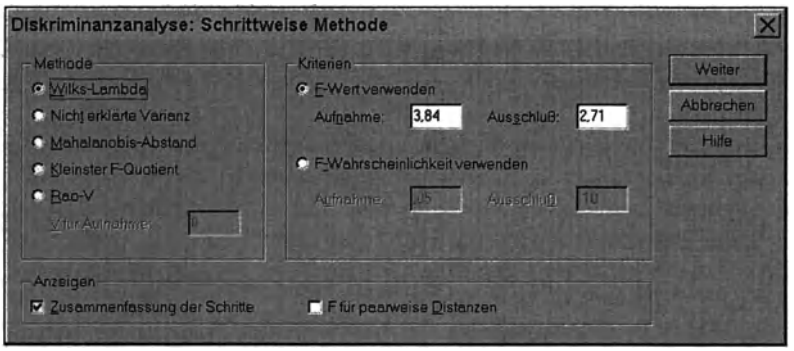


Abb. 20.7. Dialogbox „Diskriminanzanalyse: Schrittweise Methode"

Tabelle 20.9. Ergebnisausgabe: Schrittweise Methode

Aufgenommene/Entfernte Variablen[a,b,c,d]

		Wilks-Lambda							
						Exaktes F			
Schritt	Aufgenommen	Statistik	df1	df2	df3	Statistik	df1	df2	Signifikanz
1	LALT	,412	1	1	216,000	308,444	1	216,00	,000
2	LAST	,352	2	1	216,000	198,261	2	215,00	,000
3	LGLDH	,336	3	1	216,000	140,834	3	214,00	,000

Bei jedem Schritt wird die Variable aufgenommen, die das gesamte Wilks-Lambda minimiert.

a. Maximale Anzahl der Schritte ist 8.

b. Minimaler partieller F-Wert für die Aufnahme ist 3.84.

c. Maximaler partieller F-Wert für den Ausschluß ist 2.71.

d. F-Niveau, Toleranz oder VIN sind für eine weitere Berechnung unzureichend.

④ *Klassifizieren.* Nach Klicken von „Klassifizieren" (⇨ Abb. 20.4) öffnet sich die in Abb. 20.9 dargestellte Dialogbox. Es bestehen folgende Wahlmöglichkeiten:

❑ *A-priori-Wahrscheinlichkeit.* Die A-priori-Wahrscheinlichkeit P(G) in Gleichung 20.5 kann vorgegeben werden:

- *Alle Gruppen gleich.* Im Fall von z. B. zwei Gruppen wird P(G)= 50 v.H. für beide Gruppen (G = 1,2) vorgegeben.
- *Aus der Gruppengröße berechnen.* Hier wird die Wahrscheinlichkeit P(G) durch den Anteil der Fälle in der Gruppe G an allen Fällen berechnet. Im Beispiel ist für die 1. Gruppe (virale Hepatitis) P(G = 1) = 57/216 = 26,147 %, da in der Datei LEBER.SAV von 218 Fällen 57 Fälle mit viraler Hepatitis vorliegen.

❑ *Anzeigen.*

- *Fallweise Ergebnisse.* Die Ergebnisausgabe kann auf eine vorzugebende Anzahl von (ersten) Fällen beschränkt werden. In Abb. 20.10 ist das Ausgabeergebnis für die ersten 6 Fälle zu sehen. Standardmäßig werden in der Ausgabe Informationen zur „höchsten" und „zweithöchsten" Gruppe gegeben. Da es sich hier um einen Zwei-Gruppenfall handelt, werden in der „höchsten" Gruppe Informationen zu Fällen mit viraler Hepatitis und in der „zweithöchsten" Gruppe zu Fällen mit anderen Lebererkrankungen gegeben. In den ersten 6 Fällen stimmt die mit dem Diskriminanzgleichungsmodell vorhergesagte Gruppe mit der tatsächlichen Gruppe überein. In der letzten Spalte wird der Diskriminanzwert aufgeführt. Für den ersten Fall beträgt dieser 2,521.

  $P(D_i \geq d/G = g) = P(D_i \geq 2{,}521/G = 1) = 0{,}866$ ist die in Gleichung 20.5 im Zähler und Nenner des Bruches aufgeführte bedingte Wahrscheinlichkeit des Auftretens des beobachteten Diskriminanzwerts für den ersten Fall bei Annahme der Zugehörigkeit zur ersten Gruppe (G = 1). Die entsprechende Wahrscheinlichkeit bei Zuordnung zur zweiten Gruppe ergibt sich als Komplement zu 1. Da die A-posteriori-Wahrscheinlichkeit (vergl. Gleichung 20.5) $P(G = 0/D_i = 2{,}521) = 0{,}996$ größer ist als

$P(G = 2 / D_i = 2{,}521) = 0{,}004$ führt das Modell für den ersten Fall zur Vorhersage der Zugehörigkeit zur ersten Gruppe (virale Hepatitis) und damit zur richtigen Zuordnung. Die für beide Gruppen aufgeführte quadrierte Distanz nach Mahalanobis misst den Abstand der einzelnen Fälle vom Zentrum der jeweiligen Gruppe. Auch an diesem Abstand kann man erkennen, dass der erste Fall der ersten Gruppe zugeordnet werden sollte.

Tabelle 20.10. Ergebnisausgabe: Fallweise Ergebnisse

Fallweise Statistiken

Fallnummer	Tatsächliche Gruppe	Höchste Gruppe					Zweithöchste Gruppe			Diskriminanzwerte
		Vorhergesagte Gruppe	P(D>d | G=g) p	P(D>d | G=g) df	P(G=g | D=d)	Quadrierter Mahalanobis-Abstand zum Zentroid	Gruppe	P(G=g | D=d)	Quadrierter Mahalanobis-Abstand zum Zentroid	Funktion 1
1	0	0	,866	1	,996	,029	1	,004	11,244	2,521
2	0	0	,378	1	,906	,777	1	,094	5,304	1,470
3	0	0	,421	1	,925	,648	1	,075	5,662	1,547
4	0	0	,733	1	,982	,116	1	,018	8,087	2,011
5	0	0	,888	1	,996	,020	1	,004	11,060	2,493
6	0	0	,623	1	,971	,241	1	,029	7,254	1,861

- *Zusammenfassende Tabelle.* In Tabelle 20.11 wird das Vorhersageergebnis der Gruppenzugehörigkeit mit der tatsächlichen Gruppenzugehörigkeit der Fälle verglichen. Insgesamt werden 11 bzw. 5 v.H. (11 von 218) der Fälle (ein Fall viraler Hepatitis und 10 Fälle anderer Lebererkrankungen) durch das Diskriminanzmodell fehlerhaft zugeordnet. Im Vergleich zu der hier angenommenen A-priori-Wahrscheinlichkeit von 50 v.H. für die Gruppenzuordnung ist die korrekte Zuordnungsquote von 95 v.H. durch das Modell beträchtlich.

Tabelle 20.11. Übersicht über das Klassifizierungsergebnis

Klassifizierungsergebnisse[a]

		GRUP1	Vorhergesagte Gruppenzugehörigkeit: virale Hepatitis	Vorhergesagte Gruppenzugehörigkeit: andere Lebererkrankung	Gesamt
Original	Anzahl	virale Hepatitis	56	1	57
		andere Lebererkrankung	10	151	161
	%	virale Hepatitis	98,2	1,8	100,0
		andere Lebererkrankung	6,2	93,8	100,0

a. 95,0% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.

- Klassifikation *mit Fallauslastung*. Die Tabelle 20.11 wird um eine „Kreuzvalidierung" ergänzt. In der Kreuzvalidierung ist jeder Fall durch die Funktionen klassifiziert, die von allen anderen Fällen außer diesem Fall abgeleitet werden.

❑ *Fehlende Werte durch Mittelwerte ergänzen*. Ob man von dieser Option Gebrauch machen soll, muss gut überlegt sein.

❑ *Kovarianzmatrix verwenden*. Bei Wahl von „Gruppenspezifisch" können sich die Ergebnisse im Vergleich zu „Innerhalb der Gruppen" unterscheiden.

❑ *Grafiken*. Es werden für die Diskriminanzwerte (D) Häufigkeitsverteilungen in Form von Histogrammen oder Streudiagrammen erstellt.

- *Kombinierte Gruppen*. Eine Grafik wird nur für den Fall mehrerer Diskriminanzfunktionen erstellt.
- *Gruppenspezifisch*. In Abb. 20.8 (entspricht Abb. 20.3) ist das Ergebnis zu sehen. Für jede Gruppe wird eine Häufigkeitsverteilung grafisch dargestellt. Die Überlagerung beider Häufigkeitsverteilungen ist deutlich sichtbar. Aus den beigefügten Angaben wird ersichtlich, dass die Mittelwerte der Diskriminanzwerte beider Gruppen sich mit $\overline{D}_1 = 2{,}35$ (virale Hepatitis) und $\overline{D}_2 = -0{,}83$ (andere Lebererkrankung) stark unterscheiden.
- *Territorien*. Diese Grafik hat nur im Fall von mehr als zwei Gruppen Bedeutung.

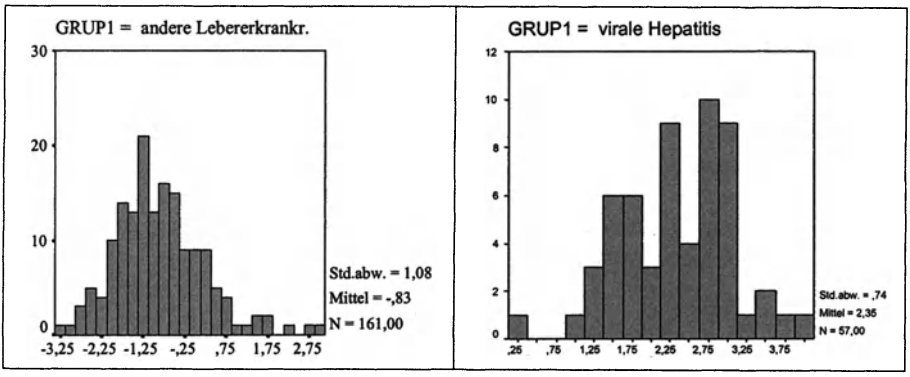


Abb. 20.8. Häufigkeitsverteilungen der Diskriminanzwerte beider Gruppen

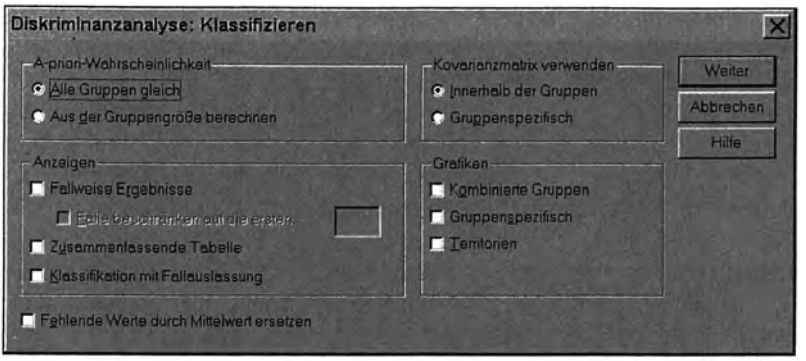


Abb. 20.9. Dialogbox „Diskriminanzanalyse: Klassifizieren"

⑤ *Speichern.* Nach Klicken der Schaltfläche „Speichern" (Abb. 20.4) öffnet sich die in Abb. 20.10 dargestellte Dialogbox. Eine Auswahl zu speichernder Variablen erfolgt durch Klicken auf die Kontrollkästchen. Den Variablen des Datensatzes werden die vorhergesagte Gruppenzugehörigkeit mit der Variable dis_, der Wert der Diskriminanzfunktion mit dis1 und die Wahrscheinlichkeiten der Gruppenzugehörigkeit mit dis2_ hinzugefügt. Des weiteren kann im XML-Format gespeichert werden.

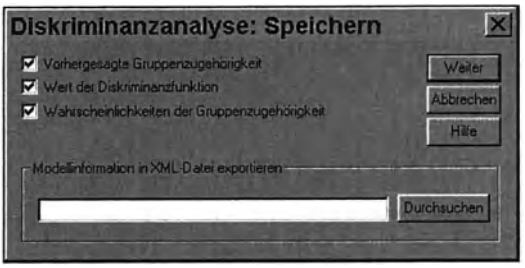


Abb. 20.10. Dialogbox „Diskriminanzanalyse: Neue Variablen speichern"

Diskriminanzanalyse für mehr als zwei Gruppen. Liegen in dem Datensatz mehr als zwei Gruppen vor, so geht man bei der Durchführung der Analyse analog zum Zwei-Gruppenfall vor. In der in Abb. 20.4 dargestellten Dialogbox muss ebenfalls die entsprechende Variable, die die Gruppenzugehörigkeit der Fälle festhält, mit ihrem Wertebereich aufgeführt werden.

Die Ergebnisse der Diskriminanzanalyse unterscheiden sich vom Zwei-Gruppenfall darin, dass nun mehr als eine Diskriminanzfunktion berechnet wird. Liegen k Gruppen vor, so werden k–1 Diskriminanzfunktionen bestimmt. Die Diskriminanzfunktionen werden derart bestimmt, dass sie orthogonal zueinander sind (die D-Achsen sind zueinander rechtwinklig). Ein zweite Diskriminanzfunktion wird derart ermittelt, dass diese einen maximalen Anteil der Streuung erklärt, die nach Bestimmung der ersten Diskriminanzfunktion als Rest verbleibt usw. Im Output

erscheinen die Diskriminanzfunktionen als Funktion 1, Funktion 2 etc. Der Eigenwertanteil einer Diskriminanzfunktion an der Summe der Eigenwerte aller Funktionen ist ein Maß für die relative Bedeutung der Diskriminanzfunktion.

In der Datei LEBER.SAV enthält die Variable GRUP2 drei Gruppen (virale Hepatitis, chronische Hepatitis, andere Lebererkrankungen). Es werden zwei Diskriminanzfunktionen berechnet. Mit Ausnahme einer Grafik wird hier aus Platzersparnisgründen auf die Wiedergabe der Ergebnisse der Diskriminanzanalyse verzichtet. In Abb. 20.11 ist ein Koordinatensystem mit den Werten beider Diskriminanzfunktionen als Achsen zu sehen. In diesem Koordinatensystem sind analog der Abb. 20. 1 die einzelnen Fälle der Datei dargestellt. Durch die Vergabe unterschiedlicher Symbole für die drei Gruppen wird deutlich, dass die drei Gruppen voneinander getrennte Punktwolken bilden. Die Lage einer jeden Punktwolke wird durch den Gruppenmittelpunkt (Zentroid) bestimmt. Diese Grafik wird angefordert, wenn in der Dialogbox 20.9 in Grafiken „Kombinierte Gruppen" gewählt wird. Wählt man „Gruppenspezifisch", so wird für jede Gruppe eine entsprechende Grafik dargestellt. Wählt man „Territorien", so entsteht ebenfalls eine Grafik mit den Diskriminanzwerten beider Funktionen. Es werden aber nicht die Fälle der drei Gruppen, sondern die Gruppenmittelwerte und Trennlinien für die drei Cluster abgebildet. Die Trennlinien (analog der Trennlinie in Abb. 20.1 für zwei Cluster im 2-Variablen-Koordinatensystem) werden durch Ziffernkombinationen dargestellt. Die Ziffernkombination 31 z. B. besagt, dass es sich um die Trennlinie zwischen Cluster 1 und 3 handelt.

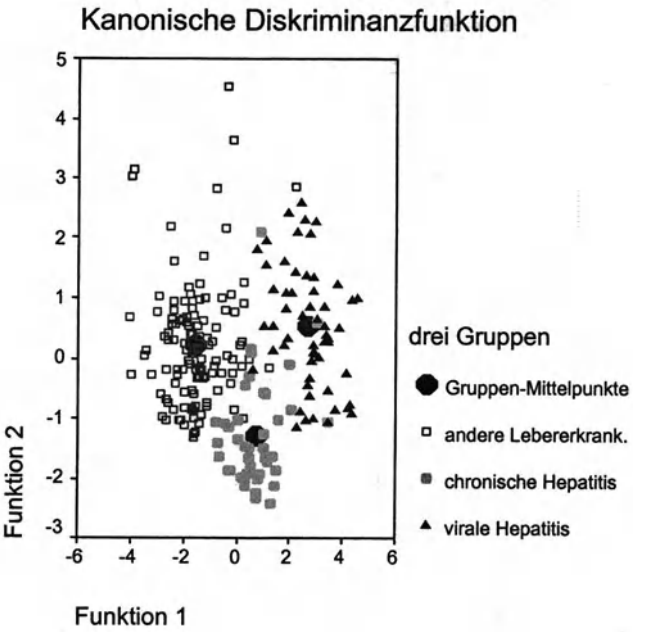


Abb. 20.11. Punktwolken im Diskriminanzraum

21 Faktorenanalyse

21.1 Theoretische Grundlagen

Oftmals kann man davon ausgehen, dass sich eine Menge miteinander korrelierter Beobachtungsvariablen (auch als Observablen oder Indikatoren bezeichnet) auf eine kleinere Menge latenter Variablen (Faktoren) zurückführen lässt. Bei der Faktorenanalyse handelt es sich um eine Sammlung von Verfahren, die es erlauben, eine Anzahl von Variablen auf eine kleinere Anzahl von Faktoren oder Komponenten zurückzuführen[1].

Mögliche Ziele einer Faktorenanalyse können sein:

- *Aufdeckung latenter Strukturen.* Es sollen hinter den Beobachtungsvariablen einfachere Strukturen entdeckt und benannt werden.
- *Datenreduktion.* Die Messwerte der Variablen sollen für die weitere Analyse durch die geringere Zahl der Werte der dahinterstehenden Faktoren ersetzt werden.
- *Entwicklung und Überprüfung eines Messinstruments.* Die Faktorenanalyse dient dazu, ein mehrteiliges Messinstrument (z.B. Test) auf Eindimensionalität zu prüfen oder von in dieser Hinsicht unbefriedigenden Teilinstrumenten zu bereinigen.

In jedem dieser Fälle kann entweder explorativ (ohne vorangestellte Hypothese) oder konfirmatorisch (Überprüfung einer vorangestellten Hypothese) verfahren werden.

Eine Faktorenanalyse vollzieht sich in folgenden Schritten:

① Vorbereitung einer Korrelationsmatrix der Beobachtungsvariablen (mitunter auch Kovarianzmatrix).
② Extraktion der Ursprungsfaktoren (zur Erkundung der Möglichkeit der Datenreduktion).
③ Rotation zur endgültigen Lösung und Interpretation der Faktoren.
④ Eventuelle Berechnung der Faktorwerte für die Fälle und Speicherung als neue Variable.

---

1 Wir besprechen hier die R-Typ Analyse. Diese untersucht Korrelationen zwischen Variablen. Die weniger gebräuchliche Q-Typ Analyse dagegen untersucht Korrelationen zwischen Fällen und dient zur Gruppierung der Fälle, ähnlich der Clusteranalyse.

Unterschiede zwischen den verschiedenen Verfahren ergeben sich in erster Linie bei den Schritten ② und ③. Sowohl für die Extraktion als auch die Rotation existieren zahlreiche Verfahren, die zu unterschiedlichen Ergebnissen führen.

Wichtige Differenzen bestehen darin, ob:

- *Unique* Faktoren angenommen werden oder nicht (⇨ unten).
- Eine *rechtwinklige* (orthogonale) oder eine *schiefwinklige* (oblique) Rotation vorgenommen wird. Ersteres unterstellt unkorrelierte, letzteres korrelierte Faktoren.

Der Kern des Verfahrens besteht in der Extraktion der Faktoren. Diese geht von der Matrix der Korrelationen zwischen den Variablen aus. In der Regel werden die Produkt-Moment-Korrelations-Koeffizienten zugrunde gelegt. Daraus ergibt sich als Voraussetzung: Vorhandensein mehrerer *normalverteilter, metrisch skalierter, untereinander korrelierte* Merkmalsvariablen X_j (j=1,...,m). Ergebnis ist: Eine geringere Zahl *normalverteilter, metrisch skalierter, nicht unmittelbar beobachtbarer* (und bei der in der Regel verwendeten orthogonalen Lösung untereinander nicht korrelierter) Variablen (Faktoren) F_p (p=1,...,k), mit deren Hilfe sich der Datensatz einfacher beschreiben lässt.

Es wird unterstellt, dass sich die beobachteten Variablen X_j als lineare Kombination der Faktorwerte F_p ausdrücken lassen (Fundamentaltheorem der Faktorenanalyse).

Der Variablenwert X_j eines Falles lässt sich aus den Faktorwerten errechnen:

$$X_j = A_{j1}F_1 + A_{j2}F_2 + ... + A_{jk}F_k \qquad (21.1)$$

F_p = gemeinsame (common) Faktoren der Variablen (p = 1...k)
A_{jp} = Konstanten des Faktors p der Variablen j

Oder, da die Faktorenanalyse mit standardisierten Werten (kleine Buchstaben stehen für die standardisierte Werte) arbeitet:

$$z_j = a_{j1}f_1 + a_{j2}f_2 + ... + a_{jk}f_k \qquad (21.2)$$

Die Koeffizienten a_{jp} werden als *Faktorladungen* bezeichnet.

Man unterscheidet in der Faktorenanalyse drei Arten von Faktoren:

- *Allgemeiner Faktor (general factor):* Die Ladungen sind für alle Variablen hoch.
- *Gemeinsamer Faktor (common factor):* Die Ladungen sind für mindestens zwei Variablen hoch.
- *Einzelrestfaktor (unique factor):* Die Ladung ist nur für eine Variable hoch.

Allgemeine Faktoren sind ein Spezialfall der gemeinsamen Faktoren. Sie interessieren nur bei einfaktoriellen Lösungen, wie sie z.B. bei der Konstruktion eindimensionaler Messinstrumente angestrebt werden. Einzelrestfaktoren sind Faktoren, die speziell nur eine Variable beeinflussen. Sie reduzieren den Erklärungswert eines Faktorenmodells (Fehlervarianz). Die Extraktionsverfahren unterscheiden sich u.a. darin, ob sie Einzelrestfaktoren in ihr Modell mit einbeziehen oder nicht. Von zentraler Bedeutung sind die gemeinsamen Faktoren. Ihr Wirken soll die Daten der Variablen erklären.

Werden Einzelrestfaktoren berücksichtigt, ändern sich die Gleichungen:

$$X_j = A_{j1}F_1 + A_{j2}F_2 + ... + A_{jk}F_k + U_j \quad (21.3)$$

Bei standardisierten Werten gilt für die Variable j:

$$z_j = a_{j1}f_1 + a_{j2}f_2 + ... + a_{jk}f_k + u_j \quad (21.4)$$

u_j = der Einzelrestfaktor (unique factor) der Variable j.

Die Berechnung der Koeffizienten (Faktorladungen) a_{jp} (j = 1,...,m; p = 1,...,k) stellt das Hauptproblem der Faktorenanalyse dar.

Umgekehrt können die Faktoren als eine lineare Kombination der beobachteten Variablen angesehen werden:

Generell gilt für die Schätzung des Faktors p aus m Variablen:

$$F_p = w_{1p}x_1 + w_{2p}x_2 + ... + w_{mp}x_m \quad (21.5)$$

w_{jp} = Factor-score Koeffizient des Faktors p der Variablen j

(In der Regel werden hier wieder nicht die Rohdaten, sondern z-transformierte Daten verwendet. Entsprechend wäre dann die Gleichung anzupassen.)

21.2 Anwendungsbeispiel für eine orthogonale Lösung

21.2.1 Die Daten

Zur Illustration wird ein fiktives Beispiel verwendet, das einerseits sehr einfach ist, da es nur zwei Faktoren umfasst, andererseits den Voraussetzungen einer Faktorenanalyse in fast idealer Weise entspricht.

Entgegen den normalen Gegebenheiten einer Faktorenanalyse seien uns die zwei Faktoren bekannt. Es handele sich um F1 (sagen wir Fleiß) und F2 (sagen wir Begabung). Beobachtbar seien sechs Variablen, z.B. die Ergebnisse von sechs verschiedenen Leistungstest V1 bis V6. Die Ergebnisse dieser Tests hängen von beiden Faktoren ab, sowohl von Begabung als auch von Fleiß, dies aber in unterschiedlichem Maße. (Zur besseren Veranschaulichung bei den graphischen Darstellungen wird hier allerdings – entgegen dem, was man in der Realität üblicherweise antrifft – die Variable V1 mit dem Faktor F1 und die Variable V2 mit dem Faktor F2 gleichgesetzt.) Schließlich wird jeder Wert einer Variablen auch noch von einem für diese Variable charakteristischen Einzelrestfaktor beeinflusst.

Die Beziehungen zwischen Faktoren, Einzelrestfaktoren und Variablen seien uns bekannt. Sie sind in den folgenden Gleichungen ausgedrückt:

$$V_1 = 0{,}8 \cdot F_1 + 0 \cdot F_2 + U_1$$
$$V_2 = 0{,}72 \cdot F_1 + 0{,}08 \cdot F_2 + U_2$$
$$V_3 = 0{,}56 \cdot F_1 + 0{,}24 \cdot F_2 + U_3$$
$$V_4 = 0{,}24 \cdot F_1 + 0{,}56 \cdot F_2 + U_4$$
$$V_5 = 0{,}08 \cdot F_1 + 0{,}72 \cdot F_2 + U_5$$
$$V_6 = 0 \cdot F_1 + 0{,}8 \cdot F_2 + 0{,}2 \cdot U_6$$

Die einzelnen Variablen werden von den Faktoren unterschiedlich stark bestimmt, wie stark ergibt sich aus den Koeffizienten der Gleichungen (den Faktorladungen). V1 wird z.B. sehr stark von F1 (Faktorladung/Gewicht = 0,8), aber gar nicht von F2 (Faktorladung = 0) beeinflusst. Außerdem – wie alle Variablen – durch einen Einzelrestfaktor mit dem Gewicht 0,2. Auch V2 und V3 werden überwiegend durch F1 bestimmt, aber z.T. auch von F2. Umgekehrt ist es bei den Variablen V6 bis V4.

Tabelle 21.1. Beispieldatensatz „LEISTUNG.SAV"

Fall	F1	F2	U1	U2	U3	U4	U5	U6	V1	V2	V3	V4	V5	V6
1	1	1	0,6	0,2	0,8	0,2	0,2	0,2	1,4	1	1,6	1	1	1
	-1,342	-1,342							-1,109	-1,680	-1,320	-1,845	-1,719	-1,505
2	1	2	0,4	0,2	0,2	0,2	0,6	0,4	1,2	1,08	1,24	1,56	2,12	2
	-1,342	-0,447							-1,334	-1,591	-1,827	-1,113	-0,456	-0,526
3	1	3	0,6	0,6	0,6	0,4	0,8	0,6	1,4	1,56	1,88	2,32	3,04	3
	-1,342	0,447							-1,109	-1,053	-0,926	-0,121	0,580	0,453
4	1	4	0,2	0,6	0,8	0,4	0,2	0,8	1	1,64	2,32	2,88	3,16	4
	-1,342	1,342							-1,559	-0,963	-0,306	0,611	0,716	1,432
5	2	1	0,2	0,8	0,4	0,4	0,4	0,2	1,8	2,32	1,76	1,44	1,28	1
	-0,447	-1,342							-0,660	-0,202	-1,095	-1,270	-1,403	-1,505
6	2	2	0,6	0,4	0,6	0,2	0,6	0,2	2,2	2	2,2	1,8	2,2	1,8
	-0,447	-0,447							-0,211	-0,560	-0,457	-0,800	-0,366	-0,722
7	2	3	0,2	0,4	0,6	0,6	0,8	0,2	1,8	2,08	2,44	2,76	3,12	2,6
	-0,447	0,447							-0,660	-0,470	-0,137	0,454	0,671	0,061
8	2	4	0,2	0,6	0,4	0,2	0,2	0,8	1,8	2,36	2,48	2,92	3,24	4
	-0,447	1,342							-0,660	-0,157	-0,081	0,663	0,806	1,432
9	3	1	0,6	0,2	0,4	0,2	0,6	0,4	3	2,44	2,32	1,48	1,56	1,2
	0,447	-1,342							0,688	-0,067	-0,306	-1,218	-1,088	-1,309
10	3	2	0,8	0,8	0,8	0,4	0,2	0,6	3,2	3,12	2,96	2,24	1,88	2,2
	0,447	-0,447							0,913	0,694	0,595	-0,225	-0,727	-0,330
11	3	3	0,2	0,2	0,4	0,6	0,8	0,6	2,6	2,6	2,8	3	3,2	3
	0,447	0,447							0,239	0,112	0,370	0,767	0,761	0,453
12	3	4	0,4	0,6	0,2	0,8	0,6	0,6	2,8	3,08	2,84	3,76	3,72	3,8
	0,447	1,342							0,463	0,650	0,426	1,760	1,347	1,236
13	4	1	0,2	0,6	0,4	0,8	0,4	0,8	3,4	3,56	2,88	2,32	1,44	1,6
	1,342	-1,342							1,137	1,187	0,482	-0,121	-1,223	-0,918
14	4	2	0,6	0,8	0,6	0,6	0,6	0,8	3,8	3,84	3,32	2,68	2,36	2,4
	1,342	-0,447							1,587	1,501	1,102	0,349	-0,186	-0,135
15	4	3	0,2	0,4	0,6	0,2	0,6	0,8	3,4	3,52	3,56	2,84	3,08	3,2
	1,342	0,447							1,137	1,143	1,439	0,558	0,652	0,649
16	4	4	0,2	0,6	0,8	0,4	0,8	0,6	3,4	3,8	4	3,6	4	3,8
	1,342	1,342							1,137	1,456	2,059	1,551	1,662	1,236

Auf Basis dieser Beziehungen wurde eine Datendatei für 16 Fälle erstellt. Sie wurde wie folgt gebildet. Die Faktoren 1 (Fleiß) und 2 (Begabung) sind metrisch skaliert und können nur die Messwerte 1, 2, 3 und 4 annehmen. Diese sind für die einzelnen Fälle bekannt. Sie wurden uniform verteilt, das heißt sind je vier mal vorhanden.[2] Die Faktoren sind völlig unkorreliert. Das wird dadurch erreicht, dass je vier Fälle auf dem Faktor 1 die Werte 1, 2 usw. haben, die vier Fälle mit demselben Wert auf Faktor 1, aber auf Faktor 2 je einmal den Wert 1, 2, 3 und 4 zugewiesen bekommen. Es werden zusätzlich für jeden Fall Werte für die Unique-Faktoren (d.h. für jede Variable einer) eingeführt. Sie sollen untereinander und mit den Faktoren unkorreliert sein. Am ehesten lässt sich dieses durch zufällige Zuordnung erreichen. Diesen Faktoren wurden mit der SPSS-Berechnungsfunktion TRUNC(RV.UNIFORM(1,5)) ganzzahlige Zufallszahlen zwischen 1 und 4 zugeordnet.

Nachdem die Faktorwerte bestimmt waren, konnten die Werte für die Variablen (V1 bis V6) mit den angegebenen Formeln berechnet werden. Die Ausgangswerte für die Faktoren und die daraus berechneten Werte der Variablen sind in Tabelle 21.1 enthalten. Die oberen Zahlen in den Zellen geben jeweils die Rohwerte, die unteren die z-Werte wieder. (Die z-Werte sind mit den Formeln für eine Grundgesamtheit und nicht, wie in SPSS üblich, für eine Stichprobe berechnet.) Die Rohdaten der Variablen sind als Datei LEISTUNG.SAV gespeichert.

Bei einer echten Analyse sind natürlich nur die Werte der Fälle auf den Variablen bekannt. Aus ihnen lässt sich eine Korrelationsmatrix für die Beziehungen zwischen den Variablen berechnen. Diese dient als Ausgangspunkt der Analyse. Die Analyse des Beispieldatensatzes müsste idealerweise folgendes leisten: Extraktion zweier Faktoren, diese müssten inhaltlich als Fleiß und Begabung interpretiert werden können; weiter Rekonstruktion der Formeln für die lineare Beziehung zwischen Faktoren und Variablen (d.h. in erster Linie: Ermittlung der richtigen Faktorladungen), Ermittlung der richtigen Faktorwerte für die einzelnen Fälle. Dies würde bei einer so idealen Konstellation wie in unserem Beispiel perfekt gelingen, wenn keine Einzelrestfaktoren vorlägen. Wirken Einzelrestfaktoren, kann die Rekonstruktion immer nur näherungsweise gelingen, die Einzelrestfaktoren selbst sind nicht rekonstruierbar.

21.2.2 Anfangslösung: Bestimmen der Zahl der Faktoren

Tabelle 21.2 zeigt die Korrelationsmatrix zwischen den Variablen (V1 bis V6). Ihr kann man bereits entnehmen, dass zwei Gruppen von Variablen (V1 bis V3 und V4 bis V6) existieren, die untereinander hoch korrelieren, d.h. eventuell durch einen Faktor ersetzt werden könnten. Allerdings fasst die Faktoranalyse nicht einfach Gruppen von Variablen zusammen, sondern isoliert die dahinterliegende latente Faktorenstruktur.

---

[2] Dies entspricht nicht der Voraussetzung der Normalverteilung, ist aber eine vernachlässigbare Verletzung der Modellvoraussetzungen.

Tabelle 21.2. Korrelationsmatrix

	V1	V2	V3	V4	V5	V6
V1	1,000	,933	,829	,360	,068	,010
V2	,933	1,000	,911	,577	,268	,245
V3	,829	,911	1,000	,728	,511	,484
V4	,360	,577	,728	1,000	,901	,886
V5	,068	,268	,511	,901	1,000	,935
V6	,010	,245	,484	,886	,935	1,000

Eine Faktorenanalyse muss dazu folgende Aufgabe lösen: Zu ermitteln sind die (unkorrelierten) Faktoren f_p (p = 1,...,k), deren Varianz *nacheinander* jeweils maximal ist. Die Varianz des Faktors f_p (s_p^2) ergibt sich aus der Summe der quadrierten Faktorladungen zwischen dem jeweiligen Faktor f_p und den Variablen x_j.

$$s_p^2 = a_{11}^2 + ... + a_{mk}^2 \qquad (21.6)$$

Die Ermittlung des ersten Faktors bedeutet die Lösung einer Extremwertaufgabe mit einer Nebenbedingung, die des zweiten Faktors eine Extremwertaufgabe mit zwei Nebenbedingungen etc.. Diese wird in der Mathematik über die Eigenvektoren/Eigenwerte einer Korrelationsmatrix gelöst. Die Faktorladungen lassen sich direkt aus den Eigenvektoren/Eigenwerten einer Korrelationsmatrix berechnen.

Allerdings entstehen zwei Probleme:

- Bei dieser Berechnung spielt die Diagonale der Korrelationsmatrix eine wesentliche Rolle. In ihr sind die Korrelationskoeffizienten der Variablen mit sich selbst durch die *Kommunalitäten* zu ersetzen. Die Kommunalität ist die durch die Faktoren erklärte Varianz einer Variablen. Bei Verwendung standardisierter Werte (Modell der Hauptkomponentenmethode) beträgt sie maximal 1. Bei Verwendung eines Modells, das keine Einzelrestfaktoren annimmt, ist die Kommunalität immer auch 1 und eine Lösung kann unmittelbar berechnet werden. Werden Einzelrestfaktoren angenommen, müssen dagegen die Kommunalitäten (die durch die gemeinsamen Faktoren erklärte Varianz) geringer ausfallen. Zur Faktorextraktion wird daher von der *reduzierten Korrelationsmatrix* ausgegangen. Das ist die Korrelationsmatrix, in der in der Diagonalen anstelle der Werte 1 die Kommunalitäten eingesetzt werden. Diese sind aber zu Beginn der Analyse nicht bekannt. Sie können nur aus den Faktorladungen gemäß Gleichung 21.7 (entspricht den Korrelationskoeffizienten zwischen Faktoren und Variablen) berechnet werden, die aber selbst erst aus der Matrix zu bestimmen sind. Es werden daher zunächst geschätzte Kommunalitäten eingesetzt und die Berechnung erfolgt iterativ, d.h. es werden vorläufige Lösungen berechnet und so lange verbessert, bis ein vorgegebenes Kriterium erreicht ist.
- Die anfängliche Lösung ist immer eine Lösung, die den formalen mathematischen Kriterien entspricht, aber – wenn es sich nicht um eine Einfaktorlösung handelt – gewöhnlich keine Lösung, die zu inhaltlich interpretierbaren Faktoren führt. Es existiert eine Vielzahl formal gleichwertiger Lösungen. Durch eine Rotation soll eine auch inhaltlich befriedigende Lösung gefunden werden. Des-

halb ist für die Klärung der meisten Aufgaben der Faktorenanalyse erst die rotierte Lösung relevant. Allerdings ändern sich die Kommunalitäten nicht durch Rotation. Daher kann für die Auswahl der Zahl der Faktoren die Ausgangslösung herangezogen werden.

Zur Faktorextraktion stehen verschiedene Verfahren zur Verfügung, die mit unterschiedlichen Algorithmen arbeiten und bei entsprechender Datenlage zu unterschiedlichen Ergebnissen führen. Wir demonstrieren das *Hauptachsen–Verfahren*, das gebräuchlichste Verfahren. Die Eigenschaften der anderen in SPSS verfügbaren Verfahren werden anschließend kurz erläutert.

Das Hauptachsenverfahren geht in seinem Modell vom Vorliegen von Einzelrestfaktoren aus. Es ist daher ein iteratives Verfahren. Wie bei allen iterativen Verfahren erfolgt die Faktorextraktion in folgenden Schritten:

① Schätzung der Kommunalitäten.
② Faktorextraktion (d.h. Berechnung der Faktorladungsmatrix).
③ Berechnung der Kommunalitäten anhand der Faktorladungen.
④ Vergleich von geschätzten und berechneten Kommunalitäten.
- Falls annähernde gleich: Ende des Verfahrens.
- Ansonsten: Wiederholung ab Schritt ②.

Als Schätzwerte für die Kommunalitäten kann zunächst jeder beliebige Wert zwischen 0 und 1 eingesetzt werden. Man kennt allerdings die mögliche Untergrenze. Sie ist gleich der quadrierten multiplen Korrelation zwischen der betrachteten Variablen und allen anderen im Set. Diese wird daher häufig bei den Anfangslösungen als Schätzwert für die Kommunalität in die Diagonale der Korrelationsmatrix eingesetzt (R^2-Kriterium). So auch in diesem Verfahren und als Voreinstellung in SPSS bei allen Verfahren (außer der Hauptkomponentenanalyse). Die auf diese Weise veränderte Korrelationsmatrix nennt man *reduzierte Korrelationsmatrix*.

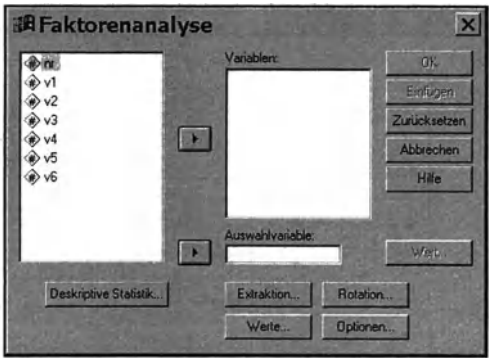


Abb. 21.1. Dialogbox „Faktorenanalyse"

Um eine Ausgangslösung für unser Beispiel zu erhalten, gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Dimensionsreduktion" und „Faktorenanalyse". Es öffnet sich die Dialogbox „Faktorenanalyse" (⇨ Abb. 21.1).
- ▷ Übertragen Sie die Variablen V1 bis V6 aus der Quellvariablenliste in das Feld „Variablen:".
- ▷ Klicken Sie auf die Schaltfläche „Extraktion". Es öffnet sich die Dialogbox „Faktorenanalyse: Extraktion" (⇨ Abb. 21.2).

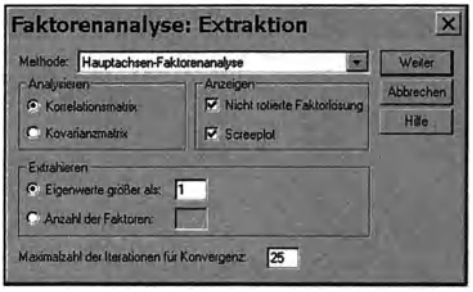


Abb. 21.2. Dialogbox „Faktorenanalyse: Extraktion"

- ▷ Klicken Sie auf den Pfeil neben dem Auswahlfenster „Methode:", und wählen Sie aus der sich öffnenden Liste „Hauptachsen-Faktorenanalyse".
- ▷ Klicken Sie auf die Kontrollkästchen „Nicht rotierte Faktorlösung" und „Screeplot" in der Gruppe „Anzeigen". (Dadurch werden die anfänglichen Faktorladungen und eine unten besprochene Grafik angezeigt.)
- ▷ Bestätigen Sie mit „Weiter" und „OK".

Tabelle 21.3: Anfängliche Faktorladungen und Kommunalitäten

Faktorenmatrix

	Faktor	
	1	2
V1	,64941	,73429
V2	,80659	,56144
V3	,91142	,31569
V4	,92391	-,32368
V5	,76258	-,59324
V6	,74007	-,62964

Kommunalitäten

	Anfänglich	Extraktion
V1	,939	,961
V2	,957	,966
V3	,926	,930
V4	,953	,958
V5	,929	,933
V6	,924	,944

Tabelle 21.3 enthält einen Teil der Ausgabe. Die Faktorenmatrix gibt die Faktorladungen der einzelnen Variablen an (bei einer Zwei-Faktoren-Lösung). Da wir noch keine Schlusslösung vorliegen haben, wären diese irrelevant, wenn sich daraus nicht die *Kommunalitäten* und die *Eigenwerte* errechnen ließen. Letztere sind für die Bestimmung der Zahl der Faktoren von Bedeutung.

Die *Kommunalität*, d.h. die gesamte durch die gemeinsamen Faktoren erklärte Varianz jeweils einer Variablen errechnet sich nach der Gleichung:

$$h_j^2 = a_{j1}^2 + a_{j2}^2 + ... + a_{jk}^2 \qquad (21.7)$$

Der unerklärte Anteil der Varianz (unique Varianz) einer Variablen j ist dann $1 - h_j^2$. Daraus ergibt sich auch der Koeffizient (Gewicht) des Einzelrestfaktors: $\sqrt{1 - h_j^2}$. Für die Variable V1 etwa gilt (nach Extraktion):

$h_1^2 = 0{,}64941^2 + 0{,}73429^2 = 0{,}961.$

Und für V4: $h_4^2 = 0{,}92391^2 + -0{,}32368^2 = 0{,}958$

Der *Eigenwert* ist der durch einen Faktor erklärte Teil der Gesamtvarianz. Der Eigenwert kann bei standardisierten Daten maximal gleich der Zahl der Variablen sein. Denn jede standardisierte Variable hat die Varianz 1. Je größer der Eigenwert, desto mehr Erklärungswert hat der Faktor. Die Eigenwerte lassen sich aus den Faktorladungen errechnen nach der Gleichung:

$$\lambda_p = a_{1p}^2 + a_{2p}^2 + ... + a_{mp}^2 \qquad (21.8)$$

Für Faktor 1 etwa gilt (bei einer Zwei-Faktorenlösung):

$$\lambda_1 = 0{,}649^2 + 0{,}807^2 + 0{,}911^2 + 0{,}924^2 + 0{,}763^2 + 0{,}740^2 = 3{,}886$$

Der Anteil der durch diesen Faktor erklärten Varianz an der Gesamtvarianz beträgt bei m Variablen:

$\frac{1}{m}\sum_{j=1}^{m} a_{j1}^2$, im Beispiel etwa für Faktor 1: $(1/6) \cdot 3{,}886 = 0{,}647$ oder 64,7%.

Sie sehen im zweiten Teil der Tabelle 21.4 „Summen von quadrierten Faktorladungen für Extraktion" in der Spalte „Gesamt" den Eigenwert 3,886 für Faktor 1. Das sind 64,674 % von der Gesamtvarianz 6, wie Sie aus der Spalte „% der Varianz" entnehmen können. Die beiden für die Extraktion benutzten Faktoren erklären zusammen 94,884 % der Gesamtvarianz. Wir haben also insgesamt ein sehr erklärungsträchtiges Modell vorliegen.

Tabelle 21.4. Erklärte Gesamtvarianz

Erklärte Gesamtvarianz

	Anfängliche Eigenwerte			Summen von quadrierten Faktorladungen für Extraktion		
Faktor	Gesamt	% der Varianz	Kumulierte %	Gesamt	% der Varianz	Kumulierte %
1	3,938	65,629	65,629	3,886	64,764	64,764
2	1,857	30,943	96,572	1,807	30,120	94,884
3	,075	1,252	97,824			
4	,072	1,202	99,026			
5	,037	,618	99,645			
6	,021	,355	100,000			

Allerdings werden anfänglich immer so viele Faktoren extrahiert, wie Variablen vorhanden sind. Alle Analyseverfahren – auch die hier verwendete Hauptachsen-Methode – bestimmen die anfängliche Lösung und die Zahl der Faktoren nach der *Hauptkomponentenmethode*. Bis dahin handelt es sich noch um keine Faktorenanalyse im eigentlichen Sinne, denn es wird lediglich eine Anzahl korrelierter Variablen in eine gleich große Anzahl unkorrelierter Variablen transformiert.

Es muss also nach dieser vorläufigen Lösung bestimmt werden, von wie vielen Faktoren die weiteren Lösungsschritte ausgehen sollen. Die vorliegende Lösung basiert deshalb auf zwei Faktoren, weil wir in der Dialogbox „Faktorenanalyse: Extraktion" (⇨ Abb. 21.2) in der Gruppe „Extrahieren" die Voreinstellung „Eigenwerte größer als: 1" nicht verändert haben. Die anfängliche Lösung (vor weiteren Iterationsschritten) mit noch 6 Faktoren sehen wir im ersten „Anfängliche Eigenwerte" überschriebenen Teil der Tabelle 21.4. Dort sehen wir für den Faktor 1 den Eigenwert 3,938, für den Faktor 2 1,857, den Faktor 3 0,075 usw.. Da der Eigenwert ab Faktor 3 kleiner als 1 war, wurden für die weitere Analyse nur 2 Faktoren benutzt.

Es sind allerdings mehrere Kriterien für die Bestimmung der Zahl der Faktoren gängig:

- *Kaiser-Kriterium*. Das voreingestellte Verfahren, nach dem Faktoren mit einem Eigenwert von mindestens 1 ausgewählt werden. Dem liegt die Überlegung zugrunde, dass jede Variable bereits eine Varianz von 1 hat. Jeder ausgewählte Faktor soll mindestens diese Varianz binden.
- *Theoretische Vorannahme* über die Zahl der Faktoren.
- Vorgabe eines *prozentualen Varianzanteils* der Variablen, der durch die Faktoren erklärt wird:
  - Anteil der Gesamtvarianz oder
  - Anteil der Kommunalität.
- *Scree-Test* (⇨ unten).
- *Residualmatrix-Verfahren*. Es wird so lange extrahiert, bis die Differenz zwischen der Korrelationsmatrix und der reduzierten Korrelationsmatrix nicht mehr signifikant ist.
- Jeder Faktor, auf dem eine Mindestzahl von Variablen hoch lädt, wird extrahiert.

Ein Scree-Plot ist die Darstellung der Eigenwerte in einem Diagramm, geordnet in abfallender Reihenfolge. Dabei geht man davon aus, dass die Grafik einem Berg ähnelt, an dessen Fuß sich Geröll sammelt. Entscheidend ist der Übergang vom Geröll zur eigentlichen Bergflanke. Diese entdeckt man durch Anlegen einer Geraden an die untersten Werte. Faktoren mit Eigenwerten oberhalb dieser Geraden werden einbezogen. Die Grundüberlegung ist, dass Eigenwerte auf der Geraden noch als zufällig interpretiert werden können. In unserem Beispiel (⇨ Abb. 21.3) liegen die Faktoren F3 bis F6 auf einer (ungefähren) Geraden, die das Geröll am Fuß des Berges markiert, während zu Faktor 2 und 1 eine deutliche Steigung eintritt. Daher würden wir auch nach diesem Kriterium eine Zwei-Faktorenlösung wählen.

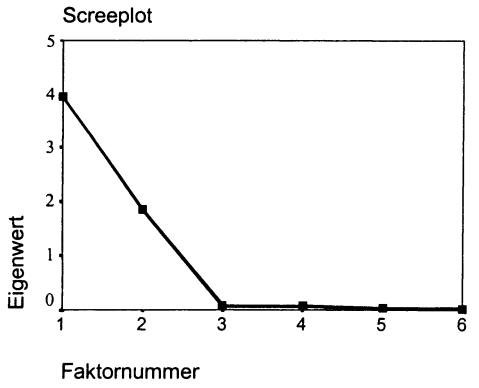


Abb. 21.3. Scree-Plot

Die Zahl der Faktoren für die iterative Lösung kann man steuern, indem man in der Dialogbox „Faktorenanalyse: Extraktion" (⇨Abb. 21.2) in der Gruppe „Extrahieren" die Voreinstellung „Eigenwerte größer als:" einen anderen Eigenwert als 1 einsetzt oder per Optionsschalter „Anzahl der Faktoren" und Eingabe einer Zahl die Zahl der Faktoren genau festlegt.

Verfügbare Extraktionsmethoden. SPSS bietet eine Reihe von Extraktionsmethoden. Die Methoden unterscheiden sich in dem Kriterium, das sie benutzen, eine gute Übereinstimmung (good fit) mit den Daten zu definieren. Sie werden hier kurz erläutert:

- *Hauptkomponenten Analyse (principal component).* Sie geht von der Korrelationsmatrix aus, mit den ursprünglichen Werten 1 in der Diagonalen. Die Berechnung erfolgt ohne Iteration.
- *Hauptachsen-Faktorenanalyse.* Verfährt wie die Hauptkomponentenanalyse, ersetzt aber die Hauptdiagonale der Korrelationsmatrix durch geschätzte Kommunalitäten und rechnet iterativ.
- *Ungewichtete kleinste Quadrate (unweighted least squares).* Produziert für eine fixierte (vorgegebene) Zahl von Faktoren eine factor-pattern Matrix, die die Summe der quadrierten Differenzen zwischen der beobachteten und der reproduzierten Korrelationsmatrix (ohne Berücksichtigung der Diagonalen) minimiert.
- *Verallgemeinerte kleinste Quadrate (generalized least squares).* Minimiert dasselbe Kriterium. Aber die Korrelationen werden invers gewichtet mit der Uniqueness (der durch die Faktoren nicht erklärte Varianz). Variablen mit hoher Uniqueness $1 - h_j^2$ wird also weniger Gewicht gegeben. Liefert auch einen χ^2-Test für die Güte der Anpassung. (Problematik wie Nullhypothesentest ⇨ Kap 13.3.)
- *Maximum Likelihood.* Produziert Parameterschätzungen, die sich am wahrscheinlichsten aus der beobachteten Korrelationsmatrix ergeben hätten, wenn diese aus einer Stichprobe mit multivariater Normalverteilung stammen. Wieder

werden die Korrelationen invers mit der Uniqueness gewichtet. Dann wird ein iterativer Algorithmus verwendet. Liefert auch einen χ^2-Test für die Güte der Anpassung.

- *Alpha-Faktorisierung.* Man sieht die Variablen, die in die Faktoranalyse einbezogen werden, als eine Stichprobe aus dem Universum von Variablen an. Man versucht einen Schluss auf die G
- *Image-Faktorisierung.* Guttman hat eine andere Art der Schätzung der common und unique Varianzanteile entwickelt. Die wahre Kommunalität einer Variablen ist nach dieser Theorie gegeben durch die quadrierte multiple Korrelation zwischen dieser Variablen und allen anderen Variablen des Sets. Diesen common part bezeichnet er als partial image.

21.2.3 Faktorrotation

Außer in speziellen Situationen (z.B. bei Einfaktorlösungen) führt die Anfangslösung der Faktorextraktion selten zu inhaltlich sinnvollen Lösungen, formal dagegen erfüllt die Lösung die Bedingungen. Das liegt daran, dass die Faktoren sukzessive extrahiert werden. So wird beim Hauptkomponenten- und Hauptachsenverfahren die Varianz der Faktoren über *alle* Variablen *nacheinander* maximiert. Daher korrelieren die Faktoren mit *allen* Variablen möglichst hoch. Deshalb tendiert der erste Faktor dazu, ein genereller Faktor zu sein, d.h. er lädt auf jeder Variablen signifikant. Alle anderen Faktoren dagegen neigen dazu bipolar zu werden, d.h. sie laden auf einem Teil der Variablen positiv, auf einem anderen negativ. Bei einer Zweifaktorsituation – wie in unserem Beispiel – lässt sich das anhand des Faktordiagramms für die unrotierte Lösung (⇨ Abb. 21.4) besonders gut verdeutlichen.

Sie erhalten dieses auf folgendem Weg: Klicken Sie in der Dialogbox "Faktorenanalyse" (⇨ Abb. 21.1) auf die Schaltfläche "Rotation". Es öffnet sich die Dialogbox "Faktorenanalyse: Rotation"(⇨ Abb. 21.5). Wählen Sie dort das Kontrollkästchen "Ladungsdiagramm(e)" aus.

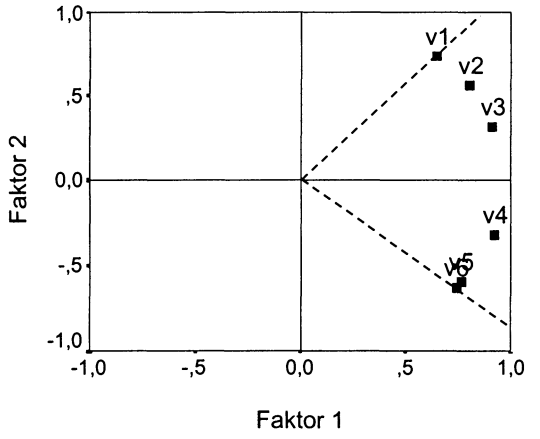


Abb. 21.4. Faktordiagramm

Das Faktordiagramm (Abb. 21.4.) ist für Zwei-Faktorenlösungen leicht zu lesen. Es enthält zwei Achsen (die senkrechte und waagrechte Linien durch die Ursprünge: 0,0), die die Faktoren darstellen. Bei orthogonalen Lösungen sind sie rechtwinklig angeordnet. In diesem Achsenkreuz sind die einzelnen Variablen durch Punkte repräsentiert. Variablen, die nahe beieinander liegen, korrelieren untereinander hoch. Je stärker eine Variablen von einem Faktor beeinflusst wird, desto näher liegt der Punkt an dessen Achse. Liegt er zum Ende der Achse, bedeutet dies, dass er alleine von diesem beeinflusst wird.

Da uns bekannt ist, dass die Variablen V1 bis V3 stark auf dem Faktor 1 laden, V1 sogar (bis auf die Wirkung des Einzelrestfaktors) mit diesem identisch ist, müsste bei der inhaltlich richtigen Lösung die Achse des Faktors 1 durch V1 laufen. Analog gilt dasselbe für die Variablen V4 bis V6 und den Faktor 2. Die zweite Achse müsste durch V6 laufen. Offensichtlich ist das nicht der Fall, sondern die Achse des Faktors 1 (es ist die horizontale Linie in der Mitte der Grafik) verläuft genau in der Mitte zwischen den Punktwolken hindurch. Das ist nach dem oben Gesagten über die Ermittlung des Faktors 1 verständlich. Er wird ja zunächst alleine ermittelt, und zwar als der Faktor, der die Varianz aller Variablen maximiert. Er muss also in der Mitte aller Variablenpunkte liegen.

Die richtige Achsenlage kann man aber erreichen, indem man die Achsen um einen bestimmten Winkel φ (Phi) um ihren Ursprung rotiert. Die Drehung φ erfolgt gegen den Uhrzeigersinn. Algebraisch bedeutet das: die Faktorladungsmatrix wird mit Hilfe einer *Transformationsmatrix* umgerechnet. In der Abbildung sind entsprechende Achsen gestrichelt eingezeichnet (sie werden nicht in dieser Weise von SPSS ausgegeben).

Dazu werden grundsätzlich zwei verschiedene Verfahren verwandt:

- *Orthogonale (rechtwinklige) Rotation.* Es wird unterstellt, dass die Faktoren untereinander nicht korrelieren. Die Faktorachsen verbleiben bei der Drehung im rechten Winkel zueinander.
- *Oblique (schiefwinklige) Rotation.* Es wird eine Korrelation zwischen den Faktoren angenommen. Entsprechend deren Größe werden die Achsen in schiefem Winkel zueinander rotiert.

Wiederum stehen mehrere Verfahren zur Verfügung, Die Methoden unterscheiden sich im benutzten Algorithmus und dem Kriterium, das sie benutzen, eine gute Übereinstimmung (good fit) mit den Daten zu definieren. Alle gehen nach irgendeinem Maximierungs- bzw. Minimierungskriterium vor und verfahren iterativ.

Die endgültige Lösung sollte sachlich bedeutsame (meaningful) Faktoren enthalten und eine einfache Struktur aufweisen. Thurstone hat einige verbreitet angewendete Regeln für eine Einfachstruktur entwickelt. Danach gilt für die Faktorladungsmatrix:

- Einzelne Variablen korrelieren möglichst nur mit einem Faktor hoch, mit allen anderen schwach (nur eine hohe Ladung in jeder Zeile der Faktorenmatrix).
- Einzelne Faktoren korrelieren möglichst entweder sehr hoch oder sehr niedrig mit den Variablen (keine mittelmäßige Ladung in den Spalten der Faktorenmatrix ⇨ Abb. 21.3).

In der Praxis lautet die Frage: Wie können wir bei einer gegebenen Zahl von Faktoren und einem festen Betrag der durch die Faktoren erklärten Varianz (oder dem festen Betrag der gesamten Kommunalitäten) die Reihen und/oder Spalten der Faktorladungsmatrix vereinfachen? Vereinfachen der Reihen heißt: In jeder Reihe sollen so viele Werte wie möglich nahe 0 sein. Vereinfachen der Spalten heißt: Jede Spalte soll so viele Werte wie möglich nahe 0 aufweisen. Beides führt zur gleichen vereinfachten Struktur. Und geometrisch ausgedrückt heißt das: 1. Viele Punkte sollten nahe den Endpunkten der Achsen liegen. 2. Eine große Zahl der Variablen soll nahe dem Ursprung liegen (nur bei mehr als zwei Faktoren). 3. Nur eine kleine Zahl von Punkten sollte von beiden Achsen abseits bleiben.

Die Rotation beeinflusst nicht die Korrelationsmatrix. Sie beeinflusst zwar die Faktorladungen, aber nicht die Kommunalitäten, d.h. den durch die Faktoren erklärten Varianzanteil einer Variablen (es ändert sich nur deren Verteilung auf die Faktoren). Vor allem ändert sie nicht den Eigenwert der Lösung insgesamt. Aber es ändern sich die durch die einzelnen Faktoren erklärten Varianzanteile (Eigenwerte). Daher ist die Anfangslösung für die Bestimmung der Zahl der Faktoren und die Beurteilung der Qualität des Modells geeignet, nicht aber zur Bestimmung inhaltlich interpretierbarer Faktoren und der Faktorladungen.

Wir führen für unser Beispiel eine Rotation nach dem am häufigsten benutzen Rotationsverfahren, der Varimax-Methode durch. Dazu verfahren Sie wie folgt:

- ▷ Führen Sie zunächst in der Dialogbox "Faktorenanalyse" (⇨ Abb. 21.1) wie oben die Auswahl der Variablen und "Faktorenanalyse: Extraktion" (⇨ Abb. 21.2) die Auswahl der Extraktionsmethode durch.
- ▷ Klicken Sie in der Dialogbox "Faktorenanalyse" auf die Schaltfläche "Rotation". Es öffnet sich die Dialogbox "Faktorenanalyse: Rotation" (⇨ Abb. 21.5).
- ▷ Wählen Sie in der Gruppe "Methode" den Optionsschalter "Varimax".
- ▷ Wählen Sie in der Gruppe "Anzeigen" die Kontrollkästchen "Rotierte Lösung" und "Ladungsdiagramm(e)".
- ▷ Bestätigen Sie mit "Weiter" und "OK".

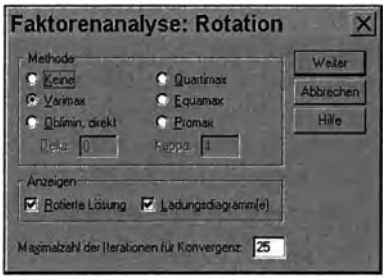


Abb. 21.5. Dialogbox "Faktorenanalyse: Rotation"

Es werden jetzt die wichtigsten Teile des Outputs besprochen. Tabelle 21.5 enthält auf der linken Seite das wichtigste Ergebnis, die rotierte Faktorenmatrix.

Die rotierte Faktorenladungsmatrix ist aus der anfänglichen Faktorenladungsmatrix durch Rotation entstanden. Die Umrechnungsfaktoren sind in der Faktortransformations-Matrix angegeben.

So ergibt sich z.B. die Faktorladung für V1 und den Faktor 1 aus:

$$(0{,}649 \cdot 0{,}723) + (0{,}734 \cdot -0{,}690) = -0{,}037$$

Die Faktorladungen haben sich also geändert. Gleichzeitig sehen wir aber in der rechten Tabelle "Kommunalitäten" in der Spalte "Extraktion" für V1 den Wert 0,961. Das ist derselbe Wert, den wir schon in Tabelle 21.3 vorfanden. Er hat sich nicht geändert, obwohl er sich durch die Summe der Quadrate der jetzt veränderten Faktorladungen ergibt.

Tabelle 21.5. Rotierte Faktormatrix, Kommunalitäten und Faktor-Transformationsmatrix

Rotierte Faktorenmatrix

	Faktor	
	1	2
V1	-,037	,980
V2	,196	,963
V3	,441	,858
V4	,892	,404
V5	,961	,097
V6	,970	,056

Kommunalitäten

	Anfänglich	Extraktion
V1	,939	,961
V2	,957	,966
V3	,926	,930
V4	,953	,958
V5	,929	,933
V6	,924	,944

Faktor-Transformationsmatrix

Faktor	1	2
1	,723	,690
2	-,690	,723

Tabelle 21.6. Erklärte Gesamtvarianz

Erklärte Gesamtvarianz

	Anfängliche Eigenwerte			Summen von quadrierten Faktorladungen für Extraktion			Rotierte Summe der quadrierten Ladungen		
Faktor	Gesamt	% der Varianz	Kumulierte %	Gesamt	% der Varianz	Kumulierte %	Gesamt	% der Varianz	Kumulierte %
1	3,938	65,629	65,629	3,886	64,764	64,764	2,895	48,248	48,248
2	1,857	30,943	96,572	1,807	30,120	94,884	2,798	46,637	94,884
3	,075	1,252	97,824						
4	,072	1,202	99,026						
5	,037	,618	99,645						
6	,021	,355	100,000						

Extraktionsmethode: Hauptachsen-Faktorenanalyse.

Tabelle 21.6 gibt die Eigenwerte der Faktoren und die Summe der Eigenwerte wieder. Sie ist in den ersten beiden Teilen identisch mit Tabelle 21.4. Neu ist der dritte, der die Eigenwerte nach der Faktorenrotation zeigt. Hier sehen wir deutliche Unterschiede. Vor der Rotation erklärte der erste Faktor mit ca. 65 % den größten Anteil der Varianz, der zweite dagegen nur ca. 30 %. Nach der Rotation erklären beide Faktoren praktisch gleich viel. Dies entspricht auch der Konstruktion unseres Beispiels. Insgesamt erklären aber beide Modelle einen gleich großen Anteil der Gesamtvarianz, nämlich 94,884 %.

Dem rotierten Ladungsdiagramm (Faktordiagramm im gedrehten Faktorbereich ⇨ Abb. 21.6) sieht man die Rotation auf den ersten Blick nicht an, denn die Faktorachsen werden aus technischen Gründen wie in der nicht-rotierten Matrix dargestellt. Statt der Achsen sind aber die Ladungspunkte rotiert, was auf dasselbe hinauskommt. Tatsächlich gehen jetzt die beiden Faktorachsen fast genau durch die Variablen V1 bzw. V6, wie es nach der Konstruktion unseres Beispiels sein muss.

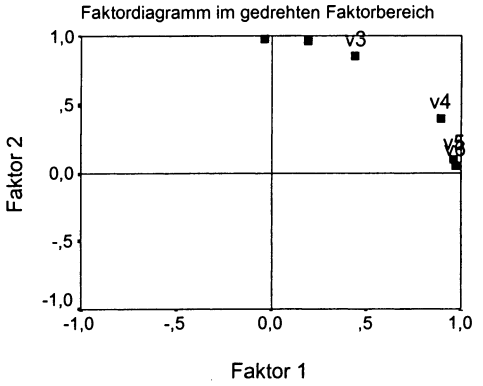


Abb. 21.6. Rotiertes Ladungsdiagramm

Für die inhaltliche Interpretation der Faktoren zieht man gewöhnlich *Leitvariablen* heran. Das sind Variablen, die auf diesem Faktor besonders hoch laden. An ihrem Inhalt erkennt man am ehesten die Bedeutung des Faktors. In unserem Beispiel lädt V6 besonders hoch auf Faktor 1, nämlich 0,97. Dagegen lädt V1 besonders hoch auf Faktor 2, nämlich 0,98. Da wir aus der Konstruktion wissen, dass V6 eine Variable ist, die Begabung misst, dagegen V1 eine, die Fleiß misst, würden wir Faktor 1 am bestem als "Begabung", Faktor 2 als "Leistung" bezeichnen. (Es ist also aufgrund der Rotationsrichtung genau umgekehrt, wie wir die Faktoren bei der Konstruktion des Beispiels benannt hatten.) Für die inhaltliche Interpretation kann es nützlich sein, in der Dialogbox "Faktorenanalyse: Optionen" (⇨ Abb. 21.13) die Kontrollkästchen "Sortiert nach Größe" und "Unterdrücken von Absolutwerten kleiner als" (mit einem Betrag zwischen 0 und 1, Voreinstellung 0,1) auszuwählen. Man sieht dann in der Tabelle "Rotierte Faktormatrix" besser, welche Variablen auf welchen Faktoren hoch laden.

Verfügbare Methoden für die orthogonale Rotation. SPSS bietet insgesamt fünf Rotationsmethoden an, davon drei für orthogonale, zwei für oblique Rotation. Die ersteren werden hier kurz erläutert.

- *Varimax.* Sie versucht, die Zahl der Variablen mit hohen Ladungen auf einem Faktor zu minimieren. Hier werden die *Spalten* der Faktorladungsmatrix simplifiziert. Ein einfacher Faktor ist einer bei dem in der Matrix der Faktorladungen in der Spalte annähernd die Werte 1 oder 0 auftreten. Dazu müssen die qua-

drierten Ladungen in der Spalte maximiert werden. (Daher der Name Varimax = Maximierung der Varianz der quadrierten Faktorladungen.)

- *Quartimax.* Das Verfahren minimiert die Zahl der zur Interpretation der Variablen notwendigen Faktoren. Das Verfahren sucht nach einer Simplifizierung der *Reihen* der Matrix. Dies ist der Fall, wenn:

$$\sum_{j=1}^{m}\sum_{p=1}^{k} a_{jp}^{4} \rightarrow \text{maximum} \qquad (21.9)$$

(Wegen der vierten Potenz in der Gleichung der Name Quartimax.) Der Mangel des Verfahrens besteht darin, dass es häufig in einem generellen Faktor resultiert mit mittleren und hohen Ladungen auf allen Variablen.

- *Equamax.* Ein Kompromiss zwischen Varimax und Quartimax. Das Verfahren versucht Reihen und einige Spalten zu vereinfachen.

21.2.4 Berechnung der Faktorwerte der Fälle

Häufig ist es sinnvoll, für jeden Fall die Werte auf den jeweiligen Faktoren (factor scores) zu berechnen. Insbesondere dient es der Vereinfachung der Beschreibung einer Analyseeinheit (Datenreduktion). Die Faktorwerte können für nachfolgenden Analysen verwendet werden.

Im Prinzip können die Faktorwerte als eine lineare Kombination der Werte der Variablen geschätzt werden.

Für den z-Wert des Falles i auf dem Faktor p ergibt sich:

$$\hat{f}_{ip} = \sum_{j=1}^{m} w_{jp} z_{ij} \qquad (21.10)$$

z_{ij} = standardisierter Wert der Variablen j für den Fall i

w_{jp} = Factor-score Koeffizient für die Variable j und den Faktor p.

Wie die Faktorladungen zur Berechnung der Variablenwerte als Gewichte benötigt werden, so werden umgekehrt zur Berechnung der Faktorwerte die *factor-score Koeffizienten* benötigt.

Aber nur bei der Hauptkomponentenanalyse können diese unter Verwendung der rotierten Faktorenmatrix genau berechnet werden. Bei allen anderen Methoden handelt es sich um über multiple Regression (⇨ Kap. 17) geschätzte Werte.

SPSS bietet drei Verfahren zur Schätzung von Faktorwerten an, die zu unterschiedlichen Ergebnissen führen. Alle drei führen zu standardisierten Faktorwerten (Mittelwert 0, Standardabweichung 1).

- *Regression (Voreinstellung).* Die Faktorwerte können korrelieren, selbst wenn die Faktoren orthogonal geschätzt wurden.
- *Bartlett.* Auch hier können die Faktorwerte korrelieren.
- *Anderson-Rubin.* Eine modifizierte Bartlett-Methode, bei der die Faktoren unkorreliert sind und eine Standardabweichung von 1 haben.

Zur Illustration arbeiten wir mit der Regressionsmethode.

SPSS gibt die *factor-score Koeffizienten* in der Tabelle "Koeffizientenmatrix der Faktorwerte" aus (⇨ Tabelle 21.7).

Wenn Sie diese Matrix erhalten möchten und die Faktorwerte der Datendatei hinzugefügt werden sollen, gehen Sie wie folgt vor:

- ▷ Klicken Sie in der Dialogbox "Faktorenanalyse" (⇨ Abb. 21.1) auf die Schaltfläche "Werte". Die Dialogbox "Faktorenanalyse: Faktorwerte" (⇨ Abb. 21.7) öffnet sich.
- ▷ Wählen Sie dort "Koeffizientenmatrix der Faktorwerte anzeigen".
- ▷ Wählen Sie "Als Variablen speichern".
- ▷ Wählen Sie in der Gruppe "Methode" eine Methode (hier: "Regression").
- ▷ Bestätigen Sie mit "Weiter" und "OK".

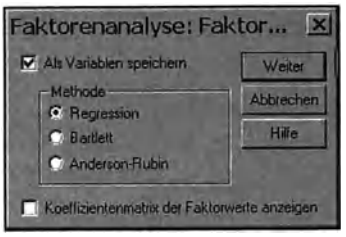


Abb. 21.7. Dialogbox "Faktorenanalyse: Faktorwerte"

Im Output finden Sie die angeforderte Matrix (⇨ Tabelle 21.7), und der Datenmatrix werden die Faktorwerte der einzelnen Fälle (hier bezeichnet als "fact1_1" und "fact2_1") angehängt. Für Fall 1 berechnet das Programm z.B. fact1_1 = -1,41212 und fact2_1 = -1,20271.

Tabelle 21.7. Koeffizientenmatrix der Faktorwerte

	Faktor	
	1	2
V1	-,152	,428
V2	-,142	,491
V3	,087	,140
V4	,427	-,056
V5	,261	-,024
V6	,342	-,065

Aus den jetzt verfügbaren Informationen ist es möglich, die Werte der Fälle, sowohl für die Variablen als auch die Faktoren, zu rekonstruieren, aus den Variablenwerten auch die Korrelationsmatrix. Allerdings wird wegen des Schätzcharakters der extrahierten Parameter dieses nur näherungsweise gelingen. Je stärker die Übereinstimmung mit den Ausgangswerten, desto besser die Lösung.

Die Faktorwerte der Fälle ergeben sich aus den Factor-score-Koeffizienten und den z-Werten der Variablen nach 21.9.

Der Wert des Faktors 1 des Falles 1 ist demnach:

$$f_{11} = -0{,}152 \cdot -1{,}109 + -0{,}142 \cdot -1{,}680 + 0{,}087 \cdot -1{,}320 + 0{,}427 \cdot -1{,}845 + 0{,}261 \cdot -1{,}719 + 0{,}342 \cdot -1{,}505 = -1{,}459$$

Dies stimmt wegen der Ungenauigkeiten bei der Schätzung der Parameter nicht genau mit unserem Ausgangswert von –1,342 überein und auch nicht mit dem durch SPSS ermittelten Wert von –1,412 (letzteres resultiert aus der unterschiedlichen Berechnung der z-Werte).

Die z-Werte der Variablen können aus den Faktorwerten und den Faktorladungen nach Gleichung 21.2 berechnet werden.

Für Fall 1 z.B. beträgt der Wert des ersten Faktors nach der Ausgabe von SPSS -1,41212, für Faktor 2 -1,20271. Die Faktorladungen entnehmen wir der Tabelle der "rotierten Faktorenmatrix". Wir wollen den Wert der Variablen V1 für Fall 1 berechnen. Für diese betragen die Faktorladungen –0,037 für Faktor 1 und 0,980 für Faktor 2. Demnach ist gemäß Gleichung 21.4:

$$z_{11} = -0{,}037 \cdot 1{,}41212 + 0{,}980 \cdot -1{,}20271 = -1{,}126$$

Das weicht natürlich etwas von dem tatsächlichen Wert -1,109 (bzw. -1,07415 nach der Berechnungsmethode von SPSS) ab. Das hängt damit zusammen, dass uns der Wert des Einzelrestfaktors unbekannt ist. Aus der Kommunalität von 0,961 für die Variable V1 können wir das Gewicht des Einzelrestfaktors nach der oben angegebenen Formel berechnen. Es beträgt $\sqrt{1-0{,}961} = 0{,}197$. Der Einzelrestfaktor beeinflusst also mit diesem Gewicht den z-Wert der Variablen. Das so berechnete Gewicht des Einzelrestfaktors entspricht recht genau dem von uns im Beispiel vorgegebenen Wert von 0,2.

Der z-Wert kann in den Rohwert der Variablen transformiert werden, wenn Mittelwert und Standardabweichung der entsprechenden Variablen bekannt sind. Das ist – anders als in der Realität – in unserem konstruierten Beispiel der Fall. Für die Variable V1 betragen sie 2,39 und 0,89. Daraus ergibt sich für V1 für den Fall 1:

$$V_{11} = 2{,}39 + (-1{,}126 \cdot 0{,}89) = 1{,}34$$

Der tatsächliche Ausgangswert V1 in unserem Beispiel war 1,4.

Auch die Tabelle "Reproduzierte Korrelationen" (Tabelle 21.8) gibt Auskunft über die Güte des Modells. Um diese zu erhalten, müssen Sie zunächst in der Dialogbox "Faktorenanalyse"(⇨ Abb. 21.1) auf die Schaltfläche "Deskriptive Statistik" klicken. In der sich öffnenden Dialogbox "Faktorenanalyse: Deskriptive Statistik" (⇨ Abb. 21.12) wählen Sie in der Gruppe "Korrelationsmatrix" die Option "reproduziert".

In der Tabelle sehen Sie im oberen Teil zunächst auf Basis des Modells reproduzierten Korrelationskoeffizienten. Im unteren Teil "Residuum" können Sie ablesen, wie stark diese von den ursprünglichen Korrelationen abweichen, z.B. weicht der Korrelationskoeffizient zwischen den Variablen V1 und V2 um $-0{,}003$, also mi-

nimal, vom ursprünglichen Korrelationskoeffizienten 0,936 ab. Bei einer guten Lösung sollten möglichst alle Residuen nahe Null liegen. Die Fußnote "a." gibt Auskunft, dass in dieser Tabelle kein einziges Residuum einen kritischen Wert von 0,05 überschreitet. Die Diagonale des oberen Teils der Tabelle enthält außerdem die reproduzierten Kommunalitäten.

Tabelle 21.8. Reproduzierte Korrelationen

		V1	V2	V3	V4	V5	V6
Reproduzierte Korrelation	V1	,961[b]	,936	,824	,362	,060	,018
	V2	,936	,966[b]	,912	,563	,282	,243
	V3	,824	,912	,930[b]	,740	,508	,476
	V4	,362	,563	,740	,958[b]	,897	,888
	V5	,060	,282	,508	,897	,933[b]	,938
	V6	,018	,243	,476	,888	,938	,944[b]
Residuum[a]	V1		-,003	,005	-,003	,009	-,008
	V2	-,003		-,001	,013	-,014	,002
	V3	,005	-,001		-,012	,004	,008
	V4	-,003	,013	-,012		,004	-,001
	V5	,009	-,014	,004	,004		-,003
	V6	-,008	,002	,008	-,001	-,003	

Extraktionsmethode: Hauptachsen-Faktorenanalyse.

a. Residuen werden zwischen beobachteten und reproduzierten Korrelationen berechnet. Es gibt 0 (,0%) nichtredundante Residuen mit Absolutwerten > 0,05.

b. Reproduzierte Kommunalitäten

21.3 Anwendungsbeispiel für eine oblique (schiefwinklige) Lösung

Zur Illustration einer schiefwinkligen Rotation wird ebenfalls ein fiktives Beispiel verwendet. Es enthält dieselben zwei Faktoren und 6 Variablen wie das Beispiel für die orthogonale Lösung. Der Unterschied besteht lediglich darin, dass für eine Korrelation der beiden Faktoren gesorgt wurde. Um dieses besser gewährleisten zu können, wurde die Zahl der Fälle auf 80 erhöht, je 20 pro Ausprägung 1, 2, 3, 4 auf dem Faktor 1. Um eine Korrelation der Faktoren zu erreichen, wurde aber nicht für eine gleiche Verteilung der Werte des Faktors 2 gesorgt, sondern die Verteilung wurde je nach Ausprägung auf Faktor 1 verändert nach dem Schema:

Faktor 1	**Faktor 2**
Wert	**Häufigkeit · Wert**
1	7 ·1, 6 · 2, 4 · 3 und 3 · 4
2	6 · 1, 7 · 2, 4 · 3 und 3 · 4
3	6 · 4, 7 · 3, 4 · 2 und 3 · 1
4	3 · 1, 4 · 2, 6 · 3 und 7 · 4

Die so erzeugten Faktoren korrelieren mit r = 0,276 miteinander. Die Variablenwerte wurden nach denselben Formeln aus den Faktorwerten und den Einzelrestfaktoren berechnet. Die Daten sind in der Datei "LEISTUNG2.SAV" gespeichert.

Am besten lässt sich die oblique Rotation wieder anhand von Faktorendiagrammen illustrieren. Nehmen wir an, wir führen für den neuen Datensatz dieselbe Faktorenanalyse wie im obigen Beispiel durch, also "Hauptachsen-Faktorenanalyse" mit der Rotationsmethode "Varimax" und lassen uns für die rechtwinklig rotierte Lösung ein Faktordiagramm ausgeben. Dann bekommen wir ein Ergebnis wie in Abb. 21.8.

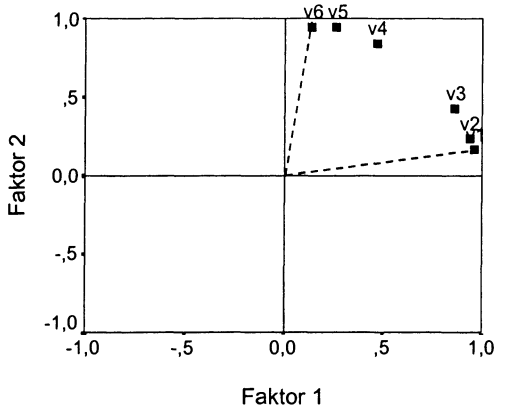


Abb. 21.8. Faktordiagramm für eine nach "Varimax" rotierte Lösung

Da wir aus der Bildung unseres Beispiels wissen, dass die Variablen V1 bzw. V6 genau den Faktoren 1 und 2 entsprechen, sehen wir, dass dieses Diagramm nicht ganz der Realität entspricht. Auch nach der Rotation liegen die Punkte für V1 und V6 nicht auf den Achsen. Beide liegen von den Achsen etwas nach innen versetzt. Die Achsen würden durch diese Punkte führen, wenn man den Winkel zwischen den Achsen etwas verändern würde, wie es die gestrichelten Linien andeuten (also nicht rechtwinklig rotieren). Wie der Winkel zu verändern ist, bleibt dem Augenmaß des Anwenders vorbehalten.

SPSS stellt für die schiefwinklige Rotation zwei Verfahren zur Verfügung:

- *Oblimin, direkt.* Bei diesem obliquen Verfahren wird die Schiefe durch einen Parameter δ (Delta) kontrolliert. Voreingestellt ist δ = 0. Das ergibt die schiefste mögliche Lösung. Die größte zulässige Zahl beträgt 0,8. Positive Werte sollten aber nicht verwendet werden. Negative Werte unter 0 führen zu zunehmend weniger schiefwinkligen Rotationen. Als Faustregel gilt, dass ca. bei –5 die Lösung nahezu orthogonal ausfällt (in unserem Beispiel eher früher).
- *Promax.* Eine schiefwinklige Rotation, die schneller als "Oblimin direkt" rechnet und daher für größere Datenmengen geeignet ist. Steuert die Schiefe über einen künstlichen Parameter Kappa. Voreingestellt ist ein Kappa von 4. Kappa-

werte sind positive Werte ab dem Mindestwert 1 bis maximal 9999. Unter 4 wird der Winkel der Achsen weiter, über 4 enger.

Wir benutzen zur Illustration die Methode "Oblimin, direkt". Einige Versuche ergaben, dass $\delta = -2{,}1$ zu einer recht guten Lösung führt. Diese erhalten Sie wie folgt:

- ▷ Wählen Sie in der Dialogbox "Faktorenanalyse: Rotation" (⇨ Abb. 21.5) in der Gruppe "Methode" den Optionsschalter "Oblimin, direkt".
- ▷ Tragen Sie in das Eingabefeld "Delta:" den Wert –2,1 ein.

Sie werden feststellen, dass SPSS bei der Voreinstellung (für die Maximalzahl der Iterationen) kein Ergebnis erzeugt und der Lauf mit der Fehlermeldung "Die Rotation konnte nicht mit 25 Iterationen konvergieren" abbricht.

- ▷ Ändern Sie deshalb im Eingabefeld "Maximalzahl der Iterationen für Konvergenz" die Zahl auf 50.

Das Faktordiagramm des Ergebnisses zeigt Abb. 21.9. Wiederum sind aus technischen Gründen nicht die Winkel zwischen den Achsen verändert (diese stehen nach wie vor senkrecht zueinander), sondern die Punktewolken sind entsprechend verschoben. Jedenfalls liegen jetzt V1 und V6 fast auf den Achsen der Faktoren 1 und 2. (Bei Verwendung von Promax würde ein Kappa von ca. 1,6 das beste Ergebnis zeigen.)

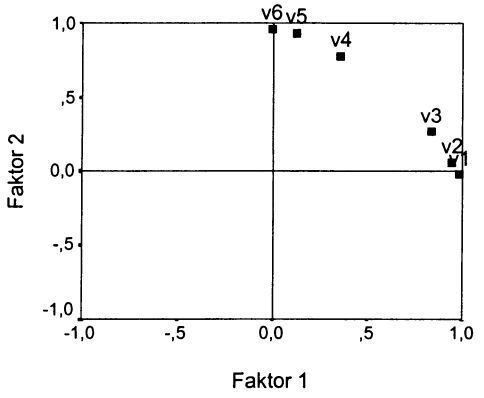


Abb. 21.9. Faktordiagramm für die rotierte Lösung "Oblimin direkt" $\delta = -2{,}1$

Auch die Ausgabe der obliquen Modelle unterscheidet sich etwas von der orthogonaler Modelle. Beide Arten von Modellen basieren auf derselben Korrelationsmatrix. Beide extrahieren für die Anfangsfaktoren die gleiche orthogonale Faktorlösung (in der Faktormatrix). Entsprechend unterscheiden sich weder die Kommunalitäten, noch die anfänglichen Eigenwerte der Faktoren. In der Tabelle "erkläre Gesamtvarianz" werden im letzten Teil der Tabelle für die rotierte Lösung bei obliquen Modellen zwar die Eigenwerte der Faktoren, aber nicht ihr Prozentanteil an

der Erklärung der Gesamtvarianz ausgegeben, weil bei schiefwinkligen Lösungen hierfür die genaue Basis fehlt.

Der Hauptunterschied tritt bei den endgültigen Faktoren nach der Rotation auf. Bei orthogonalen Lösungen erscheint die "Rotierte Faktormatrix". Sie enthält die Faktorladungen. Diese sind sowohl Regressionskoeffizienten für die Gleichungen, in denen Variablenwerte aus den Faktorwerten geschätzt werden, als auch Korrelationskoeffizienten zwischen Faktoren und Variablen. Bei obliquen Lösungen werden dagegen zwei Tabellen ausgegeben. Die erste heißt *"Mustermatrix"*. Sie enthält die Regressionskoeffizienten, d.h. gibt nur die direkten Wirkungen der Faktoren auf die Variable wieder, nicht die indirekten. Sie sind als Gewichte bei der Schätzung der Variablenwerte relevant. Die zweite heißt *"Strukturmatrix"*. Sie gibt die Korrelation zwischen Faktoren und Variablen an, also die direkte und indirekte Wirkung.

Tabelle 21.9. Muster-, Strukturmatrix und Korrelationsmatrix für die Faktoren bei obliquer Rotation

Mustermatrix[a]

	Faktor	
	1	2
V1	,985	-,024
V2	,945	,054
V3	,836	,268
V4	,357	,779
V5	,123	,932
V6	-,005	,959

a. Die Rotation ist in 33 Iterationen konvergiert.

Strukturmatrix

	Faktor	
	1	2
V1	,977	,308
V2	,963	,373
V3	,926	,550
V4	,619	,899
V5	,438	,974
V6	,318	,957

Korrelationsmatrix für Faktor

Faktor	1	2
1	1,000	,337
2	,337	1,000

Wir sehen z.B. in der Mustermatrix, dass der direkte Beitrag des Faktors 2 zur Variablen V1 negativ ist, nämlich –0,024. Alle Beiträge, direkte und indirekte zusammen, dagegen sind positiv, nämlich 0,308.

Da die Faktoren korreliert sind, liefert die oblique Lösung zusätzlich auch eine "Korrelationsmatrix für Faktor" genannte Tabelle, in der Korrelationskoeffizienten zwischen den Faktoren angegeben sind (⇨ Tabelle 21.9 rechts). Die hier extrahierten Faktoren korrelieren 0,337, was deutlich über den durch unsere Konstruktion gegeben "wahren Korrelationkoeffizienten" von 0,276 liegt.

21.4 Ergänzende Hinweise

21.4.1 Faktordiagramme bei mehr als zwei Faktoren

Normalerweise wird man eher Datensätze haben, bei denen mehr als zwei Faktoren extrahiert werden. Bei Anforderung von "Ladungsdiagrammen" gibt das Pro-

gramm dann ein dreidimensionales "Faktordiagramm im rotierten Raum" mit den ersten drei Faktoren als Achsen aus. Abb. 21.10 zeigt ein solches Diagramm für die Daten einiger Variablen zur Kennzeichnung von Stadtteileigenschaften, die für Hamburg erhoben wurden (Datei: VOLKSZ1). Es ist eine Hauptkomponentenanalyse mit Varimaxrotation durchgeführt worden, die zu vier Faktoren führte. Die ersten drei sind im Diagramm aufgenommen. Solche dreidimensionalen Diagramme sind häufig schwer zu lesen. Dann ist es zu empfehlen, sie in mehrere zweidimensionale Faktordiagramme umzuwandeln.

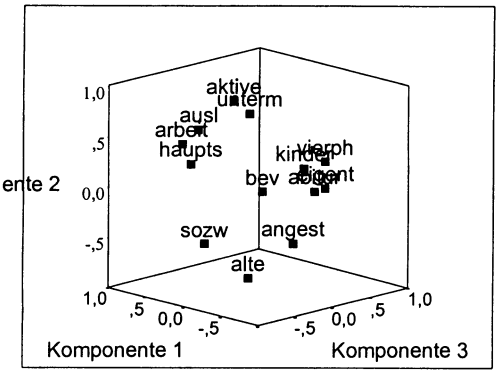


Abb. 21.10. Dreidimensionales Faktordiagramm im rotierten Raum

- ▷ Doppelklicken Sie dazu auf das dreidimensionale Diagramm. Dann öffnet sich der "Diagramm-Editor".
- ▷ Wählen Sie "Galerie" und "Streudiagramm". Es öffnet sich die Dialogbox "Streudiagramme".

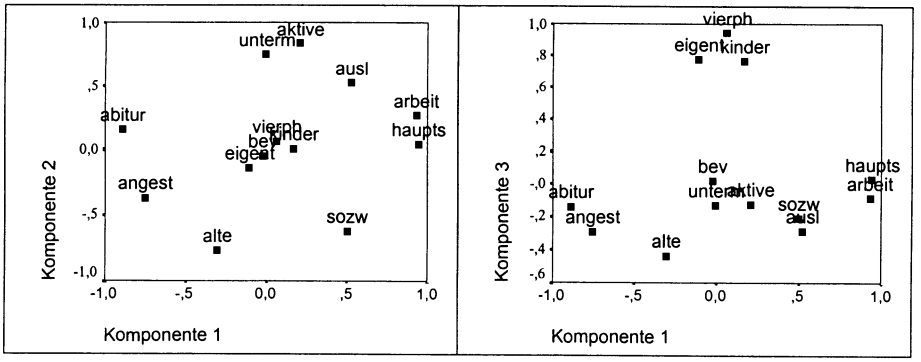


Abb. 21.11. Zwei zweidimensionale Faktordiagramme im rotierten Raum

- ▷ Klicken Sie auf "Einfach" und "Ersetzen". Es öffnet sich die Dialogbox "Einfaches Streudiagramm: Angezeigte Daten".
- ▷ Wählen Sie dort aus der Liste der Faktoren jeweils denjenigen aus, der die X-Achse und denjenigen, der die Y-Achse des neuen zweidimensionalen Diagramms darstellen soll.
- ▷ Bestätigen Sie mit "OK", und schließen Sie den Editor.

In Abb. 21.11 ist dies für zwei der im Beispiel sechs möglichen Kombinationen dargestellt.

Noch anschaulicher kann es sein, wenn die bivariaten Punktdiagramme für alle Paare der ausgewählten Faktoren in einem Matrixdiagramm dargestellt werden (⇨ Kap. 26.11.2).

21.4.2 Deskriptive Statistiken

Klickt man in der Dialogbox "Faktorenanalyse" auf die Schaltfläche "Deskriptive Statistik", öffnet sich die Dialogbox "Faktorenanalyse: Deskriptive Statistiken" (⇨ Abb. 21.12).

Abb. 21.12. Dialogbox "Faktorenanalyse: Deskriptive Statistiken"

Dort kann eine Reihe weiterer Statistiken angefordert werden. In der Gruppe *"Statistik"* können folgende Kontrollkästchen markiert werden:

- ❒ *Anfangslösung* (Voreinstellung). Gibt die anfängliche Kommunalitäten und, in der Tabelle "erklärte Gesamtvarianz", die anfänglichen Eigenwerte aus.
- ❒ *Univariate Statistiken.* Es werden für die Variablen Mittelwerte, Standardabweichungen und Fallzahlen ausgegeben.

Die Optionen der Gruppe *"Korrelationsmatrix"* dienen zum größten Teil der Diagnostik, d.h., es geht darum, ob die Voraussetzungen für eine Faktorenanalyse gegeben sind. Das ist nur dann der Fall, wenn erstens die Variablen zumindest mit einem Teil der anderen Variablen korrelieren, zweitens die Variablen möglichst vollständig durch die anderen Variablen erklärt werden. Die Qualität der Lösung ergibt sich dagegen u.a. aus dem Grad der Übereinstimmung zwischen Ausgangswerten und Schätzwerten (ersichtlich z.B. aus der reproduzierten Korrelationsma-

trix oder Residuen ⇨ Tabelle 21.8). In der Gruppe "Korrelationsmatrix" gibt es folgende Wahlmöglichkeiten:

- ☐ *Koeffizienten.* Ergibt die Korrelationsmatrix der Variablen.
- ☐ *Signifikanzniveaus.* Einseitige Signifikanzniveaus der Korrelationskoeffizienten in der Korrelationsmatrix der Variablen. Wird diese zusätzlich zu Koeffizienten angewählt, erscheinen sie im unteren Teil einer Tabelle, in deren oberen Teil die Korrelationskoeffizienten stehen.
- ☐ *Determinante.* Die Determinante der Korrelationskoeffizientenmatrix. Wird gewöhnlich unter der Korrelationsmatrix angegeben.
- ☐ *Inverse.* Die Inverse der Matrix der Korrelationskoeffizienten.
- ☐ *Reproduziert.* Die aus den Faktorlösungen geschätzte Korrelationsmatrix. Residuen, d.h. die Differenzen zwischen geschätzten und beobachteten Korrelationskoeffizienten werden im unteren Teil ebenfalls angezeigt. Die Diagonale enthält die reproduzierten Kommunalitäten.
- ☐ *Anti-Image.* Ergibt eine Doppeltabelle. Die obere enthält die Matrix *der Anti-Image-Kovarianz*en. Das sind die negativen Werte der partiellen Kovarianzen. Die untere Teiltabelle zeigt die Matrix der Anti-Image-Korrelationen. Darunter versteht man die negativen Werte der partiellen Korrelationskoeffizienten. Beide können als Test für die Strenge der Beziehungen zwischen den Variablen verwendet werden. Wenn die Variablen gemeinsame Faktoren teilen, ist die partielle Korrelation gering, wenn der Effekt der anderen Variablen ausgeschaltet wird. Also sollten bei einem geeigneten Modell die Werte außerhalb der Diagonale in den beiden Matrizen möglichst klein (nahe Null) sein. In der Diagonalen der unteren Tabelle werden MSA-Werte (messure of sampling adequacy) angezeigt. Das ist ein Maß für die Angemessenheit der einzelnen Variablen in einem Faktorenmodell. Die Variablen i sollten einerseits hoch mit anderen Variablen j korrelieren, andererseits weitgehend durch die anderen erklärt werden. Daher sollte die einfache Korrelation mit anderen Variablen hoch, die partielle aber gering sein. MSA stellt die einfachen und die partiellen Korrelationen ins Verhältnis. Ist die Summe der quadrierten partiellen Korrelationskoeffizienten im Vergleich zu der Summe der quadrierten einfachen Korrelationskoeffizienten gering, nimmt es den Wert 1 an. Ein MSA-Wert nahe 1 für eine Variable j zeigt die Angemessenheit der Variablen an.

$$MSA_j = \frac{\sum r_{ij}^2}{\sum r_{ij}^2 + \sum a_{ij}} \qquad (21.11)$$

r_{ij} = einfacher Korrelationskoeffizient zwischen zwei Variablen i und j

a_{ij} = partieller Korrelationskoeffizient zwischen zwei Variablen i und j

- ☐ *KMO und Bartlett-Test auf Sphärizität.*
  - *Kaiser-Meyer-Olkin Maß* (KMO). Es dehnt das MSA gemäß Gleichung 21.11 zur Prüfung der Angemessenheit der Daten auf die Gesamtheit der Variablen aus. Es überprüft, ob die Summe der quadrierten partiellen Korrelationskoeffizienten zwischen Variablen im Vergleich zu der Summe der

quadrierten Korrelationskoeffizienten zwischen den Variablen klein ist. Die partiellen Korrelationskoeffizienten sollten klein sein, denn sie entsprechen den (durch die Faktoren) nicht erklärten Teil der Varianz. KMO ist ein zusammenfassendes Maß dafür:

$$KMO = \frac{\sum\sum r_{ij}^2}{\sum\sum r_{ij}^2 + \sum\sum a_{ij}} \quad (21.12)$$

Korrelationen von Variablen mit sich selbst werden nicht berücksichtigt. Daher ist $i \neq j$. KMO kann Werte zwischen 0 und 1 annehmen. Kleine Werte geben an, dass die partiellen Korrelationskoeffizienten groß sind. Dann ist die Variablenauswahl ungeeignet. Werte unter 0,5 gelten als inakzeptabel, von 0,5 bis unter 0,6 als schlecht, von 0,6 bis unter 0,7 als mäßig, von 0,7 bis unter 0,8 als mittelprächtig, von 0,8 bis unter 0,9 als recht gut und über 0,9 als fabelhaft.

- Bartl*ett-Test auf Sphärizität*. Er prüft, ob die Korrelationskoeffizienten der Korrelationsmatrix insgesamt signifikant von 0 abweichen (das ist relevant, wenn die Daten einer Stichprobe entstammen). Denn sinnvoll ist eine Faktorenanalyse nur dann, wenn zwischen den Variablen und zumindest einigen anderen Variablen tatsächlich Korrelationen existieren. Ergebnis ist ein Chi-Quadrat-Wert. Bei einer signifikanten Abweichung von der Einheitsmatrix (einer Matrix mit ausschließlich Korrelationskoeffizienten = 0), gelten die Voraussetzungen für eine Faktorenanalyse als gegeben. (zu dem Grenzen solcher Tests ⇨ Kap. 13.3.)

Tabelle 21.10. KMO- und Bartlett-Test

KMO- und Bartlett-Test

Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin.		,738
Bartlett-Test auf Sphärizität	Ungefähres Chi-Quadrat	126,177
	df	15
	Signifikanz nach Bartlett	,000

21.4.3 Weitere Optionen

❒ *Faktorenanalyse Extraktion* (⇨ Abb. 21.2). Außer den bereits besprochenen sind noch folgende Optionen relevant:

- *Kovarianzmatrix*. Die Auswahl dieser Option in der Gruppe "Analysieren" bewirkt, dass die Faktorenextraktion von der Kovarianzmatrix und nicht von der Korrelationsmatrix ausgeht.
- *Maximalzahl der Iterationen für die Konvergenz*. Durch Ändern des Wertes in diesem Eingabefeld (Voreinstellung: 25) bestimmt man, wieviele Iterationsschritte maximal durchgeführt werden. Um die Rechenzeit zu reduzieren,

sollte man die Zahl klein halten. Führt die Berechnung bei der angegebenen Zahl der Iterationen nicht zu einem Ergebnis, muss sie heraufgesetzt werden.

❒ *Faktorenanalyse: Optionen* (Abb. 21.13)

- *Fehlende Werte".* Hier wird die Behandlung der fehlenden Werte während der Analyse festgelegt. Möglich sind: *"Listenweiser Fallausschluss", "Paarweiser Fallausschluss", "Durch Mittelwert ersetzen"* (ein fehlender Wert wird durch den Mittelwert aller anderen Fälle ersetzt).
- *Anzeigeformat für Koeffizienten.*
  - *Sortiert nach Größe.* Sortiert die Faktorenmatrix-, die Mustermatrix und die Strukturmatrix so, dass jeweils die Variablen, die auf demselben Faktor hoch laden, zusammen stehen.
  - *Unterdrücken von Absolutwerten kleiner als* (Voreinstellung: 0,1). In denselben Matrizen werden keine Werte, die unter dem angegebenen Wert liegen, ausgewiesen. (Mögliche Werte können zwischen 0 und 1 betragen.)

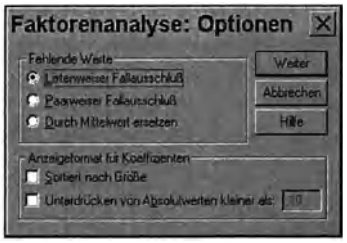


Abb. 21.13. Dialogbox " Faktorenanalyse Optionen"

Weitere Möglichkeiten bei Verwenden der Befehlssyntax.

❒ Es können Konvergenzkriterien für die Iteration bei der Extraktion und Rotation festgelegt werden (voreingestellt sind 0,001 für die Extraktion und 0,0001 für die Rotation).
❒ Es können einzelne Diagramme für die rotierten Faktoren angefordert werden.
❒ Die Anzahl der zu speichernden Faktorwerte kann festgelegt werden.
❒ Bei der Methode der Hauptachsen-Faktoranalyse können Diagonalwerte (Kommunalitäten) für die Korrelationsmatrix angegeben werden.
❒ Die Korrelationsmatrizen oder Matrizen der Faktorladungen können für eine spätere Analyse gespeichert werden.
❒ Korrelationsmatrizen oder Matrizen der Faktorladungen können eingelesen und analysiert werden.

22 Nichtparametrische Tests

22.1 Einführung und Überblick

Nichtparametrische versus parametrische Tests. Nichtparametrische Tests (auch verteilungsfreie Tests genannt) ist ein Sammelbegriff für eine Reihe von statistischen Tests für ähnliche Anwendungsbedingungen. Sie kommen grundsätzlich in folgenden Situationen zur Anwendung:

- Die zu testenden Variablen haben Ordinal- oder Nominalskalen, so dass parametrische Tests (Tests mit Annahmen über die Verteilung der Variablen), wie z.B. der t-Test zur Prüfung auf Differenz von Mittelwerten zweier Verteilungen, der Test eines Korrelationskoeffizienten auf Signifikanz u.ä. nicht angewendet werden dürfen.
- Die zu testenden Variablen haben zwar ein metrisches Skalenniveau (Intervall- oder Rationalskala), aber die Datenlage gibt Anlass für die Annahme, dass die zugrundeliegenden Verteilungen nicht normalverteilt sind. Dieses gilt für die Verteilung der Grundgesamtheit und aber insbesondere für die Stichprobenverteilung einer Prüfgröße bei kleinen Stichprobenumfängen, da hier der zentrale Grenzwertsatz nicht anwendbar ist.

Derartige Situationen sind im sozialwissenschaftlichen Bereich recht häufig anzutreffen. Nichtparametrische Tests werden auch verteilungsfreie Tests genannt, weil sie keine Annahme über zugrundeliegende Verteilungen benötigen. Insofern sind parameterfreie Tests weniger restriktiv bezüglich ihrer Anwendungsbedingungen als parametrische Tests. So wird z.B. für den parametrischen t-Test vorausgesetzt, dass die zwei Zufallsstichproben aus Grundgesamtheiten mit Normalverteilungen stammen, die eine gleiche Varianz haben. Dem Vorteil wenig restriktiver Anwendungsbedingungen steht aber der Nachteil gegenüber, dass nichtparametrische Tests nicht so trennscharf sind wie parametrische, und zwar gerade deshalb, weil Annahmen über die Verteilung nicht einfließen.

Nichtparametrische Test basieren auf Rangziffern oder Häufigkeiten der Variablen. Die Verwendung von Rangziffern stellt gegenüber der Verwendung von Variablenwerten ein Verlust von Informationen dar. Dieser Informationsverlust bedingt die schwächere Trennschärfe des Tests.

Als Leitlinie für die Frage, ob ein parametrischer oder nichtparametrischer Test verwendet werden soll, kann folgendes gelten:

- Sind die Anwendungsbedingungen für die Verwendung eines parametrischen Tests erfüllt, so sollte man diesen verwenden, da er bezüglich der beiden Hypo-

thesen trennschärfer ist. Das bedeutet, dass in höherem Maße der parametrische Test zu richtigen Ergebnissen hinsichtlich der Annahme bzw. Ablehnung der H_0-Hypothese führt, wenn sie richtig bzw. falsch ist.

- Wenn parametrische Tests aufgrund des Skalenniveaus der Variablen oder weil keine Normalverteilung angenommen darf, nicht zur Anwendung kommen können, so sollte ein nichtparametrischer Test eingesetzt werden. Bei Verwendung eines (trennschärferen) parametrischen Tests besteht die Gefahr, dass ein falsches Testergebnis resultiert.

Unterscheidungskriterien für nichtparametrische Tests. Die Tests unterscheiden sich durch die Anzahl der verwendeten Stichproben, durch das Skalenniveau der Variablen und/oder die Frage, ob die verwendeten Stichproben unabhängig voneinander sind oder nicht. Bei der Anzahl der Stichproben werden ein, zwei oder mehr als zwei (allgemein k) Stichproben unterschieden.

Stichproben sind unabhängig voneinander, wenn die Messwerte einer Stichprobe unabhängig von den Messwerten der anderen Stichprobe sind. Wird beispielsweise eine Zufallsstichprobe von Befragten erhoben zur Messung von Meinungen zu verschiedenen Themen, so können die beiden Befragtengruppen Männer und Frauen in der Stichprobe als voneinander unabhängige Einzelstichproben aufgefasst werden. Mit einem Test kann dann geprüft werden, ob die beiden Gruppen sich hinsichtlich einer Meinung unterscheiden oder nicht. Tests für unabhängige Stichproben können auch für klinische Studien eingesetzt werden, in denen Individuen zufällig einer von zwei Behandlungen zugeordnet werden (Lehmann, 1975).

Abhängige bzw. verbundene Stichproben entstehen in der Regel in einer experimentellen Versuchsanordnung. Der typische Fall ist, dass man prüfen will, ob eine Maßnahme oder Aktivität wirksam ist oder nicht und deshalb eine Experiment- und eine Kontrollgruppe bildet (matched pairs). Damit aber die Messung der Wirksamkeit einer Maßnahme nicht durch andere Einflussgrößen gestört bzw. überlagert wird, wählt man im 2-Stichprobenfall (im k-Stichprobenfall sinngemäß) jeweils Paare für die Experimentier- und Kontrollgruppe aus. Die Paare werden derart gebildet, dass sich ein Paar hinsichtlich wichtiger sonstiger relevanter Einflussfaktoren nicht unterscheidet (englisch: matching). Damit sollen andere wichtige Einflussfaktoren kontrolliert (konstant gehalten) werden. Geht es z.B. darum, den Lernerfolg einer neuen Lehrmethode für ein Fach zu prüfen, so werden Schülerpaare derart ausgewählt, dass sich ein Paar nicht hinsichtlich relevanter Einflussfaktoren auf das Lernergebnis (wie Fleiß, Intelligenz etc.) unterscheidet. Bei einem derartigen Stichprobenkonzept hat man es mit einer verbundenen Stichprobe zu tun, da der Lernerfolg eines Schülers in einer Gruppe nicht mehr unabhängig ist von dem eines Schülers in der anderen Gruppe. Welche Person eines Paares jeweils in die Experimentier- oder Kontrollgruppe kommt, kann ausgelost werden.

Um eine besondere Form verbundener Stichproben handelt es sich, wenn es sich um den Vergleich von „vorher"- und „nachher"-Konstellationen bei gleichen Personen handelt. Soll beispielsweise geprüft werden, ob ein spezielles Augentraining die Sehfähigkeit verbessert, so wird die Sehfähigkeit bei einer Gruppe von Personen vor und nach dem Training gemessen.

Eine weitere Form verbundener Stichproben liegt vor, wenn beispielsweise jeweils mehrere Mitglieder einer Familie (z.B. Ehepaare) in Befragungen einbezogen werden. Meinungsäußerungen von Ehepartnern sind nicht voneinander unabhängig.

Die k-Stichproben-Tests erlauben zu prüfen, ob es Unterschiede zwischen mehreren Stichproben gibt oder nicht. Es wird dabei aber nicht aufgedeckt, zwischen welchen der k Stichproben diese Unterschiede bestehen.

Überblick über die Tests in SPSS. Aus der Übersicht in Abb. 22.1 kann man entnehmen, welche nichtparametrischen Tests von SPSS bereitgestellt werden. Die Reihenfolge orientiert sich an der im Menü „Nichtparametrische Tests" in SPSS. Es wird im Überblick kurz angeführt, welchen Testzweck die einzelnen Tests haben, welches Messniveau für die Variablen erforderlich ist, um wieviel Stichproben es sich handelt und ob es sich um ein Design von unabhängigen oder verbundenen Stichproben handelt.

Exakte Tests. Für die Basismodulanwendungen von SPSS werden bei den einzelnen Tests Prüfgrößen berechnet und für diese werden approximativ theoretische Prüfverteilungen verwendet. Aber nicht immer sind die Bedingungen dafür gegeben, dass die Verteilung der Prüfgrößen hinreichend durch die theoretischen Verteilungen approximiert werden dürfen. SPSS für Windows bietet daher ab der Version 6.1.2 in Ergänzung zum Basismodul das Modul „Exakte Tests" an. Nach Installation dieses Moduls steht in den Dialogboxen zur Durchführung nichtparametrischer Tests zusätzlich eine Schaltfläche „Exakt..." zur Verfügung. Durch Klicken auf die Schaltfläche kann man die Dialogbox „Exakte Tests" öffnen und zwischen zwei Verfahren zur Durchführung exakter Tests wählen (ausführlicher: ⇨ Kap. 29).

22.2 Tests für eine Stichprobe

22.2.1 Chi-Quadrat-Test (Anpassungstest)

Der Chi-Quadrat-Test ist schon im Zusammenhang mit der Kreuztabellierung behandelt worden (⇨ Kap. 10.2). Dort geht es um die Frage, ob zwei nominalskalierte Variable voneinander unabhängig sind oder nicht (Chi-Quadrat-Unabhängigkeitstest).

Hier geht es darum, ob sich für eine Zufallsstichprobe eine (nominal- oder gruppierte ordinalskalierte) Variable in ihrer Häufigkeitsverteilung signifikant von erwarteten Häufigkeiten der Grundgesamtheit unterscheidet (Anpassungs- bzw. „Goodness of Fit"-Testtyp). Die erwarteten Häufigkeiten können z.B. gleichverteilt sein oder einer anderen Verteilung folgen. SPSS erlaubt neben der Prüfung auf Gleichverteilung auch die Prüfung, ob es sich um eine Normalverteilung oder Poisson-Verteilung handelt.

Das folgende Beispiel bezieht sich auf Befragungsdaten der Arbeitsgruppe Wahlforschung an der Hamburger Hochschule für Wirtschaft und Politik zur Vorhersage der Wahlergebnisse für die Bürgerschaft der Freien und Hansestadt Hamburg im Herbst 1993. Unter anderem wurde gefragt, welche Partei zur Bürger-

Tabelle 22.1. Übersicht über nichtparametrische Tests von SPSS

Testname	Mess-niveau*	Testzweck	Anzahl der Stichproben	Stichpro-bendesign#
1. Chi-Quadrat	n	Empirische gleich erwartete Häufigkeit?	1	-
2. Binomial	d	Empirische Häufigkeit binomialverteilt?	1	-
3. Sequenzanalyse	d	Reihenfolge der Variablenwerte zufällig?	1	-
4. Kolmogorov-Smirnov	o	Empirische Verteilung gleich theoretischer?	1	-
5. Mann-Whitney U	o	2 Stichproben aus gleicher Verteilung?	2	u
6. Moses	o	2 Stichproben aus gleicher Verteilung?	2	u
7. Kolmogorov-Smirnov Z	o	2 Stichproben aus gleicher Verteilung?	2	u
8. Wald-Wolfowitz	o	2 Stichproben aus gleicher Verteilung?	2	u
9. Kruskal-Wallis H	o	k Stichproben aus gleicher Verteilung?	k	u
10. Median	o	2 oder k Stichproben aus Verteilung mit gleichem Median?	2 bzw. k	u
11. Jonckheere-Terpstra	o	k Stichproben aus gleicher Verteilung. Für geordnete Verteilungen	k	u
12. Wilcoxon	o	2 verbundene Stichproben aus gleicher Verteilung?	2	v
13. Vorzeichen	o	2 verbundene Stichproben aus gleicher Verteilung?	2	v
14. McNemar	d	2 Stichpr. verändert im Vorher/Nachher-Design	2	v
15. Marginale Homogenität	n	2 Stichpr. verändert im Vorher/Nachher-Design	2	v
16. Friedman	o	k verbundene Stichpr. aus gleicher Verteilung?	k	v
17. Kendall's W	o	k verbundene Stichpr. aus gleicher Verteilung?	k	v
18. Cochran Q	d	k verbundene Stichpr. mit gleichem Mittelwert?	k	v

* n = nominal, o = ordinal, d = dichotom

\# u = unabhängig, v = verbunden

schaftswahl 1991 gewählt worden ist. Die Verteilung dieser Variable - mit PART_91 bezeichnet - mit den Werten 1 bis 7 (für die Parteien SPD, CDU, Grüne/GAL, F.D.P., Republikaner und Sonstige; der Wert 6 kommt nicht vor) soll mit den tatsächlichen Wahlergebnissen in 1991 für diese Parteien verglichen und getestet werden, ob sich ein signifikanter signifikanter Unterschied in den Verteilungen ergibt. Ergibt sich ein signifikanter Unterschied, so könnte das als ein Indikator dafür gesehen werden, dass die Stichprobe nicht hinreichend repräsentativ ist. Die Hypothese H_0 lautet also, die Stimmenverteilung auf die Parteien in der Stichprobe entspricht dem Ergebnis der Bürgerschaftswahl. Entsprechend lautet die H_1-Hypothese, dass die Verteilungen signifikant unterschiedlich sind. Nach dem Einlesen der Datei WAHLEN2.SAV gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Nichtparametrische Tests" ▷ „Chi-Quadrat···". Es öffnet sich dann die in Abb. 22.1 wiedergegebene Dialogbox.
- ▷ Aus der Quellvariablenliste wird die Testvariable PART_91 in das Eingabefeld „Testvariablen:" übertragen. Sollen für weitere Variablen Tests durchgeführt werden, so sind auch diese zu übertragen.
- ▷ Die gewählte Option „Aus den Daten" in der Auswahlgruppe „Erwarteter Bereich" bedeutet, dass der gesamte Wertebereich der Variablen (hier: 1 bis 7) für den Test benutzt wird. Soll nur ein Teilwertebereich für den Test ausgewertet werden, so kann dieses mit der Option „Angegebener Bereich verwenden" geschehen, indem man in das Eingabefeld „Minimum" den kleinsten (z.B. 1) und in das Eingabefeld „Maximum" den größten Wert (z.B. 4) eingibt.
- ▷ In „Erwartete Werte" kann man aus zwei Optionen auswählen. „Alle Kategorien gleich" wird man wählen, wenn die gemäß der Hypothese H_0 erwarteten Häufigkeiten der Kategorien der Variablen (hier die Parteien) gleich sind (Gleichverteilung). Für unser Beispiel ist die Option „Werte" relevant. In das Eingabefeld von Werte gibt man die gemäß der H_0-Hypothese erwarteten Häufigkeiten für die Kategorien (Parteien) ein. Wichtig ist, dass sie in der Reihenfolge entsprechend der Codierung der Variable, beginnend mit dem kleinsten Wert (hier: 1 für SPD), eingegeben werden. Mit „Hinzufügen" werden die jeweils in das Werte-Eingabefeld eingetragenen Häufigkeiten nacheinander in das darunter liegende Textfeld übertragen. Die erwarteten Werte können sowohl als prozentuale als auch absolute Häufigkeiten eingegeben werden. Die in der Abb. 22.1 sichtbaren Eintragungen ergeben sich daraus, dass bei der Bürgerschaftswahl 1991 die SPD 48,0 %, die CDU 35,1 %, die Grünen/GAL 7,2 %, die FDP 5,4 % und Sonstige 3,1 % Stimmenanteile erhalten haben (da in der Datei für den codierten Wert 5 (für Republikaner) keine Fälle enthalten sind, darf man den Stimmenanteil der Republikaner nicht angeben, weil sonst von SPSS der Test mit einer Fehlermeldung abgebrochen wird). Hat man sich bei schon eingegebenen Werten vertan, so kann man sie markieren und mittels „Entfernen" aus dem Textfeld entfernen.

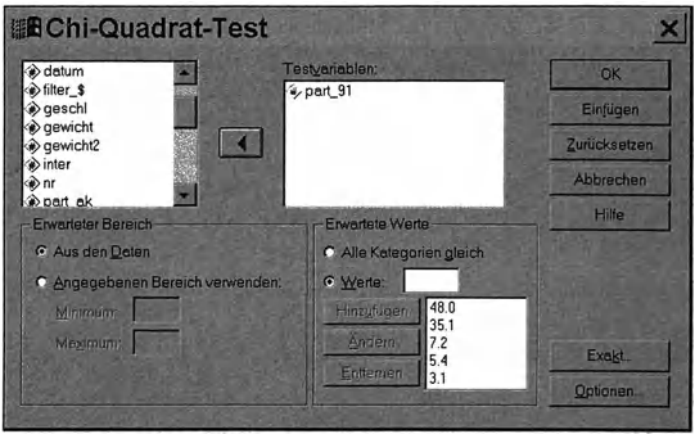


Abb. 22.1. Dialogbox „Chi-Quadrat-Test"

In Tabelle 22.2 ist das Ergebnis des Chi-Quadrat-Tests niedergelegt. Für die Parteien werden die empirischen („Beobachtetes N") und erwarteten („Erwartete Anzahl") Häufigkeiten sowie die Abweichungen dieser („Residuum") aufgeführt. Die erwarteten Häufigkeiten unter H_0 ergeben sich durch Multiplikation der Fallanzahl mit dem Stimmenanteil für eine Partei. Werden mit f_i die empirischen und mit e_i die erwarteten Häufigkeiten einer Kategorie bezeichnet, so ergibt sich für die Prüfgröße Chi-Quadrat (die Summierung erfolgt über die Kategorien i = 1 bis k (hier: k = 5)

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} = 19{,}32 \qquad (22.1)$$

Aus der Formel wird ersichtlich, dass die Testgröße χ^2 umso größer wird, je stärker die Abweichungen zwischen beobachteten und empirischen Häufigkeiten sind. Ein hoher Wert für χ^2 ist folglich ein Ausdruck für starke Abweichungen in den Verteilungen. Je größer der χ^2-Wert ist, umso unwahrscheinlicher ist es, dass die Stichprobe aus der Vergleichsverteilung stammt. Die Prüfgröße χ^2 ist asymptotisch chi-quadrat-verteilt mit k-1 Freiheitsgraden (df = degrees of freedom). Für eine gegebene Anzahl von Freiheitsgraden und einem Signifikanzniveau α (Irrtumswahrscheinlichkeit die H_0-Hypothese abzulehnen, obwohl sie richtig ist) lassen sich aus einer Chi-Quadrat-Verteilungs-Tabelle kritische Werte für χ^2 entnehmen. Für fünf Kategorien in unserem Beispiel ist df = 4. Bei einem Signifikanzniveau von α = 0,05 (5 % Irrtumswahrscheinlichkeit) und df = 4, ergibt sich aus einer tabellierten Chi-Quadrat-Verteilung für χ_{krit}^2 = 9,488. Der empirische Wert von χ^2 fällt in den Ablehnungsbereich der H_0-Hypothese, da er mit 19,32 größer ist als der kritische. „Asymptotische Signifikanz" (= 0,001) ist die Wahrscheinlichkeit, bei df = 4 ein $\chi^2 \geq 19{,}32$ zu erhalten. Ohne Einblick in eine χ^2-Tabelle zu nehmen, ergibt sich daraus, dass bei einem Signifikanzniveau von 5 % (α =

0,05) die H_0-Hypothese abzulehnen ist (0,05 > 0,001). Die Stimmenverteilung auf die Parteien in der Stichprobe entspricht demnach nicht der tatsächlichen für 1991.

Tabelle 22.2. Ergebnisausgabe eines Chi-Quadrat-Tests

Altes Parteienvotum

	Beobachtetes N	Erwartete Anzahl	Residuum
SPD	243	218,1	24,9
CDU	122	159,5	-37,5
Grüne/Gal	47	32,7	14,3
FDP	27	24,5	2,5
Sonstige	10	14,1	-4,1
Gesamt	449		

Statistik für Test

	PART_91
Chi-Quadrat[a]	19,320
df	4
Asymptotische Signifikanz	,001

a. Bei 0 Zellen (,0%) werden weniger als 5 Häufigkeiten erwartet. Die kleinste erwartete Zellenhäufigkeit ist 14,1.

Optionen. Durch Klicken auf „Optionen" öffnet sich die in Abb. 22.2 dargestellte Dialogbox mit der optionale Vorgaben festgelegt werden können:

- *Statistik.* Mit „Deskriptive Statistik" können das arithmetische Mittel (mean), die Standardabweichung (Std. Dev) sowie das Minimum und das Maximum angefordert werden. Mit „Quartile" werden der Wert des 25., 50. (= Median) und 75. Perzentils berechnet.
- *Fehlende Werte.* Mit „Fallausschluss Test für Test" werden beim Testen mehrerer Variablen die fehlenden Werte jeweils für die einzelne Testvariable und mit „Listenweiser Fallausschluss" für alle Tests ausgeschlossen.

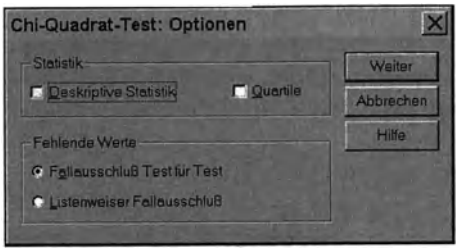


Abb. 22.2. Dialogbox „Chi-Quadrat-Test: Optionen"

Exakte Tests. Sollte man verwenden, wenn die Anwendungsbedingungen einen asymptotischen Chi-Quadrat-Test verbieten (⇨ Kap. 29).

Anwendungsbedingungen. Für den asymptotischen Chi-Quadrat-Test sollten folgende Anwendungsbedingungen beachtet werden: im Falle von df = 1 sollte $e_i \geq 5$ für alle Kategorien i sein. Für df > 1 sollte $e_i \leq 5$ für nicht mehr als 20 % der Kategorien i und $e_i \geq 1$ für alle i sein.

Warnung. Der Chi-Quadrat-Test führt zu unsinnigen Ergebnissen, wenn die Fälle mit einer Variablen gewichtet werden, deren Werte Dezimalzahlen sind (z.B. 0,85, 1,20). Bei wiederholten Berechnungen ergeben sich unterschiedliche Zahlen für

die beobachteten Fälle und dementsprechend werden verschiedene Signifikanzniveaus angeführt.

22.2.2 Binomial-Test

Eine Binomialverteilung ist eine Wahrscheinlichkeitsverteilung für eine diskrete Zufallsvariable, die nur zwei Werte annimmt (dichotome Variable). Mit Hilfe der Binomialverteilung lässt sich testen, ob ein prozentualer Häufigkeitsanteil für eine Variable in der Stichprobe mit dem der Grundgesamtheit vereinbar ist. Das oben verwendete Beispiel zur Wahlvorhersage (WAHLEN2.SAV, ⇨ Kap 22.2.1) soll dieses näher erläutern. Geprüft werden soll, ob der prozentuale Männeranteil in der Stichprobe mit dem in der Grundgesamtheit - alle Wahlberechtigten für die Hamburger Bürgerschaft - vereinbar ist oder nicht. Dazu gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Nichtparametrische Tests ▷", „Binomial···". Es öffnet sich die in Abb. 22.3 dargestellte Dialogbox.
- ▷ Aus der Quellvariablenliste wird die Variable GESCHL in das Eingabefeld von „Testvariablen:" übertragen. Sollen mehrere Variablen getestet werden, so sind diese alle in das Variableneingabefeld zu übertragen.
- ▷ In „Dichotomie definieren" bestehen alternative Auswahlmöglichkeiten:
  - „Aus den Daten" ist zu wählen, wenn - wie es in diesem Beispiel der Fall ist - die Variable dichotom ist.
  - „Trennwert" ist zu wählen, wenn eine nicht-dichotome Variable mit Hilfe des einzugebenden Trennwertes dichotomisiert wird. Beispielsweise lässt sich die Variable ALTER durch „Trennwert" = 40 in eine dichotome Variable verwandeln: bis einschließlich 40 haben alle Befragten den gleichen Variablenwert und ab 41 einen anderen Wert.
- ▷ In das Eingabefeld „Testanteil:" wird der Anteilswert gemäß H_0-Hypothese für die Grundgesamtheit in dezimaler Form eingegeben. Die Männerquote für die Wahlberechtigten für die Bürgerschaft beträgt 48,3 % (einzugeben ist 0,483).

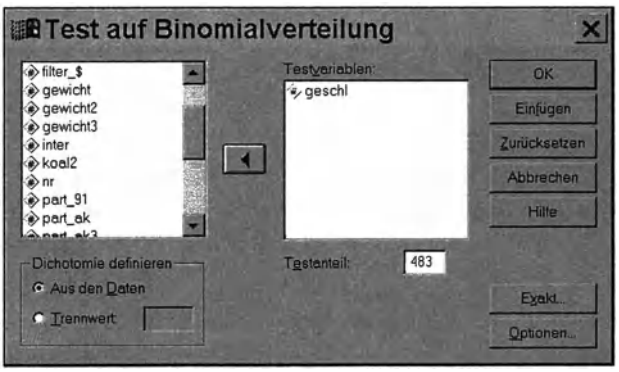


Abb. 22.3. Dialogbox „Binomial-Test"

In Tabelle 22.3 ist das Ergebnis des Binomial-Tests niedergelegt. Die empirische Männerquote („Beobachteter Anteil") beträgt 0,468571 im Vergleich zur vorgege-

benen Quote (0,483). Da der Stichprobenumfang hinreichend groß ist, wird die Binomialverteilung durch eine Normalverteilung approximiert. Der Test kann dann vereinfachend mittels der standardnormalverteilten Variable Z vorgenommen werden. Ergebnis ist, dass unter der H_0-Hypothese (eine Männerquote von 0,483 für die Wahlberechtigten) eine Wahrscheinlichkeit („Asymptotische Signifikanz, 1-seitig") von 0,268 besteht, dass die Männerquote gleich bzw. kleiner als die beobachtete ist. Bei einem Signifikanzniveau von 5 % (α = 0,05) wird wegen 0,268 > 0,05 die Hypothese H_0 nicht verworfen.

Optionen. ⇨ Erläuterungen zu Abb. 22.2.

Tabelle 22.3. Ergebnisausgabe des Binomial-Tests

Test auf Binomialverteilung

		Kategorie	N	Beobachteter Anteil	Testanteil	Asymptotische Signifikanz (1-seitig)
GESCHL	Gruppe 1	männlich	246	,468571	,483	,268[a,b]
	Gruppe 2	weiblich	279	,531		
	Gesamt		525	1,000		

a. Nach der alternativen Hypothese ist der Anteil der Fälle in der ersten Gruppe < ,483.

b. Basiert auf der Z-Approximation.

22.2.3 Sequenz-Test (Runs-Test) für eine Stichprobe

Dieser Test ermöglicht es zu prüfen, ob die Reihenfolge der Werte einer Variablen in einer Stichprobe (und damit die Stichprobe) zufällig ist (H_0-Hypothese). Angewendet wird der Test z.B. in der Qualitätskontrolle und bei Zeitreihenanalysen.

Im folgenden Beispiel für eine Stichprobe mit einem Umfang von 20 sei eine (dichotome) Variable mit nur zwei Ausprägungen (hier dargestellt als + und −) in einer Reihenfolge gemäß Tabelle 22.4 erhoben. Diese Stichprobe hat eine Sequenz (runs) von 8, da achtmal gleiche (positive bzw. negative) Werte aufeinander folgen.

Tabelle 22.4. Beispiel für acht Sequenzen bei einem Stichprobenumfang von 20

++	- - -	+	- -	+ + + +	-	+ + +	- - - -
1	2	3	4	5	6	7	8

Wären die Merkmalswerte „+" bzw. „-" z.B. Zahl bzw. Wappen bei 20 aufeinander folgenden Würfen mit einer Münze, so kann die Sequenz der Stichprobe Hinweise hinsichtlich der „Fairness" der Münze geben, die durch Feststellung einer „Wappen-Quote" in der Stichprobe von ca. 50 % verdeckt bleiben würde. Die Erfassung von Sequenzen beschränkt sich nicht auf schon im Stadium der Messung dichotome Variablen, da Messwerte von Variablen in dichotome verwandelt werden können, indem festgehalten wird, ob die Messwerte kleiner oder größer als ein bestimmter Messwert (z.B. das arithmetische Mittel) sind.

Die Stichprobenverteilung der Anzahl von Sequenzen (= Prüfgröße) ist bekannt. Für große Stichproben ist die Prüfgröße approximativ standardnormalverteilt.

Beispiel. Im folgenden soll getestet werden, ob die Stichprobe für die Wahlprognose (Datei WAHLEN2.SAV, ⇨ Kap. 22.2.1) zufällig ist. Als Testvariable wird das Alter der Wähler gewählt. Zur Durchführung des Tests gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Nichtparametrische Tests ▷„ „Sequenzen...". Es öffnet sich die in Abb. 22.4 dargestellte Dialogbox.
- ▷ Aus der Quellvariablenliste wird die Variable ALTER in das Eingabefeld „Testvariablen:" übertragen. Zur Dichotomisierung der Variablen stehen im Feld „Trennwert" vier Optionen zur Verfügung:
  - *Median:* Zentralwert.
  - *Modalwert:* häufigster Wert.
  - *Mittelwert :* arithmetisches Mittel.
  - *Anders:* vom Anwender vorgegebener Wert.
- ▷ In unserem Beispiel wird „Median" gewählt. Dadurch erhält die Variable ALTER zur Ermittlung der Sequenz nur zwei Merkmalsausprägungen: kleiner als der Median und größer bzw. gleich dem Median.

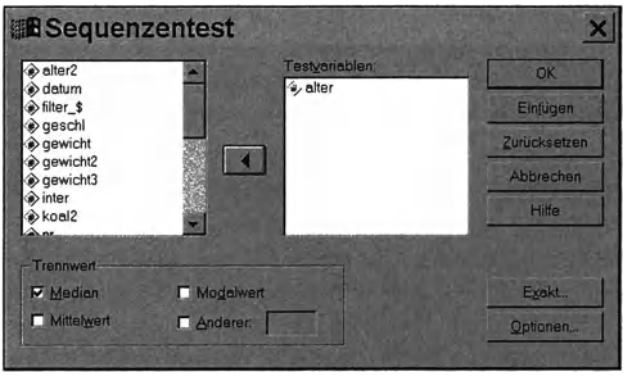


Abb. 22.4. Dialogbox „Sequenzanalyse"

In Tabelle 22.5 ist das Ergebnis des Tests zu sehen. Bei einem Stichprobenumfang in Höhe von 529 werden 158 Sequenzen („Runs") ermittelt. 261 Befragte haben ein Alter kleiner und 268 größer bzw. gleich als der Median in Höhe von 51 Jahren. Für den Z-Wert der standardisierten Normalverteilung in Höhe von 9,354 ergibt sich die zweiseitige asymptotische Wahrscheinlichkeit in Höhe von 0,000. Die Anzahl der Sequenzen ist derart niedrig, dass die H_0-Hypothese (die Reihenfolge der Befragten ist zufällig) abgelehnt wird (wegen Irrtumswahrscheinlichkeit $\alpha = 0{,}05 > 0{,}000$).

Optionen. ⇨ Erläuterungen zu Abb. 22.2.
Exakter Test. ⇨ Kap. 29.

Tabelle 22.5. Ergebnisausgabe eines Sequenzen-Tests

Sequenzentest

	ALTER
Testwert[a]	51,00
Fälle < Testwert	261
Fälle >= Testwert	268
Gesamte Fälle	529
Anzahl der Sequenzen	158
Z	-9,354
Asymptotische Signifikanz (2-seitig)	,000

a. Median

22.2.4 Kolmogorov-Smirnov-Test für eine Stichprobe

Wie der oben angeführte Chi-Quadrat-Test und der Binomial-Test hat auch der Kolmogorov-Smirnov-Test die Aufgabe zu prüfen, ob die Verteilung einer Stichprobenvariable mit einer theoretischen Verteilung übereinstimmt oder nicht (Anpassungstest). Im Vergleich zum χ^2-Test hat der Kolmogorov-Smirnov-Test den Vorteil, dass er auch für kleine Stichproben angewendet werden kann. Für kleine Stichproben ist meistens nicht gewährleistet, dass 20 % der Zellen eine erwartete Häufigkeit von mindestens 5 haben. Der Test kann aber nur für stetige Variablen angewendet werden.

Dieser Test basiert auf der kumulierten empirischen sowie kumulierten erwarteten (theoretischen) Häufigkeitsverteilung. Die größte Differenz (D_{max}) zwischen beiden kumulierten Verteilungen und der Stichprobenumfang gehen in die Prüfgröße Z nach Kolmogorov-Smirnov ein ($KS - Z = \sqrt{n} * D_{max}$). Aus Tabellen kann man für einen gegebenen Stichprobenumfang n kritische Werte für D_{max} bei einem vorgegebenem Signifikanzniveau entnehmen (Siegel, 1956, S. 251).

Für die Befragung zur Wahlprognose für die Bürgerschaftswahl im Herbst 1993 (Datei WAHLEN2.SAV, ⇨ Kap. 22.2.1) soll geprüft werden, ob das Alter der Befragten vereinbar ist mit der Hypothese H_0: die Stichprobe stammt aus einer Grundgesamtheit mit normalverteiltem Alter (es wird hier ignoriert, dass die Grundgesamtheit der Wahlberechtigten tatsächlich nicht normalverteilt ist). Das Alter hat ein metrisches Messniveau. Der Kolmogorov-Smirnov-Test ist aber auch für ordinalskalierte Variablen anwendbar. Sie gehen wie folgt vor:

- ▷ Klicken Sie die Befehlsfolge „Analysieren", „Nichtparamametrische Tests ▷„K-S bei einer Stichprobe...". Es öffnet sich die in Abb. 22.5 dargestellte Dialogbox.
- ▷ Die Testvariable ALTER wird in das Eingabefeld „Testvariablen" übertragen.
- ▷ Die Testverteilung ist in diesem Beispiel die Normalverteilung. Daher wird in „Testverteilung" diese ausgewählt. Als alternative theoretische Testverteilungen sind die Gleich-, die Poisson- und Exponentialverteilung wählbar.

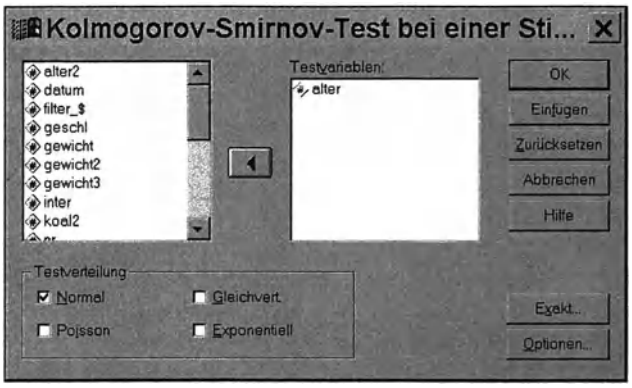


Abb. 22.5. Dialogbox „Ein-Stichproben-Kolmogorov-Smirnov-Test"

In Tabelle 22.6 ist das Ergebnis des Tests zu sehen. Das durchschnittliche Alter der Befragten beträgt 51,07 Jahre mit einer Standardabweichung von 18,48. Mit „Extremste Differenzen" wird bei „Absolut"(und „Positiv") D_{max} = 0,076 angeführt. Die größte negative Abweichung beträgt –0,42. Es ist $KS - Z = \sqrt{n} * D_{max} =$ $= \sqrt{529} * 0{,}0762 =$ 1,753. Die zweiseitige (asymptotische) Wahrscheinlichkeit beträgt 0,004. Bei einem Signifikanzniveau von 5 % ($\alpha = 0{,}05$) wird wegen 0,004 < 0,05 die Hypothese H_0 (das Alter ist normalverteilt) abgelehnt.

Optionen. ⇨ Erläuterungen zu Abb. 22.2.
Exakter Test. ⇨ Kap. 29.

Tabelle 22.6. Ergebnisausgabe des Kolmogorov-Smirnov-Tests zur Prüfung auf Normalverteilung

Kolmogorov-Smirnov-Anpassungstest

		ALTER
N		529
Parameter der Normalverteilung[a,b]	Mittelwert	51,07
	Standardabweichung	18,48
Extremste Differenzen	Absolut	,076
	Positiv	,076
	Negativ	-,042
Kolmogorov-Smirnov-Z		1,753
Asymptotische Signifikanz (2-seitig)		,004

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

22.3 Tests für 2 unabhängige Stichproben

Die folgenden Tests prüfen, ob eine Variable in zwei unabhängig voneinander erhobenen Stichproben aus einer gleichen Grundgesamtheit stammen.

22.3.1 Mann-Whitney U-Test

Dieser Test ist die Alternative zum parametrischen t-Test für den Vergleich von zwei Mittelwerten von Verteilungen (zentrale Tendenz bzw. Lage), wenn die Voraussetzungen für den t-Test nicht erfüllt sind: es liegt keine metrische Skala vor und/oder die getestete Variable ist nicht normalverteilt. Der Test prüft auf Unterschiede hinsichtlich der zentralen Tendenz von Verteilungen. Voraussetzung für den Mann-Whitney-Test ist, dass die getestete Variable mindestens ordinalskaliert ist. Bei dem Test werden nicht die Messwerte der Variablen, sondern Rangplätze zugrundegelegt. An einem folgenden Beispiel sei das Test-Verfahren zunächst erläutert. Es werden zwei Schülergruppen A und B eines Jahrgangs mit unterschiedlichen Methoden in Mathematik unterrichtet. Schülergruppe B mit $n_1 = 5$ Schülern wird mit einer neuen Methode und die Kontroll-Schülergruppe A mit $n_2 = 4$ Schülern mit der herkömmlichen Methode unterrichtet. Zum Abschluss des Experiments werden Klausuren geschrieben. In der Tabelle 22.7 sind die Ergebnisse für beide Gruppen in erreichten Punkten aufgeführt.

Tabelle 22.7. Erreichte Leistungsergebnisse für zwei Testgruppen

A	21	14	10	24	
B	17	22	18	23	26

Geprüft werden soll, ob die Schülergruppe B eine bessere Leistung erbracht hat. Wegen der kleinen Stichproben und der ordinalskalierten Variable eignet sich hierfür der Mann-Whitney-Test. Da die beiden Gruppen als zwei unabhängige Stichproben aus Grundgesamtheiten interpretiert werden, lassen sich folgende Hypothesen gegenüberstellen:

- H_0-Hypothese: die Variable hat in beiden Grundgesamtheiten die gleiche Verteilung.
- H_1-Hypothese für die hier relevante einseitige Fragestellung: die Variable ist in der Grundgesamtheit B größer als in A.

Zur Prüfung der Nullhypothese werden die Werte beider Stichproben in aufsteigender Reihenfolge unter Aufzeichnung der Gruppenherkunft zusammengefasst (⇨ Tabelle 22.8). Aus der Reihenfolge von Werten aus den beiden Gruppen wird eine Testvariable U nach folgendem Messverfahren ermittelt: Es wird zunächst gezählt, wie viele Messwerte aus der Gruppe B vor jedem Messwert aus der Gruppe A liegen. U ist die Anzahl der Messwerte aus der Gruppe B, die insgesamt vor den Messwerten der Gruppe A liegen. Vor dem Messwert 10 der Gruppe A liegt kein Messwert der Gruppe B. Für den Messwert 14 der Gruppe A gilt gleiches. Vor dem

Messwert 21 der Gruppe A liegen zwei Messwerte der Gruppe B usw. Durch Addition erhält man

$$U = 0 + 0 + 2 + 4 = 6. \tag{22.2}$$

Tabelle 22.8. Rangordnung der Leistungsergebnisse

Messwerte	10	14	17	18	21	22	23	24	26
Gruppe	A	A	B	B	A	B	B	A	B
Rangziffer	1	2	3	4	5	6	7	8	9

Des weiteren kann U′ ermittelt werden. Zur Ermittlung von U′ wird nach gleichem Schema gezählt, wie viele Messwerte der Gruppe A vor den Messwerten der Gruppe B liegen. Es ergibt sich

$$U' = 2 + 2 + 3 + 3 + 4 = 14. \tag{22.3}$$

Der kleinere Wert der beiden Auszählungen ist die Prüfvariable U. Wegen $U' = n_1 * n_2 - U$ und $U = n_1 * n_2 - U'$ lässt sich der kleinere Wert nach einer Auszählung leicht ermitteln. Der mögliche untere Grenzwert für U ist 0: alle Werte von A liegen vor den Werten von B. Insofern sprechen sehr kleine Werte von U für die Ablehnung der Hypothese H_0. Die Stichprobenverteilung von U ist unter der Hypothese H_0 bekannt. Für sehr kleine Stichproben ($n_1, n_2 < 8$) gibt es Tabellen. Aus diesen kann man die Wahrscheinlichkeit - für H_0 ein U gleich/kleiner als das empirisch bestimmte U zu erhalten - entnehmen (Siegel, 1956). Für unser Beispiel mit $n_1 = 4$, $n_2 = 5$ und $U = 6$ ergibt sich eine Wahrscheinlichkeit von $P = 0{,}206$. Wenn das Signifikanzniveau auf $\alpha = 0{,}05$ festgelegt wird, kann die Hypothese H_0 nicht abgelehnt werden, da $0{,}206 > 0{,}05$ ist. Für große Stichproben ist die standardisierte Testgröße U approximativ standardnormalverteilt.

Von *Wilcoxon* ist für gleiche Anwendungsbedingungen ein äquivalenter Test vorgeschlagen worden. Der Test von Wilcoxon ordnet ebenfalls die Werte der zusammengefassten Stichproben nach der Größe. Dann werden Rangziffern vergeben: der kleinste Wert erhält die Rangziffer 1 der nächstgrößte die Rangziffer 2 usw. (⇨ Tabelle 22.8). Schließlich werden für die Fälle einer jeden Gruppe die Rangziffern addiert. Wenn beide Gruppen die gleiche Verteilung haben, so sollten sie auch ähnliche Rangziffernsummen haben. Im obigen Beispiel ergibt sich für Gruppe A eine Rangsumme in Höhe von 16 und für B eine in Höhe von 29. Da die Rangziffernsummen in die Größen U bzw. U′ überführt werden können, führen beide Tests zum gleichen Ergebnis.

Nicht unproblematisch ist es, wenn Mitglieder verschiedener Gruppen gleiche Messwerte haben (im angelsächsischen Sprachraum spricht man von *ties*). Wäre z.B. der größte Messwert der Gruppe B auch 24, so wären für diese Fälle zwei Rangfolgen (zuerst A oder zuerst B) möglich mit unterschiedlichen Ergebnissen für die Höhe von U. Diesen Sachverhalt muss das Testverfahren natürlich berücksichtigen. Im Fall gleicher Messwerte wird zur Ermittlung von Rangziffernsummen das arithmetische Mittel der Rangordnungsplätze als Rangziffer verge-

ben: z.B. würden beim Messwert 24 für beide Gruppen die Rangordnungsplätze 8 und 9 belegt werden und der Mittelwert 8,5 als Rangziffer zugeordnet.

Im folgenden Anwendungsbeispiel wird auf den Datensatz ALLBUS90.SAV zurückgegriffen. Untersucht werden soll, ob die Einstellung zur Treue in einer Partnerschaft von Frauen und Männern gleich oder unterschiedlich bewertet wird. Die Befragungen von Männern und Frauen können als zwei unabhängige Stichproben angesehen werden. Die Messwerte „1" bis „4" der ordinalskalierten Variablen TREUE erfassen die Antworten „sehr schlimm" bis „gar nicht schlimm" auf die Frage nach der Bedeutung eines „Seitensprungs". Die Variable TREUE ist ordinalskaliert. Zum Testen der Hypothese mit dem Mann-Whitney U-Test gehen Sie wie folgt vor:

- ▷ Klicken Sie die Befehlsfolge „Analysieren", „Nichtparametrische Tests ▷„ „Zwei unabhängige Stichproben···". Es öffnet sich die in Abb. 22.6 dargestellte Dialogbox.
- ▷ Von den in „Welche Tests durchführen?" auswählbaren Tests wird der Mann-Whitney U-Test durch Anklicken ausgewählt.
- ▷ Aus der Quellvariablenliste wird die Testvariable TREUE in das Eingabefeld „Testvariablen" übertragen.
- ▷ Danach wird die Variable GESCHL, die die zwei unabhängigen Stichproben (Gruppen) definiert, in das Eingabefeld von „Gruppenvariable" übertragen. Sie erscheint dort zunächst als „geschl(? ?)".
- ▷ Durch Anklicken von „Gruppen definieren..." öffnet sich die in Abb. 22.7 dargestellte Dialogbox. In die Eingabefelder werden die Variablenwerte „1" und „2" der Variablen GESCHL zur Bestimmung der beiden Gruppen Männer und Frauen eingetragen. Mit „Weiter" und „OK" wird die Testprozedur gestartet.

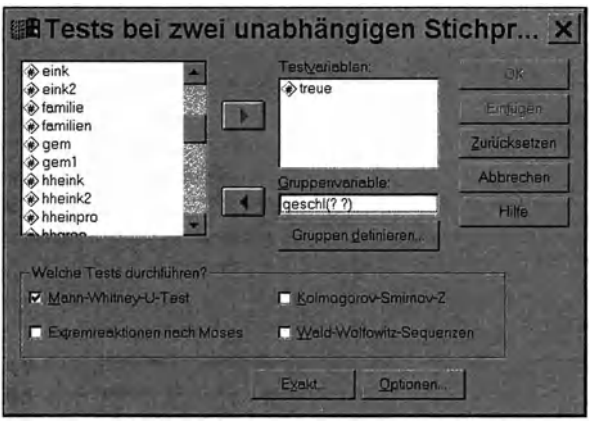


Abb. 22.6. Dialogbox „Tests bei zwei unabhängigen Stichproben"

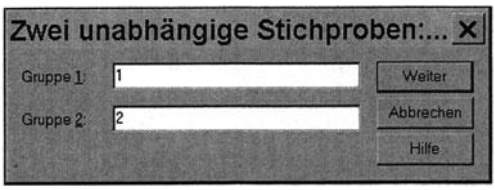


Abb. 22.7. Dialogbox „Zwei unabhängige Stichproben: Gruppen definieren"

Aus der Ergebnisausgabe (⇨ Tab. 22.9) kann man entnehmen, dass es insgesamt 153 gültige Fälle gibt mit 74 männlichen und 79 weiblichen Befragten. „Rangsumme" gibt die Rangziffernsumme und „Mittlerer Rang" die durchschnittliche Rangziffernsumme für jede Gruppe an. „Wilcoxon-W" = 5394,5 ist die kleinste der Rangziffernsummen (in Vers. 6.0 wird fehlerhaft die größere angegeben). „Mann-Whitney-U" (= 2234,5) ist die Prüfgröße des Tests. Da für große Stichproben ($n_1 + n_2 \geq 30$) die Verteilung der Prüfgröße U durch eine Standardnormalverteilung approximiert werden kann, wird mit Z = -2,609 der empirische Wert der Standardnormalverteilung angegeben. Dem Z-Wert entspricht die zweiseitige Wahrscheinlichkeit von 0,009. Da diese Wahrscheinlichkeit kleiner ist als ein für den Test angenommenes 5-%- Signifikanzniveau ($\alpha = 0{,}05$), wird die H_0-Hypothese einer gleichen Verteilung abgelehnt. Die Einstellung von Männer und Frauen ist demnach verschieden.

Der Test kann auch für die einseitige Fragestellung (H_1-Hypothese: Frauen bewerten einen Seitensprung als schlimmer als Männer) angewendet werden. Die durchschnittliche Rangziffernsumme für Frauen ist kleiner. Kleinere Rangziffern implizieren eine höhere Ablehnung eines Seitensprungs (sehr schlimm ist mit „1", gar nicht schlimm mit „4" codiert). Die einseitige exakte Signifikanz kann mit „Exakt Test" berechnet werden.

Optionen. ⇨ Erläuterungen zu Abb. 22.3.
Exakter Test. ⇨ Kap. 29.

Tabelle 22.9. Ergebnisausgabe des Mann-Whitney U-Tests

Ränge

	GESCHL	N	Mittlerer Rang	Rangsumme
TREUE	MAENNLICH	74	86,30	6386,50
	WEIBLICH	79	68,28	5394,50
	Gesamt	153		

Statistik für Test [a]

	TREUE
Mann-Whitney-U	2234,500
Wilcoxon-W	5394,500
Z	-2,609
Asymptotische Signifikanz (2-seitig)	,009

a. Gruppenvariable: GESCHL

22.3.2 Moses-Test bei extremer Reaktion

Dieser Test eignet sich dann, wenn man erwartet, dass bei experimentellen Tests unter bestimmten Testbedingungen manche Personen stark in einer Weise und andere Personen stark in einer entgegengesetzten Weise reagieren. Insofern stellt der Test auf Unterschiede in den Streuungen der Verteilungen ab.

Die Messwerte von zwei Vergleichsgruppen A und B (einer Kontroll- und einer Experimentiergruppe) werden in eine gemeinsame aufsteigende Rangfolge gebracht und erhalten Rangziffern. Unter der H_0-Hypothese (die Stichproben A und B kommen aus einer gleichen Grundgesamtheit) kann man erwarten, dass sich die Messwerte in der Kontroll- und Experimentiergruppe gut mischen. Unter der Hypothese H_1 (die Stichproben stammen aus unterschiedlichen Grundgesamtheiten bzw. unter den Testbedingungen haben die Testpersonen reagiert) kann man für die Experimentiergruppe sowohl relativ mehr höhere als auch niedrigere Messwerte erwarten. Der Test von Moses prüft, ob sich die Spannweite der Rangziffern (höchster minus kleinster plus eins) der Kontrollgruppe von der aller Probanten unterscheidet.

Beispiel. Es soll geprüft werden, ob sich die Einstellung zur Treue (hinsichtlich ihrer Streuung) bei jungen (18-29-jährige) und älteren (60-74-jährige) Menschen unterscheidet (Datensatz ALLBUS90.SAV). Vermutet wird, dass bei älteren eine höhere Variation in der Einstellung zur Treue besteht. Testvariable ist TREUE und Gruppenvariable ist ALT2 in der die Altersgruppen codiert sind. Zur Durchführung des Tests geht man wie in Kap. 22.3.1 erläutert vor. Im Unterschied dazu wird der Test von Moses sowie „1" und „4" als Gruppen der Gruppenvariable ALT2 gewählt.

In Tabelle 22.10 ist die Ergebnisausgabe niedergelegt. Es werden in der ersten Tabelle die gültigen Fallzahlen für beide Altersgruppen und in der zweiten Tabelle die Spannweite für die Kontrollgruppe (= Gruppe 1) sowie das exakte Signifikanzniveau für die einseitige Fragestellung („Signifikanz") angegeben. Die Spannweite und das Signifikanzniveau wird auch unter Ausschluss von Extremwerten bzw. Ausreißern („getrimmte Kontrollgruppe") aufgeführt. Als Testergebnis kann festgehalten werden, dass die H_0-Hypothese - die Altersgruppen unterscheiden sich nicht hinsichtlich ihrer Einstellung zur Treue - abgelehnt wird, da der Wert von „Signifikanz" (0,00 bzw. 0,021) kleiner ist als ein vorgegebenes Signifikanzniveau von z.B. 5 % ($\alpha = 0{,}05$).

Optionen. ⇨ Erläuterungen zu Abb. 22.2.
Exakter Test. Da immer exakt berechnet wird, erübrigt sich „Exakte Tests".

Tabelle 22.10. Ergebnisausgabe des Tests von Moses

Häufigkeiten

	ALT2	N
TREUE	18 - 29 JAHRE (Kontrolle)	36
	60 - 74 JAHRE (Experimente	33
	Gesamt	69

Statistik für Test[a,b]

		TREUE
Beobachtete Spannweite der Kontrollgruppe	N	57
	Signifikanz (1-seitig)	,000
Spannweite der getrimmten Kontrollgruppe	N	57
	Signifikanz (1-seitig)	,021
Ausreißer an beiden Enden entfernt		1

a. Moses-Test

b. Gruppenvariable: ALT2

22.3.3 Kolmogorov-Smirnov Z-Test

Dieser Test hat die gleichen Anwendungsvoraussetzungen wie der Mann-Whitney U-Test: zwei unabhängige Zufallsstichproben, das Messniveau der Variable ist mindestens ordinalskaliert. Auch die H_0-Hypothesen entsprechen einander: beide Stichproben stammen aus Grundgesamtheiten mit gleicher Verteilung.

Im Vergleich zum Mann-Whitney U-Test prüft der Test jegliche Abweichungen der Verteilungen (zentrale Tendenz, Streuung etc.; deshalb auch Omnibus-Test genannt). Soll lediglich geprüft werden, ob sich die zentrale Tendenz der Verteilungen unterscheidet, so sollte der Mann-Whitney U-Test bevorzugt werden.

Analog zum Kolmogorov-Smirnov-Test für den 1-Stichprobenfall (⇨ Kap. 22.2.4) basiert die Prüfgröße auf der maximalen Differenz (D_{max}) zwischen den kumulierten Häufigkeiten der beiden Stichprobenverteilungen. Wenn die Hypothese H_0 gilt (die Verteilungen unterscheiden sich nicht) so kann man erwarten, dass die kumulierten Häufigkeiten beider Verteilungen nicht stark voneinander abweichen. Ist D_{max} größer als unter der Hypothese H_0 zu erwarten ist, so wird H_0 abgelehnt.

Zur Anwendung des Kolmogorov-Smirnov Z-Tests im 2-Stichprobenfall wird wie zur Durchführung des Mann-Whitney U-Tests (⇨ Kap. 22.3.1) vorgegangen. Im Unterschied dazu wird aber der Kolmogorov-Smirnov Z-Test gewählt. Ein Test auf Unterschiede zwischen Männern und Frauen in der Einstellung zur Treue führt zu zwei Ausgabetabellen. In der ersten (hier nicht aufgeführten) Tabelle wird die Häufigkeit der Variable TREUE nach dem Geschlecht untergliedert (⇨ Tabelle 22.9 links). In der zweiten Tabelle steht das Testergebnis (⇨ Tabelle 22.11).

Als größte (positive) Differenz D_{max} der Abweichungen in den kumulierten Häufigkeiten wird 0,17961 ausgewiesen. Aus der Differenz ergibt sich nach Kolmogorov und Smirnov für die Prüfgröße Z = 1,11 gemäß Gleichung 22.4.

$$KS - Z = D_{max}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = 0{,}17961\sqrt{\frac{74 \cdot 79}{74 + 79}} = 1{,}11 \qquad (22.4)$$

Von Smirnov sind Tabellen entwickelt worden, in denen den Z-Werten zweiseitige Wahrscheinlichkeiten zugeordnet sind. Dem Wert Z = 1,11 entspricht die zweiseitige Wahrscheinlichkeit 0,17. Eine maximale absolute Differenz gemäß der bestehenden kann demnach mit einer Wahrscheinlichkeit von 17 % auftreten. Legt man das Signifikanzniveau auf $\alpha = 0{,}05$ fest, so kann wegen 0,17 > 0,05 die Hypothese H_0 (es gibt keinen Unterschied in der Einstellung zur Treue) nicht abgelehnt werden.

Führt man aber einen exakten Test mit dem Monte Carlo-Verfahren durch, so ergibt sich eine (2-seitige) Signifikanz = 0,038 (⇨ Tabelle 22.11). Demgemäß würde die Hypothese H_0 abgelehnt werden. Hier zeigt sich, dass man nicht immer auf die Ergebnisse asymptotischer Tests vertrauen kann.

Optionen. ⇨ Erläuterungen zu Abb. 22.2.
Exakter Test. ⇨ Kap. 29.

Tabelle 22.11. Ergebnisausgabe des Kolmogorov-Smirnov Z-Tests für zwei Stichproben

Statistik für Test[a]

		TREUE
Extremste Differenzen	Absolut	,180
	Positiv	,180
	Negativ	,000
Kolmogorov-Smirnov-Z		1,110
Asymptotische Signifikanz (2-seitig)		,170

a. Gruppenvariable: GESCHLECHT

Statistik für Test[b]

			TREUE
			,170
Monte-Carlo-Signifikanz(2-seitig)	Signifikanz		,038[a]
	99%-Konfidenzintervall	Untergrenze	,033
		Obergrenze	,043

a. Basiert auf 10000 Stichprobentabellen mit einem Startwert von 334431365.

b. Gruppenvariable: GESCHLECHT, BEFRAGTE<R>

22.3.4 Wald-Wolfowitz-Test

Der Wald-Wolfowitz-Test testet die H_0-Hypothese - beide Stichproben stammen aus gleichen Grundgesamtheitsverteilungen - gegen die Hypothese verschiedener Verteilungen in jeglicher Form (zentrale Lage, die Streuung etc., deshalb auch Omnibus-Test genannt). Er ist eine Alternative zum Kolmogorov-Smirnov Z-Test. Vorausgesetzt werden mindestens ein ordinales Skalenniveau sowie zwei unabhängige Stichproben.

Ganz analog zum Mann-Whitney U-Test werden die Messwerte beider Stichproben in eine Rangordnung gebracht, wobei mit dem kleinsten Wert begonnen wird. Dann wird - anolog zum Sequenzen-Test für eine Stichprobe - die Anzahl der Se-

quenzen gezählt. Es handelt sich also um einen Sequenzen-Test in Anwendung auf den 2-Stichprobenfall.

Am Beispiel zur Erläuterung des Mann-Whitney U-Tests (⇨ Tabelle 22.7) kann dieses gezeigt werden. Die Anzahl der Sequenzen beträgt 6 (⇨ Tabelle 22.12). Im Fall von Bindungen (gleiche Messwerte in beiden Gruppen) wird der Mittelwert der Ränge gebildet.

Tabelle 22.12. Beispiel zur Ermittlung von Sequenzen

Messwerte	10	14	17	18	21	22	23	24	26
Gruppe	A	A	B	B	A	B	B	A	B
Sequenz	1.	1.	2.	2.	3.	4.	4.	5.	6.

Das Beispiel Einstellung zur Treue aus dem Datensatz ALLBUS90.SAV eignet sich nicht für den Test, weil die Variable TREUE nur vier Werte hat und es deshalb zu viele Bindungen (ties) gibt. Es wird das Beispiel aus der Abb. 22.7 zur Berechnung mit SPSS genommen (Datei MATHE.SAV). Die Vorgehensweise entspricht - bis auf die Auswahl des Tests - der in Kapitel 22.3.1 erläuterten. In Tabelle 22.13 ist die Ergebnisausgabe zu sehen.

Für Stichprobengrößen $n_1 + n_2 \leq 30$ wird ein einseitiges exaktes Signifikanzniveau berechnet. Für Stichproben > 30 wird eine Approximation durch die Standardnormalverteilung verwendet. In der ersten Ausgabetabelle (hier nicht aufgeführt) werden die Häufigkeiten für die Gruppen genannt. In der zweiten Ausgabetabelle (Tabelle 22.13) werden die Z-Werte mit der damit verbundenen einseitigen Wahrscheinlichkeit für die Anzahl der exakten Sequenzen [bzw. minimale und maximale Anzahl im Fall von Bindungen (ties)] angegeben. Sind die ausgewiesenen Wahrscheinlichkeiten kleiner als das gewählte Signifikanzniveau (z.B. α = 0,05), so wird die Hypothese H_0 abgelehnt. Da „Exakte Signifikanz" mit 0,786 größer ist als α =0,05, wird H_0 (kein Unterschied in den Mathematik-Lehrmethoden) angenommen.

Optionen. ⇨ Erläuterungen zu Abb. 22.2.
Exakter Test. ⇨ Kap. 29.

Tabelle 22.13. Ergebnisausgabe des Wald-Wolfowitz-Tests

Statistik für Test[b,c]

		Anzahl der Sequenzen	Z	Exakte Signifikanz (1-seitig)
PUNKTE	Exakte Anzahl der Sequenzen	6[a]	,763	,786

a. Es wurden keine Bindungen zwischen Gruppen gefunden.

b. Test nach Wald-Wolfowitz

c. Gruppenvariable: METHODE

22.4 Tests für k unabhängige Stichproben

Bei diesen Tests wird in Erweiterung der Fragestellung für den Fall von zwei unabhängigen Stichproben geprüft, ob sich k (drei oder mehr) Gruppen (Stichproben) unterscheiden oder nicht. Es wird die H_0-Hypothese (alle Gruppen stammen aus der gleichen Grundgesamtheit) gegen die H_1-Hypothese (die Gruppen entstammen aus unterschiedlichen Grundgesamtheiten) geprüft. Die übliche parametrische Methode für einer derartige Fragestellung ist der F-Test der einfaktoriellen Varianzanalyse. Voraussetzung dafür aber ist, dass die Messwerte unabhängig voneinander aus normalverteilten Grundgesamtheiten mit gleichen Varianzen stammen. Des weiteren ist Voraussetzung, dass das Messniveau der abhängigen Variablen mindestens intervallskaliert ist. Wenn die untersuchte Variable ordinalskaliert ist oder die Annahme einer Normalverteilung fragwürdig ist, sind die folgenden nichtparametrischen Tests einsetzbar.

22.4.1 Kruskal-Wallis H-Test

Der Kruskal-Wallis-Test eignet sich gut zur Prüfung auf eine unterschiedliche zentrale Tendenz von Verteilungen. Er ist eine einfaktorielle Varianzanalyse für Rangziffern. Die Messwerte für die k Stichproben bzw. Gruppen werden in eine gemeinsame Rangordnung gebracht. Aus diesen Daten wird die Prüfgröße H wie folgt berechnet:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} R_i^2 / n_i - 3(n+1) \qquad (22.5)$$

R_i = Summe der Rangziffern der Stichprobe i

n_i = Fallzahl der Stichprobe i

n = Summe des Stichprobenumfangs aller k Gruppen

Für den Fall von Bindungen (englisch: ties), wird die Gleichung mit einem Korrekturfaktor korrigiert (⇨ Bortz/Lienert/Boehnke, S. 223). Die Prüfgröße H ist approximativ chi-quadrat-verteilt mit k-1 Freiheitsgraden.

Beispiel. Mit dem Datensatz ALLBUS90.SAV soll untersucht werden, ob die Einstellung zur Treue in einer Partnerschaft unabhängig vom Alter ist. Die Personen verschiedener Altersgruppen (codiert in der Variable ALT2) können als vier unabhängige Stichproben angesehen werden. Zum Testen der Hypothese wird der Kruskal-Wallis H-Test wie folgt angewendet:

- ▷ Klicken Sie die Befehlsfolge „Analysieren", „Nichtparametrische Tests ▷„ „K unabhängige Stichproben···". Es öffnet sich die in Abb. 22.8 dargestellte Dialogbox.
- ▷ In „Welche Tests durchführen?" wird „Kruskal-Wallis H" angeklickt.
- ▷ Aus der Quellvariablenliste wird die Testvariable TREUE in das Eingabefeld „Testvariablen" übertragen.
- ▷ Danach wird die Variable ALT2, deren Altersgruppen als unabhängige Stichproben anzusehen sind, in das Eingabefeld von „Gruppenvariable" übertragen. Sie erscheint dort zunächst als „alt2(? ?)".

▷ Durch Anklicken von „Bereich definieren“ öffnet sich die in Abb. 22.9 dargestellte Dialogbox. In die Eingabefelder „Minimum“ und „Maximum“ werden die Variablenwerte „1“ und „4“ der Variablen ALT2 zur Bestimmung des Wertebereichs der Variable ALT2 eingetragen. Der Test prüft dann auf Unterschiede für die ersten vier Altersgruppen. Mit „Weiter“ und „OK“ wird die Testprozedur gestartet.

In Tabelle 22.14 ist die Ergebnisausgabe des Tests zu sehen. „Mittlerer Rang“ gibt die durchschnittlichen Rangziffern und „N“ die Fallzahlen der vier Altersgruppen an. Der Wert der approximativ chi-quadrat-verteilten Prüfgröße ist mit 5,64 kleiner als ein aus einer Chi-Quadrat-Tabelle für $k-1=3$ Freiheitsgrade (df) bei einer Irrtumswahrscheinlichkeit von $\alpha = 0{,}05$ entnehmbarer kritischer Wert von 7,82 [in Vers. 6.0. werden zwei empirische Prüfwerte aufgeführt: ein korrigierter zur Berücksichtigung von „ties“ (d.h. gleicher Rangziffern für verschiedene Fälle) und einer ohne diese Berücksichtigung]. Demnach wird die Hypothese H_0 (es gibt für die Altersgruppen keinen Unterschied in der Einstellung zur Treue) angenommen. Diese Schlussfolgerung ergibt sich auch aus dem angegebenem Signifikanzniveau 0,13 („Asymptotische Signifikanz“), das die mit $\alpha = 0{,}05$ vorzugebene Irrtumswahrscheinlichkeit übersteigt.

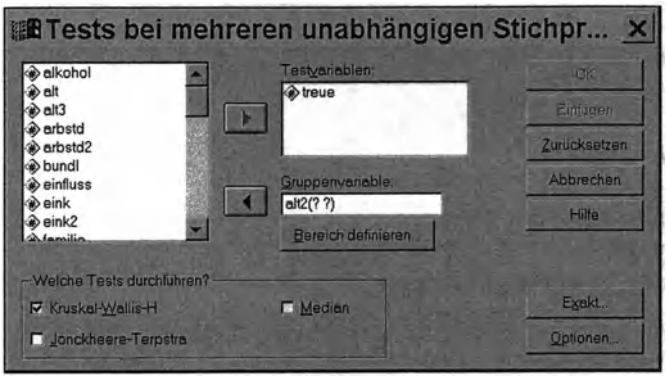


Abb. 22.8. Dialogbox „Tests bei mehreren unabhängigen Stichproben“

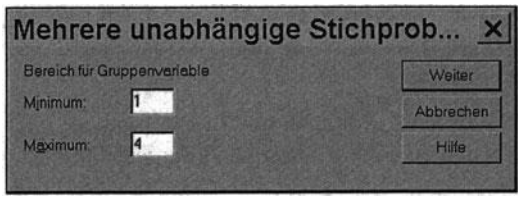


Abb. 22.9. Dialogbox „Mehrere unabh. Stichproben: Bereich definieren“

Optionen. ⇨ Erläuterungen zu Abb. 22.4.
Exakter Test. ⇨ Kap. 29.

Tabelle 22.14. Ergebnisausgabe des Kruskal-Wallis H-Tests

Ränge

	ALTER, BEFRAGTE<R>,	N	Mittlerer Rang
TREUE	18 - 29 JAHRE	32	68,31
	30 - 44 JAHRE	45	66,68
	45 - 59 JAHRE	26	87,50
	60 - 74 JAHRE	38	67,09
	Gesamt	141	

Statistik für Test[a,b]

	VERHALTENSBEURTEILUNG: SEITENSPRUNG
Chi-Quadrat	5,640
df	3
Asymptotische Signifikanz	,130

a. Kruskal-Wallis-Test

b. Gruppenvariable: ALTER, BEFRAGTE

22.4.2 Median-Test

Auch der Median-Test verlangt, dass die Variable mindestens ordinalskaliert ist. Geprüft wird, ob die Stichproben aus Grundgesamtheiten mit gleichen Medianen stammen.

Der Test nutzt nur Informationen über die Höhe eines jeden Beobachtungswertes im Vergleich zum Median. Daher ist er ein sehr allgemeiner Test.

Bei diesem Testverfahren wird zunächst für die Messwerte aller k Gruppen der gemeinsame Median bestimmt. Im nächsten Schritt wird jeder Messwert als kleiner bzw. größer als der gemeinsame Median eingestuft und für alle Gruppen werden die Häufigkeiten des Vorkommens von kleiner bzw. größer als der Median ausgezählt. Es entsteht für k Gruppen eine 2*k-Häufigkeitstabelle. Falls $n > 30$ ist, wird aus der Häufigkeitstabelle eine Chi-Quadrat-Prüfgröße ermittelt und für $k-1$ Freiheitsgrade ein approximativer Chi-Quadrat-Test durchgeführt. Für kleinere Fallzahlen wird mit Fischer's exact Test die genaue Wahrscheinlichkeit berechnet.

Das folgende Anwendungsbeispiel aus dem Datensatz ALLBUS90.SAV ist das gleiche wie in Kap. 22.4.1: es soll geprüft werden, ob die Einstellung zur Treue in einer Partnerschaft unabhängig vom Alter ist. Zum Testen der Hypothese geht man wie dort beschrieben vor. Im Unterschied dazu wird aber der Test „Median" durch Klicken gewählt.

In Tabelle 22.15 ist die Ergebnisausgabe dargestellt. Da $k = 4$ ist, wird eine 2*4-Häufigkeitstabelle dargestellt. In der ersten Ausgabetabelle werden für die vier Altersgruppen die Häufigkeiten für die Variable TREUE mit den Ausprägungen größer als der Median („GT Median" = greater than median) und gleich-kleiner als der Median („LE Median" = less equal median) aufgeführt. Mit „Chi-Quadrat" = 7,473 wird der ermittelte empirische Chi-Quadrat-Wert ausgewiesen. Für $k-1 = 3$ Freiheitsgrade („df") und einem Signifikanzniveau von 5 % ($\alpha = 0{,}05$) ergibt sich aus einer Chi-Quadrat-Tabelle ein kritischer Wert von 7,82. Da der empirische Wert kleiner ist als der kritische, wird die Hypothese H_0 (die Einstellung zur Treue ist unabhängig vom Alter) angenommen. Dieses Testergebnis ergibt sich einfacher auch daraus, dass die von SPSS ausgewiesene „Signifikanz" (= 0,058) größer ist als die gewählte in Höhe von $\alpha = 0{,}05$.

Optionen. ⇨ Erläuterungen zu 22.4.1.
Exakter Test. ⇨ Kap. 29.

Tabelle 22.15. Ergebnisausgabe des Median-Tests

Häufigkeiten

		ALTER, BEFRAGTE<R>, KATEGORISIERT			
		18 - 29 JAHRE	30 - 44 JAHRE	45 - 59 JAHRE	60 - 74 JAHRE
TREUE	> Median	11	16	17	15
	< = Median	21	29	9	23

Statistik für Test[b]

	VERHALTENSBEURTEILUNG: SEITENSPRUNG
N	141
Median	2,00
Chi-Quadrat	7,473[a]
df	3
Asymptotische Signifikanz	,058

a. Bei 0 Zellen (,0%) werden weniger als 5 Häufigkeiten erwartet. Die kleinste erwartete Zellenhäufigkeit ist 10,9.

b. Gruppenvariable: ALTER

22.4.3 Jonckheere-Terpstra-Test

Dieser Test ist nur nach Installation des SPSS-Moduls „Exakt Test" verfügbar.

Weder der Kruskal-Wallis- noch der Median-Test sind geeignet, Annahmen über die Richtung des Unterschiedes zwischen den Gruppen zu prüfen. In manchen Untersuchungen (speziell bei experimentellen Untersuchungsdesigns) hat man die Situation, dass die Wirkungen mehrerer Aktivitäten oder Maßnahmen simultan geprüft werden sollen und eine Rangfolge in der Wirkungsrichtung angenommen werden kann. In unserem Anwendungsbeispiel haben wir oben geprüft, ob mit wachsendem Alter die Einstellung zur Treue unterschiedlich ist. Es kam zur Annahme der H_0-Hypothese: kein Unterschied. Geht man aber davon aus, dass mit wachsendem Alter die Einstellung zur Treue sich in eine Richtung verändert (je höher das Alter, umso größer wird die Wertschätzung von Treue), kann man mit dem Jonckheere-Terpstra-Test eine bessere Trennschärfe zum Testen auf Unterschiede der Altersgruppen in der Einstellung zur Treue erzielen. Der Test ermöglicht ein Testen von geordneten Alternativen. Ein anderes Beispiel dafür wäre, wenn für mehrere Versuchsgruppen die Wirkung eines Medikaments mit jeweils einer höheren Dosis geprüft wird.

Zum Testen der Hypothese geht man wie in Kap. 22.4.1 beschrieben vor. Im Unterschied dazu wird aber der Test „Jonckheere-Terpstra" durch Klicken gewählt. In Tabelle 22.16 ist die Ergebnisausgabe dargestellt. Für die 141 gültigen Fälle („N") wird die empirische („beobachtete") Testgröße J-T, ihr Mittelwert, ihre Standardabweichung, die standardisierte Testgröße J-T (Differenz von J-T

zum Mittelwert dividiert durch die Standardabweichung) sowie ein asymptotisches 2-seitiges Signifikanzniveau ausgewiesen. Da der Wert von „Asymptotische Signifikanz (2-seitig)" mit 0,603 größer ist als ein vorzugebendes Signifikanzniveau von z.B. 0,05 (α = 0,05 %) wird die H_0-Hypothese (ein Zusammenhang zwischen der Einstellung zur Treue und dem Alter besteht nicht) angenommen. Damit werden die Ergebnisse in Kap. 22.4.1 und 22.4.2 (Kruskal-Wallis- und Median-Test) bestätigt.

Optionen. ⇨ Erläuterungen zu Abb. 22.2.
Exakter Test. ⇨ Kap. 29.

Tabelle 22.16. Ergebnisausgabe des Jonckheere-Terpstra-Test

Jonckheere-Terpstra-Test[a]

	TREUE
Anzahl der Stufen in ALTER, KATEGORISIERT	4
N	141
Beobachtete J-T-Statistik	3813,500
Mittelwert der J-T-Statistik	3678,000
Standardabweichung der J-T-Statistik	260,173
Standardisierte J-T-Statistik	,521
Asymptotische Signifikanz (2-seitig)	,603

a. Gruppenvariable: ALTER, KATEGORISIERT

22.5 Tests für 2 verbundene Stichproben

Bei diesem Testtyp möchte man prüfen, ob eine Maßnahme oder Aktivität wirksam ist oder nicht und bildet zwei Stichprobengruppen: eine Experiment- und eine Kontrollgruppe (matched pairs, ⇨ Kap. 22.1).

Die Grundhypothese (auch H_0-Hypothese genannt) postuliert, dass keine Unterschiede zwischen beiden Gruppen bestehen. Mit dieser Hypothese wird die Wirkung einer Maßnahme (z.B. die Wirksamkeit eines Medikaments oder der Erfolg einer neuen Lehr- oder Lernmethode) nicht anerkannt. Die Gegenthese H_1 geht von der Wirksamkeit aus.

22.5.1 Wilcoxon-Test

Der Test eignet sich, wenn Unterschiede in der zentralen Tendenz von Verteilungen geprüft werden sollen. Der Test beruht auf Rängen von Differenzen in den Variablenwerten. Der Wilcoxon-Test ist dem Vorzeichen(Sign)-Test (⇨ Kap. 22.5.2) vorzuziehen, wenn die Differenzen aussagekräftig sind.

Im folgenden wird zur Anwendungsdemonstration ein Beispiel aus dem Bereich der Pädagogik gewählt. Zur Überprüfung einer neuen Lehrmethode werden Schü-

lerpaare gebildet, die sich hinsichtlich ihres Lernverhaltens und ihrer Lernfähigkeiten gleichen. Eine Schülergruppe mit jeweils einem Schüler der Paare wird nach der herkömmlichen Lehrmethode (Methode A genannt) und die andere Gruppe mit dem zweiten Schüler der Paare nach der neuen (Methode B genannt) unterrichtet. Die Lernergebnisse wurden bei Leistungstests in Form von erreichten Punkten erfasst und als Variable METH_A und METH_B als SPSS-Datei gespeichert (⇨ Abb. 22.10, Datei LEHRMETH.SAV). Geprüft werden soll, ob die beiden Methoden sich unterscheiden oder nicht.

Es handelt sich hier um ordinalskalierte Variablen, wobei aber Differenzen von Variablenwerten eine gewisse Aussagekraft haben.

	nr	meth_a	meth_b	meth_c
1	1	11	14	9
2	2	15	13	17
3	3	12	14	13
4	4	14	15	16
5	5	12	14	15
6	6	13	13	17

Abb. 22.10. Ausschnitt aus den Daten des Anwendungsbeispiels

Bei dem Testverfahren werden im ersten Schritt die Differenzen der Messwerte für die Paare berechnet. Im nächsten Schritt werden die absoluten Differenzen (also keine Vorzeichenbeachtung) in eine gemeinsame Rangziffernreihen-Ordnung gebracht. Haben Paare gleiche Messwerte, so werden diese Fälle aus der Analyse ausgeschlossen. Schließlich werden diesen Rangziffern die Vorzeichen der Differenzen zugeordnet. Unter der Hypothese H_0 (kein Unterschied der beiden Methoden) kann man erwarten, dass aufgetretene große Differenzen sowohl durch die Methode A als auch durch die Methode B bedingt sind. Summiert man jeweils die positiven und negativen Rangziffern, so ist unter H_0 zu erwarten, dass die Summen sich zu Null addieren. Unter H_1 wäre dementsprechend zu erwarten, dass sich die Summen unterscheiden. Von Wilcoxon liegen Tabellen vor, aus denen man für die Prüfgröße (die kleinere der Rangziffernsummen) für ein vorgegebenes Signifikanzniveau von z.B 5 % ($\alpha = 0{,}05$) kritische Werte entnehmen kann (Siegel, 1956, S. 79 f.).

Zum Testen, ob die Lehrmethoden A und B unterschiedlichen Erfolg haben oder nicht, kann der Wilcoxon-Test wie folgt angewendet werden:

- ▷ Klicken Sie die Befehlsfolge „Analysieren", „Nichtparametrische Tests ▷„ „Zwei verbundene Stichproben···". Es öffnet sich die in Abb. 22.11 dargestellte Dialogbox.
- ▷ In „Welche Tests durchführen?" wird „Wilcoxon" angeklickt.
- ▷ Aus der Quellvariablenliste werden die Variablen METH_A und METH_B nacheinander angeklickt. Sie erscheinen dann im Fenster „Aktuelle Auswahl" als die gewählten Variablen. Anschließend wird durch Klicken auf den Pfeil-

schalter das gewählte Variablenpaar in das Eingabefeld „Ausgewählte Variablenpaare:“ übertragen. Mit „OK“ wird die Testprozedur gestartet.

In Tabelle 22.17 ist die Ergebnisausgabe des Tests zu sehen. In der ersten Tabelle werden für die negativen (METH_B < METH_A) und positiven (METH_B > METH_A) Rangziffern die Summe, die Durchschnitte („Mittlerer Rang“) und Fallzahlen („N“) aufgeführt. In einem Fall sind die Messwerte gleich. Dieser Fall wird als „Bindungen“ ausgewiesen (METH_B = METH_A). Die negative Rangsumme ist mit 59 am kleinsten. Aus der Tabelle von Wilcoxon (Siegel, 1956, S. 254) ergibt sich z.B. für ein Signifikanzniveau von 5 % (bei einem zweiseitigen Test) und für n = 19 ein kritischer Wert von 46 für die kleinere Rangziffernsumme. Da der empirische Wert mit 59 diesen übersteigt, wird die Hypothese H_0 angenommen. Die Differenz der Rangziffernsummen ist nicht hinreichend groß, um einen Unterschied der Methoden zu begründen.

Für Stichprobenumfänge n > 25 kann die Tabelle von Siegel nicht genutzt werden. Da die Prüfgröße der kleinere Rangziffernsumme in derartigen Fällen approximativ normalverteilt ist, kann der Test mit Hilfe der Standardnormalverteilung durchgeführt werden. Von SPSS wird der empirische Z-Wert der Standardnormalverteilung sowie das zugehörige zweiseitige Signifikanzniveau ausgegeben. Da dieses (zweiseitige) Signifikanzniveau („Signifikanz = 0,138“ für „Z = -1,483“) das vorgegebene $\alpha = 0{,}05$ übersteigt, kann auch hieraus der Schluss gezogen werden, dass die Hypothese H_0 (keine signifikanten Unterschiede der Lehrmethoden) angenommen wird.

Optionen. ⇨ Erläuterungen zu Abb. 22.2.
Exakter Test. ⇨ Kap. 29.

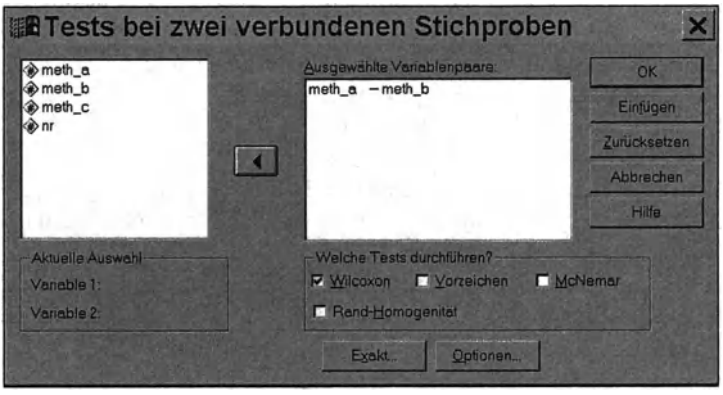


Abb. 22.11. Dialogbox „Tests bei zwei verbundenen Stichproben“

Tabelle 22.17. Ergebnisausgabe des Wilcoxon-Tests

Ränge

		N	Mittlerer Rang	Rangsumme
METH_B - METH_A	Negative Ränge	6[a]	9,83	59,00
	Positive Ränge	13[b]	10,08	131,00
	Bindungen	1[c]		
	Gesamt	20		

a. METH_B < METH_A

b. METH_B > METH_A

c. METH_A = METH_B

Statistik für Test[b]

	METH_B - METH_A
Z	-1,483[a]
Asymptotische Signifikanz (2-seitig)	,138

a. Basiert auf negativen Rängen.

b. Wilcoxon-Test

22.5.2 Vorzeichen-Test

Der Vorzeichen-Test (englisch: sign) stützt sich - wie der Wilcoxon-Test (⇨ Kap. 22.5.1) - auf Differenzen von Messwerten zwischen Paaren von Gruppen bzw. im „vorher-nachher"-Stichprobendesign. Im Unterschied zum Wilcoxon-Test gehen nur die Vorzeichen der Differenzen, nicht aber die Größen der Differenzen in Form von Rangziffern in das Testverfahren ein. Dieser Test bietet sich immer dann an, wenn (bedingt durch die Datenlage) die Höhe der Differenzen nicht aussagekräftig ist. Fälle, bei denen die Differenzen der Paare gleich Null sind, werden nicht in das Testverfahren einbezogen. Gezählt wird die Anzahl der positiven und die Anzahl der negativen Differenzen.

Unter der Hypothese H_0 (keine unterschiedliche Wirkung einer Maßnahme bzw. Aktivität) ist zu erwarten, dass die Fallzahlen mit positiven und negativen Vorzeichen etwa gleich sein werden. Für $n < 25$ kann die Wahrscheinlichkeit für die Häufigkeit der Vorzeichen mit Hilfe der Binomialverteilung berechnet werden.

Zur Durchführung des Vorzeichen-Tests geht man wie beim Wilcoxon-Test vor (⇨ Kap. 22.5.1). Im Unterschied dazu wird aber der Vorzeichen-Test in der Dialogbox der Abb. 22.11 angeklickt. Für das obige Beispiel des in Abb. 22.10 ausschnittsweise dargestellten Datensatzes (Datei LEHRMETH.SAV) ergibt sich die in Tabelle 22.18 dargestellte Ergebnisausgabe. Es werden in der ersten Tabelle die Fallzahlen mit negativen und positiven Differenzen angeführt. Die Wahrscheinlichkeit für das Auftreten von sechs negativen und 13 positiven Vorzeichen wird mit 0,167 („Exakte Signifikanz") angegeben. Bei einem Signifikanzniveau von 5 % ($\alpha = 0{,}05$) für den Test wird die Hypothese H_0 angenommen, da $0{,}167 > 0{,}05$ ist. Das Testergebnis entspricht dem von Wilcoxon.

Für große Fallzahlen ($n > 25$) wird die Binomialverteilung durch die Normalverteilung approximiert. SPSS gibt in diesen Fällen (wie bei dem Wilcoxon-Test) den Z-Wert der Standardnormalverteilung sowie die zugehörige Wahrscheinlichkeit an.

Optionen. ⇨ Erläuterungen zu Abb. 22.2.
Exakter Test. ⇨ Kap. 29.

Tabelle 22.18. Ergebnisausgabe des Vorzeichen-Tests

Häufigkeiten

		N
METH_B - METH_A	Negative Differenzen[a]	6
	Positive Differenzen[b]	13
	Bindungen[c]	1
	Gesamt	20

a. METH_B < METH_A

b. METH_B > METH_A

c. METH_A = METH_B

Statistik für Test[b]

	Ergebnis Lehrmethode B - Ergebnis Lehrmethode A
Exakte Signifikanz (2-seitig)	,167[a]

a. Verwendetete Binomialverteilung.

b. Vorzeichentest

22.5.3 McNemar-Test

Der McNemar-Test eignet sich für ein „vorher-nachher"-Testdesign mit dichotomen Variablen und testet Häufigkeitsunterschiede. Anhand eines Beispiels sei der Test erklärt. Um zu prüfen, ob zwei Aufgaben den gleichen Schwierigkeitsgrad haben, können diese nacheinander Probanten zur Lösung vorgelegt werden. Das Ergebnis in Form von Häufigkeiten kann in einer 2*2-Tabelle festgehalten werden. Die Häufigkeiten n_A und n_D in Tabelle 22.19 erfassen Veränderungen im Lösungserfolg durch den Wechsel der Aufgaben. Die Häufigkeiten n_C und n_B geben die Fälle mit gleichem Lösungserfolg an. Je weniger sich diese Häufigkeiten unterscheiden, um so wahrscheinlicher ist es, dass die H_0-Hypothese (durch den Wechsel der Aufgaben tritt keine Veränderung im Lösungserfolg ein) zutrifft. Die Wahrscheinlichkeit kann mit Hilfe der Binomialverteilung berechnet werden.

Zur Anwendungsdemonstration wird der ausschnittsweise in Abb. 22.12 dargestellte Datensatz genutzt (Datei TESTAUFG.SAV).

In dem Datensatz sind Lösungsresultate für von Studierenden bearbeitete Testaufgaben erfasst. Die Variablen AUFG1 und AUFG2 sind nominalskalierte Variable mit dichotomen Ausprägungen: Der Variablenwert „1" steht für „Aufgabe nicht gelöst" und „2" für „Aufgabe gelöst". Zur Durchführung des Tests geht man wie bei den anderen Tests bei zwei verbundenen Stichproben vor (⇨ Kap. 22.5.1). Im Unterschied dazu wird der McNemar-Test angeklickt.

Tabelle 22.19. 4-Felder Tabelle zur Erfassung von Änderungen

	Aufgabe 2	
Aufgabe 1	nicht gelöst	gelöst
gelöst	n_A	n_B
nicht gelöst	n_C	n_D

	nr	aufg1	aufg2	aufg3
1	1	2	1	2
2	2	1	1	1
3	3	2	1	2
4	4	1	1	2
5	5	2	1	1
6	6	2	1	2

Abb. 22.12. Ausschnitt aus den Daten des Anwendungsbeispiels (TESTAUFG.SAV)

In der folgenden Tabelle 22.20 wird das Ausgabeergebnis für den Test dokumentiert. Die Ausgabeform entspricht der Tabelle 22.19 bei einer Fallzahl von 15. Die Wahrscheinlichkeit wird wegen der kleinen Fallzahl ($n < 25$) auf der Basis einer Binomialverteilung ausgegeben.

Für große Fallzahlen wird approximativ ein Chi-Quadrat-Test durchgeführt. Testergebnis ist, dass die H_0-Hypothese (kein Unterschied im Schwierigkeitsgrad der Aufgaben) angenommen wird, da die angeführte zweiseitige Wahrscheinlichkeit („Exakte Signifikanz") das Signifikanzniveau von 5 % ($\alpha = 0{,}05$) übersteigt.

Optionen. ⇨ Erläuterungen zu Abb. 22.2.

Tabelle 22.20. Ergebnisausgabe des McNemar-Tests

1 & 2

	AUFG2	
AUFG1	1	2
1	3	3
2	8	1

Statistik für Test[b]

	1 & 2
N	15
Exakte Signifikanz (2-seitig)	,227[a]

a. Verwendetete Binomialverteilung.

b. McNemar-Test

22.5.4 Rand-Homogenitäts-Test

Dieser Test ist nur nach Installation des SPSS-Moduls „Exakt Test" verfügbar.

Dieser Test ist eine Verallgemeinerung des McNemar-Tests. Anstelle von zwei (binären) Kategorien (vorher - nachher) werden mehr als zwei Kategorien berücksichtigt. Dabei muss es sich um geordnete Kategorien handeln. Auf die theoretischen Grundlagen des Tests kann hier nicht eingegangen werden (⇨ Kuritz, Landis und Koch, 1988).

Beispiel. Ein Arzt verabreicht 25 Personen ein Präparat zur Erhöhung der allgemeinen Leistungsfähigkeit und im Abstand von drei Monaten ein Placebo. Anstelle der binären Kategorien „Wirkung" - „keine Wirkung" (codiert mit „1" und „2"), der einen McNemar-Test auf Prüfung der Wirksamkeit ermöglichen würde, werden die Merkmale „keine Wirkung", „geringe Wirkung" und „starke Wirkung" (kodiert mit „1", „2" und „3") erfasst. Anstelle einer 2*2-Kreuztabelle

(⇨ Tabelle 22.19) für den McNemar-Test würde nun eine 3*3-Kreuztabelle entstehen.

In Abb. 22.13 ist ein Ausschnitt aus der Datendatei PATIENT.SAV zu sehen.

patient	praepara	placebo
1	1	2
2	2	1
3	3	2
4	2	1
5	3	1

Abb. 22.13. Ausschnitt aus den Daten des Anwendungsbeispiels (PATIENT.SAV)

In Tabelle 22.21 ist die Ergebnisausgabe zu sehen. Außerhalb der Diagonalen der in der SPSS-Ausgabe nicht aufgeführten 3*3-Kreuztabelle gibt es 15 Fälle. Der empirische Wert für die Prüfgröße beträgt 35. Die Prüfgröße (MH = marginale Homogenität) hat einen Durchschnittswert von 29,5 mit einer Standardabweichung von 2,29. Daraus ergibt sich die standardisierte Prüfgröße in Höhe von 2,40. Da die ausgegebene 2-seitige Wahrscheinlichkeit („Asymptotische Signifikanz") kleiner ist als ein vorzugebendes Signifikanzniveau von z.B. $\alpha = 0{,}05$, kann von der Wirksamkeit des Präparats ausgegangen werden.

Tabelle 22.21. Ergebnisausgabe des Rand-Homogenitäts-Test

Rand-Homogenitätstest

	1 & 2
Unterschiedliche Werte	*3*
Fälle außerhalb der Diagonalen	*15*
Beobachtete MH-Statistik	*35,000*
Mittelwert der MH Statistik	*29,500*
Standardabweichung der MH-Statistik	*2,291*
Standardisierte MH-Statistik. MH Statistic	*2,400*
Asymptotische Signifikanz (2-seitig)	*,016*

22.6 Tests für k verbundene Stichproben

Bei diesen Testverfahren geht es um die simultane Prüfung von Unterschieden zwischen drei und mehr Stichproben bzw. Gruppen, wobei es sich um abhängige bzw. verbundene Stichproben handelt (⇨ Kap. 22.1). Die H_0-Hypothese lautet, dass die Stichproben aus identischen Grundgesamtheiten stammen.

22.6.1 Friedman-Test

Der Friedman-Test ist eine Zwei-Weg-Varianz-Analyse für Rangziffern zur Prüfung der Frage, ob die Stichproben aus einer gleichen Grundgesamtheit kommen. Es handelt sich um einen allgemeinen Test, der auf Unterschiede prüft ohne aufzudecken, um welche Unterschiede es sich handelt.

Der Test wird am Beispiel der Prüfung von drei Lehrmethoden auf den Lernerfolg von drei Studentengruppen demonstriert (⇨ Datensatz der Abb. 22.10, LEHRMETH.SAV). Die drei Stichprobengruppen wurden dabei aus Sets von jeweils drei Studenten mit gleicher Fähigkeiten, Lernmotivation u.ä. zusammengestellt, um die Wirkung der Lehrmethoden eindeutiger zu messen. In Tabelle 22.22 werden die Messwerte der ersten vier Zeilen (Sets) aus dem Datensatz der Abb. 22.10 angeführt. Im Testverfahren werden für jede Reihe (Zeile) der Tabelle Rangziffern vergeben. Unter der Hypothese H_0 - kein Unterschied im Erfolg der Methoden - verteilen sich die Rangziffern auf die drei Spalten zufällig, so dass auch die spaltenweise aufsummierten Rangziffernsummen sich kaum unterscheiden. Der Friedman-Test prüft, ob sich die Rangziffernsummen signifikant voneinander unterscheiden. Zum Testen dient folgende Prüfgröße:

$$\chi^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \qquad (22.6)$$

n = Anzahl der Fälle (= Anzahl der sets)
k = Anzahl der Variablen (= Spaltenanzahl = Anzahl der Gruppen)
R_j = Summe der Rangziffern in der Spalte j, d.h. der Gruppe j

Die Prüfgröße ist asymptotisch chi-quadrat-verteilt mit k-1 Freiheitsgraden.

Tabelle 22.22. Messwerte und Rangziffern der ersten vier Sets des Datensatzes

Methode	**Meth. A**		**Meth. B**		**Meth. C**	
Set	*Messwert*	*Rangziffer*	*Messwert*	*Rangziffer*	*Messwert*	*Rangziffer*
Set 1	11	2	14	1	9	3
Set 2	15	2	13	3	17	1
Set 3	12	3	14	1	13	2
Set 4	14	3	15	2	16	1
Rangsumme R		10		7		7

Zum Testen der Hypothese - haben die Lehrmethoden A, B und C unterschiedlichen Erfolg oder nicht - gehen Sie wie folgt vor:

- ▷ Klicken Sie die Befehlsfolge „Analysieren", "Nichtparamametrische Tests ▷„ „K verbundene Stichproben...". Es öffnet sich die in Abb. 22.14 dargestellte Dialogbox.
- ▷ In „Welche Tests durchführen?" wird „Friedman" angeklickt.
- ▷ Aus der Quellvariablenliste werden die Variablen METH_A, METH_B und METH_C markiert und durch Klicken auf den Pfeilschalter in das Eingabefeld „Testvariablen" übertragen. Mit „OK" wird die Testprozedur gestartet.

In Tabelle 22.23 ist die Ergebnisausgabe zu sehen. Es werden die durchschnittliche Rangziffernsumme („Mittlerer Rang"), die Fallzahl („N"), der empirische Wert der Prüfgröße Chi-Quadrat mit der Anzahl der Freiheitsgrade („df") sowie der zugehörigen Wahrscheinlichkeit („Signifikanz") angegeben. Wird für den Test z.B. ein Signifikanzniveau von 5 % ($\alpha = 0{,}05$) gewählt, so wird die H_0-Hypothese abgelehnt, da $0{,}006 < 0{,}05$ ist. Dieses Testergebnis wird plausibel, wenn man feststellt, dass der Wilcoxon-Test (⇨ Kap.22.5.1) ergibt, dass sich die Methode C signifikant sowohl von A als auch von B unterscheidet.

Statistiken. Durch Klicken auf die Schaltfläche „Statistiken..." öffnet sich eine (Unter-)Dialogbox in der deskriptive statistische Maßzahlen (Mittelwert, Standardabweichung) sowie Quartile (25., 50. 75. Perzentil) angefordert werden können.

Exakter Test. ⇨ Kap. 29.

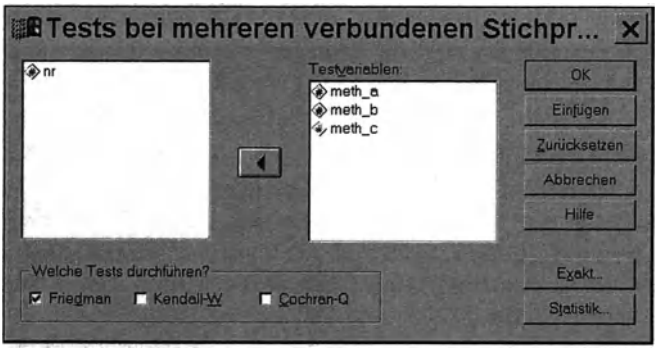


Abb. 22.14. Dialogbox „Tests bei mehreren verbundenen Stichproben"

Tabelle 22.23. Ergebnisausgabe des Friedman-Tests

Ränge

	Mittlerer Rang
Ergebnis Lehrmethode A	1,60
Ergebnis Lehrmethode B	1,85
Ergebnis Lehrmethode C	2,55

Statistik für Test[a]

N	20
Chi-Quadrat	10,347
df	2
Asymptotische Signifikanz	,006

a. Friedman-Test

22.6.2 Kendall's W-Test

Der Test ist dem von Friedman äquivalent. Er beruht im Unterschied zum Friedman-Test auf dem Maß W. Kendall's Koeffizient der Konkordanz W ist ein Maß für die Stärke des Zusammenhangs von mehr als zwei ordinalskalierten Variablen. Er misst, in welchem Maße Rangziffern für k Gruppen übereinstimmen.
Es sei angenommen, drei Lehrer bewerten die Klassenarbeit von 20 Schülern. Für die Klassenarbeiten der Schüler entsteht pro Lehrer eine Rangfolge in Form von Rangziffern, wobei für die beste Arbeit die Rangziffer 1 vergeben worden sei. Zur Bestimmung des Maßes W werden die Rangziffern für jeden Schüler aufsummiert. Aus diesen Summen wird das Ausmaß der unterschiedlichen Bewertung deutlich. Bewerten alle drei Lehrer die Arbeiten gleich, so hat der beste Schüler von allen Lehrern die Rangziffer 1, der zweitbeste die Rangziffer 2 usw. erhalten. Daraus ergeben sich die Rangziffersummen 3, 6, 9 usw. Die Unterschiedlichkeit - die Variation - der Rangziffernsummen ist demgemäß ein Maß für die Übereinstimmung der Bewertung. In Gleichung 22.7 ist das Maß W definiert. Im Zähler des Bruches steht die Variation der Rangziffernsumme in Form ihrer quadratischen Abweichung vom Mittelwert. Im Nenner steht diese Variation für den Fall völlig gleicher Bewertung: er reduziert sich dann auf den im Nenner angegebenen Ausdruck.

$$W = \frac{\sum_{j=1}^{n}(R_j - \frac{\sum_{j=1}^{n} R_j}{n})^2}{(1/12)k^2(n^3 - n)} \qquad (22.7)$$

R_j = Rangziffernsumme der Objekte oder Individuen j
k = Anzahl der Sets von Bewertungen bzw. Bewerter
n = Anzahl der bewerteten Objekte bzw. Individuen

Aus der Formel ergibt sich, dass mit der Höhe von W das Ausmaß der Übereinstimmung bei der Rangziffernvergabe wächst. W kann zwischen 0 und 1 liegen.

Für Stichprobenumfänge größer sieben ist $k(n-1)W$ annähernd chi-quadratverteilt mit $n-1$ Freiheitsgraden (Siegel, 1956, S. 236). Zur praktischen Demonstration wird der in Kap. 22.5.1 benutzte Datensatz (⇨ Abb. 22.10) in anderer Interpretation verwendet. Es soll sich bei den Variablen jetzt um Bewertungen von Schülerarbeiten durch drei Lehrer A, B und C handeln. Dafür wurden die Variablen in LEHR_A, LEHR_B und LEHR_C umbenannt (Datei LEHRER.SAV). Die Durchführung des Tests entspricht der Vorgehensweise in Kap. 22.6.1 mit dem Unterschied, dass nun „Kendall W" angeklickt wird.

In Tabelle 22.24 ist die Ergebnisausgabe des Tests zu sehen. Sie unterscheidet sich nicht von der in Tabelle 22.23, so dass auf eine Erläuterung verzichtet werden kann. Da „Signifikanz" mit 0,006 kleiner ist als das gewählte Signifikanzniveau von z.B. 5 % ($\alpha = 0{,}05$), wird die Hypothese H_0 - die Bewertungen stimmen überein - abgelehnt.

Statistiken. ⇨ Kap. 22.6.1.
Exakter Test. ⇨ Kap. 29.

Tabelle 22.24. Ergebnisausgabe des Kendall's W-Test

Ränge

	Mittlerer Rang
LEHR_A	1,60
LEHR_B	1,85
LEHR_C	2,55

Statistik für Test

N	20
Kendall-W[a]	,259
Chi-Quadrat	10,347
df	2
Asymptotische Signifikanz	,006

a. Kendalls Übereinstimmungskoeffizient

22.6.3 Cochran Q-Test

Dieser Test entspricht dem McNemar-Test mit dem Unterschied, dass er für mehr als zwei dichotome Variablen (z.B. „2" = Erfolg einer Aktivität bzw. eines Einflusses, „1" = nicht Erfolg) angewendet werden kann.

Die Prüfgröße Q wird - ausgehend von der Datenmatrix - aus den Häufigkeiten des Eintretens von „Erfolg" ermittelt. Q ist wie folgt definiert:

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} ss_j^2 - \left(\sum_{J=1}^{k} ss_j\right)^2\right]}{k\sum_{i=1}^{n} zs_i - \sum_{i=1}^{n} zs_i^2} \qquad (22.10)$$

ss_j = Spaltensumme für Variable j (Häufigkeit des Erfolges, also z.B. von „2")

zs_i = Zeilensumme für den Fall i (Häufigkeit des Erfolges, also z.B. von „2")

k = Anzahl der Stichproben (Variablen)

Q ist asymptotisch chi-quadrat-verteilt mit k – 1 Freiheitsgraden.

Das folgende Beispiel verwendet die Variablen aus dem in Abb. 22.12 ausschnittsweise dargestellten Datensatz (Datei TESTAUFG.SAV). In den Variablen AUFG1, AUFG2 und AUFG3 ist erfasst, ob drei verschiedene Aufgaben von Studenten gelöst worden sind oder nicht. Zur Anwendung des Tests geht man wie in Abschnitt 22.6.1 beschrieben vor mit dem Unterschied, dass der Cochran Q-Test angeklickt wird.

In Tabelle 22.25 wird die Ergebnisausgabe dargestellt. Für die drei Variablen werden die Häufigkeiten des Auftretens der Werte „2" (= Aufgabe gelöst) und „1" (= Aufgabe nicht gelöst) aufgelistet. Es wird die Zahl der Fälle („N"), Cochrans Q, die Zahl der Freiheitsgrade (df) sowie das Signifikanzniveau für den Test angegeben. Da das ausgegebene Signifikanzniveau mit 0,076 größer ist als ein z.B. mit 5 % (α = 0,05) gewähltes, wird die Hypothese H_0 - der Lösungserfolg und somit der Schwierigkeitsgrad der Aufgaben unterscheiden sich nicht - beibehalten.

Statistiken. ⇨ Kap. 22.6.1.
Exakter Test. ⇨ Kap. 29.

Tabelle 22.25. Ergebnisausgabe des Cochran Q-Tests

Häufigkeiten

	Wert	
	1	2
AUFG1	6	9
AUFG2	11	4
AUFG3	5	10

Statistik für Test

N	15
Cochrans Q-Test	5,167[a]
df	2
Asymptotische Signifikanz	,076

a. 2 wird als Erfolg behandelt.

23 Reliabilitätsanalyse

Das Menü „Reliabilitätsanalyse" dient zur Konstruktion und Überprüfung sogenannter *Summated Rating- oder (Likert-) Skalen*. Das sind Messinstrumente, die mehrere gleichwertige Messungen additiv zusammenfassen. Wie die Messungen entstehen, ist gleichgültig: Ob es um mehrere Fragen (Items) geht oder ob mehrere Richter dasselbe beurteilen etc. (dies sind alles Variablen). Und es ist auch gleichgültig, auf wen oder was sich die Messungen beziehen, auf Individuen, Objekte, Partikel etc. (Fälle). Der Sinn dieser Zusammenfassung mehrerer gleichwertiger Messungen besteht darin, die Zuverlässigkeit (Reliabilität) der Messung einer Variablen zu erhöhen. Bei im Prinzip nur sehr ungenau messbaren Sachverhalten (z.B. Einstellungen) ist die (ungewichtete oder gewichtete) Summe (oder der Durchschnitt) der Werte mehrerer gleichwertiger Messungen ein besserer Schätzwert des „wahren Wertes" als das Ergebnis einer einzigen Messung.

So misst man z.B. die bekannte A-Skala die Variable „Autoritarismus" dadurch, dass Probanden zu 13 Aussagen (Statements) auf einer 6-stufigen Rating-Skala Stellung beziehen. Ein solches Statement samt zugehöriger Ratingskala war folgendes:

Nicht auf Gesetz und Verfassung kommt es an, sondern einzig und allein auf den Menschen	Zustimmung			Ablehnung		
	stark	mittel	schwach	schwach	mittel	stark
	+3	+2	+1	-1	-2	-3

Die Werte dieser einzelnen Stellungnahmen werden zu einem Gesamtwert (Totalscore) aufsummiert. (Überwiegend wird als Teilmessinstrument eine solche mehrstufige Ratingskala verwendet und als Messniveau mindestens Intervallskalenniveau angenommen. Man kann aber auch vom Ordinalskalenniveau ausgehen, sollte dann aber mit rangtransformierten Daten arbeiten. Auch dichotome z.B. Ja/Nein Messungen sind möglich. Für die letztgenannten Fälle stehen einige spezielle Auswertungsvarianten zur Verfügung).

Man geht davon aus, dass sich ein gemessener Wert aus dem „wahren Wert" w und einem „Fehler" e zusammensetzt nach der Formel:

$$X = w + e$$

Dann gibt die Zuverlässigkeit (Reliabilität) an, wie genau im Durchschnitt in einer Population der beobachtete Wert dem „wahren Wert" entspricht. Der entsprechende Reliabilitätskoeffizient wird in der klassischen Form definiert als:

$R = 1 - \frac{\sigma_e^2}{\sigma_o^2}$, wobei σ_e^2 = Fehlervarianz und σ_o^2 = Varianz der beobachteten Werte.

Die „wahren" Werte sind in der Regel unbekannt, daher kann die Zuverlässigkeit eines Messinstruments auch faktisch nicht durch Vergleich der Messwerte mit den wahren Werten geprüft werden. Anstelle dessen tritt die Konsistenzzuverlässigkeit, d.h. ein Instrument gilt dann als umso zuverlässiger, je stärker die Ergebnisse der verschiedenen Teilmessungen übereinstimmen.

Summated Rating-Skalen sollen also die Zuverlässigkeit der Messung erhöhen. Und das Menü „Reliabilitätsanalyse" unterstützt deren Nutzung auf zweierlei Weise:

- *Konstruktion der Skala* durch Auswahl geeigneter Teilmessinstrumente (Items) aus einem vorläufigen Itempool.
- *Überprüfung der Zuverlässigkeit der Skala.*

23.1 Konstruktion einer Likert-Skala - Itemanalyse

Die Konstruktion einer Likert-Skala beginnt mit dem Sammeln eines Pools geeigneter Items. Die Items sollen dieselbe Variable messen und im Prinzip gleich schwer sein, d.h. denselben Mittelwert und dieselbe Streuung aufweisen. Außerdem sind trennscharfe Items von Vorteil, d.h. solche, deren Werte in der Population hinreichend streuen und bei denen extreme Fälle möglichst stark differierende Ergebnisse erbringen. Das Ganze wird häufig unter dem Begriff „Konsistenzzuverlässigkeit" zusammengefasst. Es sind aber unterschiedliche strenge Modelle der Zuverlässigkeit in Gebrauch. Im einfachsten und verbreitetsten Falle wird lediglich eine hohe Korrelation zwischen den einzelnen Items verlangt.[1]

Dies so entstandene, vorläufige sehr umfangreiche, Messinstrument wird bei einer Testpopulation angewandt. Aufgrund der dabei gewonnenen Ergebnisse werden z.T. in mehreren Schritten die geeignetesten Items des Pools für eine wesentlich kürzere Endfassung des Instruments ausgewählt. Dabei macht man sich die Tatsache zu Nutze, dass im Prinzip ein Instrument umso zuverlässiger wird, je mehr Messungen es zusammenfasst. Das Gesamtergebnis des Ausgangspools kann daher als geeigneter Prüfpunkt für die Qualität der Einzelmessungen (Items) herangezogen werden.

Beispiel. Die 13 Items der A-Skala sollen auf Basis der Ergebnisse einer Testpopulation von 32 Personen noch einmal auf ihre Qualität überprüft werden (Datei: A-SKALA-ITEMS.SAV). Die Items werden in einer Analyse auf ihre Brauchbarkeit unterzogen. Verschiedene Verfahren werden im folgenden erörtert.

Item-zu-Totalscore-Korrelation und Cronbachs Alpha. Als Hauptkriterium für die Brauchbarkeit eines Items gilt die Korrelation der Messwerte dieses Items mit denen der Gesamtmessung (Totalscore). Sie wird erfasst durch die Item-zu-Total-

[1] In Wissenschaften wie der Psychologie oder den Sozialwissenschaften sind höhere Ansprüche auch kaum zu realisieren, so auch in unserem Beispiel.

score- oder die Item-zu-Rest-Korrelation. Ähnliches erkennt man, berechnet man einen Zuverlässigkeitskoeffizienten (z.B. Cronbachs Alpha) unter Ausschluss des geprüften Items und vergleicht man dessen Wert mit dem Ergebnis unter Einschluss des Items.

Zu weiteren Prüfungen kann man die Korrelationsmatrix und deskriptive Statistiken (Mittelwerte Streuungsmaße) der Items heranziehen.

Um eine Itemanalyse durchzuführen, gehen Sie wie folgt vor:

- ▷ Laden Sie die Datei A-SKALA-ITEM.SAV.
- ▷ Wählen Sie die Befehlsfolge „Analysieren", „Skalieren" und „Reliabilitätsanalyse". Die Dialogbox „Reliabilitätsanalyse" öffnet sich (⇨ Abb. 23.1).
- ▷ Übertragen Sie die zu analysierenden Variablen (hier S2 bis S17) in das Feld „Items".
- ▷ Klicken Sie auf die Schaltfläche „Statistik". Die Dialogbox „Reliabilitätsanalyse: Statistik" öffnet sich (⇨ Abb. 23.2).
- ▷ Wählen Sie im Feld „Deskriptive Statistiken für" die Option „Skala, wenn Item gelöscht".
- ▷ Bestätigen Sie mit „Weiter" und „OK". Das Ergebnis sehen Sie in Tabelle 23.1.

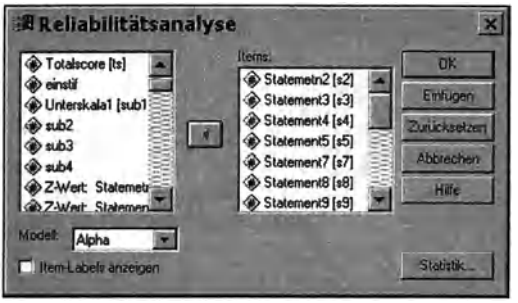


Abb. 23.1. Dialogbox „Reliabilitätsanalyse"

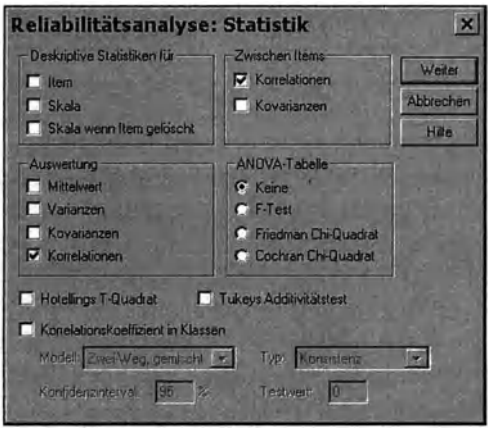


Abb. 23.2. Dialogbox „Reliabilitätsanalyse: Statistik"

Tabelle 23.1. Ausgabeergebnis einer Reliabilitätsanalyse

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
S2	26,6250	164,0484	,1366	,8599
S3	25,6875	142,0282	,4976	,8447
S4	26,2500	143,9355	,7444	,8324
S5	25,6250	141,3387	,4849	,8461
S7	25,1250	132,6290	,6182	,8362
S8	25,2813	129,3700	,7383	,8257
S9	26,5000	157,8065	,5346	,8476
S11	26,4688	146,3216	,7447	,8343
S12	26,6250	156,2419	,4855	,8474
S13	24,3125	140,9315	,4673	,8481
S14	25,9688	155,9022	,3416	,8523
S16	26,2188	144,4345	,5163	,8429
S17	25,6875	135,5121	,5910	,8381

Für die Itemanalyse sind die Spalten „Corrected Item-Total Correlation" und „Alpha if Item Deleted" am wichtigsten. Ersteres ist der Item zu Rest-Korrelationskoeffizient. Diese schwanken zwischen 0,1366 und 0,7447. Am besten sind Items mit hohem Koeffizient, also hier etwa S4, S8 und S11. Ganz schlecht ist S2 mit sehr geringem Koeffizient. Cronbachs Alpha ist eigentlich ein Koeffizient zur Beurteilung der Reliabilität der Gesamtskala. Wird allerdings der Koeffizient der Gesamtskala damit verglichen, wie er ausfiele, wenn das Statement gestrichen würde, gibt dies auch Aufschluss über die Qualität des Statements. Und zwar ist ein Statement dann besonders schlecht, wenn sich die Gesamtreliabilität verbessert. Man würde es dann streichen. In unserem Beispiel würde sich die Streichung jedes der Statements auf Alpha negativ auswirken, außer S2. Wird dieses gestrichen, verbessert sich die Gesamtreliabilität von 0,8535 auf 0,8599. Man sollte dieses Statement auf jeden Fall streichen.

Zur Prüfung weiterer Kriterien, insbesondere des Kriteriums gleicher Schwere der Items kann man verschiedene deskriptive Statistiken heranziehen.

Mittelwerte und Streuungen der Items. Betrachten wir als nächstes Mittelwerte und Streuungen der Statements.

Hierzu wählen wir im Fenster „Reliabilitätsanalyse: Statistik" im Feld „Deskriptive Statistiken für" die Option „Item" und im Feld „Auswertung" die Optionen „Mittelwert", „Varianzen". Ersteres liefert Mittelwerte und Standardabweichungen für die einzelnen Items. Letzteres dagegen den Durchschnitt der Mittelwerte und den Durchschnitt der Varianzen aller Items sowie weitere Kennzahlen wie Minimum und Maximum aller Mittelwerte und aller Varianzen. Betrachten wir erst die Durchschnittswerte. Im Durchschnitt liegt der Mittelwert der Items bei

2,1563, der Durchschnitt der Varianzen bei 2,7515. Ideal wäre es, wenn bei einer 7-stufigen Skala der Durchschnittswert 4 betrüge. Dies ist nicht so. (Die Skala ist range-restringiert.) Die Streuung sollte möglichst groß sein.

Vor allem aber sollte beides bei allen Items möglichst gleich ausfallen. In dieser Hinsicht sind die Items alles andere als ideal. Die Durchschnittswerte schwanken zwischen 1,4063 und 3,7188, die Standardabweichungen zwischen 0,5796 und 4,7974. Falls es nicht möglich ist, bessere Items zu konstruieren (was in unserem Beispiel der Fall ist), sollte man zumindest daran denken, die unterschiedliche Schwere der Items durch eine z-Transformation auszugleichen.

Korrelationen und Kovarianzen. Wählt man z.B. in der Dialogbox „Reliabilitätsanalyse: Statistik" im Feld „Zwischen Items" die Option „Korrelationen", im Feld „Auswertung" ebenfalls die Option „Korrelationen", so führt die erste Option zur Ausgabe einer Korrelationsmatrix zwischen allen Items, die zweite zur Ausgabe zusammenfassender Werte wie dem Mittelwert der Korrelationen zwischen Items und der höchsten und niedrigsten Korrelation.

Am besten inspiziert man zuerst die zusammenfassenden Angaben. Die mittlere Korrelation zwischen den Variablen ist 0,3265. Dies ist ein ausreichender Wert. Allerdings schwanken die Korrelationen zwischen –0,0086 (Minimum) und 0,7956 (Maximum). Das weckt den Verdacht, dass zumindest ein Item nicht in die Skala passt. Die nähere Inspektion der Korrelationsmatrix weist Statement 2 als problematisch aus. Es korreliert mit den anderen Statements insgesamt sehr niedrig.

23.2 Reliabilität der Gesamtskala

Reliabilität stellt eine ganze Reihe von Verfahren zur Prüfung der Qualität der Gesamtskala zur Verfügung:

- *Verschiedene Reliabilitätskoeffizienten.* Dies sind Maße, die den Grad der Korrelation der Items untereinander schätzen. Ein Wert von 1 steht für perfekte Reliabilität, von 0 für vollständig fehlende. Es existiert keine Konvention für die Höhe des Reliabilitätskoeffizienten, ab dem eine Skala als hinreichend zuverlässig angesehen wird. Mindestwerte von 0,7 oder 0,8 werden häufig empfohlen.
- *Verschiedene Arten der Varianzanalyse.* Sie dienen der Überprüfung der Frage, ob die Schwankung der Messergebnisse zwischen den Items als noch zufallsbedingt angesehen werden können.
- *Verschiedene weitere Tests.* Dienen der Überprüfung verschiedener weiterer Kriterien wie Gleichheit der Mittelwerte und Additivität der Skala.

Dabei werden unterschiedliche Aspekte der Zuverlässigkeit und Modelle mit unterschiedliche strengen Bedingungen getestet. Wir legen – wie üblich – den Schwerpunkt auf die klassischen Reliabilitätskoeffizienten.

23.2.1 Reliabilitätskoeffizienten-Modell

Die zur Berechnung der Zuverlässigkeit der Gesamtskala ausgewählten Reliabilitätskoeffizienten fordert man in der Dialogbox „Reliabilitätsanalyse" über die Auswahlliste „Modell" an. Zur Wahl stehen:

Cronbachs Alpha. Der heute gebräuchlichste Reliabilitätskoeffizient ist Cronbachs Alpha. Sie erhalten ihn, wenn im Fenster „Reliabilitätsanalyse" im Auswahlfeld „Modell" „Alpha" ausgewählt ist (Voreinstellung). Es handelt sich um eine Schätzung der Reliabilität, die auf der Korrelation *aller* Items untereinander beruht, nach der Formel:

$$\alpha = \frac{a}{a-1} \cdot \left[1 - \frac{a}{a+2b}\right]$$

a = Zahl der Items
b = die Summe der Korrelationen der Items untereinander

Im Beispiel führt dessen Anforderung zu folgender Ausgabe:

```
Reliability Coefficients

N of Cases =      32,0                    N of Items = 13

Alpha =     ,8535
```

Alpha fällt mit 0,8535 gut aus. Nach diesem Kriterium ist die Skala hinreichend zuverlässig.

Split Half. Dies ist das klassische Verfahren. Die Skala wird in zwei Hälften geteilt und die Gesamtscores der Skalenhälften werden miteinander korreliert. Wählt man dieses Modell, ergibt sich folgende Ausgabe mit mehreren Koeffizienten:

```
Reliability Coefficients

N of Cases =      32,0                    N of Items = 13

Correlation between forms =     ,7956     Equal-length Spearman-Brown =     ,8862

Guttman Split-half =            ,8742     Unequal-length Spearman-Brown =   ,8867

  7 Items in part 1                         6 Items in part 2

Alpha for part 1 =              ,7598     Alpha for part 2 =                ,6952
```

Correlation between Forms ist der Split-Half-Zuverlässigkeitskoeffizient, der die Korrelation zwischen den beiden Skalenhälften wieder gibt (= 0,7956). Da aber ja jeweils nur die Hälfte der Items in jeder Teilskala ist, unterschätzt dieser Koeffizient die Zuverlässigkeit des Gesamtinstruments. Das wird bei *Spearman-Brown* berücksichtig gemäß der Formel:

$$r_n = \frac{n \cdot r}{1 + (n-1)r}$$, wobei n = Zahl der Items

Da die Skala mit 13 Items eine ungerade Zahl von Items umfasst, ist die Variante für ungleiche Längen zuständig (Unequal-length Spearman-Brown). Der Wert be-

trägt 0,8876. *Guttmans Koeffizient* (Guttman Split-half), eine andere korrigierte Variante, beträgt 0,8742.

Außerdem ist getrennt für jede Hälfte ein Alpha berechnet (Alpha for part 1 bzw. 2). Das kann man benutzen, um die Gleichwertigkeit der Hälften zu beurteilen. Für die erste beträgt es 0,7598, für die zweite 0,6952. Der erste Hälfte ist also etwas besser gelungen.

Guttman. Dieses Modell berechnet eine Serie von 6 durch Guttman für unterschiedliche Varianten des Modells entwickelte Reliabilitätskoeffizienten. Der Koeffizient mit dem höchsten Wert gibt die Mindestreliabilität der Skala an.

```
Reliability Coefficients    13 items

  Lambda 1 =  ,7879    Lambda 2 =  ,8695    Lambda 3 =  ,8535
  Lambda 4 =  ,8742    Lambda 5 =  ,8523    Lambda 6 =  ,9258
```

In der Beispielsausgabe beträgt der höchste Wert Lambda 6 = 0,9258. Er gibt nach Guttman die wahre Reliabilität der Skala an.

Parallel. Es wird ein Modell mit bestimmten Annahmen erstellt und geprüft, ob die Daten mit diesen Annahmen übereinstimmen (goodness of fit) und zugleich ein korrigierter Reliabilitätskoeffizient berechnet. Dieses Modell beruht auf relativ strengen Annahmen der Äquivalenz, nämlich der Annahme gleicher wahrer Varianz im Set der gemessenen Fälle und gleicher Irrtumsvarianz über die verschiedenen Messungen. Diese sind im humanwissenschaftlichen Bereich nur selten zu erfüllen.

```
Test for Goodness of Fit of Model          Parallel

     Chi-square =      204,9030      Degrees of Freedom =       89
     Log of determinant of unconstrained matrix =       2,510909
     Log of determinant of constrained matrix   =      10,264374
     Probability =   ,0000

         Parameter Estimates

     Estimated common variance =       2,7515
                Error variance =       1,8999
                 True variance =        ,8516
     Estimated common inter-item correlation =     ,3095

     Estimated reliability of scale   =  ,8535
     Unbiased estimate of reliability =  ,8630
```

Entscheidend ist hier das Ergebnis eines Chi-Quadrat-Tests, der die Übereinstimmung der Daten mit den Modellannahmen prüft. Ist dies gegeben, darf kein signifikantes Ergebnis auftreten. Die Ergebnisausgabe zeigt aber, dass die Daten diesem Modell nicht gut entsprechen. Die durch den Chi-Quadrat-Test ermittelten Abweichungen sind signifikant.

Streng parallel. Es wird ein noch strengeres Modell angenommen (zusätzlich Annahme gleicher Mittelwerte der Items) und ebenfalls die Übereinstimmung getes-

tet. Der Output entspricht im Aufbau dem des parallelen Modells. Zusätzlich wird der gemeinsame Mittelwert (Common Mean) der Items geschätzt. Die Ergebnisse unterscheiden sich dagegen, insbesondere wird durch die strengeren Annahmen die Reliabilität etwas niedriger eingeschätzt.

Hinweis. Die Auswahl der Modelle wirkt sich auch auf die Ausgabe bei der Anforderung von Optionen aus den Gruppen „Deskriptiven Statistiken" und „Auswertung" aus. Beim Modell „Split-Half" werden bei Anforderung von „Auswertung", „Korrelationen" auch die entsprechenden Werte für die beiden Skalenhälften ausgegeben, ebenso beim Modell Guttman. Dasselbe gilt bei „Deskriptive Statistik" für „Skala". Bei den Modellen „Guttmann", „Parallel" und „Streng parallel" enthält die Ausgabetabelle bei der Option „Skala wenn Item gelöscht" in der letzten Spalte statt „Alpha wenn Item gelöscht", „Squared multiple Korrelation". Beim Modellen „Split-Half" werden bei Anforderung von „Auswertung", „Mittelwert" und „Varianzen" die Werte auch für die Skalenhälften ausgegeben, ebenso beim Modell Guttman.

3.2.2 Weitere Statistik-Optionen

In der Dialogbox „Reliabilitätsanalyse: Statistik" können weitere Statistiken angefordert werden.

ANOVA Tabellen.[2] Hier werden drei Arten von Tests angeboten, die dazu dienen zu prüfen, ob sich die Mittelwerte der Items signifikant unterscheiden:

- *F-Test.* Es wird eine Varianzanalyse für wiederholte Messung durchgeführt. Dies ist das klassische Vorgehen, setzt aber mindestens intervallskalierte Daten voraus. Ist dies nicht gegeben, sollte einer der beiden anderen Tests verwendet werden.
- *Friedman Chi-Quadrat.* Chi-Quadrat nach Friedman und Konkordanzkoeffizient nach Kendall. Ersteres ersetzt F für Rangdaten.
- *Cochrans Chi-Quadrat.* Für dichotomisierte Daten. Cochrans Q ersetzt in der ANOVA-Tabelle das F.
- *Hotellings T-Quadrat.* Ein weiterer multivariater Test zur Überprüfung der Hypothese, dass alle Items der Skala den gleichen Mittelwert haben.
- *Tukeys Additivitätstest.*[3]

[2] Idealerweise unterscheiden sich die Mittelwerte der Messungen nicht. Da im Basismodul von SPSS keine Varianzanalyse bei Messwiederholung angeboten wird, kann man diese Prozedur aus „Reliability" auch allgemein für diese verwenden.

[3] Eine etwas ausführliche Darstellung finden Sie auf den Internetseiten zum Buch (⇨ Anhang B).

24 Multidimensionale Skalierung

24.1 Theoretische Grundlagen

Grundkonzept. Das Verfahren der Multidimensionale Skalierung (MDS) wird in erster Linie als ein exploratives Verfahren angewendet. Analysedaten sind Ähnlichkeits- bzw. Unähnlichkeitsmaße (Distanzmaße)[1] von Paaren von Objekten (⇨ Kap. 16.3), die sich u.a. aus Urteilsbildungen von Personen ergeben. Die Aufgabe der MDS besteht darin, die Objekte als Punkte in einem möglichst niedrigdimensionalen (zwei- bzw. höchstens dreidimensionalen) Koordinatensystem (die Achsen werden Dimensionen genannt) darzustellen. Dabei sollen die Abstände zwischen den Objekten im Koordinatensystem so gut wie möglich den Ähnlichkeiten (bzw. Unähnlichkeiten) der Objekte entsprechen. Ähnliche Objektpaare sollen also nahe beieinander liegen und unähnliche einen hohen Abstand haben (⇨ Abb. 24.6 für ein Beispiel). Man interpretiert die Konstruktion einer derartigen räumlichen Darstellung von Objekten (auch Konfiguration genannt) als Abbildung des Wahrnehmungsraums von Personen. Diesem liegt die Vorstellung zugrunde, dass Personen bei Ähnlich(Unähnlich)keitsurteilen sich an nicht messbaren Kriterien (Dimensionen) orientieren. Eine MDS stellt sich die Aufgabe, diese Dimensionen aufzudecken.

Ein Beispiel aus der Marktforschung soll zur Erläuterung und für die praktische Anwendung dienen. Ausgangsdaten seinen Urteile von einzelnen Verbrauchern (Versuchspersonen) über die Ähnlichkeitseinschätzung von 11 Zahncrememarken (Objekten). Die Daten können dadurch gewonnen werden, dass einige Verbraucher für jede Kombination von Markenpaaren eine Einschätzung der Ähnlichkeit der Marken aus ihrer persönlichen Verbrauchersicht abgeben und auf einer Ratingskala mit den Werten von z.B. 1 bis 9 einordnen (Ratingverfahren). Dabei soll 1 eine sehr hohe und 9 eine sehr schwache Ähnlichkeit bedeuten (bzw. 1 sowie 9 bedeuten eine sehr kleine bzw. sehr hohe Unähnlichkeit bzw. Distanz). Auf diese Weise erhält man für jeden der Verbraucher eine Matrix von Unähnlichkeitsmaßen (Distanzmaßen) für alle Markenpaarkombinationen. Bei 11 Marken muss jeder der Verbraucher insgesamt 55 Ähnlichkeitsurteile fällen.[2]

---

[1] Man nennt derartige Maße auch Proximitäten. Dazu gehören auch Korrelationskoeffizienten.

[2] Bei einer alternativen Datenerhebungsmethode werden die 55 Paarvergleiche nach der Ähnlichkeit in eine Rangordung von 1 bis 55 gebracht (1 = am ähnlichsten, 55 = am unähnlichsten). Das Markenpaar mit dem Rangplatz 1 erhält die Kodierung 1 und das Markenpaar mit dem Rangplatz 55 die Kodierung 55. Beide Formen der Datenerhebung führen zu einer quadratischen und symmetrischen Datenmatrix. Die Vorgehensweise bei einer MDS mit SPSS unterscheidet sich daher nicht. Werden die Daten mit einer weiteren Erhebungsmethode, dem Ankerpunktverfahren, erho-

Bei diesen Formen der Messung von Ähnlichlichkeitsbeziehungen zwischen Objekten entstehen ordinalskalierte Daten (die Messwerte bilden eine Rangordnung, die Abstände der Messwerte haben keine Aussagekraft). Zur Auswertung derartiger Daten wird eine nichtmetrische (ordinale) MDS herangezogen.

In Abb. 24.1 ist die quadratische und symmetrische (daher sind nur Werte unterhalb der Diagonalen eingetragen) Matrix mit den Unähnlichkeitsmesswerten für die Markenpaare als Datenmatrix von SPSS für Windows zu sehen.[3] Der Messwert von z.B 8,5 für das Markenpaar Meridol und Signal bedeutet, dass die Marken sich wenig ähnlich sind. Der Messwert 2,7 für das Markenpaar Signal und Colgate hingegen weist aus, dass die Marken als ähnlich eingeschätzt worden sind.

Die Messdaten sind für das einfachste Modell der MDS aufbereitet: Die in jeder Zelle der Matrix enthaltenen Distanzmaße sind Mittelwerte der für die einzelnen Verbraucher gewonnenen Unähnlichkeitsdaten.

Um die Zielrichtung einer MDS zu konkretisieren, werfen wir nun einen Blick auf die SPSS-Ergebnisausgabe (die Lösungskonfiguration einer nichtmetrischen MDS) in Abb. 24.6. Die von den Verbrauchern eingeschätzten Ähnlichkeiten der Objekte sind durch ihre Abstände in einem zweidimensionalen Koordinatensystem mit den Achsen Dimension 1 und Dimension 2 dargestellt. Da es sich hier um ordinalskalierte Daten handelt, ist die Rangordnung der Abstände in der Grafik so gut wie möglich der Rangordnung der Ähnlichkeitsmesswerte in der Datenmatrix angepasst worden. Wenn die Anpassung der Abstände an die Daten gut gelingt, wird das Beziehungsgeflecht der Ähnlichkeiten (Unähnlichkeiten) der Objekte durch die räumliche Darstellung leichter überschaubar und im Sinne der Abbildung eines Wahrnehmungsraumes interpretierbar. Im Koordinatensystem nahe beieinander liegende Marken (z.B. Signal und Colgate) zeigen deren hohe Ähnlichkeit (und damit auch Austauschbarkeit) und voneinander entfernt liegende Marken (z.B. Signal und Meridol) zeigen das Ausmaß ihrer Unähnlichkeit aus Verbrauchersicht.

Gütemaße. Wir bezeichnen im folgenden mit ∂_{ij} die Unähnlichkeitsmaße in der Datenmatrix und mit d_{ij} die Abstände (Distanzen)[4] der Objekte in der Konfiguration für die Objektpaare i und j. Bei der Ermittlung einer Lösungskonfiguration (d.h. bei einer Anpassung der Rangordnung von d_{ij} an die von ∂_{ij} durch Verschieben der Punkte) werden die Distanzen d_{ij} tatsächlich nicht direkt an die Unähnlichkeitsmaße (Distanzen) ∂_{ij} angepasst, sondern an eine Hilfsvariable $\hat{d}_{ij}$ (sie wird Disparität genannt). Da für alle Objektpaare i und j die Werte von $\hat{d}_{ij}$ die

---

ben, so entsteht eine asymmetrische Datenmatrix, die eine spezielle Vorgehensweise bei einer MDS mit SPSS erfordert (⇨ Kap. 24.2.2).

[3] SPSS verlangt in der Datenmatrix für die MDS Unähnlichkeitsmaße (Distanzmaße). Enthält die Datenmatrix Ähnlichkeitsmaße oder Merkmalsvariable der Objekte, so müssen diese vorher in Distanzmaße transformiert werden. Merkmalsvariable können in der Dialogbox „Multidimensionale Skalierung" („Distanzen aus Daten erzeugen") oder auch ebenso wie Ähnlichkeitsmaße im Menü Distanzen (⇨ Kap. 16.3) transformiert werden. Ähnlichkeitsmaße können auch per Syntax transformiert werden.

[4] Sie werden als Euklidische Distanzen berechnet (⇨ Kap16.3).

gleiche Rangordnung bekommen[5] wie die Unähnlichkeitsmesswerte ∂_{ij}, entspricht eine Anpassung der Abstände von d_{ij} an die Disparitätswerte $\hat{d}_{ij}$ einer ordinalen Anpassung an ∂_{ij}. Die Werte von $\hat{d}_{ij}$ werden im Prozess des Lösungsverfahren der MDS außerdem so festgelegt, dass die Abweichungen von d_{ij} so klein wie möglich sind. Abweichungen von d_{ij} von $\hat{d}_{ij}$ sind Ausdruck einer mangelnden Anpassung der Distanzen in der Konfiguration an die Unähnlichkeitsmaße. Darauf basieren die Stressmaße[6], die die Güte der Lösungskonfiguration (gemessen an der perfekten Lösung) messen. Das Stressmaß zur Beurteilung der Anpassungsgüte einer Konfiguration nach Kruskal ist wie folgt definiert:

$$S = \frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2} \qquad (1)$$

Je größer die Abweichungen zwischen d_{ij} und $\hat{d}_{ij}$ sind (d.h. je schlechter die Anpassung der Lösungskonfiguration an die Unähnlichkeitsdaten ist), umso höher wird das Stressmaß. Der Ausdruck im Nenner dient der Normierung. Für den Fall $d_{ij} = \hat{d}_{ij}$ für alle Objektpaare wird S gleich Null (eine perfekte Lösung).

Von Kruskal sind Stresswertbereiche für S gemäß Tabelle 24.1 zur Beurteilung der Güte von MDS-Lösungen als Richtlinie vorgeschlagen worden. Diese sollte man nur als Anhaltspunkte zu Rate ziehen, da S auch von der Anzahl der Objekte und der Anzahl der Dimensionen abhängig ist.

Tabelle 24.1. Stresswertbereiche zur Gütebeurteilung einer MD-Lösung

$S \geq 0{,}2$	Schlechte Übereinstimmung
$0{,}2 \geq S \geq 0{,}1$	Befriedigende Übereinstimung
$0{,}1 \geq S \geq 0{,}05$	Gute Übereinstimmung
$0{,}05 \geq S \geq 0{,}025$	Hervorragende Übereinstimmung
$0{,}025 \geq S \geq 0{,}00$	Perfekte Übereinstimmung

Ein weiteres Gütemaß ist RSQ (entspricht R^2 in der Regressionsanalyse). Dieses wird unten bei der Erläuterung der Ausgabe des Anwendungsbeispiels erklärt.

Festlegen der Anzahl der Dimensionen. Bevor der Anpassungsprozess im Lösungsverfahren der MDS beginnt, muss die Anzahl der Dimensionen der Konfiguration durch den Anwender bestimmt werden. Es ist klar, dass sich mit einer höheren Anzahl von Dimensionen eine bessere Anpassung der Abstände in der

[5] In der Sprache der Mathematik: $\hat{d}_{ij}$ ergibt sich durch monotone Transformation (d.h eine Transformation ohne Änderung der Reihenfolge der Werte) von ∂_{ij}. Bei metrischen Daten werden lineare Transformationen vorgenommen: $\hat{d}_{ij} = a + b\,\partial_{ij}$ bei intervall- und $d_{ij} = b\,\partial_{ij}$ bei rationalskalierten Daten.

[6] Stress = Standardized residual sum of squares.

Lösungskonfiguration an die Unähnlichkeitsmaße erreichen lässt (d.h. mit einer höheren Dimension wird der Stresswert kleiner). Es soll aber unter der Nebenbedingung einer möglichst guten Anpassung die kleinstmögliche Anzahl von Dimensionen gewählt werden. Man sollte daher bei der praktischen Arbeit die MDS mit unterschiedlicher Anzahl von Dimensionen durchführen, um die beste Lösung zu bekommen.

Das Lösungsverfahren. Das Stressmaß (im Programm SPSS allerdings ein leicht modifiziertes nach Young, S-Stress genannt) dient auch dazu, ausgehend von einer festgelegten Anzahl von Dimensionen und einer ausgewählten Startkonfiguration (d.h. einer Anfangsverteilung der Objekte im Lösungsraum), sich der perfekten Lösung in iterativen Berechnungsschritten zu nähern. Dabei werden mit einem hier nicht erläuterten Algorithmus [s. Backes u.a. (2001)] in einem iterativen Optimierungsprozess die Objekte Schritt für Schritt im Konfigurationsraum verschoben, um die Abstände in der Konfiguration den Unähnlichkeitsmaßen anzupassen. In dem Lösungsverfahren wird der Stresswert also Schritt für Schritt verkleinert (minimiert).

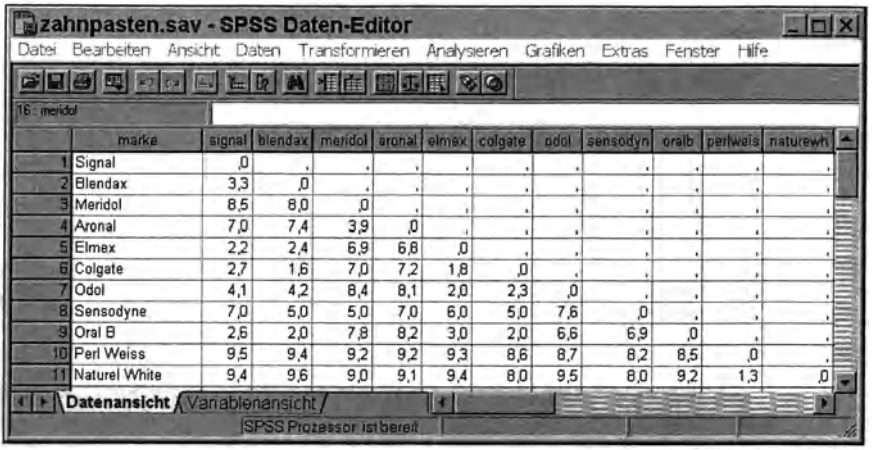
Unterschiedliche Modelle der MDS. In den meisten Anwendungen wird - wie oben dargelegt – eine (quadratische und symmetrische) Matrix von Unähnlichkeitsmaßen bzw. Distanzen von Objektpaaren analysiert. Beruhen die Unähnlichkeitsdaten auf Befragungen mehrerer Personen, so werden Durchschnitte gebildet zur Herstellung der zu analysierenden Matrix. Dabei wird unterstellt, dass die Messwerte der verschieden Personen vergleichbar sind. Je nach Datenlage kann eine nichtmetrische oder metrische MDS angewendet werden. Die Praxis der MDS hat gezeigt, dass sich metrische und nichtmetrische MDS-Lösungen angewendet auf metrische Daten kaum unterscheiden. Daher wird vorwiegend die nichtmetrische MDS eingesetzt.

SPSS kann aber auch Modellvarianten bearbeiten. Diese sollen mit ihren spezifischen Datenkonstellationen und ihren Besonderheiten und Annahmen in Kap. 24.2.2 im Zusammenhang mit der SPSS-Anwendung nur kurz behandelt werden.

24.2 Praktische Anwendung

24.2.1 Ein Beispiel einer nichtmetrischen MDS

In Abb. 24.1 sind die durchschnittlichen Unähnlichkeitsmaßzahlen von Befragten zur vergleichenden Bewertung von 11 Zahncrememarken als Analysedatenmatrix von SPSS für Windows zu sehen. Insgesamt gibt es 55 Ähnlichkeitsurteile, da jeweils immer Paare von Zahncremen mit einer Ratingskala von 1 bis 9 bewertet werden. Nur die Zellen unterhalb der Diagonale enthalten Werte, da es sich um eine quadratische und symmetrische Datenmatrix handelt. Die Diagonalwerte haben den Wert 0 (alternativ könnten diese auch einen fehlenden Wert anzeigen).
Bevor Sie das Verfahren MDS starten, sollten Sie sicherstellen., dass im Menü „Optionen" (Aufruf durch die Befehlsfolge „Bearbeiten, „Optionen") im Register „Allgemein" für die Variablenliste „Datei" eingeschaltet ist. Mit der Einstellung „Alphabetisch" werden die Ergebnisse falsch.


zahnpasten.sav - SPSS Daten-Editor

Datei Bearbeiten Ansicht Daten Transformieren Analysieren Grafiken Extras Fenster Hilfe

16 : meridol

	marke	signal	blendax	meridol	aronal	elmex	colgate	odol	sensodyn	oralb	perlweis	naturewh
1	Signal	,0	.	.	.	.	.	.	.	.	.	.
2	Blendax	3,3	,0	.	.	.	.	.	.	.	.	.
3	Meridol	8,5	8,0	,0	.	.	.	.	.	.	.	.
4	Aronal	7,0	7,4	3,9	,0	.	.	.	.	.	.	.
5	Elmex	2,2	2,4	6,9	6,8	,0	.	.	.	.	.	.
6	Colgate	2,7	1,6	7,0	7,2	1,8	,0	.	.	.	.	.
7	Odol	4,1	4,2	8,4	8,1	2,0	2,3	,0	.	.	.	.
8	Sensodyne	7,0	5,0	5,0	7,0	6,0	5,0	7,6	,0	.	.	.
9	Oral B	2,6	2,0	7,8	8,2	3,0	2,0	6,6	6,9	,0	.	.
10	Perl Weiss	9,5	9,4	9,2	9,2	9,3	8,6	8,7	8,2	8,5	,0	.
11	Naturel White	9,4	9,6	9,0	9,1	9,4	8,0	9,5	8,0	9,2	1,3	,0

Datenansicht | Variablenansicht

SPSS Prozessor ist bereit

Abb. 24.1. Matrix der Unähnlichkeitsdaten in der Datenansicht von SPSS für Windows

Zur Durchführung der MDS gehen Sie nach Laden der Datei ZAHNPASTEN.SAV wie folgt vor:

- ▷ Wählen Sie per Mausklick die Befehlsfolge "Analysieren", "Skalieren“ und „Multidimensionale Skalierung“. Es öffnet sich die in Abb. 24.2 dargestellte Dialogbox.
- ▷ Übertragen Sie die Variablen aus der Quellvariablenliste in das Feld "Variablen" (hier die Zahncrememarken). Achten Sie darauf, dass alle Variablen in der gleichen Reihenfolge wie in der Analysedatenmatrix in das Feld „Variablen“ übertragen werden, da sonst falsche Lösungen entstehen.
- ▷ Im Feld „Distanzen“ sind die Optionen „Daten sind Distanzen“ und die „Form“ (der Analysedatenmatrix) „quadratisch und symmetrisch“ voreingestellt und werden so belassen.
- ▷ Klicken der Schaltfläche „Modell“ öffnet die in Abb. 24.4 dargestellte Unterdialogbox. Im Feld „Messniveau“ ist die gewünschte Option „Ordinalskala“ voreingestellt. Für den Fall, dass gleiche Werte (Bindungen bzw. ties) in der Datenmatrix enthalten sind, kann man die Option „gebundene Beobachtungen lösen“ wählen (man nimmt dann an, dass die Werte Intervalle repräsentieren). Als „Skalierungsmodell“ ist die hier passende Option „Euklidischer Abstand“ schon voreingestellt. Für „Konditionalität“ (der zu analysierenden Datenmatrix) ist mit „Matrix“ die hier richtige Auswahl ebenfalls voreingestellt. Dabei geht es um die Frage, welche Werte in der Datenmatrix vergleichbar sind. Auch für die Anzahl der „Dimensionen“ wird die Voreinstellung „Minimum 2“ sowie „Maximum 2“ übernommen. Mit Klicken von „Weiter“ kommt man zur Dialogbox zurück.
- ▷ Klicken auf die Schaltfläche „Optionen“ öffnet die in Abb. 24.5 dargestellte Unterdialogbox. Wir wählen in „Anzeigen“ die Option „Gruppendiagramme“. Im Feld „Kriterien“ sind Voreinstellungen für den in iterativen Schritten ablaufenden Prozess der MDS-Lösungsfindung zu sehen. „S-Stress-Konvergenz“ besagt, dass das iterative Lösungsverfahren abgebrochen wird, wenn die Verrin-

gerung des S-Stresswertes nach Young kleiner wird als 0,001. Auch wenn der S-Stresswert kleiner als 0,005 wird, stoppt das Berechnungsverfahren. Schließlich sind maximal 30 Iterationsschritte voreingestellt. Diese Vorgaben kann man durch Überschreiben ändern. Wir übernehmen die Voreinstellungen. Mit „Weiter" kommt man zur Dialogbox zurück und mit „OK" startet man die MDS.

Abb. 24.2. Dialogbox „Multidimensionale Skalierung"

Wahlmöglichkeiten

① *Distanzen.*

- *Daten sind Distanzen.* Die Schaltfläche „Form" öffnet eine Unterdialogbox (⇨ Abb. 24.3) zur Angabe der Form der zu analysierenden Datenmatrix. „Quadratisch und symmetrisch" entspricht unserem Beispiel. Die anderen Optionen („Quadratisch und asymmetrisch" sowie „Rechteckig") sind für Modellvarianten der MDS bzw. für andere Datenmatrizen relevant (⇨ Kap. 24.2.2).

Abb. 24.3. Dialogbox „Multidimensionale Skalierung: Form"

- *Distanzen aus Daten erzeugen.* Die Schaltfläche „Maß" öffnet eine Unterdialogbox zum Transformieren von Variablen. Wenn in der Datenmatrix Eigenschaftsvariablen von Objekten vorliegen (diese können metrisch, binär oder auch Häufigkeitsdaten sein), so können diese hier in Distanzen (d.h. Unähnlichkeitswerte) transformiert werden. Da das Menü „Distanzen" auch diese Möglichkeit bietet, sei auf Kapitel 16.3 verwiesen.

② *Individuelle Matrizen für:*. Falls Eigenschaftsvariable in Distanzen transformiert werden sollen, kann man hier eine Variable zur Gruppenidentifizierung übertragen. Für jede Gruppe wird eine Distanzmatrix berechnet.

③ *Schaltfläche Modell.* Sie öffnet die in Abb. 24.4 dargestellte Unterdialogbox.

- Messniveau. Neben ordinalskalierten Variablen können auch intervall- oder rationalskalierte Daten analysiert werden. Wenn viele gleiche Werte (ties) in der Datenmatrix vorkommen, sollte man die Option „Gebundene Beobachtungen lösen" wählen.
- *Skalierungsmodell.* „Euklidischer Abstand" ist die Standardeinstellung. Die Entfernung von zwei Punkten im Koordinatensystem der Konfiguration berechnet sich als Euklidischer Abstand (⇨ Kap. 16.3). Die Optionen „Euklidischer Abstand mit individuellen gewichteten Differenzen" ist nur für das Modell INDSCAL relevant (⇨ Kap. 23.2.2).
- *Konditionalität.* Hier geht es um die Vergleichbarkeit der in der Matrix stehenden Unähnlichkeitsmaße (bzw. Rangziffern). Bei „Matrix" sind alle Werte einer Matrix vergleichbar (nicht aber die verschiedener Matrizen). Bei „Zeile" nur die Werte einer Zeile. Die Option „Zeile" wird bei asymmetrischen Matrizen gewählt (⇨ Kap. 23.2.2). Die Option „Unkonditional" ist zu wählen, wenn bei Messwiederholungen die Distanzen aller Matrizen vergleichbar sind.
- *Dimensionen.* Man wählt hier die Anzahl der Achsen der Konfigurationslösung.

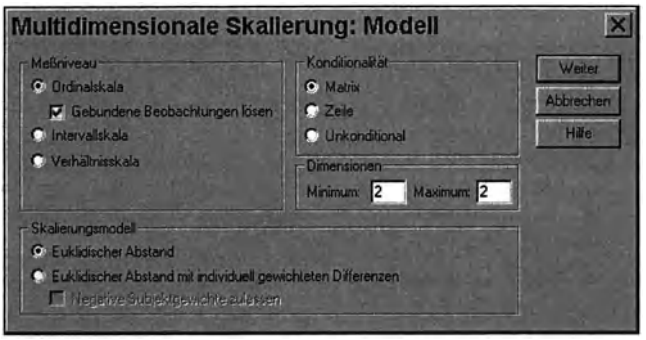


Abb. 24.4. Dialogbox „Multidimensionale Skalierung: Modell"

④ *Schaltfläche Optionen.* Sie öffnet die Unterdialogbox in Abb. 24.5.

- *Anzeigen.* Man kann aus den angebotenen Optionen für die Ergebnisausgabe auswählen. „Gruppendiagramme" sollte man immer wählen.
- *Kriterien.* Hier kann man wählen, wann der Iterationsprozess abgebrochen werden soll.
- *Distanzen kleiner als fehlend behandeln.* Als Standardeinstellung ist die Ziffer 0 als fehlender Wert eingetragen. Man kann diese Voreinstellung überschreiben.

Abb. 24.5. Dialogbox „Multidimensionale Skalierung: Optionen"

In Tabelle 24.2 ist ein erster Teil der Ergebnisausgabe zu sehen. Nach sieben Iterationsschritten wird das Optimierungsverfahren zur Erzielung einer MDS-Konfiguration abgebrochen, da die Verringerung des Stresswertes nach Young kleiner als der voreingestellte Grenzwert von 0,005 ist. Für die MDS-Lösung wird ein Stresswert nach Young in Höhe von 0,08093 erzielt. Der Stresswert nach Kruskal beträgt 0,0992. Gemäß der Güterichtlinien in Tabelle 24.1 wird damit eine gute Anpassung der Distanzen in der Konfiguration an die Ähnlichkeitsmaße erreicht.

Als ein weiteres Gütemaß wird RSQ = 0,9626 (entspricht R^2 in der Regressionsanalyse) ausgeben, das die gute Anpassung bestätigt. Es handelt sich dabei um das Quadrat des Korrelationskoeffizienten (⇨ Kap. 16) zwischen den Disparitäten $\hat{d}_{ij}$ und den (euklidischen) Distanzen d_{ij}. Damit wird ausgewiesen, dass in der Lösungskonfiguration 96,26 % der Variation von d_{ij} der Variation der Unähnlichkeitsmaße ∂_{ij} entsprechen. Anschließend werden für jede Marke die Koordinaten (diese sind z-transformiert) im zweidimensionalen Lösungsraum aufgeführt..

In Abb. 24.6 ist die Lösungskonfiguration zu sehen. Die Grafik ist gestaucht dargestellt: eine Einheit auf der waagerechten Achse in cm gemessen entspricht nicht einer Einheit auf der senkrechten Achse. Dadurch sind auch die Abstände zwischen den Marken verzerrt dargestellt. Durch Kopieren der Grafik in Word und verändern der Höhe der Grafik im Vergleich zur Breite kann man dieses korrigieren.

Die Konfiguration zeigt, wie eine Marke im Vergleich zu anderen wahrgenommen wird. Die Abstände zwischen den Marken zeigen die Ähnlichkeit der Marken aus der Sicht der Verbraucher. Kleine Abstände weisen eine hohe Ähnlichkeit und damit Austauschbarkeit aus Verbrauchersicht aus. Die Marken Odol, Oral B, Signal, Colgate und Blendax liegen alle relativ eng beieinander in einem Cluster und werden als ähnlich eingeschätzt. Weitere Cluster bilden einerseits die Marken Aronal und Meridol (Marken mit zahnmedizinischem Anspruch) und die Marken Perlweiss und Nature White (Zahnweißwirkung, Entfernen von Raucherbelag). Die Marken eines Clusters haben hohe Abstände zu den Marken eines anderen Clusters und zeigen damit, dass die Verbraucher unterschiedliche Pro-

duktprofile bei der Ähnlichkeitseinschätzung sehen. Die Marke Sensodyne liegt etwa zwischen den Markencluster mit Blendax und anderen Marken sowie dem Cluster mit den Marken Meridol und Aronal.

Tabelle 24.2. Ergebnisausgabe: Iterationsschritte und Gütemaße für die MDS

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Iteration	S-stress	Improvement
1	,13800	
2	,09419	,04382
3	,08768	,00651
4	,08492	,00276
5	,08316	,00176
6	,08190	,00125
7	,08093	,00097

Iterations stopped because S-stress improvement is less than ,001000

Stress and squared correlation (RSQ) in distances
RSQ values are the proportion of variance of the scaled data (disparities)
in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances. Stress values are Kruskal's stress formula 1.For matrix

Stress = ,09920 RSQ = ,96266

Configuration derived in 2 dimensions

Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension 1	2
1	SIGNAL	1,0159	,3559
2	BLENDAX	,9673	,2110
3	MERIDOL	-,3459	-1,6012
4	ARONAL	-,2328	-1,5663
5	ELMEX	,7710	-,3100
6	COLGATE	,4848	,3349
7	ODOL	,9961	,7020
8	SENSODYN	-,1735	-,6141
9	ORALB	,7597	,6382
10	PERLWEIS	-2,0684	1,0662
11	NATUREWH	-2,1742	,7833

In der Regel wird man versuchen, die gefundenen Dimensionen im Sinne des Wahrnehmungsraumes von Verbrauchern zu interpretieren. Da eine MDS-Lösung nur die relative Lage der Marken zueinander im Lösungsraum bestimmt, ist eine Drehung der Achsen zulässig[7] und zur Erleichterung der Interpretation häufig hilfreich. Man wird die Achsen für eine Interpretation so drehen, dass vom Koordinatenschnittpunkt am weitesten entfernt liegende Objekte bzw. Objektcluster auf

[7] Dieses ist bedingt durch die Darstellung von Euklidischen Distanzen in der Konfiguration.

bzw. nahe an den Achsen liegen. Nun kann man versuchen aus der Kenntnis der Objekte (bzw. Objektcluster), den Achsen eine Bedeutung zu geben. Hier soll zur Demonstration eine Interpretation versucht werden.

Da die Marken Meridol und Aronal auch eine zahnmedizinische Wirkung (Gesunderhaltung des Mundraumes, Schutz vor ungesunden Bakterien) versprechen und auch die Marke Sensodyne eine zahngesunderhaltende Wirkung verspricht (Schutz vor Schmerzgefühl an den Zähnen), könnte man die senkrechte (bzw. um ca. 45 Grad nach rechts gedrehte) Achse als „Gesunderhaltung" im Wahrnehmungsraum von Verbrauchern deuten. Aronal und Meridol werden relativ stark mit einer medizinischen Wirkung im Vergleich zu den anderen Marken wahrgenommen. Die Marken Perlweiss und Nature White stehen für eine stark reinigende Wirkung, so dass es nahe liegt, die waagerechte (bzw. um ca. 45 Grad nach rechts gedrehte) Achse als „Reinigungskraft" im Wahrnehmungsraum zu deuten. Der Reinigungseffekt dieser Marken wird sehr stark wahrgenommen im Vergleich zu den Marken im Cluster mit Blendax und den anderen Marken. Perlweiss und Nature White haben auf der senkrechten Achse den höchsten Abstand zu Aronal und Meridol. Die Marken Perlweiss und Nature werden aus der Wahrnehmung „Gesunderhaltung" am wenigsten gut eingeschätzt. Dieses könnte man damit erklären, dass die stark reinigende Wirkung auch ein gewisses Risiko birgt, den Zahnschmelz zu schädigen.

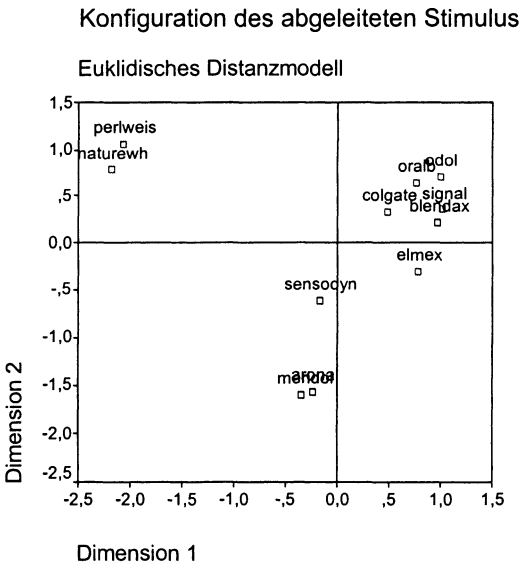


Abb. 24.6. MDS-Lösungskonfiguration für 11 Zahncrememarken

Standardmäßig werden einige ergänzende Grafiken erzeugt. Das hier nicht aufgeführte Diagramm mit der Überschrift „Streudiagramm mit linearer Anpassung" ist hier irrelevant, da ordinale Daten analysiert werden und daher eine nichtlineare (nämlich monotone) Transformation der Daten erfolgt (⇨ Fußnote 5). Es soll aber darauf hingewiesen werden, dass die waagerechte Achse falsch beschriftet ist

(anstelle „Disparitäten" muss wie bei der Grafik „Streudiagramm mit nichtlinearer Anpassung" „Beobachtungen" stehen).

In den Streudiagrammen in Abb. 24.7 sind auf den waagerechten Achsen die Unähnlichkeitsmaße (Beobachtungen genannt) und auf der senkrechten die Distanzen (linke Grafik) sowie die Disparitäten (rechte Grafik) dargestellt. Jeder Punkt (insgesamt 55) in einem Diagramm ist ein Objektpaar. Verbindet man die Punkte in der rechten Grafik, so entsteht eine monoton ansteigende Kurve, da die Rangordung der Werte beider Variablen sich entsprechen (in der SPSS-Ausgabe heißt die Grafik „Transformations-Streudiagramm", üblich ist der Name Shephard-Diagramm). Für die Grafik auf der linken Seite stellt man sich am besten vor, dass die rechte Grafik sie überlagert. Dann kann man erkennen, in welchem Maße es (senkrechte) Abweichungen zwischen der tatsächlichen und der gewünschten perfekten Anpassung der Konfiguration an die Daten gibt.

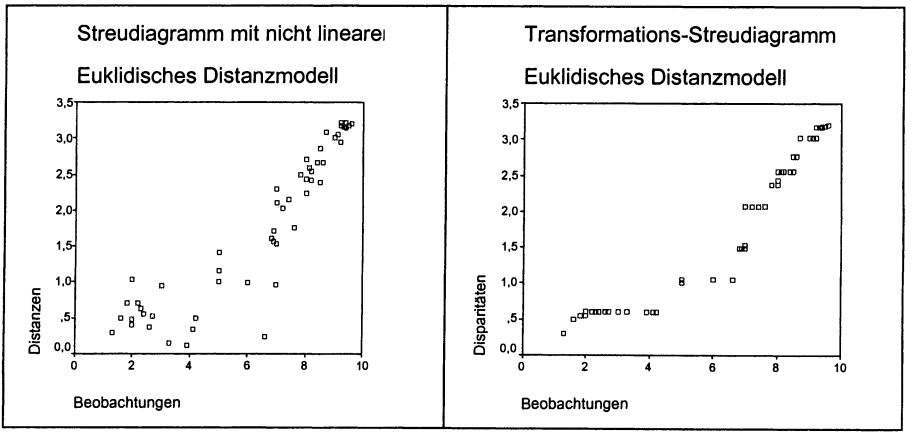


Abb. 24.7. Ergänzende Grafiken zur MDS

24.2.2 MDS bei Datenmatrix- und Modellvarianten

MDS bei einer durch die Ankerpunktmethode entstandenen Datenmatrix.
In unserem obigen Anwendungsbeispiel ist die Datenmatrix durch das Ratingverfahren entstanden. Die Urteiler vergleichen alle Objektpaare und stufen die Ähnlichkeit auf einer Skala (z.B von 1 bis 9) ein.

Bei der Messung von Ähnlichkeiten von Objekten mit dem Ankerpunktverfahren geht man bei der Messung von Ähnlichkeiten (Unähnlichkeitsdaten) anders vor. Jedes Objekt dient bei Paarvergleichen jeweils als Vergleichsobjekt. Unser Anwendungsbeispiel soll dazu dienen, das Verfahren näher zu erläutern. Nimmt man z.B. die Marke Signal als ersten Ankerpunkt, so wird diese Marke mit den anderen 10 verglichen und der Grad der Ähnlichkeit (Unähnlichkeit) in eine Rangfolge gebracht, um Rangplätze zu vergeben (Rang 1 = am ähnlichsten, Rang 10 = am unähnlichsten). Die Rangziffern sind die Kodierungswerte. Nächster Ankerpunkt wäre dann Blendax. Nun vergleicht man Blendax mit den anderen Marken und vergibt wieder Rangplätze von 1 bis 10 usw. Auf diese Weise entsteht für jede Urteilsperson eine 11*11-Matrix (mit 0 in der Diagonalen). Bei diesem Bewer-

tungsverfahren wird in der Regel eine asymmetrische Matrix entstehen, da bei dem Vergleich von Ankerpunkt-Marke i mit Marke j und bei dem Vergleich von Ankerpunkt-Marke j mit Marke i sich unterschiedliche Rangplätze ergeben können. Zudem handelt es sich um eine konditionale Matrix bei der nur die Werte eines Ankerpunkts (d.h. jeweils die einer Zeile) vergleichbar sind. Wenn z.B. 20 Personen an diesem Verfahren der Datenerhebung beteiligt sind, so werden im Dateneditor die 20 11*11-Matrizen nacheinander ohne Leerzeile eingegeben. In der Diagonalen kann der Wert 0 eingetragen werden.

Bei der Auswertung einer derartigen Datenmatrix mit SPSS wählt man in der Unterdialogbox „Multidimensionale Skalierung: Form" (⇨ Abb. 24.3) die Option „quadratisch und asymmetrisch" und in der Dialogbox „Multidimensionale Skalierung: Modell" (⇨ Abb. 24.4) für „Konditionalität" „Zeile".

MDS bei Messwiederholungen (RMDS).
Als Daten werden ebenfalls Unähnlichkeitsmaße untersucht. Im Unterschied zur einfachen MDS werden die Messdaten einzelner Personen aber nicht durch Durchschnittsbildung aggregiert, sondern alle Datenmatrizen von Ähnlichkeitsurteilern sind Datengrundlage der MDS. Es wird dabei unterstellt, dass die Ähnlichkeitsmaße verschiedener Personen vergleichbar sind (gleicher Wahrnehmungsraum). Die Datenmatrizen werden in der SPSS-Datenmatrix hintereinander eingegeben. In unserem Beispiel nähme die erste Matrix wie bei der einfachen MDS die Zeilen 1 bis 11 ein. In Zeile 12 bis 22 schließt sich die 2. Matrix an usw. Die Optionen für „Form" und „Modell" sind die gleichen wie bei der einfachen MDS (für den Fall, dass die Distanzen aller Matrizen vergleichbar sind, ist die Option „Unkonditional" zu wählen). SPSS erkennt die neue Datenkonstellation. Es wird mit den Standardeinstellungen eine Konfiguration erstellt. Im Unterschied zur einfachen MDS werden zu jeder Matrix Stresswerte ausgegeben. Es kann nun geprüft werden, ob die unterschiedlichen Stresswerte mit der Annahme einer Konfiguration (eines Wahrnehmungsraumes) konpatibel ist.

MDS bei Messwiederholungen bei individueller Gewichtung (INDSCAL).
Diese Modellvariante der MDS gestattet es, analog der MDS mit Messwiederholungen mehrere individuelle Matrizen von Unähnlichkeitsmaßen (bzw. Distanzen) zu analysieren (die Anordnung der Daten im Dateneditor ist wie im Fall von RMDS). Dabei wird unterstellt, dass die unterschiedlichen Ähnlichkeitsurteilsbildungen einzelner Personen zwar aus einem allen gemeinsamen Wahrnehmungsraum kommen, aber durch unterschiedliche individuelle Gewichtungen der Dimensionen entstehen. Wenn z.B. alle Urteilspersonen einen gemeinsamen Wahrnehmungsraum für Automobile haben (mit den Dimensionen Wirtschaftlichkeit und Prestige), so ist vorstellbar, dass die individuelle Gewichtung der einzelnen Dimensionen bei der Urteilsbildung verschieden ist. Bei dieser MDS wird einerseits aus den Daten eine gemeinsame Konfiguration wie im Fall einer einfachen MDS erzeugt. Andererseits werden auch für jeden einzelnen Urteilsgeber individuelle Konfigurationen erstellt. Die individuellen Konfigurationen ergeben sich aber im Unterschied zum Modell bei Messwiederholungen durch eine individuelle Gewichtung der Achsen der gemeinsamen Konfiguration. Eine individuelle Konfiguration entsteht durch Dehnung bzw. Stauchung der Achsen der gemeinsamen

Konfiguration (durch Multiplikation der Koordinaten mit individuellen Gewichten).

In der Dialogbox „Multidimensionale Skalierung: Modell" (⇨ Abb. 24.4) wird für Skalierungsmodell" die Option „Euklidischer Abstand mit individuell gewichteten Differenzen" gewählt (Konditionalität „Matrix" wird beibehalten). In der Dialogbox von „Form" (⇨ Abb. 24.3) wird die Voreinstellung „Matrix" übernommen.

Modell der multidimensionalen Entfaltung (MDU) (Unfolding).
In den bisher besprochenen Modellvarianten werden quadratische Matrizen (auch die per Ankerpunktmethode gewonnene Matrix ist quadratisch) mit Ähnlichkeitsurteilen von Objektpaaren analysiert. Die Matrizen haben sowohl in den Zeilen als auch in den Spalten Objekte (Objekt*Objekt). Es können aber auch rechteckige Matrizen analysiert werden. In diesen stehen in den Zeilen Urteilspersonen und in den Spalten Objekte (Subjekt*Objekt). Diese rechteckigen Matrizen entstehen durch das Untersuchungsdesign. Diese Datenmatrizen sind zeilenkonditional. In der Dialogbox „Form" wird „rechteckig" (im Fall von Messwiederholungen, d.h. mehreren Matrizen, gibt man die Anzahl der Zeilen der Matrix an) und in der Dialogbox „Modell" wird für Konditionalität „Zeile" gewählt.

In der Lösungskonfiguration werden die Objekte zusammen mit den Subjekten in einer Konfiguration dargestellt. Die dargestellten Subjekte sind dabei als „Idealpunkte" der Subjekte hinsichtlich der betrachteten Objekte zu interpretieren (z.B. das ideale Auto, die ideale Zeitschrift etc. eines Subjekts).

25 Interaktive Grafiken erzeugen und gestalten

Die ab SPSS Version 8.0 eingeführten interaktiven Grafiken unterscheiden sich in mehrfacher Weise von den herkömmlichen SPSS-Grafiken (Standarddiagramme). Das Erstellen und Überarbeiten vollzieht sich auf eine neue menügestützte Weise. Grafiken lassen sich dynamisch verändern: Die auf den Achsen abgebildeten Variablen können ausgetauscht, zweidimensionale können in dreidimensionale, ungruppierte in gruppierte Grafiken verändert werden (und umgekehrt), die Präsentationsform der Grafik kann modifiziert werden. Im Unterschied zu herkömmlichen Grafiken wird das Überarbeiten von Grafiken im Ausgabefenster und nicht im Diagramm-Editor (⇨ Abb. 25.1) vorgenommen

Die Grundtypen der interaktiven Grafiken in der Version 11 (Balken-, Linien-, Kreisdiagramm etc.) sind bis auf ein paar Ausnahmen (⇨ Kap. 25.1, neue Grafikgrundtypen) auch in herkömmlicher Weise erstellbar (⇨ Kap. 26). Aber nicht alle herkömmlichen Grafiktypen können auch als interaktive Grafiken erzeugt werden. Hoch-Tief-, Pareto-, Regelkarten-, P-P- , Q-Q- sowie Autokorrelations- und Kreuzkorrelationsdiagramme gibt es in Version 11 nur als herkömmliche Grafiken. Neu und damit für viele Anwender interessant ist aber, dass man auf der Basis der Grafikgrundtypen zu neuen Grafikarten und zu optisch neuen Gestaltungsformen kommen kann:

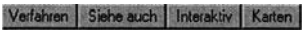
- Jeder Grafiktyp (mit Ausnahme des einfachen und gruppierten Kreisdiagrammes) ermöglicht eine echte dreidimensionale Darstellung, d.h. es können Daten in einem 3-Achsenraum dargestellt werden.
- Für die Grafiktypen können - soweit es die Daten erlauben und sinnvoll erscheinen lassen - Mischformen dargestellt werden. So lässt sich z.B. ein Balkendiagramm zur Darstellung der Durchschnittswerte der Einkommen für Männer und Frauen durch ein Fehlerbalkendiagramm oder durch ein Streudiagramm (oder beides) ergänzen.
- Für jeden Variablentyp lassen sich Grafiksets für Teilgesamtheiten, die durch Werte einer (oder mehrerer) Feldvariablen definiert werden, darstellen.
- Für jeden Grafiktyp (mit Ausnahme des Streudiagrammes) ist für zweidimensionale Grafiken ein 3D-Effekt erzielbar.
- Verschiedene Beleuchtungseffekte sowie beliebige Drehungen von dreidimensionalen Grafiken und zweidimensionalen Grafiken mit 3D-Effekt um ihre Achsen ermöglichen starke optische Wirkungen.
- Eine Vielzahl von weiteren optischen Gestaltungsmöglichkeiten wird angeboten: Die Beschriftungsmöglichkeiten von Daten sind erweitert, verschiedene Balkenformen sind möglich, neue Füllmuster für Grafikelemente (z.B. für Flä-

chen) stehen zur Auswahl, es gibt viele neue Symbole zur Darstellung der Datenpunkte in Streudiagrammen, Legenden können auf beliebige Stellen der Grafik verschoben werden, neue beliebige Textelemente können der Grafik hinzugefügt werden etc. Ein Vorzug ist auch, dass Grafiken Erläuterungen zu der Datendarstellung enthalten.

Im folgenden wird in die Technik des Erstellens und Modifizierens interaktiver Grafiken eingeführt. Dabei können wir uns auf exemplarische Darstellungen beschränken, da die Vorgehensweise bei den unterschiedlichen Grafiktypen ähnlich ist.

Um einen der interaktiven Grafiktypen zu erstellen, gibt es prinzipiell zwei Wege:

- Im Ausgabefenster wird die Befehlsfolge „Einfügen", „Interaktive 2D-Grafik" (oder „Interaktive 3D-Grafik") aufgerufen. Im Ausgabefenster öffnet sich nun ein leerer Grafikrahmen (⇨ Abb. 25.9). Auf der oberen Leiste des Rahmens wird nun das Symbol  geklickt. Es öffnet sich die Dialogbox „Variablen für Grafik zuweisen". Nun werden die Koordinatenachsen der Grafik mit Variablen versorgt, indem man die gewünschten Variablen aus der Quellvariablenliste in die Achsenfelder der Grafik übergibt. Dafür wird eine Variable mit gedrückter linker Maustaste in ein Achsenfeld gezogen. Klicken auf das Symbol  in der oberen Leiste des Grafikerstellungsrahmens öffnet eine Palette zur Auswahl eines Grafiktyps. Klickt man nun z.B. ein Balkendiagramm an, so wird dieses sofort erzeugt. Dieses Vorgehen bei der Grafikerstellung wird im folgenden nicht vertieft.
- Es wird die Befehlsfolge „Grafik", „Interaktive Grafik" aufgerufen. Es öffnet sich das Menü mit den verfügbaren interaktiven Grafiktypen (Balken-, Liniendiagramm etc.). Nun wählt man einen Grafiktyp aus. Es öffnet sich eine Dialogbox zur Erstellung der ausgewählten Grafik. Den Feldern der Koordinatenachsen der Grafik können nun die darzustellenden Variablen zugewiesen werden. Dazu wird jeweils eine Variable der Quellvariablenliste mit gedrückter linker Maustaste herübergezogen. Die Grafik erscheint nun im Ausgabefenster. Die erzeugte Grafik kann jetzt in vielfältiger Weise modifiziert sowie in eine Präsentationsform (Layout) gebracht werden. Anhand exemplarischer Beispiele soll die grundlegende Technik des Erstellens, der dynamische Wechsel zu Grafiken mit anderen Formen und mit zusätzlichen Variablen sowie die Layoutgestaltung dargestellt werden.

Zur Nutzung des zweiten (vermutlich meist verwendeten) Weges zur Grafikerstellung wird im Menü „Grafiken" mit der Wahl von „Galerie" ein Hilfesystem angeboten. Zunächst öffnet sich das Hilfe-Fenster für herkömmliche Diagramme („Hauptgalerie der Diagramme"). Nach Klicken der Schaltfläche „Interaktiv" der Schaltflächenleiste 

öffnet sich das Hilfe-Fenster „Galerie für interaktive Diagramme" (Abb. 25.1). Durch Klicken auf das Symbol eines Diagrammtyps gelangt man zur sehr anschaulichen Hilfe. Probieren sie es aus.

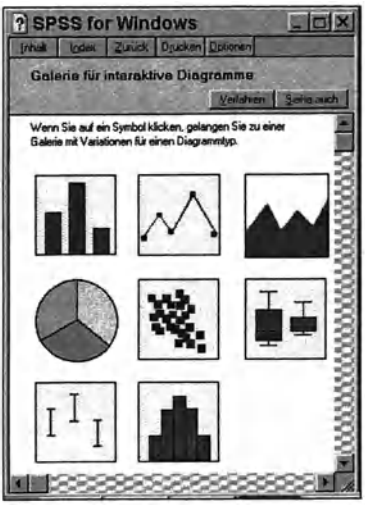


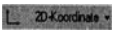
Abb. 25.1. Hilfe-Fenster „Galerie für interaktive Grafiken"

25.1 Interaktive Grafiken erzeugen

Grundlegendes. Am Beispiel interaktiver Balkendiagramme soll die Vorgehensweise dargelegt werden. Nach der Befehlsfolge „Grafiken", „Interaktiv" wird aus der geöffneten Liste der gewünschte Grafikgrundtyp gewählt (hier: Balken...). Es öffnet sich die in Abb. 25.2 links dargestellte Dialogbox „Balkendiagramm erstellen". In dieser Dialogbox kann man eine Reihe von Registerkarten anwählen. Jede Registerkarte dient der Festlegung weiterer Grafikelemente. Akuell geöffnet ist die Dialogbox „Variablen zuweisen". In dieser werden die in der Grafik darzustellenden Variablen festgelegt.

Die Anzeige in der Quellvariablenliste interaktiver Grafiken unterscheidet sich von der bei statistischen Auswertungsprozeduren oder herkömmlichen Grafiken. Durch Symbole werden drei Variablentypen angezeigt: ist das Symbol für kategoriale (ordinale bzw. nominalskalierte), für metrische und für systemeigene Standardvariablen ($CASE = Fall, $COUNT = absolute Häufigkeit, $PCT = prozentuale Häufigkeit). Dieses ist insofern bedeutsam, als manchen Achsen einer Grafik nur ein bestimmter Variablentyp zugeordnet werden kann (z.B. ist für die X-Achse eines Histogramms nur eine metrische Variable zulässig). Es ist aber leicht möglich, die Anzeige einer Variablen von einer kategorialen in eine metrische (und umgekehrt) zu ändern. Dazu klickt man mit der rechten Maustaste auf eine Variable in der Quellvariablenliste zum Öffnen eines Kontextmenüs. Mit „Metrisch" bzw. „Kategorial" kann der Wandel in der Anzeige vorgenommen werden. Mit den Befehlen des Kontextmenüs kann man außerdem die Reihenfolge der Variablen in der Quellvariablenliste ändern (Sortierung nach dem Variablennamen, nach der Reihenfolge in der Datei bzw. nach den drei Variablentypen) sowie bewirken, dass entweder die Variablennamen oder die Variablenlabel in der Quellvariablenliste angezeigt werden. Das bei Auswertungsprozeduren und her-

kömmlichen Grafiken mit der rechten Maustaste aktivierbare Kontextmenü zum Abruf von Information über Variablen ist bei interaktiven Grafiken nicht verfügbar.

Zweidimensionale Grafiken erstellen. Klicken auf den Pfeil von 

öffnet eine Drop-Down-Liste zur Auswahl einer Balkendiagrammart (zweidimensional, zweidimensional mit 3D-Effekt oder dreidimensional). Nach Wahl von 3D-Effekt (⇨ Abb. 25.2 links) können den Achsen durch Ziehen mit der Maus Variablen aus der Quellvariablenliste zugewiesen werden (Variable mit linker Maustaste festhalten und herüberziehen). Mit den Variablen $PCT (prozentuale Häufigkeit) auf der Y-Achse und SCHUL (Schulabschlüsse) auf der X-Achse wird ein Balkendiagramm dargestellt, das die prozentualen Häufigkeiten der Schulabschlüsse darstellt. Zur Darstellung absoluter Häufigkeiten wählt man für die Y-Achse $COUNT anstelle von $PCT.

In unserer grafischen Darstellung sollen die Häufigkeiten der Schulabschlüsse nach dem Geschlecht untergliedert werden (gruppiertes Balkendiagramm). Dafür überträgt man die Variable GESCHL in eines der Eingabefelder von „Legendenvariablen". Wählt man „Farbe", so werden die Balken für Männer und für Frauen in verschiedenen Farben (⇨ Abb. 25.2 rechts) und bei „Muster" in verschiedenen Füllmustern dargestellt. Man kann auch beide Eingabefelder von „Legendenvariablen" mit jeweils einer Variable versorgen. Dann werden die Häufigkeiten eines jeden Schulabschlusses nach beiden Variablen untergliedert grafisch dargestellt (durch verschiedene Farben und Füllmuster). Lässt man die Eingabefelder von „Legendenvariablen" leer, so wird auf die Untergliederung und damit auf Darstellung als gruppiertes Balkendiagramm verzichtet. Durch Klicken auf den Pfeil von 

und Wahl von 

kann man anstelle eines gruppierten Balkendiagramms auch ein gestapeltes erstellen.
Unterhalb der Legende (mit Legendentitel „Geschlecht") in Abb. 25.2 wird die *Erläuterung* „Balken zeigen Prozent" (mit „Balken" als Titel) angezeigt.

Für Punkt-, Linien-, Band-, Verbundlinien- und Streudiagramme gibt es neben „Farbe" und „Muster" auch „Größe" als Eingabefeld für „Legendenvariablen". Die durch Symbole (z.B. Kreise oder Kreuze) dargestellten Datenpunkte der Gruppen (z.B. Männer und Frauen) in diesen Diagrammen erhalten unterschiedliche Größen (anstelle unterschiedlicher Farben bzw. Muster).

In das Eingabefeld von „Feldvariablen" (⇨ Abb. 25.2 links) kann man eine (oder auch mehrere) kategoriale Variablen eintragen. Dieses bewirkt, dass für jeden Variablenwert der eingetragenen Variablen eine Grafik gebildet wird (⇨ unten).

Für zweidimensionale Grafiken ohne 3D-Effekt lässt sich mit den Schaltern  und  festlegen, ob die Balken in der Grafik senkrecht oder waagerecht dargestellt werden sollen.

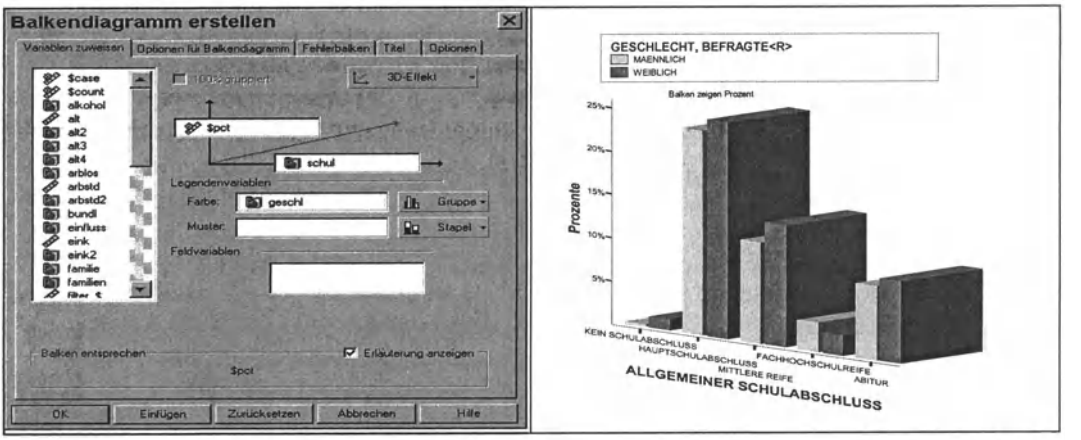


Abb. 25.2. Interaktives gruppiertes Balkendiagramm für prozentuale Häufigkeiten einer Variablen

Metrische oder kategoriale Variable auf der Y-Achse darstellen. Auf die Y-Achse kann anstelle absoluter ($COUNT) bzw. prozentualer Häufigkeiten ($PCT) auch eine kategoriale (nominal- oder ordinalskalierte) oder metrische Variable übertragen werden. Überträgt man eine kategoriale Variable, so wird für die Balkenhöhe der Modalwert dargestellt. Überträgt man eine metrische Variable, z.B. die Variable EINK (Nettoeinkommen), auf die Y-Achse, so zeigt sich in der Dialogbox unter „Balken entsprechen" ein Auswahlfeld mit einer Drop-Down-Liste zur Auswahl aus einer ganzen Reihe von Auswertungsfunktionen für die Variable EINK (⇨ Abb. 25.3 links). Standardmäßig wird mit „Mittelwert" das durchschnittliche Nettoeinkommen der Befragten für jeden Schulabschluss als Balkenhöhe abgebildet. In unserem Beispiel wird nach dem Geschlecht der Befragten untergliedert. Mit der Registerkarte „Fehlerbalken" können den Balken Fehlerbalken um den Mittelwert hinzugefügt werden.

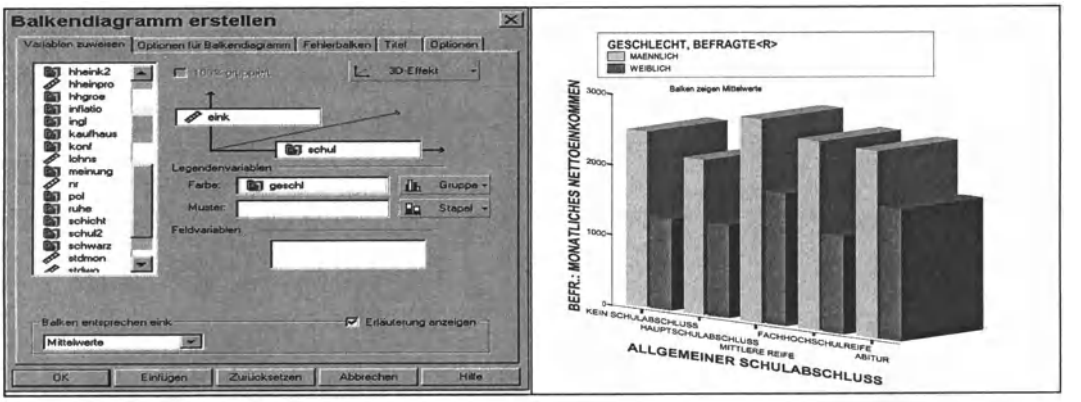


Abb. 25.3. Interaktives gruppiertes Balkendiagramm für Durchschnittswerte einer metrischen Variablen

Dreidimensionale Grafiken erstellen. Wählt man in der Dialogbox „Balkendiagramm erstellen" 3D-Koordinate (⇨ Abb. 25.4 links), so kann eine dreidimensionale Grafik erstellt werden. In unserem Beispiel wird für die Y-Achse die Variable $PCT (prozentuale Häufigkeiten), für die X_1-Achse die Variable SCHUL (Schulabschlüsse) und für die X_2-Achse die Variable GEM1 (Gemeindegrößenklassen mit „1" = unter 50 Tsd. Einwohner und „2" = 50 Tsd. und mehr Einwohner) übertragen. Da in das Eingabefeld „Farbe" die Variable GESCHL übertragen worden ist, entsteht eine dreidimensionale Grafik mit einer Untergliederung der dargestellten Häufigkeiten nach dem Geschlecht (⇨ Abb. 25.4 rechts). Auch für dreidimensionale Grafiken gilt, dass auf der Y-Achse mittels einer Auswertungsfunktion (z.B. Mittelwert) berechnete Werte einer metrischen Variablen dargestellt werden können.

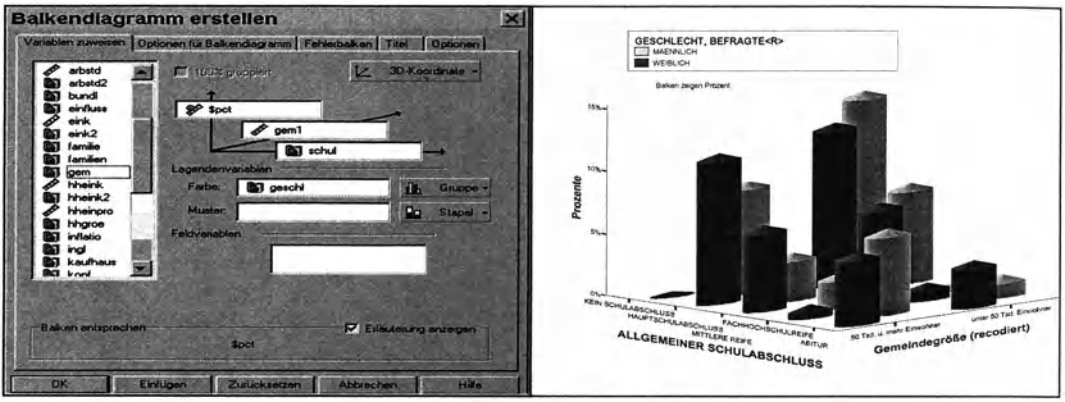


Abb. 25.4. Interaktives dreidimensionales Balkendiagramm für Häufigkeiten von Variablen

Werte von Fällen darstellen. Sollen die Werte von Fällen grafisch dargestellt werden, so wird auf die X-Achse die Standardvariable $CASE übertragen und im hier gezeigten Beispiel auf die Y-Achse die metrische Variable HHGROE (Haushaltsgröße) (⇨ Abb. 25.5 links). In Abb. 25.5 rechts ist die 2D-Grafik für die ersten zehn Fälle zu sehen.

Hinweis. Anders als bei herkömmlichen Grafiken ist die Fallauswahl „Nach Zeit- oder Fallbereich" im Menü „Daten", „Fälle auswählen" für interaktive Grafiken nicht verfügbar.

Feldvariablen: Separate Grafiken für jede Kategorie erstellen. Im folgenden Beispiel wird das durchschnittliche Einkommen von Männern und Frauen in Balkendiagrammen verglichen. Dabei soll für jeden Gemeindegrößentyp (bis 1999, 2000-4999 Einwohner etc.), d.h. für Teilgesamtheiten der Datendatei, eine derartige Grafik erzeugt werden. Dazu wird die Variable GEM (Gemeindegrößenklassen) in das Eingabefeld von „Feldvariablen" gezogen. In Abb. 25.6 ist links die Dialogbox und rechts das Grafikergebnis zu sehen.

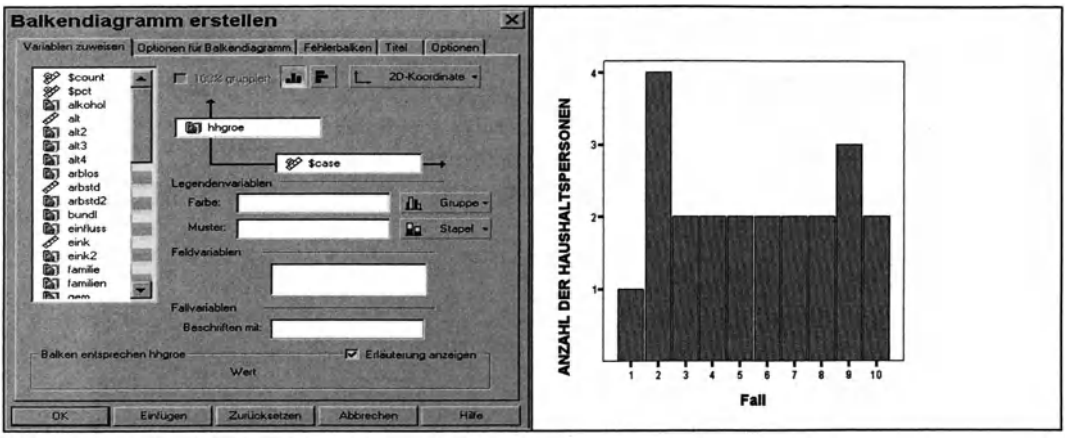


Abb. 25.5. Interaktives Balkendiagramm zur Darstellung der Werte von Fällen

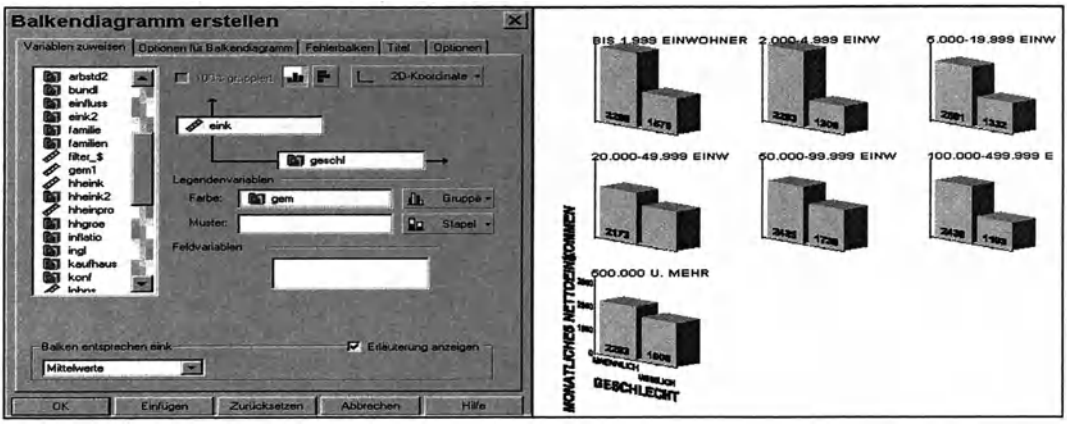


Abb. 25.6. Interaktives Balkendiagramm mit Feldvariable

Registerkarten: Festlegen von weiteren Grafikeigenschaften. Durch Klicken auf Registerkarten in der Dialogbox „Balkendiagramm erstellen"(⇨ Abb. 25.6) können weitere Elemente der Grafik festgelegt werden. In Abb. 25.7 links ist die Registerkarte „Optionen für Balkendiagramm" zu sehen. In „Form" ist die Form der Balken wählbar. In „Balkengrundlinie" kann durch Auswahl von „Benutzerdefiniert" und Überschreiben des Standardwertes 0 die Grundlinie der Balken auf einen anderen Wert als 0 festgelegt werden. Mit den Kontrollkästchen „Anzahl" bzw. „Wert" von „Beschriftung" kann man festlegen, ob für die Balkenhöhe die Fallzahlen, die Werte (z.B. Prozentwerte oder Werte von Berechnungsfunktionen) oder beides angezeigt werden soll.

In Abb. 25.7 rechts ist die Registerkarte „Fehlerbalken"zu sehen. Für diese Registerkarte sind nur dann Festlegungen möglich (Optionen aktiv geschaltet), wenn auf der Y-Achse der Mittelwert einer metrischen Variablen abgebildet wird. Das Kontrollkästchen „Fehlerbalken anzeigen" erlaubt es, die Darstellung von Fehler-

balken ein- bzw. auszuschalten. Für das grafisch dargestellte Konfidenzintervall gibt es drei Auswahlmöglichkeiten: „Konfidenzintervall für den Mittelwert", „Standardabweichung" und „Standardfehler des Mittelwerts" (⇨ Kap. 26.10.1). Das voreingestellte Konfidenzintervall von 95 % kann verändert werden. Mit den Auswahlmöglichkeiten von „Form" lässt sich die Fehlerbalkenform und mit „Richtung" die Richtung der Fehlerbalken bestimmen.

Die Registerkarte „Titel" dient dazu, die Grafik mit Titel, Untertitel und Fußnote zu versehen.

Die Registerkarte „Optionen" bietet eine Auswahl verschiedener Grafiklayouts (insbesondere für die Farben).

Für andere Grafiktypen sieht die Dialogbox zur Erzeugung der Grafik ähnlich aus. Die Registerkarten „Variablen zuweisen", „Titel" und „Optionen" sind immer vorhanden. Spezifisch sind Registerkarten, die dem jeweiligen Grafiktyp entsprechen (z.B. „Punkte und Linien" bei Punkte- und Liniendiagrammen, „Kreisdiagramme" bei Kreisdiagrammen etc.).

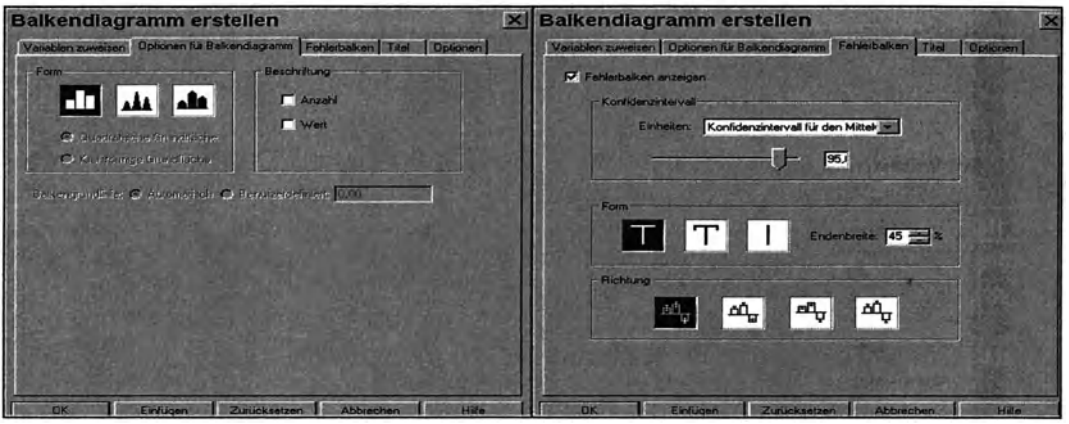


Abb. 25.7. Registerkarten „Balken" und „Fehlerbalken" interaktiver Balkendiagramme

Neue Grafikgrundtypen. Punkt- und Banddiagramme sind ab Version 8.0 als Grundtypen neu. Sie sind aber (wie auch das Verbundliniendiagramm ⇨ Kap. 26.3.3) Varianten eines Liniendiagramms. In Punktdiagrammen werden die Datenpunkte auf der Y-Achse nicht durch Linien verbunden. Ein Banddiagramm ist ein Liniendiagramm, in dem die Linien in einer 3D-Darstellung als Band erscheinen. Man kann nach der Erzeugung dieser Diagrammtypen bei entsprechender Datenlage durch Bearbeitung der Grafik zu den anderen Diagrammtypen wechseln.

Interaktive Kreisdiagramme erweitern das Grafiktypenangebot durch gestapelte (= gruppierte) und geplottete Kreisdiagramme (⇨ Abb. 25.8 und 25.9). In einem Kreisdiagramm kann in das Zuweisungsfeld von „Auswertungsvariable:" auch eine metrische Variable übertragen werden. Für die Kreissegmente können dann Auswertungsfunktionen („Summen", „Quadratsummen" usw.) ausgewählt werden, die sich von denen bei herkömmlichen Kreisgrafiken unterscheiden.

Neu ist auch, dass für Histogramme die kumulierten Häufigkeiten dargestellt werden können. Bei Streudiagrammen haben sich die Optionen für Projektionslinien erweitert.

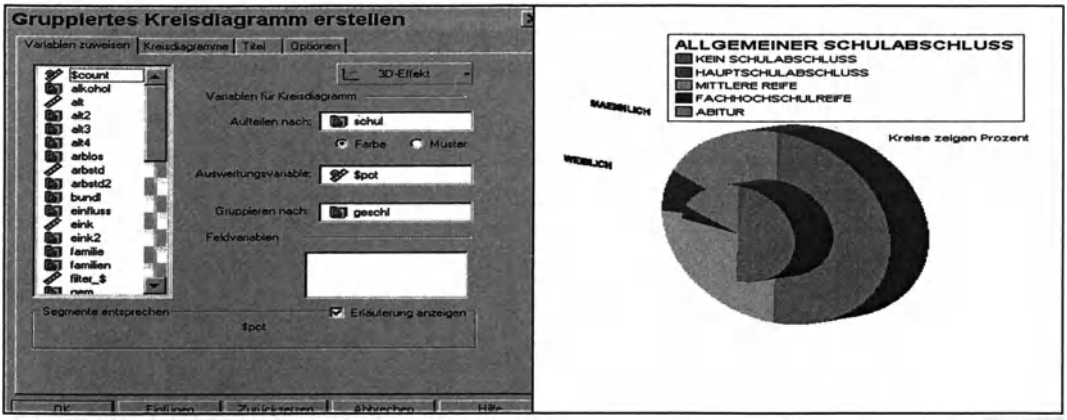


Abb. 25.8. Interaktives gruppiertes Kreisdiagramm mit 3D-Effekt

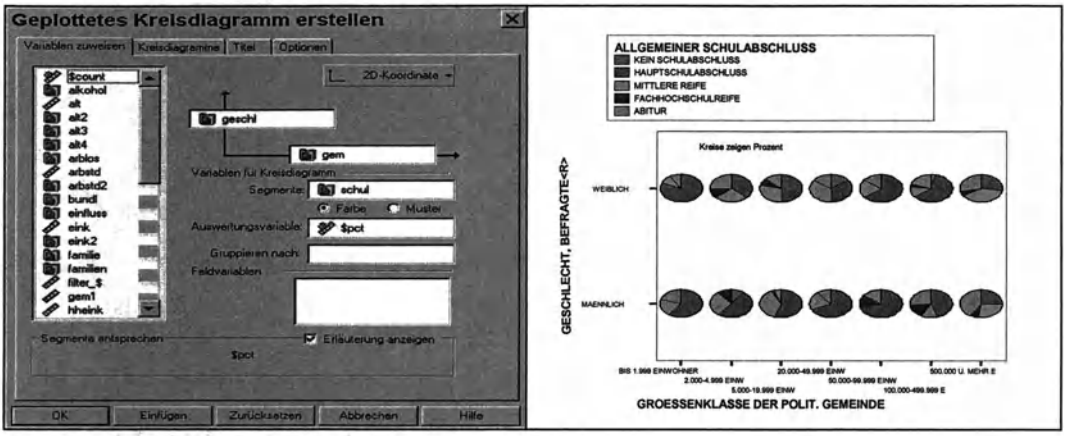


Abb. 25.9. Interaktives geplottetes Kreisdiagramm

25.2 Interaktive Grafiken verändern und gestalten

25.2.1 Grundlegende Grafikveränderungen

Grafik zum Überarbeiten aktiv schalten. Im Unterschied zu herkömmlichen Grafiken erfolgt eine Überarbeitung der Grafik im Ausgabefenster („Viewer"). Damit aber eine Grafik überarbeitet werden kann, muss sie durch Doppelklicken auf die Grafik für eine Bearbeitung aktiv geschaltet werden (Alternative 1: über das Menü „Bearbeiten" im Viewer: Nach Markieren der Grafik: „SPSS interaktives Grafikobjekt" wählen und mit „Bearbeiten" für die Bearbeitung aktivieren;

Alternative 2: Mit dem Cursor auf die Grafik zeigen, Klicken der rechten Maustaste, „SPSS interaktives Grafikobjekt" wählen und mit „Bearbeiten" für die Bearbeitung aktivieren). Die Grafik erhält dann einen Rahmen mit (beweglichen) Symbolleisten auf dem oberen und dem linken Rand (⇨ Abb. 25.10). Die Bearbeitung und Überarbeitung der Grafik kann bei Aktivschaltung mit Hilfe der Symbole auf den Symbolleisten, über nun verfügbare Befehle der Menüs oder durch Auswahl von Befehlen aus Kontextmenüs, die sich bei Klicken mit der rechten Maustaste öffnen, erfolgen.

Abb. 25.10. Interaktive Grafik nach Aktivschaltung zum Überarbeiten (durch Doppelklick) im Ausgabefenster

Variablenzuweisung ändern. Ausgangsgrafik sei die in Abb. 25.10 bzw. Abb. 25.3 dargestellte. Klicken auf das Symbol (oder über Menü: „Bearbeiten", „Variablen zuweisen"; oder über Kontextmenü: rechter Mausklick außerhalb des Datenbereichs der Grafik, „Variablen zuweisen") öffnet die Dialogbox „Variablen für Grafik zuweisen" (⇨ Abb. 25.11 links). Durch Herüberziehen anderer Variablen auf die Achsen kann die Grafik verändert werden. Dabei können alte Variablen durch Überlagerung verdrängt werden. Außerdem können den Eingabefeldern von „Legendenvariablen" („Farbe:, „Muster:", „Größe:") bzw. „Feldvariablen" Variablen zugeführt (oder weggenommen) werden, um gruppierte Diagramme bzw. Diagramme für Teilgesamtheiten zu erhalten (oder aufzuheben). In Abb. 25.11 ist die Legendenvariable GESCHL von „Farbe:" nach „Muster:" gezogen worden.

Nach Klicken auf den Pfeil von 3D-Effekt und Auswahl von 3D-Koordinate kann eine zweidimensionale Grafik durch Herüberziehen einer Variablen auf die dritte Achse in eine dreidimensionale überführt werden. Sobald die Variablenzuordnung geändert ist, entsteht die modifizierte Grafik. Auch kann der 3D-Effekt ein- oder ausgeschaltet werden.

Die Legende sowie der Text „Balken zeigen Mittelwerte“ kann beliebig verschoben werden, indem man mit der Maus darauf zeigt, mit der linken Maus festhält und dann zieht.

Registerkarten am rechten Rand (⇨ Abb. 25.11 links) ermöglichen weitere Festlegungen. Bei geeigneter Grafik (Extremwerte und Ausreißer im Boxplot- oder Datenpunkte im Streudiagramm) erlaubt die Registerkarte „Fälle“, eine Beschriftung der Fälle mittels einer Variablen einzuführen oder auszublenden. Die Beschriftung wird aber nur dann angezeigt, wenn die Beschriftung eingeschaltet ist. Die Registerkarte „Kreisdiagramme“ ermöglicht es, Kreisdiagrammen neue Variablen zuzuweisen.

Hinweis. Interaktive Grafiken im Ausgabefenster können einen unterschiedlichen Status haben. Daraus ergeben sich unterschiedliche Konsequenzen hinsichtlich der Möglichkeit, durch Neuzuordnung von Variablen die Grafik zu modifizieren. Wenn die Datendatei geöffnet ist und die Daten seit der Grafikerzeugung im Daten-Editor nicht verändert worden sind, besteht eine Verbindung der Grafik zu den Daten. In der Statusanzeige am unteren Rand wird die Grafik als interaktive angezeigt. Der Grafik können neue Variablen aus der Datendatei zugewiesen werden. Wenn auf der Registerkarte „Interaktiv“ der Dialogbox „Optionen“ (wird über das Menü „Bearbeiten“, „Optionen“ aufgerufen) „Daten mit Diagramm speichern“ eingeschaltet ist und im Daten-Editor Daten verändert wurden, so wird die Grafik in der Statusanzeige als „Interaktive Grafik (von Daten getrennt)“ angezeigt. Es besteht nur noch eine Verbindung zu den in der Grafik verwendeten Variablen. Eine Neuzuweisung von Variablen beschränkt sich auf diese. Ist auf der Registerkarte „Interaktiv“ die Option „Nur zusammengefasste Daten speichern“ gewählt, und werden Daten nach Erzeugung der Grafik verändert, so wird die Grafik statisch. In der Statusanzeige wird dieses angezeigt. Es können dann lediglich Layouteigenschaften der Grafik (Farbe, Füllmuster etc.) verändert werden, nicht aber die Zuweisung von Variablen.

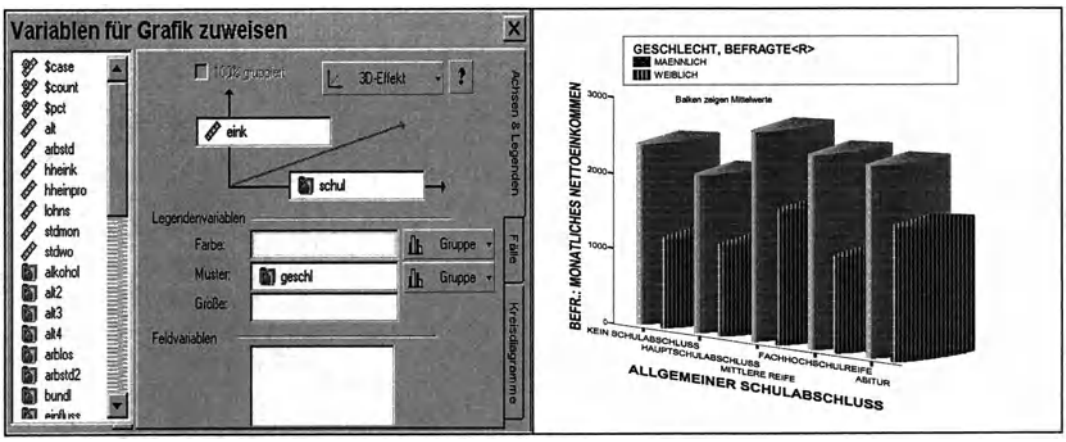


Abb. 25.11. Verschiebung der Variablen GESCHL von „Farbe:“ nach „Muster:“

Die Auswertungsfunktion ändern. Im Balkendiagramm der Abb. 25.11 wird als Balkenhöhe das durchschnittliche Einkommen (untergliedert nach Geschlecht) dargestellt. Die Auswertungsfunktion der Variable EINK auf der Y-Achse ist der Mittelwert. Diese Auswertungsfunktion kann man verändern. Da man dieses am besten mit Hilfe des Diagramm-Managers bewerkstelligt, sei auf Kap. 25.2.3 (⇨ Durch Daten dargestellte Grafikelemente überarbeiten) verwiesen.

Grafiken zu gemischten Grafiken verändern. Durch Hinzufügen von Grafikgrundtypen zu einer Grafik entstehen neue Mischformen. Im folgenden Beispiel soll das in Abb. 25.3 dargestellte Balkendiagramm (mit 3D-Effekt) um die grafische Darstellung der Streuung der Einzelwerte für jeden Schulabschluss ergänzt werden. Bei aktiv geschalteter Grafik klickt man auf das Symbol  . Es öffnet sich eine Palette mit den einfügbaren Grafikgrundtypen sowie einfügbaren Grafikelementen wie z.B. eine Regressionsgrade. In unserem Beispiel wird  (Punktwolke) gewählt (oder über Menü: „Einfügen", „Punktwolke"). Das Balkendiagramm wird um ein Streuungsdiagramm ergänzt (⇨ Abb. 25.12 links, (bei Ausschaltung des 2D-Effekts).

Dabei ist zu beachten, dass der hinzugefügte Grafiktyp für die Daten geeignet sein muss. In unserem Beispiel könnte man auch ein Fehlerbalken- oder ein Boxplotdiagramm einfügen. Ein Liniendiagramm zur Darstellung z.B. eines Mittelwerts könnte um eine Linie zur Darstellung des Medians (oder einer anderen Auswertungsfunktion) ergänzt werden. Um eine derartige Ergänzung zu bekommen, wird im ersten Schritt zum Liniendiagramm ein weiteres Liniendiagramm hinzugefügt („Punkt-Linie"). Im zweiten Schritt wird die Auswertungsfunktion von Mittelwert auf Median geändert. Streudiagrammen können „Mittelwertanpassung", „Regressionsanpassung" oder „LLR-Glättung hinzugefügt werden. Hätte man aber z.B. ein Balkendiagramm zur Darstellung der absoluten oder prozentualen Häufigkeiten der Schulabschlüsse, so ließe sich zwar eine Punktwolke einfügen. Da aber die Daten die Darstellung einer Punktwolke gar nicht erlauben, bleibt diese Modifikation jedoch ohne sichtbare Wirkung.

In Abb. 25.12 rechts ist ein zweites Beispiel dargestellt: Die Balken stellen die Mediane der Einkommen für Schulabschlüsse dar. Ergänzt ist die Darstellung um Fehlerbalken, die den 2-Sigma-Streuungsbereich um den Mittelwert der Einkommen grafisch darstellen.

Eine eingefügte Grafik kann auch wieder gelöscht werden (⇨ Kap. 25.2.3).

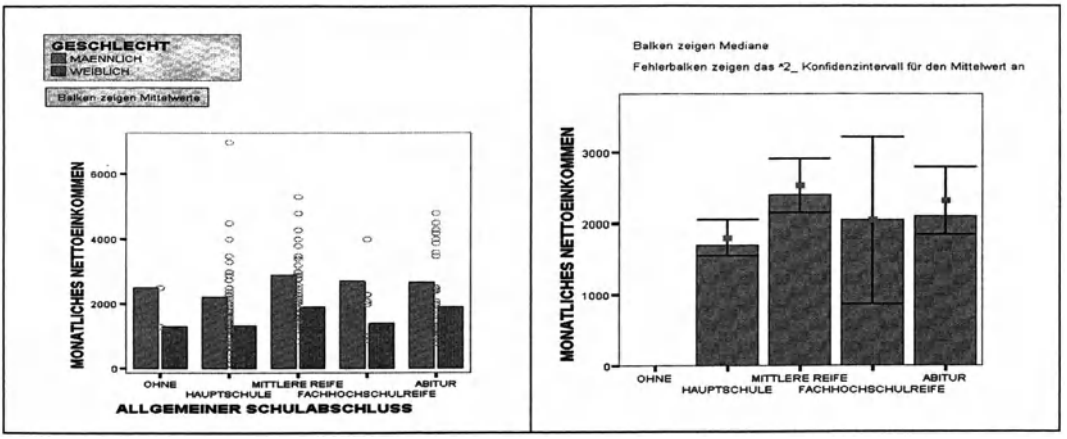


Abb. 25.12. Gemischte interaktive Diagramme

25.2.2 Grafiklayout gestalten

Überblick über alternative Vorgehensweisen bei der Layoutgestaltung. SPSS bietet verschiedene Möglichkeiten an, um eine interaktive Grafik in eine ansprechende Präsentationsform zu. Dabei ist stets eine aktiv geschaltete Grafik Voraussetzung (⇨ Kap. 25.2.1).

Diagramm-Manager. Für eine komfortable Layoutgestaltung bietet sich der Diagramm-Manager an (⇨ Kap. 25.2.3). Aber nicht alle Möglichkeiten sind abgedeckt: Für einzelne durch Daten bestimmte Grafikelemente (z.B. einen einzelnen Balken eines einfachen Balkendiagramms, einen einzelnen Punkt eines Streudiagramms) können Layouteigenschaften wie Farbe und Füllmuster eines Balkens, Symbolart, -größe und -farbe von Datenpunkten sowie Labelbeschriftungen nicht per Diagramm-Manager geändert werden. Denn mit dem Diagramm-Manager vorgenommene Änderungen werden stets auf alle Balken bzw. alle Datenpunkte angewendet. Für die Änderung von Layouteigenschaften einzelner Balken, einzelner Datenpunkte etc. kann man entweder die Symbole auf der Symbolleiste oder den Befehl „Eigenschaften" im Kontextmenü verwenden. Die Anzeige von Labels (z.B. die Fallnummer) für nur einen (oder auch mehrere) zu markierenden Datenpunkt ist nur über den Befehl „Eigenschaften" möglich.

Menü Format. Man nutzt aus den Optionen im Menü „Format" entweder „Diagrammeigenschaften" mit den acht Registerkarten auf der Dialogbox zur Festlegung einer ganzen Palette von Layouteigenschaften (⇨ unten: Diagrammeigenschaften festlegen) oder die jeweiligen Befehle des Menüs „Format", um einzelne Grafikteile („Text", „Rahmen", „Achse", „Legende", „Erläuterung", „Gitterlinien", „Datenbereich", „Grafikelemente") zu überarbeiten. Die Befehle „Text" und „Rahmen"sind nur dann aktiv, wenn zuvor ein Text (Titel, Fußnote, Achsenbeschriftungen, Legende, Erläuterung etc.) bzw. ein Textelement, das einen Rahmen erhalten kann (Legende, Erläuterung), durch Anklicken mit der linken Maustaste markiert wurde (⇨ unten: Markieren von Grafikobjekten). Der Aufruf der Befehle öffnet Dialogboxen, die Festlegungen zum Layout des jeweiligen Gra-

fikelements erlauben. Es handelt sich dabei um Dialogboxen, die mit wenigen Ausnahmen auch bei der Arbeit mit dem Diagramm-Manager geöffnet werden. Daher soll diese Vorgehensweise hier nicht weiter vertieft und auf Kap. 25.2.3 verwiesen werden.

Symbole der Symbolleiste. Zur Festlegung von Füllmuster, Füll- und Rahmenfarben für Grafikobjekte, von Symbolmuster und -größen für Grafikpunkte, von Linienmuster und -stärken können auch die Symbole der Symbolleiste auf dem Rahmen der aktiven Grafik eingesetzt werden (⇨ unten: Symbolleiste... verwenden).

Kontextmenü. Auch diese Vorgehensweise bei der Layoutgestaltung soll nur kurz angesprochen werden. Wird ein Grafikobjekt (z.B. ein Balken, ein Kreissegment, ein Text, eine Achse, die Legende etc.) mit der rechten Maustaste angeklickt, so öffnet sich ein Kontextmenü mit Befehlen, die je nach angeklicktem Objekt verschieden sind. Wird z.B. in einem gruppierten Balkendiagramm (⇨ Abb. 25.3) ein Balken mit der rechten Maustaste angeklickt, so wird das in Abb. 25.13 dargestellte Kontextmenü geöffnet.

Einige der Befehle führen Operationen aus. So führen die Befehle in der ersten Gruppe dazu, dass der angeklickte Balken oder eine Balkengruppe ausgewählt und markiert und somit die Markierung einer Balkengruppe erleichtert wird. Wird z.B. die Erläuterung (bzw. die Legende) angeklickt, so kann mit dem Kontextmenübefehl „Legende (bzw. Erklärung) ausblenden" die Anzeige ausgeblendet werden.

Andere Befehle des Kontextmenüs öffnen Dialogboxen zur Layoutbestimmung, die auch mit dem Diagramm-Manager geöffnet werden können (z.B. „Balken" in Abb. 25.13). Der letzte Befehl im Kontextmenü der Abb. 25.13 „Eigenschaften" öffnet die Dialogbox „Balkeneigenschaften", in der Füllfarbe und -muster, Rahmen und Labels von Balken festgelegt werden können. Es fällt auf, dass in beiden Dialogboxen („Balken" und „Balkeneigenschaften") Layouteigenschaften der Balken (Farbe, Füllmuster, Rahmen und Label) festgelegt werden können. Der Unterschied besteht darin, dass bei der Dialogbox „Balken" diese Eigenschaften auf alle Balken übertragen werden, bei der Dialogbox „Balkeneigenschaften" hingegen nur auf die markierten Balken. Beachten Sie aber bitte den Hinweis zu „Symbolleiste... verwenden".

Menü Bearbeiten. Anstelle des Kontextmenüs können auch die Befehle des Menüs „Bearbeiten" verwendet werden. Dazu muss das Grafikobjekt zuvor markiert werden.

Diagrammformatvorlage. Für eine Layoutgestaltung kann auch eine Diagrammformatvorlage erstellt und angewendet werden (⇨ unten: Grafikformatvorlage erstellen und anwenden).

Hinweis. Dialogboxen zur Überarbeitung eines Grafikobjekts (z.B. Balken, Achsen etc.) öffnen sich auch, wenn man sie mit der linken Maustaste doppelt anklickt.

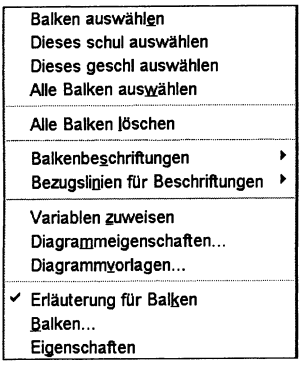


Abb. 25.13. Kontextmenü nach Klicken eines Balkens mit rechter Maustaste

Markieren von Grafikelementen. Für Layoutgestaltungen mit der Symbolleiste (oder alternativ mit dem Menü „Format" oder dem Kontextmenü) müssen Grafikobjekte (z.B. Balken eines Balkendiagramms, ein Kreissegment, ein Titel oder ein anderer Text, eine Achse, eine Achsenbeschriftung, eine Regressionsgerade etc.) zunächst markiert werden. Dazu wird ein Objekt mit der linken (bei Arbeiten mit dem Kontextmenü mit der rechten) Maustaste angeklickt. Die Markierung des Objekts wird meistens durch einen Rahmen angezeigt wie bei der Markierung eines Balkens: . Markierte Linien und Achsen werden durch das Symbol angezeigt. Sollen mehrere Objekte gleichzeitig (z.B. mehrere Balken, Kategorien- und Skalenachse, Untertitel und Fußnote etc.) markiert werden, wird mit dem linken Mausklick die Strg-Taste (oder Umschalttaste) gedrückt und festgehalten. Für durch Daten dargestellte Grafikelemente (Balken, Linien, Kreissegmente etc.) ist ein Markieren z.B. einer Balkengruppe (z.B. beide Balken der Kategorie „Abitur" in Abb. 25.2) mit dem Befehl „Dieses schul auswählen" des Kontextmenüs in Abb. 22.13 manchmal komfortabler (⇨ oben: Kontextmenü).
Ist eine Gruppe von z.B. Balken durch eine Legendenvariable definiert, markiert man sie, indem man die Gruppe in der Legende anklickt.

Layouteigenschaften einzelner Balken, Datenpunkte etc. ändern. Um Farbe, Füllmuster und andere Layouteigenschaften von Grafikelementen zu ändern, die durch Daten definiert sind, kann man die einzelnen Balken etc. markieren und dann die Symbole auf der Symbolleiste nutzen (⇨ unten: Symbolleiste... verwenden). Ein zweiter Weg führt über den Befehl „Eigenschaften" des Kontextmenüs (⇨ oben: Kontextmenü).

3D-Effekt einer zweidimensionalen Grafik ein- und ausschalten. Bei einer zweidimensionalen Grafik ohne 3D-Effekt lässt sich dieser bei aktiv geschalteter Grafik über das Menü „Ansicht" und „3D-Effekt" einschalten. Umgekehrt wird über „Ansicht", „2D-Koordinaten" der 3D-Effekt aufgehoben.

3D-Palette benutzen. Nach Doppelklicken auf eine dreidimensionale Grafik oder eine zweidimensionale Grafik mit 3D-Effekt erscheint mit dem Rahmen auch die „3D"-Palette (⇨ Abb. 25.10). Nach Klicken auf das Symbol kann man aus einem Angebot von unterschiedlichen Beleuchtungseffekten für die Grafik aus-

wählen. Zeigen mit der Maus auf den Drehschalter und Drehen durch Festhalten der linken Maustaste ermöglicht es, die Grafik um die senkrechte Achse zu drehen. Analog kann mit dem senkrecht angeordneten Drehschalter die Grafik um ihre waagerechte Achse gedreht werden. Durch Klicken auf die Symbole bzw. kann auf die Standardeinstellung für die Lage der Grafik zurückgesetzt werden. Der Schalter mit dem Symbol ermöglicht es, die Grafik permanent um ihre Achsen zu drehen. Die „3D"-Palette wird durch Klicken auf ausgeblendet und kann mit Klicken auf und Auswahl von „3D-Palette" oder mit dem Befehl „3D-Palette" des Menüs „Ansicht" wieder eingeblendet werden.

Ausrichtung der Grafik verändern. Klicken auf das Symbol und seinen Gegenpart vertauscht die Grafikachsen und damit die Ausrichtung von 2D-Grafiken.

Grafikformatvorlage erstellen und anwenden. Ab SPSS 8.0 wird eine Reihe von Grafikformatvorlagen bereitgestellt, die sich durch das Layout (Farb-, Muster-, Text-, Achsengestaltung, Symbole etc.) unterscheiden. Die Vorlagen haben die Endung .clo und sind im Unterverzeichnis „Looks" des Programmverzeichnisses abgelegt.

Bei Erzeugen der Grafik kann auf der Registerkarte „Optionen" der Dialogbox eine Vorlage gewählt werden. Verzichtet man bei der Grafikerzeugung per „Optionen" auf die Wahl einer Vorlage, so wird das Layout der Standardeinstellung des Systems genommen.

Man kann eine Vorlage auf eine schon erstellte Grafik anwenden, in dem man nach Aktivieren der Grafik die Befehlsfolge „Format", Diagrammvorlagen" klickt. Es öffnet sich die in Abb. 25.14 links dargestellte Dialogbox „Diagrammvorlage". Aus den angezeigten Vorlagen wird nun die gewünschte Vorlage ausgewählt. Nach Klicken auf die Schaltfläche „Zuweisen" wird die Vorlage auf die Grafik angewendet.

Eine verfügbare Vorlage kann auch als Voreinstellung gewählt werden. Alle erzeugten Grafiken übernehmen dann automatisch das Layout der Vorlage. Dazu wird auf der Registerkarte „Interaktiv" von „Optionen" des Menüs „Bearbeiten" die Vorlage bestimmt.

Um eine eigene Diagrammvorlage zu erstellen, wird zunächst eine Grafik mit dem gewünschten Layout angefertigt. Dieses kann auf zweifache Weise geschehen. Im einfachsten Fall ändert man das Layout z.B. eines Balkendiagramms oder eines Histogramms mit dem Diagramm-Manager (⇨ 25.2.3) in ein gewünschtes. Um das gefertigte Layout als Vorlage zu speichern, klickt man „Format", Diagrammvorlagen"und öffnet damit die Dialogbox „Diagrammvorlage (⇨ Abb. 25.14 links). Dann wählt man „<Eigenschaften des aktuellen Diagramms> (bei älteren Versionen <"Wie angezeigt>"), klickt auf die Schaltfläche „Speichern unter...", vergibt einen Dateinamen und speichert die Vorlage im Verzeichnis „Looks". Das Layout einer derartig erstellten Vorlage hat insofern nur einen beschränkten Anwendungsbereich, als nur das Layout von Grafikelementen des bearbeiteten Grafikgrundtyps enthalten ist. Um eine Grafikvorlage zum Anwenden auf unterschiedliche Grafikgrundtypen vorzubereiten, wird die Dialogbox „Diagrammeigenschaften" (⇨ Abb. 25.14 rechts) mit der Befehlsfolge „Format" „Diagrammeigenschaften" geöffnet. Dieses geht nur, wenn im Ausgabefenster eine

erzeugte Grafik aktiv geschaltet ist. Nun kann man alle möglichen Layouteigenschaften von Grafikgrundtypen mittels der acht Registerkarten der Dialogbox bestimmen (⇨ unten: Diagrammeigenschaften festlegen). Gleichzeitig werden auch die Layouteigenschaften auf die Grafik im Ausgabefenster übertragen. Insofern ist die Grafikgestaltung mittels „Diagrammeigenschaften" eine Alternative zur Anwendung des Diagramm-Managers (⇨ Kap. 25.2.3).

Um eine Vorlage zu erstellen, kann man auch eine der verfügbaren Diagrammformatvorlagen nach eigenen Vorstellungen ändern und überarbeiten. Nach „Format", „Diagrammvorlagen" wählt man eine der aufgelisteten Vorlagen und klickt „Vorlage bearbeiten". Mit den Registerkarten der dann geöffneten Dialogbox „Diagrammeigenschaften" (⇨ Abb. 25.14 rechts bezieht sich auf die Vorlage „Dante") können die gewünschten Layouteigenschaften vorgenommen werden. Anschließend wird die bearbeitete Vorlage unter einem zu vergebenden Namen gespeichert.

Legende und Erläuterung gestalten. Der Text der Erläuterung und der Legende (Ausnahme: Legendentitel) kann nicht verändert werden, wohl aber deren Eigenschaften.

Legende und Erläuterung können mit einem Rahmen und mit Füll- und Farbmuster versehen werden. Durch Klicken auf die Legende (bzw. Erläuterung) wird diese markiert (sichtbar an einem Markierungsrahmen). Mit der Befehlsfolge „Format", „Rahmen" öffnet sich die Dialogbox „Rahmen". Nun können gewünschte Festlegungen zu Farbe, Füllmuster etc. vorgenommen werden und/oder der Rahmen weggelassen werden.

Legende und Erläuterung können ausgeblendet werden. Klickt man mit der rechten Maustaste auf die Legende (bzw. Erläuterung), so wird diese mit einem Rahmen markiert und es öffnet sich ein Kontextmenü. Mit Klicken von „Legende" (bzw. „Erläuterung") ausblenden, verschwindet die Legende (bzw. Erläuterung).

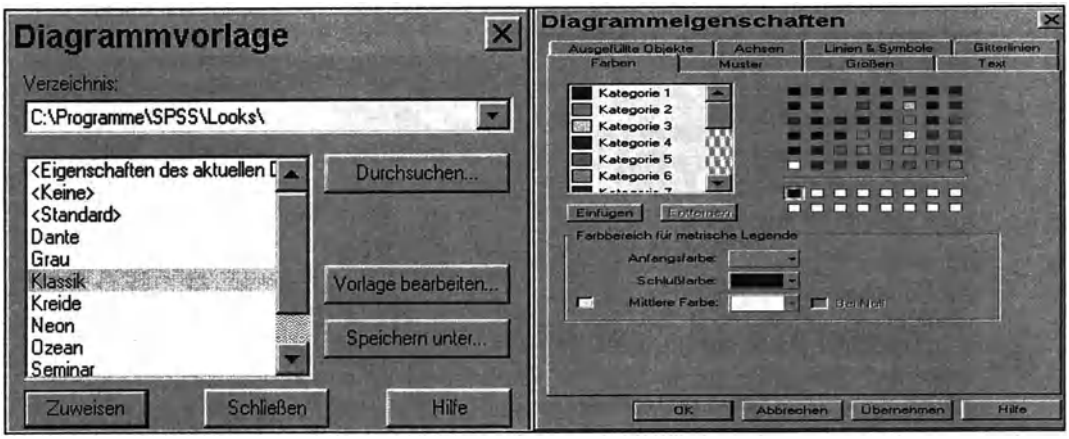


Abb. 25.14. Dialogboxen: „Diagrammvorlage" und „Diagrammeigenschaften"

Titeltexte und andere Beschriftungen überarbeiten. Bei der Erzeugung einer Grafik kann mit der Registerkarte „Titel" der Dialogbox zur Erstellung einer Grafik die Grafik mit Überschriften (Titel und Untertitel) und einer Fußnote (= Erklärung) versehen werden. Eine Grafik kann aber auch nachträglich mit Titel, Untertitel und Erklärung versorgt werden. Mit den Befehlen „Titel" (oder „Untertitel" bzw. „Erklärung") des Menüs „Einfügen", die bei aktivierter Grafik verfügbar sind, werden die Texte „Titel" (bzw. „Untertitel", „Erklärung") in die Grafik eingefügt. Um nun einen dieser Texte in einen gewünschten abzuändern, doppelklickt man mit der rechten Maustaste auf den Text. Es wird nun eine Markierung durch einen Rahmen angezeigt (z.B. Titel) und der Cursor in den Text eingefügt. Jetzt kann man den Text ändern. Auf gleiche Weise kann man Variablenlabel (Legendenüberschriften und Achsentitel) editieren. Die Texte können auch ausgeblendet werden (⇨ Kap. **25**.2.3).

Man kann auch einen beliebigen Text einfügen. Dazu klickt man auf das Symbol  und anschließend auf eine Stelle der Grafik. Es wird der Cursor sichtbar. Nun kann ein Text eingeben werden.

Sobald man bei aktivierter Grafik mit dem Cursor Texte bzw. Beschriftungen [Titeltexte, Fußnote, Achsenbeschriftungen (Achsentitel, Wertelabel, Skalenwerte), Erläuterung, Legendentexte, Fallzahlen, Datenwerte] markiert, wird auf der Symbolleiste 

aktiv geschaltet. Man kann die Schriftart und -größe ändern, auch fette oder kursive Schrift wählen.

Wird nach Markierung eines der verschiedenen Texte bzw. Beschriftungen die Befehlsfolge „Format" „Text" geklickt, so öffnet sich die Dialogbox „Text", die auch eine farbliche Gestaltung des Textes ermöglicht (= Alternative zur Verwendung der Symbolleiste, ⇨ unten).

Die Lage der Titel, Fußnote, Legende (auch Achsentitel und Kreissegmentbeschriftungen in 2D-Grafiken) kann verschoben werden. Man markiert den Text (Einfachklick mit linker Maustaste) und zieht ihn bei gedrückter linker Maustaste an die gewünschte Stelle. Will man die Texte wieder an ihre ursprüngliche Position bringen, so klickt man auf das Symbol  .

Symbolleiste zum Festlegen von Farben, Füll-, Symbol- und Linienmustern etc. verwenden. Mit  kann man Grafikobjekte [z.B. durch Daten dargestellte Balken, Kreissegmente, Punkte etc., Linien verschiedenster Art (Datenpunkte verbindende Linien, Achsen, deren Teilstriche, Verbindungslinien, Legenden- und Erläuterungsrahmen, Gitterlinien etc.) Texte bzw. Beschriftungen (Titeltexte, Fußnote, Legendentexte, Achsenbeschriftungen etc.] mit einer Füllfarbe versehen.

Um ein Objekt mit einer Füllfarbe zu versehen, markiert man zuerst das Objekt (oder mehrere: dann auch Shifttaste drücken), in dem man es mit der linken Maustaste anklickt (die Markierung wird durch einen Rahmen sichtbar). Mit Klicken auf  wird eine Palette geöffnet, aus der man die gewünschte Farbe auswählt. Diese Vorgehensweise bei der Vergabe einer Füllfarbe wird analog auch bei den nachfolgend beschriebenen Änderungen genutzt.

Mit  kann man Grafikobjekte (z.B. Balken, Kreissegmente, Legenden- und Erläuterungsrahmen) farblich umrahmen.

Mit  können Grafikobjekte (s.o.) mit einem Füllmuster versehen werden.

Mit kann man Datenpunkte (in Streuungsdiagrammen, Ausreißer und Extremwerte in Boxplots, mittlere Symbole von Fehlerbalken) mit anderen Symbolformen darstellen.

dient der Festlegung der Größe der durch Symbole dargestellten Datenpunkte in Streu- und Liniendiagrammen sowie Boxplots.

und erlauben es, für Linien einer Grafik (Linien einer Liniengrafik, aber auch Achsen, Gitterlinien, Rahmen von Legende und Erläuterung, Fehlerbalken, Whisker und Medianlinien eines Boxplots, Vorhersagelinien, Bezugslinien) Linienstile sowie die Stärke der Linien festzulegen.

dient dazu, Bezugslinien (Linien zwischen Grafikobjekten und ihren Beschriftungen, z.B. einem Kreissegment und dem Label der dargestellten Kategorie) zu verändern.

Hinweis. Änderungen der Farbe, des Musters, der Symbole, des Linienstils und der Größe sind in der angesprochenen Vorgehensweise nicht möglich, wenn die Farbe, das Muster usw. von den Daten kontrolliert wird. Am Beispiel eines gruppierten Balkendiagramm wie in Abb. 25.1 mit der Farbe als Legendenvariable sei dieses erklärt. Im Unterschied zu einem einfachen Balkendiagramm kann man nicht einen oder mehrere Balken markieren und dann mit die Füllfarbe ändern. Die Farbe der Balken einer Gruppe (z.B. der Männer) ist durch die Farbreihenfolge der Kategorien der Variable GESCHL festgelegt. Es ist daher nur möglich, alle Balken einer Gruppe farblich zu ändern. Dazu klickt man auf die Farbe einer Kategorie in der Legende. Dadurch werden alle Balken dieser Kategorie markiert. Wählt man mit eine andere Füllfarbe, so wird diese auf die Balken übertragen. Andererseits kann man in unserem Beispiel in der beschriebenen Weise aber wohl für einen einzelnen (auch für mehrere) Balken mit ein Füllmuster vergeben, da die Daten die Farbe aber nicht das Füllmuster kontrollieren.

Diagrammeigenschaften festlegen. Layoutmerkmale einer Grafik lassen sich mit Hilfe der Dialogbox „Diagrammeigenschaften" (Abb. 25.14 rechts) festlegen, die mit der Befehlsfolge „Format", „Diagrammeigenschaften" aufgerufen wird. Damit die auf den Registerkarten gemachten Festlegungen auf die aktive Grafik angewendet werden, wird auf der Dialogbox die Schaltfläche „Übertragen" oder „OK" geklickt. Die Dialogbox „Diagrammeigenschaften" kann auch zur Erstellung einer Diagrammformatvorlage genutzt werden (⇨ oben).

Registerkarte „Farben". Oben (Abb. 25.14 rechts) wird die Farbreihenfolge für Kategorien einer Kategorienachse (bis zu 16 Farben) festgelegt. Um z.B. die Farbe der ersten Kategorie zu ändern, klickt man auf „Kategorie 1" und dann auf eine der Farben aus der Farbpalette. Mit „Entfernen" bzw. „Einfügen" können Kategorienummern entfernt bzw. eingefügt werden. Unten wird die „Anfangs-„ und „Schlussfarbe" (und eventuell „Mittlere Farbe") für eine metrische Variable bestimmt. Mit dem Anstieg der Werte der Variable geht ein gradueller Übergang von der Anfangs- zur Schlussfarbe einher. Bei Wahl von „Mittlerer Farbe" wird für den mittleren Wertebereich der Variablen die hier gewählte Farbe in den Farbübergang einbezogen. Wenn die Werte der metrischen Variable sowohl positiv als auch negativ sind, so führt die Wahl der Option „Bei Null" dazu, dass vom Wert Null ausgehend die gewählte mittlere Farbe in die Anfangs- (negative Werte) und die Schlussfarbe (positive Werte) übergeht.

Registerkarte „Ausgefüllte Objekte". Für Grafikobjekte (Balken, Histogramme, Kreissegmente usw.), die aus einer Drop-Down-Liste ausgewählt werden können, kann man die Füllfarben und -muster sowie für Rahmen dieser Objekte Füllfarben und -muster sowie die Rahmenstärke bestimmen.

Registerkarte „Muster". Analog zu den Farben für Kategorien können hier Muster festgelegt werden. Je nach Wahl des Optionsschalters können Füllmuster, Symbolmuster oder Linienstile bestimmt werden.

Registerkarte „Achsen". Für Achsenlinien können Muster, Farbe und Stärke und für die verschiedene Teilstriche auf den Achsen Form, Farbe, Lage und Größe bestimmt werden.

Registerkarte „Größen". Für Symbole zur Darstellung von Punkten in Grafiken kann die Größe und für Linien die Linienstärke festgelegt werden.

Registerkarte „Linien & Symbole". Für verschiedene Linien in Grafiken („Fehlerbalken", „Medianlinie" usw.) können Muster, Farbe und Stärke der Linien bestimmt werden. Für durch Symbole dargestellte Punkte [in einer Punktwolke (= Streudiagramm), Box-Ausreißer bzw. -extremwerte usw.] kann man Symbolmuster, -farbe und -größe festlegen, für Bezugslinien (= Verbindungslinie zwischen Grafikobjekt und seiner Beschriftung) Muster und Farbe.

Registerkarte „Text". Für die verschiedenen Textarten einer Grafik (Diagrammtitel, Untertitel usw.) kann man Schriftart, -größe und -farbe sowie die Ausrichtung des Textes (links-, rechtsbündig oder zentriert) bestimmen.

25.2.3 Grafiklayout mit dem Diagramm-Manager gestalten

Grundlegendes. Eine ganze Reihe von Layoutgestaltungen können mit dem Diagramm-Manager vorgenommen werden. Klicken auf das Symbol  (oder mit Menü: „Bearbeiten", „Diagramm-Manager" oder über Kontextmenü: Rechtsklick auf Stelle außerhalb des Datenbereichs, „Diagramm-Manager") öffnet die Dialogbox „Diagramm-Manager" (⇨ Abb. 25.15 links). In der Dialogbox werden in einem Fenster (analog einem Verzeichnisbaum mit Verzeichnissen und darin enthaltenen Dateien) Diagrammteile mit ihren einzelnen Bestandteilen aufgeführt:

- *Zeichnungsfläche.* Sie enthält den Datenbereich (der Hintergrund der Balken, Linien, etc.) sowie die Achsen der Grafik. Dabei kann es sich um Kategorien- oder/und Skalenachsen handeln
- *Text.* Enthält Titel (eventuell auch Untertitel) und die Diagrammerläuterung (z.B. „Balken zeigen Mittelwerte" in Abb. 25.15 links).
- *Legende.* Enthält in unserem Beispiel „Farblegende" (⇨ Abb. 25.15 links). Es kann aber auch „Muster"- und/oder „Größenlegende" enthalten sein, wenn in der Grafik weitere Untergliederungen vorgenommen worden sind.
- *Elemente.* Enthält den Daten darstellenden Grafikgrundtyp (hier: „Balken" und „Punktwolke"). In Streudiagrammen können auch eine Regressions- oder Mittelwertanpassung oder/und LLR-Glättung enthalten sein.

Alle Grafikbestandteile können nach Markieren mit der Maus (in Abb. 25.15 links ist „Balken" markiert) bearbeitet werden. Klicken des Optionsschalters „Ausblenden" führt dazu, dass der Grafikteil ausgeblendet wird. Klicken auf die Schaltfläche „Löschen" löscht den Grafikteil. Klicken auf die Schaltfläche „Bearbeiten" öffnet eine Dialogbox zur Auswahl von Gestaltungsmöglichkeiten für den gewählten Grafiktyp. In Abb. 25.15 rechts ist z.B. die Dialogbox „Balken" zu sehen, die sich öffnet, wenn nach Markieren des Elements „Balken" in Abb. 25.15 links die Schaltfläche „Bearbeiten" geklickt wird. Im folgenden werden einige Gestaltungsmöglichkeiten aufgezeigt.

Datenbereich gestalten. Markieren von „Datenbereich" und Klicken auf „Bearbeiten" öffnet die Dialogbox „Datenbereich" (Abb. 25.16. links). Man kann für den Hintergrund der Grafik die Farbe wählen und ein Füllmuster einfügen. Des weiteren kann man die Größe der Grafik bestimmen (mit oder ohne Beibehaltung der Seitenproportionen) und für die Achsenbeschriftung sowie die Achsentitel wählen, ob diese in der Achsen- oder Bildschirmebene liegen sollen.

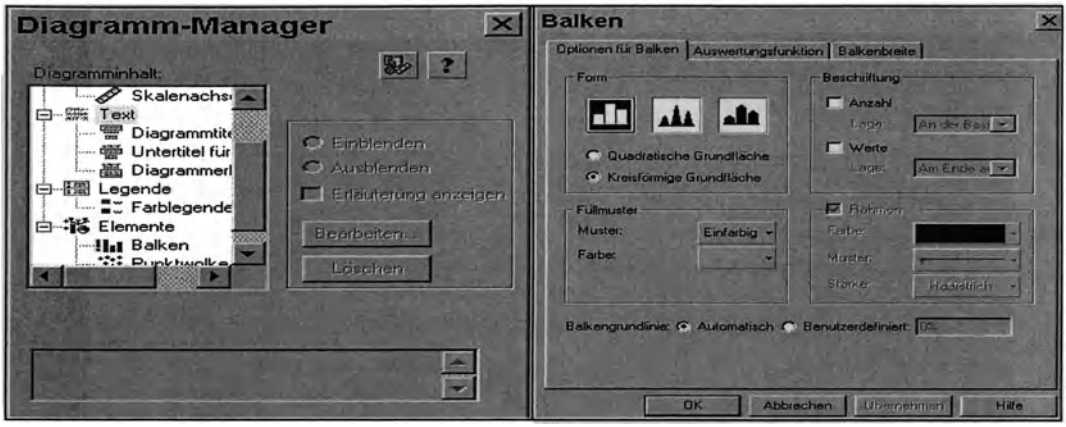


Abb. 25.15. Dialogboxen „Diagramm-Manager" und „Balken"

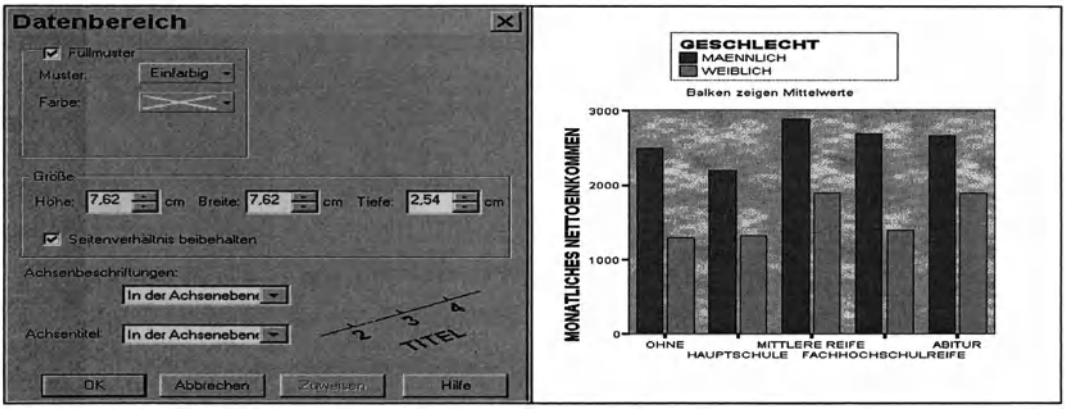


Abb. 25.16. Vergabe eines Füllmuster für den Datenbereich einer interaktiven Grafik

Kategorien- und Skalenachse überarbeiten. Markieren von „Kategorien"- bzw. „Skalenachse" und Klicken auf „Bearbeiten" öffnet die entsprechende Dialogbox (Abb. 25.17 links und rechts). Beide Dialogboxen verfügen über Registerkarten. Eine Reihe von Eigenschaften der Achsen können bestimmt werden: Art der Teilstriche (Form, Farbe, Lage, Größe), Art der Achsenlinien (Muster, Stärke, Farbe), Art der Achsenbeschriftungen (Anordnung und Häufigkeit), Anzeige und Ausrichtung von Achsentiteln, Anzeige und Art von Gitterlinien, Aufbau der Skala metrischer Achsen etc.

Diagrammtitel, Untertitel, Diagrammerläuterung überarbeiten. Markieren eines dieser Textarten und Klicken auf „Bearbeiten" öffnet die Dialogbox „Text" mit den Registerkarten „Schriftart" und „Numerisches Format" Mit der Registerkarte „Schriftart" kann Schriftart und Größe der Texte festgelegt werden. Dafür kann man aber auch das Textbearbeitungstool auf der Symbolleiste verwenden (⇨ oben: Titeltexte und andere Beschriftungen bearbeiten). Auf der Registerkarte „Numerisches Format" kann man je nach Format der Variable (numerisch, Datum, Uhrzeit, Währung) eine gewünschte Einstellung bestimmen.

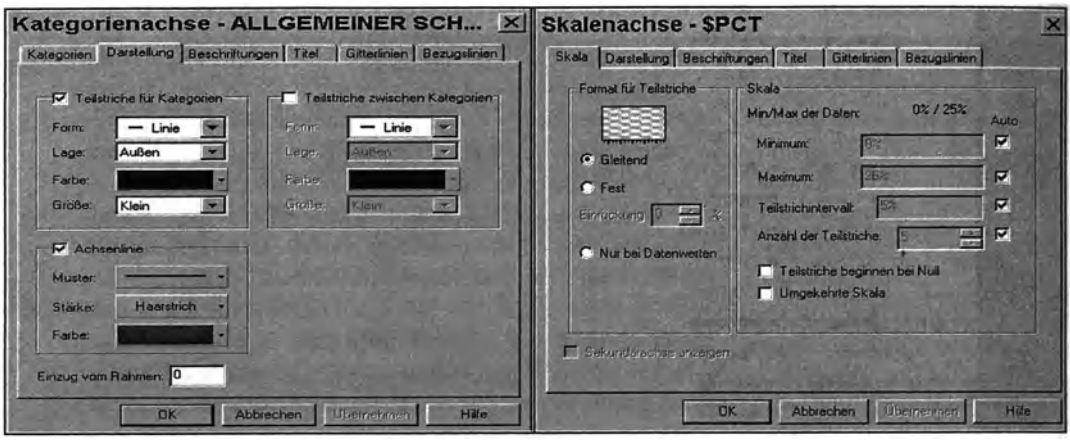


Abb. 25.17. Dialogboxen „Kategorien-„ und „Skalenachse"

Legenden überarbeiten. Je nach Zuweisung von Variablen zu „Farbe", „Muster" und „Größe" von „Legendenvariablen" bei der Erzeugung von Grafiken entstehen Farb-, Muster- bzw. Größenlegenden, die überarbeitet werden können. Markieren einer Legende und Klicken auf „Bearbeiten" öffnet die entsprechende Dialogbox für die Legendenart. In Abb. 25.18 links und rechts sind die Dialogboxen für die Farb- und Musterlegende zu sehen. Um z.B. die Farbe für die Kategorie „MAENLICH" zu ändern, wird auf der Registerkarte „Farben" diese Kategorie angeklickt und aus der Farbpalette „Kategorienfarbe" eine Farbe gewählt. Außerdem kann für die Legende, die Musterart (eine Box, ein Kreis etc.) und -größe geändert sowie eine Rahmenfarbe gewählt werden. Die Registerkarte „Titel" erlaubt es, die Anzeige des Legendentitels zu unterbinden und Lage und Ausrichtung des Legendentitels zu bestimmen. Auf der Registerkarte „Optionen" können im Fall einer kategorialen Legende mit mehr als einer Spalte die Einträge in Zeilen

(erst in Zeilen, dann in Spalten) oder in Spalten (erst in Spalten, dann in Zeilen) geordnet werden. Alle Einträge können auf die Höhe des größten Eintrags festgelegt werden (bei Legenden mit mehreren Spalten nützlich).

Die Bearbeitung einer Muster- oder Größenlegende ist in analoger Weise vorzunehmen.

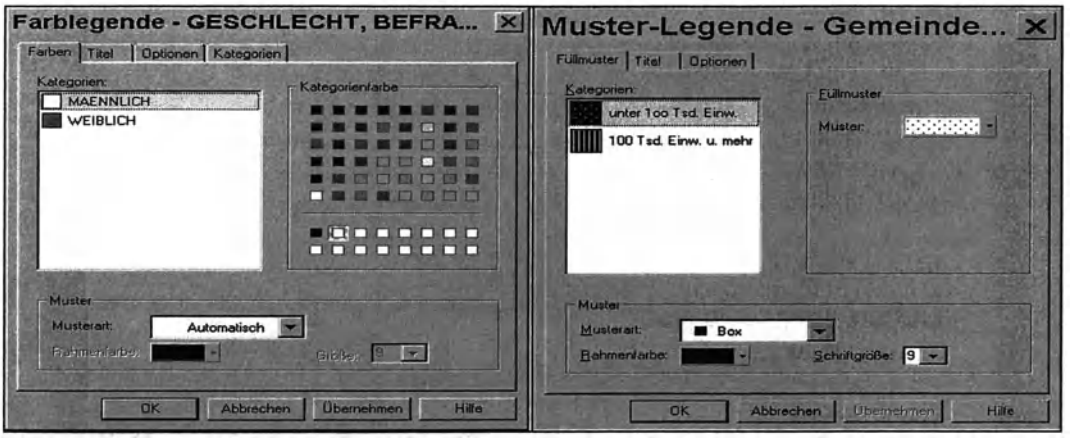


Abb. 25.18. Dialogboxen „Farb- und Muster-Legende"

Durch Daten dargestellte Grafikelemente überarbeiten. Je nach erzeugtem Grafikgrundtyp und eventuell eingefügten weiteren Grundtypen (bzw. eingefügter Regressions- und Mittelwertanpassung oder LLR-Glättung, LLR = lokale Regressionsfunktion) enthält die Rubrik „Elemente" im Fenster des Diagrammmanagers ein oder mehrere dieser Grafikelemente, die überarbeitet werden können. In Abb. 25.15 links ist die Dialogbox „Diagramm-Manager" für eine gemischte Grafik zu sehen. Sie zeigt „Balken" und „Punktwolke als Grafikelemente an. Im Unterschied zu den oben besprochenen Grafikteilen (Datenbereich, Achsen, Titel, Legenden usw.) sind diese Grafikelemente durch Daten dargestellt. Die Rubrik „Elemente" kann als Grafikelemente also Balken, Linie, Kreis, Fehlerbalken usw. enthalten. Um diese zu überarbeiten, markiert man das Grafikelement (z.B. Balken) mit der Maus und klickt auf die Schaltfläche „Bearbeiten". Es öffnet sich dann eine entsprechende Dialogbox zur Bearbeitung. Beachten Sie aber, dass Eigenschaften wie Farbe, Füllmuster etc. von einzelnen Balken, Datenpunkten, einer Regressionslinie etc. mit Hilfe der Symbole auf der Symbolleiste oder mit dem Befehl „Eigenschaften" des Kontextmenüs geändert werden.

Balken überarbeiten. In Abb. 25.15 ist rechts die Dialogbox „Balken" aufgeführt. Auf der Registerkarte „Optionen für Balken" lässt sich die Form, Füllmusterart und-farbe, Muster, Farbe und Stärke von Umrahmungen der Balken festlegen. Außerdem kann festgelegt werden, ob die Fallzahlen („Anzahl") und/oder die auf der Y-Achse dargestellten Werte angezeigt werden sollen.

Die Registerkarte „Auswertungsfunktion" ermöglicht es, die Auswertungsfunktion einer metrischen Variable auf der Y-Achse zu ändern. Insofern kann hier eine grundlegende Grafikänderung, die nicht das Layout betrifft, vorgenommen

werden. So könnte man z.B. die in Abb. 25.3 dargestellten Mittelwerte in z.B. Mediane verändern.

Auf der Registerkarte „Balkenbreite" können Balken- und Gruppenbreite festlegt werden.

Linien und Punkte überarbeiten. In Abb. 25.19 links ist die Dialogbox „Punkte und Linien" zu sehen. Sie dient der Überarbeitung von Punkt-, Linien-, Band- und Verbundliniendiagrammen. Die Registerkarte „Optionen" erlaubt es festzulegen, wie die in der Grafik dargestellten Datenpunkte und die Datenpunkte verbindenden Linien dargestellt werden sollen. Hinsichtlich der Registerkarte „Auswertungsfunktion" sei nach oben verwiesen (⇨ Balken überarbeiten). Auf der Registerkarte „Beschriftungen" wird festgelegt ob und welche Beschriftungen die Grafik für Punkte bzw. Linien enthalten soll. Die Registerkarte „Verbundlinien" dient dazu, die Anzeige von Verbundlinien sowie ihre Form, Farbe und Stärke festzulegen.

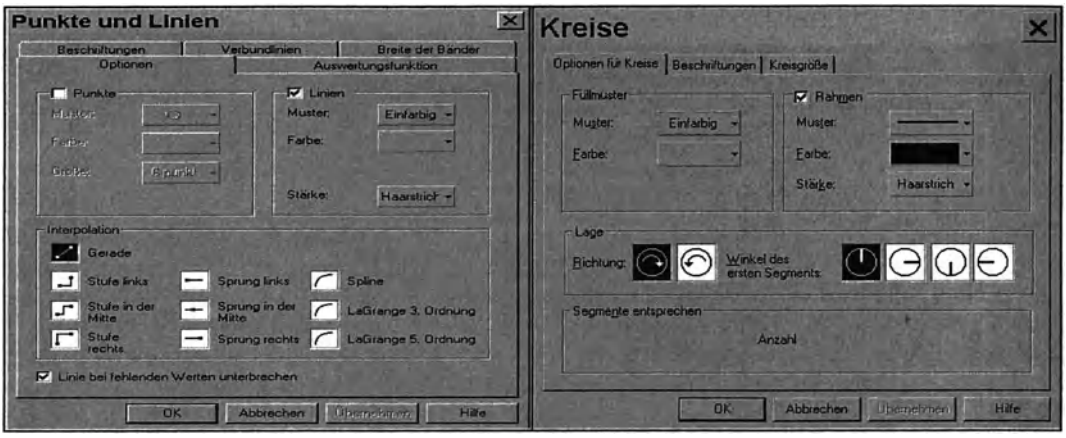


Abb. 25.19. Dialogboxen „Punkte und Linien" und „Kreise"

Kreise überarbeiten. In Abb. 25.19 rechts ist die Dialogbox „Kreise" dargestellt. Auf der Registerkarte „Optionen" werden Gestaltungsmöglichkeiten zu Füllmuster, Rahmen und der Lage von Kreissegmenten geboten. Die Registerkarte „Beschriftungen" erlaubt unterschiedliche Arten und Formen für Beschriftungen der Kreissegmente und bei gruppiertem Kreis auch für die Gruppen. Die Registerkarte „Kreisgröße" erlaubt es, den Kreisdurchmesser sowie die Tiefe in einer 3D-Darstellung festzulegen.

Boxplots überarbeiten. Abb. 25.20 links zeigt die Dialogbox „Boxen". Auf der Registerkarte „Box-Optionen" können für die Boxplots sowie deren Rahmen und die Medianlinie Füllmuster und Farbe, für die Rahmen und Medianlinie auch die Stärke dieser Linien bestimmt werden. In der 3D-Darstellung kann zwischen eckigen und runden Boxplots gewählt werden. Die Fallhäufigkeiten können angezeigt oder ausgeblendet werden.

Die Registerkarten „Whiskers“ bzw. „Ausreißer und Extremwerte“ ermöglichen Layoutgestaltungen für die Whisker-Linien bzw. die Ausreißer und Extremwerte.

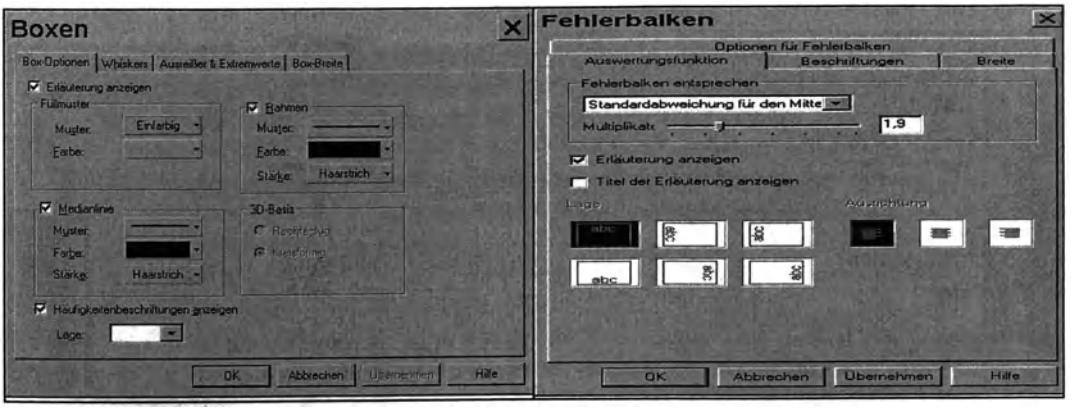


Abb. 25.20. Dialogboxen „Boxen“ und „Fehlerbalken“

Fehlerbalken überarbeiten. In Abb. 25.20 rechts ist die Dialogbox „Fehlerbalken“ zu sehen. Mit den Wahlmöglichkeiten auf den Registerkarten können eine Reihe von Layoutgestaltungen für Fehlerbalken vorgenommen werden. Auf der Registerkarte „Optionen für Fehlerbalken“ kann u.a. für die Fehlerbalken ein anderer Streuungsausdruck zur Darstellung von Fehlerbalken gewählt werden.

Histogramme überarbeiten. In Abb. 25.21 ist links die Dialogbox „Histogramm“ abgebildet. Für die Histogramme kann Farbe und Muster, für die die Rahmen der Histogramme sowie für eine eventuell darüber gelegte Normalverteilungskurve außer Farbe auch Strichmuster und -stärke festgelegt werden. Durch Klicken der Schaltfläche „Intervallfunktion“ öffnet sich eine Unterdialogbox, die es ermöglicht, die Intervallanzahl oder Intervallbreite sowie den Anfangspunkt der Grafik auf der X-Achse festzulegen.

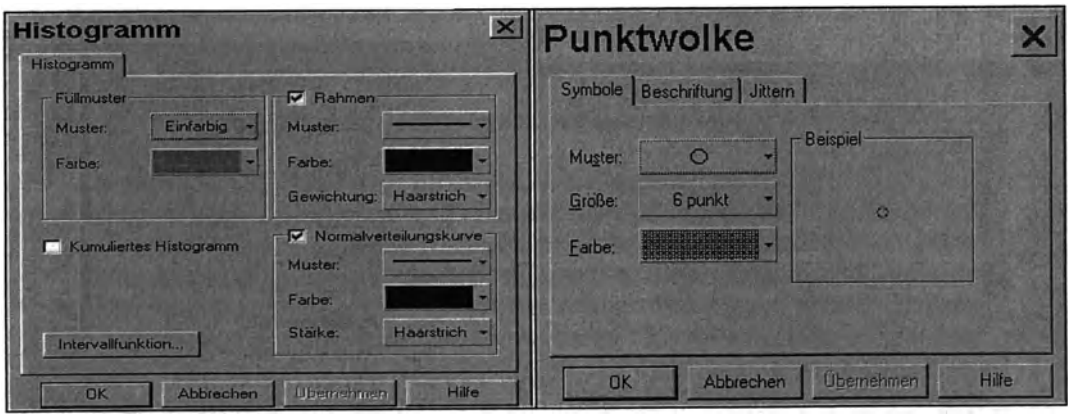


Abb. 25.21. Dialogboxen „Histogramm“ und „Punktwolke“

Punktwolke überarbeiten. In Abb. 25.21 rechts ist die Dialogbox „Punktwolke" abgebildet. Auf den Registerkarten kann man eine Reihe von Layoutgestaltungen vornehmen. Auf der Registerkarte „Jittern" kann man festlegen, in welchem Ausmaß Datenpunkte, die übereinanderliegen in der Grafik nebeneinander dargestellt werden.

Mittelwert-, Regressions- und LLR-Linien überarbeiten. Streudiagramme können - auch für Untergruppen - mit Mittelwert-, Regressions- und/oder LLR-Linien (local linear regression) versehen werden. Die Mittelwertlinie kann mit einem Konfidenz- und die Regressionslinie einem Vorhersageintervall angezeigt werden. Für die LLR-Glättunglinie können die Bestimmungsparameter verändert werden.

26 Herkömmliche Grafiken erzeugen

26.1 Einführung und Übersicht

Grafiken werden mit dem Menü „Grafiken" erzeugt. Sie erscheinen dann wie jede andere Ausgabe im Ausgabefenster (vor der SPSS-Version 7.5 im Grafikkarussell). Wie jede andere Ausgabe kann eine erzeugte Grafik gedruckt, gelöscht, in die Zwischenablage kopiert (z.B. zum Transfer in die Textverarbeitung, (⇨ Kap. 26.9) sowie als Bestandteil der gesamten Ausgabe in einer Ausgabedatei (Navigator-Dokument) gespeichert werden. Auch ein Export in ein anderes Grafikformat ist möglich (⇨ Kap. 27.1).

Ab der Version 8.0 können neben den herkömmlichen (Standarddiagramme) auch interaktive Grafiken erzeugt werden. Das Erstellen von Grafiken unterscheidet sich bei herkömmlichen und interaktiven Grafiken. Interaktive Grafiken können in dynamischer Weise im Ausgabefenster modifiziert werden und bieten zusätzliche Möglichkeiten der Grafikerstellung und -überarbeitung. Herkömmliche und interaktive Grafiken sind zum großen Teil vom gleichen Grafiktyp. Allerdings decken interaktive Diagrammtypen nicht alle herkömmlichen Grafiken ab. QQ-, PP-, Pareto-, Regelkarten-, Sequenz-, Autokorrelations-, Kreuzkorrelationsdiagramme und ROC-Kurven gibt es nur als herkömmliche Grafiken. Andererseits gibt es als interaktive Grafiken auch einige wenige neue Grafikgrundtypen (Punkt-, Banddiagramm). Interessante Neuerungen sind des weiteren, dass interaktive Grafiken als echte dreidimensionale Grafiken erstellbar sind und für alle zweidimensionalen ein 3D-Effekt möglich ist. Herkömmliche Grafiken bieten in Ausnahmefällen im Vergleich zu interaktiven aber auch Vorzüge in den Darstellungsmöglichkeiten: überlagerte Streudiagramme sind als interaktive nicht verfügbar.

Für eine herkömmliche Grafik gilt, dass sie erst durch Übergabe der Grafik in ein spezielles Fenster - den Diagramm-Editor - überarbeitet und in eine präsentationsreife Form gebracht werden kann. Danach kann sie gedruckt oder über die Zwischenablage in ein Textverarbeitungsprogramm übergeben werden. Möglich ist auch ein Export der Grafik in ein neues Grafikformat.

In diesem Kapitel wird im einzelnen auf das Erstellen der verschiedenen herkömmlichen Grafiken eingegangen. Das Überarbeiten von herkömmlichen Grafiken in eine Präsentationsform (Layoutgestaltung) wird in Kapitel 27 behandelt. In Kap. 25 wird das Erzeugen und Modifizieren von interaktiven Grafiken dargestellt. Dabei beschränken wir uns auf die grundsätzlichen Aspekte. Tabelle 26.1 zeigt diese Zuordnungen in einer Übersicht.

Tabelle 26.1. Darstellung der Grafikerzeugung und -bearbeitung in Kapiteln

Kapitel	Inhalte
25	Interaktive Grafiken: Erzeugen, Layout gestalten
26	Herkömmliche Grafiken: Erzeugen
27	Herkömmliche Grafiken: Layout gestalten

Durch die Befehlsfolge „Grafiken", „Galerie" öffnet sich ein spezielles Hilfefenster (⇨ Abb. 26.1) zur sehr guten Anleitung für die Grafikerstellung. Probieren Sie es aus, in dem Sie auf das Symbol eines Diagrammtyps klicken.

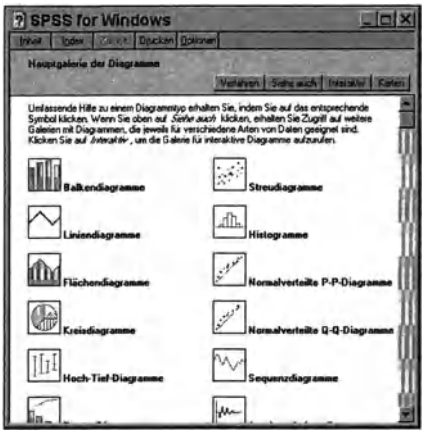


Abb. 26.1. Hilfefenster zur Erzeugung herkömmlicher Grafiken

26.2 Balkendiagramme erzeugen

Um ein Balkendiagramm zu erstellen, öffnet man die in Abb. 26.1a dargestellte Dialogbox durch Klicken der Befehlsfolge

▷ „Grafiken", „Balken...".

Als Balkendiagrammformen sind ein *einfaches*, ein *gruppiertes* oder ein *gestapeltes* Balkendiagramm wählbar. Dabei können für die Grundachse des Balkendiagramms alternativ drei Grafikdaten abgebildet (ausgewertet) werden: „...Kategorien einer Variablen", „...verschiedene Variablen", oder „Werte einzelner Fälle" (⇨ Abb. 26.1a).

Je nach Datenlage und gewünschtem Diagramm kann also jeder Balkendiagramm-Typ mit jeder Auswertungsform durch Anklicken kombiniert werden. Im folgenden werden einige dieser verschiedenen Balkendiagrammformen anhand von Beispielen aus dem ALLBUS90-Datensatz erläutert.

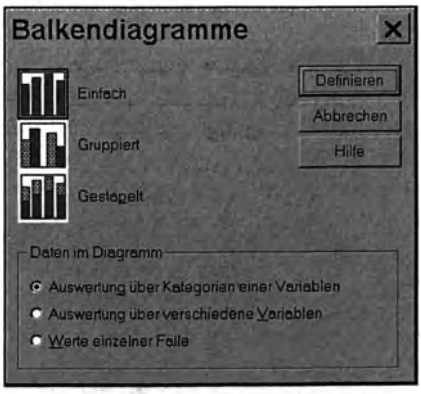


Abb. 26.1a. Dialogbox zur Auswahl eines Balkendiagramms

26.2.1 Einfaches Balkendiagramm

Auswertung über Kategorien einer Variablen. Nach der Befehlsfolge „Grafiken", „Balken..." wird die Auswahlkombination „Einfach" und „Auswertung über Kategorien einer Variablen" angeklickt. Nach Klicken von „Definieren" öffnet sich die in Abb. 26.2 links dargestellte Dialogbox. Die Abbildung zeigt ein Beispiel zur Grafikdefinition und die resultierende Grafik. Die Variable SCHUL mit verschiedenen Schulabschlüssen als Kategorien wurde aus der Quellvariablenliste in das Eingabefeld „Kategorienachse:" übertragen. Die Balkenhöhe entspricht hier der prozentualen Häufigkeit, da „% der Fälle" gewählt worden ist.

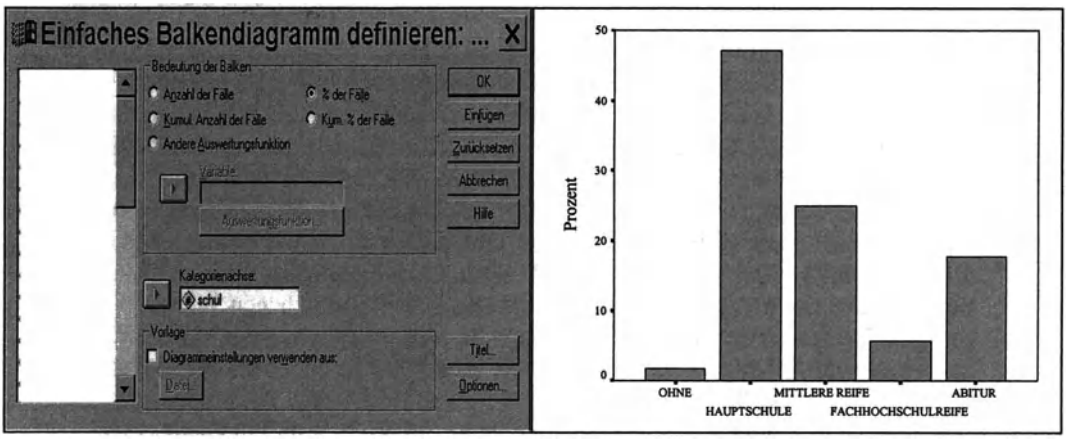


Abb. 26.2. Häufigkeiten von Schulabschlüssen der Befragten

Zur Darstellung der Höhe der Balken („Bedeutung der Balken") sind folgende alternative Optionen gegeben:

- Die Häufigkeit der Kategorien in verschiedener Form:
  - *Anzahl der Fälle.*

- *% der Fälle.*
- *Kum. Anzahl Fälle* (kumulierte absolute Häufigkeit).
- *Kum % Fälle* (kumulierte prozentuale Häufigkeit).

❑ *Andere Auswertungsfunktion.* Es kann z.B. der Mittelwert, der Median etc. einer weiteren Variablen als Balkenhöhe gewählt werden. In der folgenden Abb. 26.3 entspricht für jede Kategorie der Schulausbildung die Höhe der Balken dem arithmetischen Mittel des Alters der Befragten. In der Dialogbox wurde „Andere Auswertungsfunktion" gewählt und die Variable ALT in das Eingabefeld „Variable" übertragen. Standardmäßig wird „MEAN" der Variable - das arithmetische Mittel - als Auswertungsfunktion vorgeschlagen.

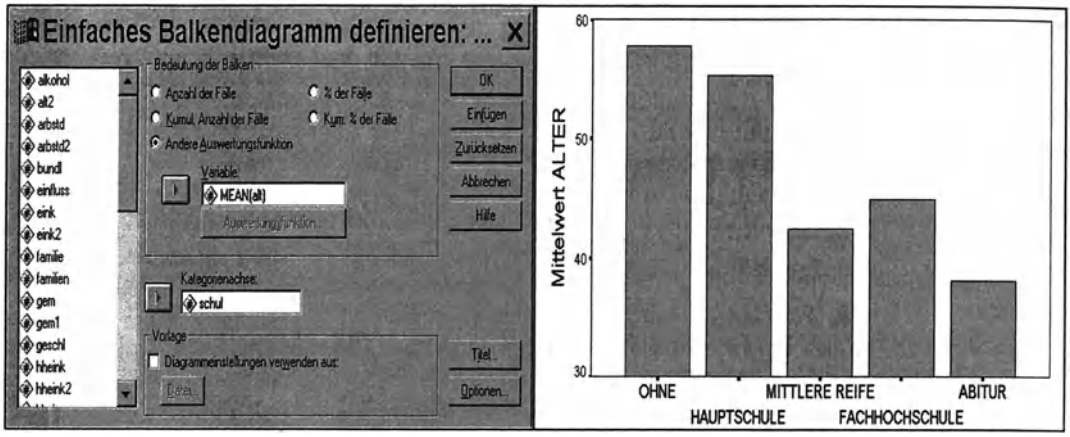


Abb. 26.3. Durchschnittliches Alter der Befragten nach Schulabschlüssen

Möchte man anstelle von „MEAN" eine andere Auswertungsfunktion wählen, so klickt man auf die Schaltfläche „Auswertungsfunktion". Es öffnet sich dann die in Abb. 26.4 dargestellte Dialogbox. Wählbar sind: der „Median", der „Modalwert", die „Anzahl der Fälle", die „Summe der Fälle", die „Standardabweichung", die „Varianz", das „Minimum", das „Maximum" sowie die „kumulierte Summe".

Des weiteren sind wählbar: „Prozent ober- bzw. unterhalb", „Anzahl ober- bzw. unterhalb" sowie „Perzentile". Wird eine dieser Möglichkeiten durch Klic??ken gewählt, so wird das Eingabefeld „Wert:" aktiv geschaltet. Nach Eingabe eines Wertes wird die gewählte Funktion auf Basis des eingegebenen Wertes ausgewertet. Beispielsweise ließe sich die Balkenhöhe für jeden Schulabschluss durch die Anzahl der Befragten mit einem Alter größer („Anzahl oberhalb") als 45 („Wert" = 45) darstellen.

Schließlich kann man auch „Prozent innerhalb" bzw. „Anzahl innerhalb" wählen. Nach der Wahl einer dieser beiden Möglichkeiten können Werte in die aktivierten Eingabefelder „Min:" und „Max:" zur Angabe von Minimum und Maximum eingetippt werden.

Anwendung auf klassifizierte Daten. Für den Fall, dass der Median oder Perzentile als Balkenhöhe für klassifizierte Daten dargestellt werden sollen, muss „Werte

sind gruppierte Mittelpunkte“ (⇨ Abb. 26.4) eingeschaltet sein, da sonst nur der Wert der Einfallsklasse als Wert dargestellt wird (⇨ Kap. 8.3.1).

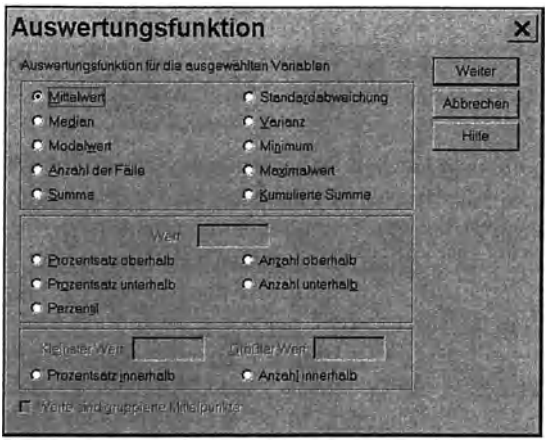


Abb. 26.4. Dialogbox zur Auswahl von Auswertungsfunktionen

Globale optionale Festlegungen. Optional können in der Dialogbox der Abb. 26.2 weitere Elemente festgelegt werden

① *Titel.* Zur Versorgung des Balkendiagramms mit Titeln und Fußnoten wird vor der Erzeugung der Grafik durch Klicken auf die Schaltfläche „Titel“ eine Dialogbox zur Titel- und Fußnotenvergabe geöffnet. Abb. 26.5 zeigt diese Dialogbox mit beispielhaften Eintragungen sowie das Ergebnis der damit ergänzten Grafik.

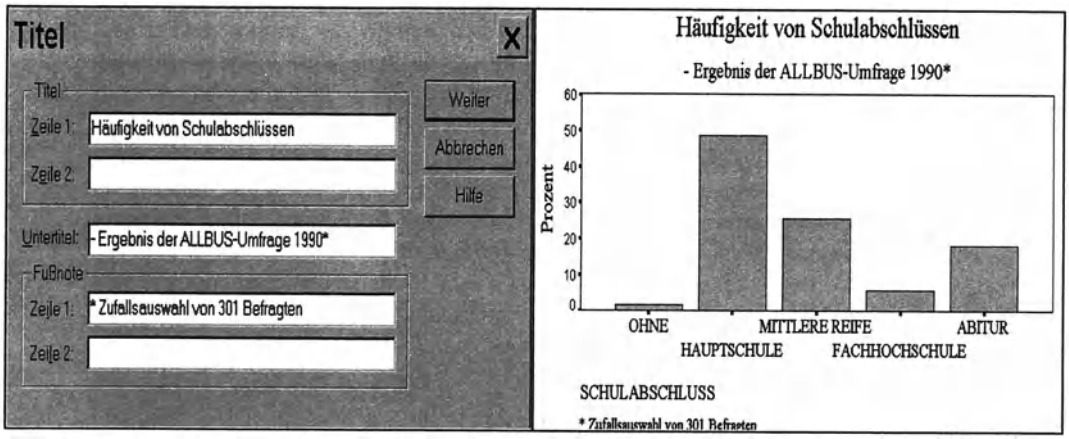


Abb. 26.5. Titel, Untertitel und Fußnote eines Balkendiagramms bestimmen

② *Optionen.* Abb. 26.6 zeigt die nach Klicken von „Optionen“ geöffnete Dialogbox zur Behandlung von fehlenden Werten bei Erstellung des Balkendia-

gramms. Im Auswahlbereich „Missing-Werte“ (fehlende Werte) sind prinzipiell zwei Alternativen gegeben:

- *Listenweiser Fallausschluss*. Grundsätzlich wird bei dieser Option ein Fall für alle Variablen ausgeschlossen, wenn eine der für die Grafik benötigten Variablen einen fehlenden Wert hat. Im obigem Beispiel ist die Option voreingestellt und kann auch nicht verändert werden, da bei der Darstellung von nur einer Variablen die andere Option keinen Sinn ergibt.
- *Fälle Variable für Variable ausschließen*. Es werden nur jeweils die Fälle von Variablen ausgeschlossen, bei denen Werte fehlen.

Des weiteren kann man durch Anklicken des Schalters „Fehlende Werte als Kategorie anzeigen“ die Voreinstellung, dass fehlende Werte im Diagramm als Kategorie auf der Kategorienachse bzw. in der Legende aufgenommen werden, ausschalten. In den Beispielen wurde so verfahren.

Die Option „Grafik mit Fallbeschriftungen anzeigen“ kann hier nicht aktiviert werden, da in dieser Grafik nicht einzelne Fälle angesprochen werden können. Möglich ist dieses nur bei Scatter- und Boxplots.

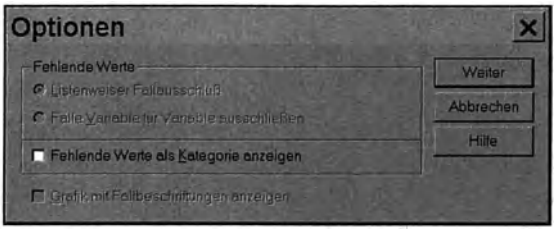


Abb. 26.6. Dialogbox „Grafik-Optionen“ zur Behandlung von fehlenden Werten

③ *Vorlage*. Beim Anfertigen einer Grafik können Titel, Fußnoten, Farben, Schriftgrößen und weitere Layoutmerkmale aus einer schon früher erstellten und gespeicherten Grafik übernommen werden. Wird in der Dialogbox zur Definition einer Grafik (⇨ Abb. 26.2) die Option „Diagrammeinstellungen verwenden aus:“ durch Mausklick gewählt, so wird die Schaltfläche „Datei“ aktiv geschaltet. Nach Mausklick auf „Datei“ öffnet sich eine Dialogbox zur Auswahl einer vorbereiteten Grafikvorlagendatei. Nach Klicken von „Öffnen“ werden die Layoutmerkmale der gewählten Grafikvorlagendatei für die aktuell zu erstellende Datei übernommen. Eine Grafikvorlagendatei wird vorbereitet, indem eine mit gewünschten Layoutmerkmalen erstellte Grafik mit der Befehlsfolge *„Datei“, „Diagrammvorlage speichern“* gespeichert wird. Standardmäßig wird für Grafiken die Endung .SCT vorgegeben.

Auswertung über verschiedene Variablen. Nach der Befehlsfolge „Grafik“, „Balken...“ wird die Auswahlkombination „Einfach“ und „Auswertung über verschiedene Variablen“ gewählt. Abb. 26.7 zeigt die nach Klicken von „Definieren“ geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition und die resultierende Grafik. In dem Beispiel sind die Mittelwerte des Alters und der Arbeitsstun-

den/Woche der arbeitenden Befragten dargestellt (Selektion mit ARBSTD > 0 im Menü „Daten"). Aus dem Quellvariablenfeld wurden die Variablen ALT und ARBSTD in das Eingabefeld „Bedeutung der Balken" übertragen. Die Funktion „mean" (Mittelwert) ist voreingestellt. Durch Anklicken von „Auswertungsfunktion" kann eine andere Berechnungsfunktion für die Balkenhöhe gewählt werden (⇨ Abb. 26.4).

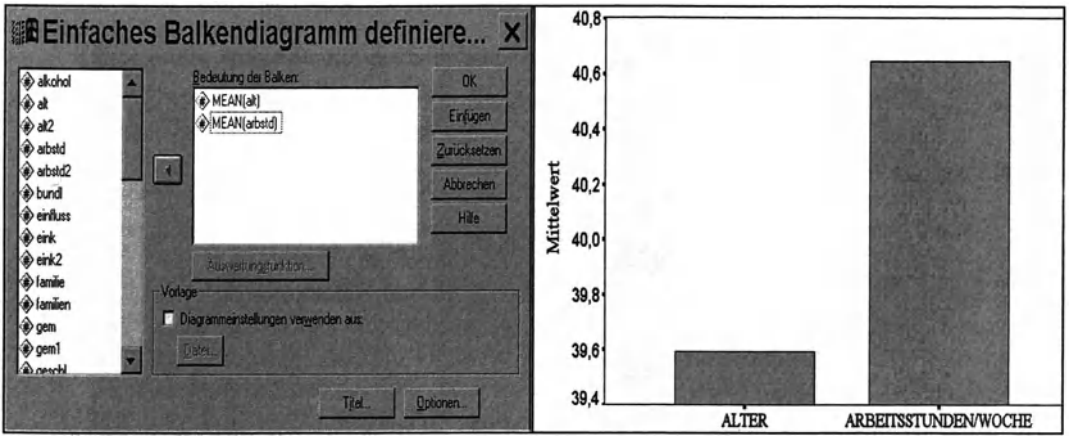


Abb. 26.7. Mittelwert von Alter und Arbeitsstunden/Woche

Werte einzelner Fälle. Nach der Befehlsfolge „Grafik", „Balken..." wird die Auswahlkombination „Einfach" und „Werte einzelner Fälle" gewählt. In der nach Klicken von „Definieren..." geöffneten Dialogbox wird die darzustellende Variable aus der Quellvariablenliste in das Eingabefeld „Bedeutung der Balken:" übertragen. Im Auswahlfeld „Achsenbeschriftung" kann „Fallnummer" oder „Variable" (dann wäre eine Variable in das Eingabefeld zu übertragen) gewählt werden. Im ersten Fall werden die Fallnummern und im zweiten Fall die Labels der Variable zur Achsenbeschriftung verwendet.

26.2.2 Gruppiertes Balkendiagramm

Auswertung über Kategorien einer Variablen. Nach der Befehlsfolge „Grafiken", „Balken...", Klicken der Auswahlkombination „Gruppiert" und „Auswertung über Kategorien einer Variablen" und Klicken von „Definieren" öffnet sich die in Abb. 26.8 dargestellte Dialogbox. Die Abbildung zeigt ein Beispiel zur Grafikdefinition und die resultierende Grafik. In dem Beispiel sind die Häufigkeiten von Schulabschlusskategorien nach dem Geschlecht der Befragten untergliedert. Als Variable für die Kategorienachse wurde wieder SCHUL und als Gruppenvariable GESCHL in die entsprechenden Eingabefelder übertragen.

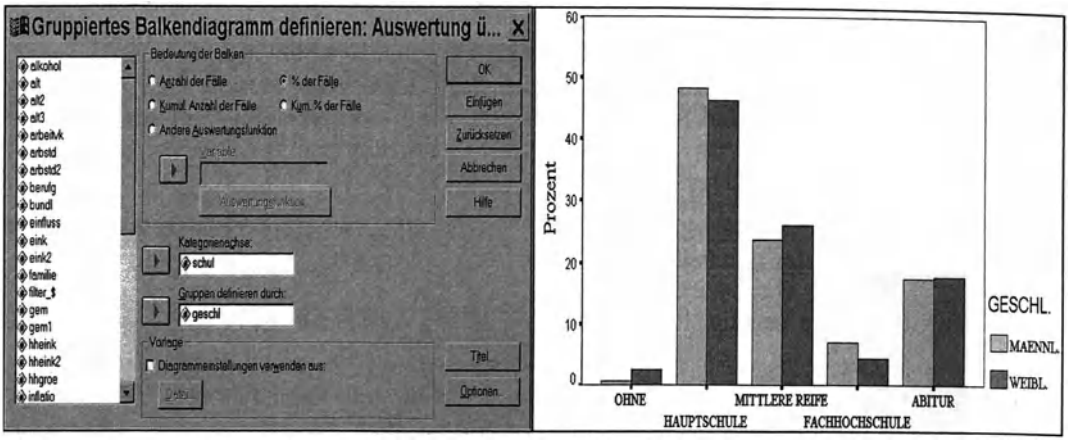


Abb. 26.8. Häufigkeit von Schulabschlüssen untergliedert nach dem Geschlecht der Befragten

Auswertung über verschiedene Variablen. Die Vorgehensweise entspricht der bei der Erstellung eines einfachen Balkendiagramms. Im Unterschied dazu wird natürlich ein gruppiertes Balkendiagramm gewählt, und es wird eine (Gruppen-) Variable in das Eingabefeld von „Kategorienachse" (= Grundachse) übertragen. Für jeden Fall werden Werte [z.B. die Mittelwerte von ALT (Alter) und ARBSTD (Arbeitsstunden)] für jede Kategorie der Gruppenvariablen (z.B. GESCHL) dargestellt.

Werte einzelner Fälle. Man geht analog zu den einfachen Balkendiagrammen vor.

26.2.3 Gestapeltes Balkendiagramm

Auswertung über Kategorien einer Variablen. Nach der Befehlsfolge „Grafiken", „Balken..." wird die Auswahlkombination „Gestapelt" und „Auswertung über Kategorien einer Variablen" angeklickt. Abb. 26.9 zeigt die durch Klicken von „Definieren..." geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition und die resultierende Grafik. Es wird das gleiche Beispiel wie in Abb. 26.12 gewählt. Als Variable auf der Kategorienachse wurde wieder SCHUL und als „Stapelvariable" GESCHL in die entsprechenden Eingabefelder übertragen.

Auswertung über verschiedene Variablen bzw. Werte einzelner Fälle. Nach der Befehlsfolge „Grafiken", „Balken..." wird die Auswahlkombination „Gestapelt" und „Auswertung über verschiedene Variablen" bzw. „Werte einzelner Fälle" angeklickt. Die nach Klicken von „Definieren..." geöffneten Dialogboxen haben die gleichen Eingabefelder und optionalen Möglichkeiten wie im Fall gruppierter Balken (⇨ Abb. 26.8). Auf Anwendungsbeispiele wird verzichtet.

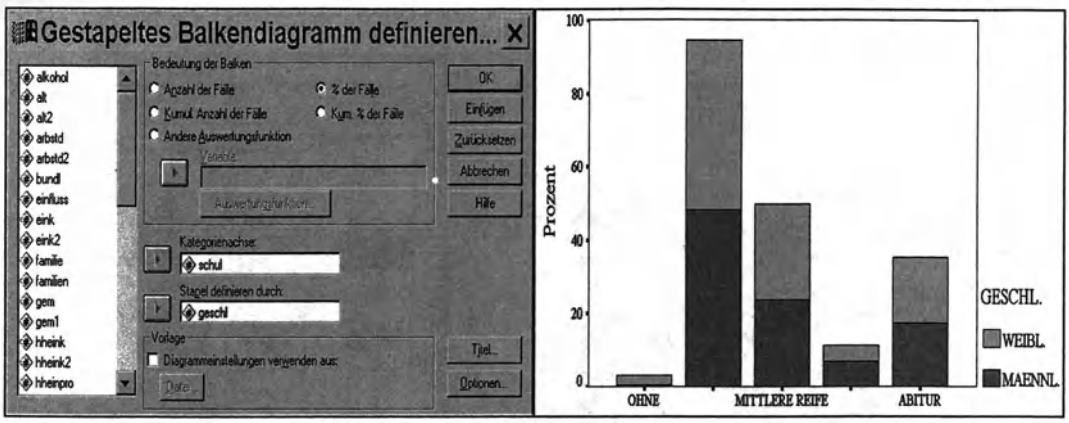


Abb. 26.9. Häufigkeit von Schulabschlüssen untergliedert nach dem Geschlecht der Befragten

26.2.4 Wahlmöglichkeiten

Für fast alle Balkendiagramme bestehen folgende Wahlmöglichkeiten:

- ❑ Wahl, was der Balkenhöhe entsprechen soll („Bedeutung der Balken").
- ❑ Versorgung mit Titel und Fußnoten („Titel").
- ❑ Form der Behandlung fehlender Werte („Optionen").
- ❑ Grafiklayout aus Vorlage übernehmen („Vorlage").

Außerdem: Überarbeitung im Diagramm-Editorfenster möglich (⇨ Kap. 27.4).

26.3 Liniendiagramme erzeugen

Um ein Liniendiagramm zu erstellen:

▹ Klicken Sie die Befehlsfolge „Grafik", „Linien...". Es öffnet sich die in Abb. 26.10 dargestellte Dialogbox.

Als Diagrammformen sind ein *einfaches, mehrfaches* oder *verbundenes* Liniendiagramm wählbar, wobei jeder Diagrammtyp - analog zu den Balkendiagrammen - auf der Grundachse des Diagramms entweder „Kategorien einer Variablen", „Verschiedene Variablen" oder „Werte einzelner Fälle" abbilden (repräsentieren) kann. Im folgenden werden einige dieser Diagrammformen anhand des ALLBUS90-Datensatzes kurz dargestellt.

26.3.1 Einfaches Liniendiagramm

Auswertung über Kategorien einer Variablen. Nach der Befehlsfolge „Grafiken", „Linie ..." wird die Auswahlkombination „Einfach" und „ Auswertung über Kategorien einer Variablen" angeklickt. Abb. 26.11 zeigt die nach Klicken von „Definieren..." geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition und die resultierende Grafik. Die Variable ARBSTD2, in der die Arbeitsstunden der erwerbstätigen Befragten klassifiziert kodiert sind, wurde aus der Quellvariablen-

liste in das Eingabefeld „Kategorienachse:“ übertragen. Die Linienhöhe entspricht hier der prozentualen Häufigkeit, da im Feld „Linie entspricht“ „% Fälle“ angeklickt wurde.

Abb. 26.10. Dialogbox „Liniendiagramm“

Zur Darstellung der Höhe der Linien sind - wie bei Balkendiagrammen - weitere Optionen möglich.

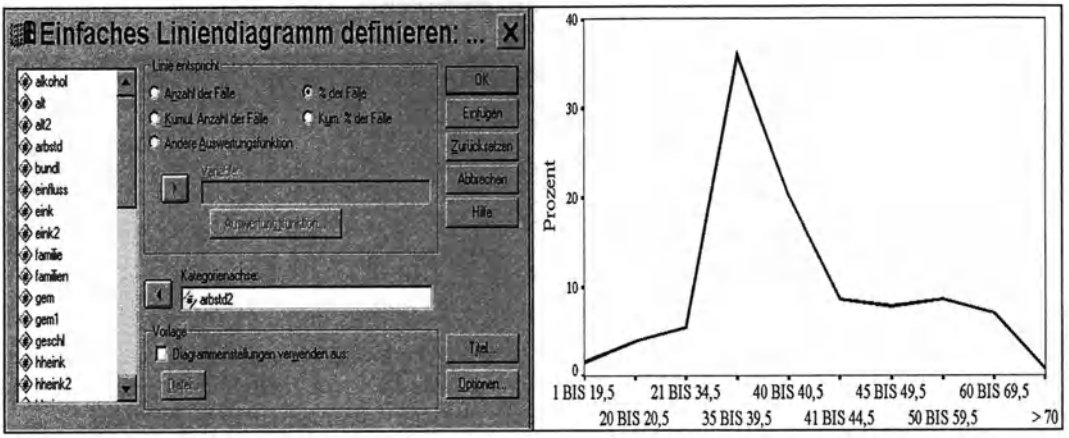


Abb. 26.11. Prozentuale Häufigkeiten der Arbeitsstunden der befragten Erwerbstätigen

Auswertung über verschiedene Variablen bzw. Werte einzelner Fälle. Nach der Befehlsfolge „Grafiken“, „Linie...“ wird die Auswahlkombination „Einfach“ und „Grafikdaten repräsentieren verschiedene Variablen“ bzw. „Werte einzelner Fälle“ angeklickt. Die nach Klicken von „Definieren...“ sich öffnenden Dialogboxen entsprechen denen für Balkendiagramme.

26.3.2 Mehrfaches Liniendiagramm

Auswertung über Kategorien einer Variablen. Nach der Befehlsfolge „Grafiken", „Linie..." wird die Auswahlkombination „Mehrfach" und „Auswertung über Kategorien einer Variablen" angeklickt. Die nach Klicken von „Definieren" geöffnete Dialogbox ähnelt der Dialogbox für gruppierte Balken. In das Eingabefeld „Linien definieren durch:" wird eine Gruppierungsvariable übertragen. Für jede Kategorie der in dieses Feld übertragenen Variablen entsteht eine Linie: also z.B. für Männer und Frauen bei der Gruppierungsvariable GESCHL.

Auswertung über verschiedene Variablen bzw. Werte einzelner Fälle. Nach der Befehlsfolge „Grafiken", „Linie..." wird die Auswahlkombination „Mehrfach" und „Auswertung über verschiedene Variablen" bzw. „Werte einzelner Fälle" angeklickt. Die Dialogboxen entsprechen denen für gruppierte Balken. Auf Beispiele wird verzichtet.

26.3.3 Verbundliniendiagramm

Auswertung über Kategorien einer Variablen. Nach der Befehlsfolge „Grafiken", „Linie..." wird die Auswahlkombination „Verbundlinie" und „Auswertung über Kategorien einer Variablen" angeklickt. Abb. 26.12 zeigt links die nach Klicken von „Definieren" geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition und rechts die resultierende Grafik.

Die Variablen SCHUL und GESCHL wurden aus der Quellvariablenliste in die Eingabefelder „Kategorienachse:" und „Punkte definieren durch:" übertragen. Die durch eine senkrechte Linie verbundenen Markierungszeichen entsprechen der prozentualen Häufigkeit, da im Auswahlfeld „Punkte entsprechen" „% Fälle" angeklickt wurde. Die Grafik entspricht in der Darstellung einem einfachen Bereichsbalkendiagramm (⇨ Abb. 26.24).

Auswertung über verschiedene Variablen bzw. Werte einzelner Fälle. Da die verbundenen Liniendiagramme den Bereichsbalkendiagrammen (⇨ Kap. 26.6) ähneln, wird auf Demonstrationsbeispiele verzichtet.

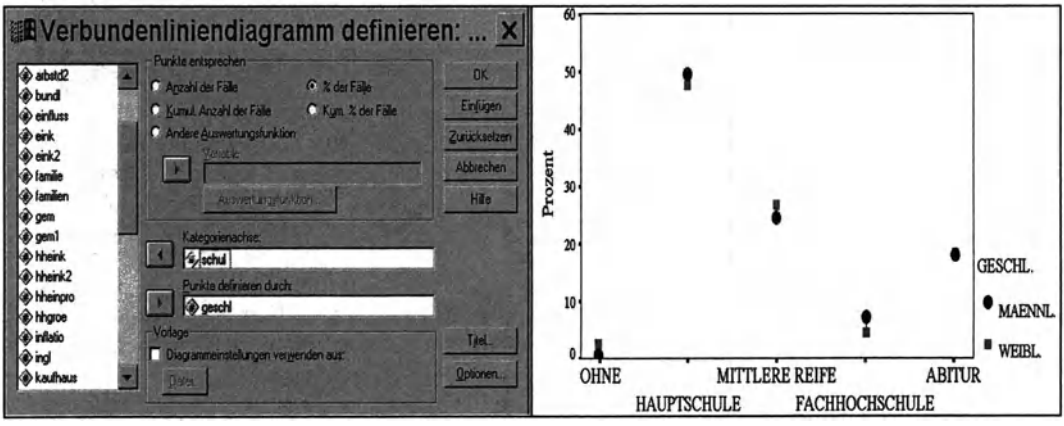


Abb. 26.12. Häufigkeitsunterschiede von Männern und Frauen für verschiedene Schulabschlüssen

26.3.4 Wahlmöglichkeiten

Für alle Liniendiagramme bestehen folgende Wahlmöglichkeiten:

- Wahl, was der Linienhöhe entsprechen soll („Linien entsprechen").
- Versorgung mit Titel und Fußnoten („Titel").
- Form der Behandlung fehlender Werte („Optionen").
- Grafiklayout aus Vorlage übernehmen („Vorlage").

Außerdem: Überarbeitung im Diagramm-Editorfenster möglich (⇨ Kap. 27.4).

26.4 Flächendiagramme erzeugen

Um ein Flächendiagramm zu erstellen, öffnet man durch Klicken der Befehlsfolge

▷ „Grafiken", „Fläche..."

die in Abb. 26.13 dargestellte Dialogbox.

Als Diagrammformen sind ein *einfaches* und ein *gestapeltes* Flächendiagramm wählbar, wobei jeder Diagrammtyp - analog zu den Balken- und Liniendiagrammen - auf der Grundachse des Diagramms entweder Kategorien einer Variablen, verschiedene Variablen oder Werte einzelner Fälle abbilden kann.

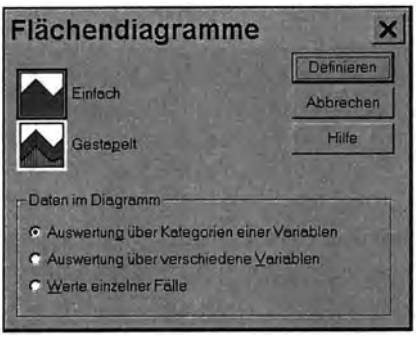


Abb. 26.13. Dialogbox zur Auswahl eines Flächendiagramms

26.4.1 Einfaches Flächendiagramm

Auswertung über Kategorien einer Variablen. Nach der Befehlsfolge „Grafiken", „Fläche..." wird die Auswahlkombination „Einfach" und „Auswertung über Kategorien einer Variablen" angeklickt. Die nach Klicken von „Definieren" geöffnete Dialogbox entspricht der für ein Liniendiagramm. Die Grafik gleicht einem Liniendiagramm, mit dem Unterschied, dass die Fläche unterhalb der Linie eingefärbt ist.

Auswertung über verschiedene Variablen bzw. Werte einzelner Fälle. Nach der Befehlsfolge „Grafiken", „Fläche..." wird die Auswahlkombination „Einfach" und „Auswertung über verschiedene Variablen" bzw. „Werte einzelner Fälle" angeklickt. Die sich öffnenden Dialogboxen entsprechen denen der Liniendia-

gramme. Auch die entstehenden Grafiken entsprechen den Liniendiagrammen, mit dem Unterschied, dass die Flächen unterhalb der Linien eingefärbt sind.

26.4.2 Gestapeltes Flächendiagramm

Auswertung über Kategorien einer Variablen. Nach der Befehlsfolge „Grafiken", „Fläche..." wird die Auswahlkombination „Gestapelt" und „Auswertung über Kategorien einer Variablen" angeklickt. Die nach Klicken von „Definieren..." geöffnete Dialogbox entspricht der für gestapelte Balkendiagramme (⇨ Abb. 26.9). In das Feld „Flächen definieren durch:" wird eine Gruppierungsvariable (z.B. GESCHL) übertragen. Für jede Kategorie der in dieses Feld übertragenen Variablen entsteht eine Fläche (z.B. eine für Männer und Frauen). Die Flächenhöhe entspricht hier der gewählten Option in „Flächen entsprechen:" (z.B. „% Fälle"). Die Grafik ähnelt einem mehrfachen Liniendiagramm. Im Unterschied zum Liniendiagramm werden die dargestellten Werte (z.B. prozentuale Häufigkeiten) aber wie im gestapelten Balkendiagramm additiv überlagert.

Auswertung über verschiedene Variablen. Nach der Befehlsfolge „Grafik", „Flächen..." wird die Auswahlkombination „Gestapelt" und „Grafikdaten repräsentieren verschiedene Variablen" angeklickt. Nach Klicken von „Definieren" wird eine Dialogbox geöffnet.

In das Eingabefeld „Flächen entsprechen" sind mindestens zwei Variablen einzutragen. Standardmäßig wird „SUM" als Auswertungsfunktion zugrunde gelegt. Der Grafiktyp ist in erster Linie für Variablen geeignet, deren Summierung über Fälle Sinn macht. Sind z.B. in einem Datensatz für die zwölf Monate eines Jahres (= Fälle) die Umsätze von drei Produkten (= drei Variablen) einer Firma enthalten, so wird ein gestapeltes Flächendiagramm aussagekräftig: Der gestapelte summierte Umsatz (summiert über alle Fälle, d.h. Monate eines Jahres) addiert sich zum Jahresumsatz, und jedes Produkt wird mit einer Fläche dargestellt.

Werte einzelner Fälle. Bei Wahl der Kombination „Gestapelt" und „Auswertung über Werte einzelner Fälle" hat die Dialogbox die gleichen Eingabefelder und Wahlmöglichkeiten wie im Fall eines gruppierten Balkendiagramms.

26.4.3 Wahlmöglichkeiten

Für fast alle Flächendiagramme bestehen folgende Wahlmöglichkeiten:

- Wahl, was der Flächenhöhe entsprechen soll („Flächen entsprechen").
- Versorgung mit Titel und Fußnoten („Titel").
- Form der Behandlung fehlender Werte („Optionen").
- Grafiklayout aus Vorlage übernehmen („Vorlage").

Außerdem: Überarbeitung im Diagramm-Editorfenster möglich (⇨ Kap. 27.4).

26.5 Kreisdiagramme erzeugen

Um ein Kreisdiagramm zu erstellen, öffnet man durch Klicken der Befehlsfolge

▷„Grafik", „Kreis..." die in Abb. 26.14 dargestellte Dialogbox.

Aus Abb. 26.14 wird ersichtlich, dass in einem Kreisdiagramm die Daten wie bei Balken-, Linien- bzw. Flächendiagrammen entweder Kategorien einer Variablen, verschiedene Variablen oder Werte einzelner Fälle abbilden können. Im folgenden wird eines der Kreisdiagramme anhand des ALLBUS90-Datensatzes dargestellt.

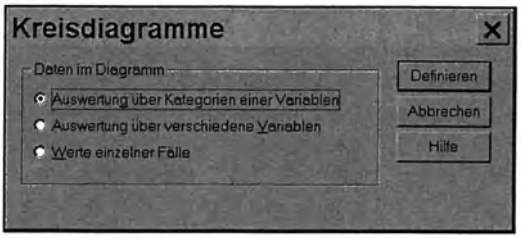


Abb. 26.14. Dialogbox „Kreisdiagramme"

Auswertung über Kategorien einer Variablen. Nach der Befehlsfolge „Grafiken", „Kreis..." wird die Auswahl „Auswertung über Kategorien einer Variablen" angeklickt. Nach Klicken von „Definieren..." öffnet sich die in Abb. 26.15 dargestellte Dialogbox. Sie enthält ein Beispiel zur Grafikdefinition sowie die resultierende Grafik. Die Variable SCHUL wurde aus der Quellvariablenliste in das Eingabefeld „Segmente definieren durch:" übertragen. Das Segment entspricht hier der prozentualen Häufigkeit von Schulabschlüssen, da im Feld „Segmente entsprechen" „% Fälle" angeklickt wurde.

Zur Darstellung der Segmentbreite sind andere Optionen möglich.

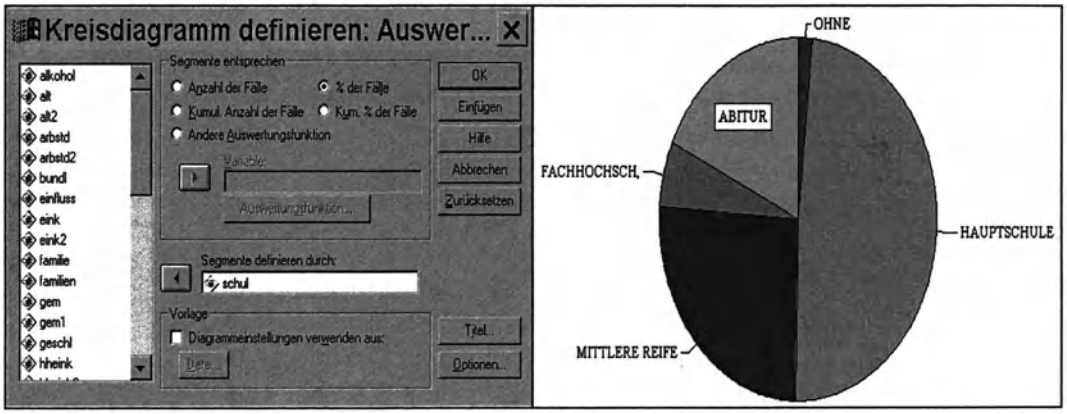


Abb. 26.15. Prozentuale Häufigkeiten von Schulabschlüssen der Befragten

Auswertung über verschiedene Variablen. Bei Wahl von „Auswerten über verschiedene Variablen" werden in das Eingabefeld „Segmente entsprechen" mehrere Variablen übertragen. Standardmäßig wird von SPSS die Auswertungsfunktion „SUM" (= Summe) eingetragen. Mit dieser Einstellung wird in einem Segment der Kreisgrafik die Summe einer Variablen abgebildet. Erfassen z.B. die Variablen UMA und UMB die regionalen Umsätze einer Firma für die Produkte A und B, so

werden bei der Auswertungsfunktion SUM im Kreisdiagramm die über die Regionen summierten Umsätze der beiden Produkte dargestellt.

Wahlmöglichkeiten. Für Kreisdiagramme bestehen folgende Wahlmöglichkeiten:

- Wahl, was der Flächenhöhe entsprechen soll („Flächen entsprechen").
- Versorgung mit Titel und Fußnoten („Titel").
- Form der Behandlung fehlender Werte („Optionen").
- Grafiklayout aus Vorlage übernehmen („Vorlage").

Außerdem: Überarbeitung im Diagramm-Editorfenster möglich (⇨ Kap. 27.4).

26.6 Hoch-Tief-Diagramme erzeugen

Bei Hoch-Tief-Diagrammen handelt es sich um Balken- bzw. Liniendiagramme. Es sind drei Formen für unterschiedliche Anwendungen zu unterscheiden:

- *Hoch-Tief-Schluss-Diagramme.* Diese Diagramme eignen sich zur Darstellung der Entwicklung von Aktien- und Währungskursen und ähnlichem im Zeitablauf. Beispielsweise kann für eine Gruppe von Gesellschaften die durchschnittlich höchste, tiefste sowie die durchschnittliche Börsenschluss-Kursnotierung für z.B. aufeinanderfolgende Tage dargestellt werden.
- *Bereichsbalkendiagramme.* In einem derartigen Diagramm können in der einfachsten Anwendung für Kategorien einer Variablen (z.B. Schulabschlüsse) die Differenzen der Häufigkeiten von zwei Gruppen (z.B. Männer und Frauen) in Form von Balken dargestellt werden.
- *Differenzliniendiagramme.* Diese ähneln den Bereichsbalkendiagrammen. Der Unterschied besteht darin, dass die Differenzen in Form von Linien dargestellt werden.

Um eines der Hoch-Tief-Diagramme zu erstellen, öffnet man durch Klicken der Befehlsfolge

▷ „Grafiken", „Hoch-Tief..."

die in Abb. 26.16 dargestellte Dialogbox.

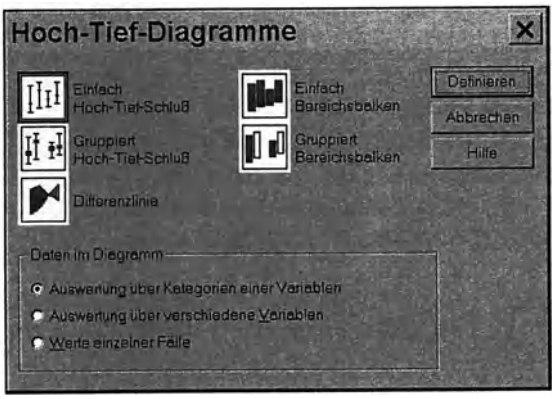


Abb. 26.16. Dialogbox zur Auswahl eines „Hoch-Tief-Diagramms"

Als Diagrammformen sind *einfache* und *gruppierte Hoch-Tief-Schluss-, einfache* und *gruppierte Bereichsbalken-* sowie *Differenzliniendiagramme* wählbar. Jeder Diagrammtyp kann - analog zu Balken- und Liniendiagrammen - unterschiedliche Daten darstellen: Kategorien einer Variablen, verschiedene Variablen oder Werte einzelner Fälle.

Im folgenden werden einige dieser verschiedenen Diagrammformen anhand von Beispielen dargestellt.

26.6.1 Einfaches Hoch-Tief-Schluss-Diagramm

Auswertung über Kategorien einer Variablen. Das folgende Übungsbeispiel bezieht sich auf Daten, die ausschnittsweise in Abb. 26.17 dargestellt sind (Datei AUTO.SAV). Für die 14. bis 18. Kalenderwoche (Variable WOCHE) sind mit der Variable KURS die höchsten, tiefsten und Börsenschlusskurse jeder Woche der drei Unternehmen BMW, Daimler-Benz und Porsche (Variable UNTERN) erfasst. Mit der Variablen HO_TI_EN wird erfasst, ob es sich um den höchsten, niedrigsten oder den Börsenschlusskurs handelt (im in Abb. 26.17 dargestellten Dateneditorfenster ist im Menü „Extras" „Werte-Labels anzeigen" aktiv geschaltet). Dargestellt werden soll die durchschnittliche Entwicklung des Aktienkurses für diese Unternehmen.

	woche	untern	ho_ti_en	kurs
1	14.	bmw	hoch	873,00
2	14.	bmw	tief	850,00
3	14.	bmw	schluß	860,00
4	14.	daimler	hoch	880,00
5	14.	daimler	tief	860,00
6	14.	daimler	schluß	869,00
7	14.	porsche	hoch	840,00
8	14.	porsche	tief	820,00
9	14.	porsche	schluß	853,00
10	15.	bmw	hoch	880,00
11	15.	bmw	tief	860,00

Abb. 26.17. Ausschnitt aus der Datendatei AUTO.SAV

Nach der Befehlsfolge „Grafiken", „Hoch-Tief..." wird die Auswahlkombination „Einfach Hoch-Tief-Schluss" und „Auswertung über Kategorien einer Variablen" angeklickt. Abb. 26.18 zeigt links die durch Klicken von „Definieren" geöffnete Dialogbox mit dem Beispiel zur Grafikdefinition und rechts die resultierende Grafik. Die Variablen WOCHE und HO_TI_EN wurden aus der Quellvariablenliste in die Eingabefelder „Kategorienachse" und „Hoch-Tief-Schluss definieren durch" übertragen. In der Auswahlgruppe „Bedeutung der Balken" wurde „Andere Funktion" gewählt. Danach wurde die Variable KURS in das Eingabefeld „Variable" übertragen. Standardmäßig wird die Funktion „MEAN" (arithmetisches Mittel) eingesetzt. Alternativ sind andere Auswertungs- und Darstellungsformen (Anzahl Fälle etc.) für die Balken möglich, für diese Daten aber nicht sinnvoll.

In der Abb. 26.18 rechts ist das Diagramm dargestellt. Aus dem Diagramm kann man die durchschnittliche Aktienkursentwicklung der Automobilunternehmen entnehmen. Die Balkenenden bilden die durchschnittlichen Höchst- und Tiefstkurse in einer Woche ab. Der durchschnittliche Börsenschlusswert wird als schwarzes Kästchen auf den Balken dargestellt. Auf die Erfassung und Darstellung des Börsenschlusswertes kann auch verzichtet werden.

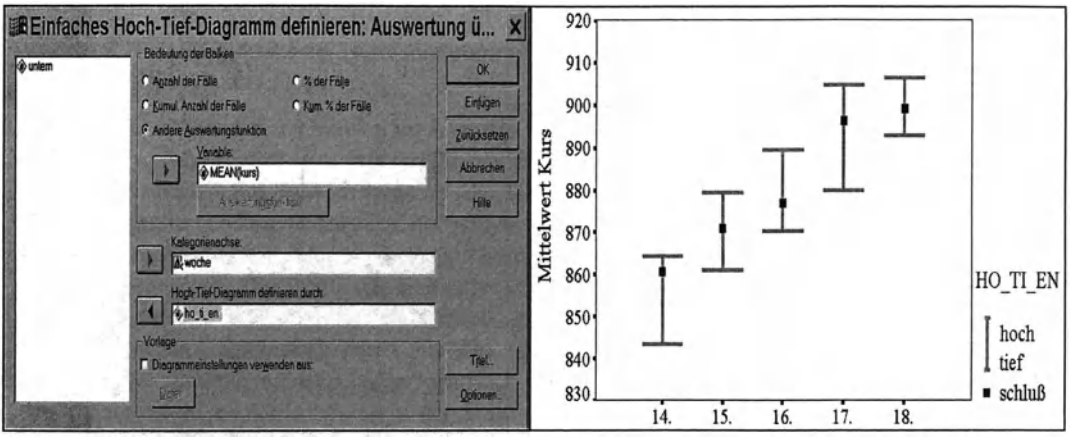


Abb. 26.18. Durchschnittliche Aktienkurse von Automobilunternehmen

Auswertung über verschiedene Variablen. Zur beispielhaften Darstellung wird auf die in Abb. 26.20 erfassten Daten verwiesen. Im Unterschied zu der dort im Dateneditorfenster ausschnittsweise abgebildeten Datendatei sind in der nun verarbeiteten Datei (AUTO1.SAV) nur die Daten der Automobilbranche enthalten (d.h. gleiche Datenorganisation, aber ohne Daten für die Bierbrauer). Dargestellt werden soll - wie im einfachen Hoch-Tief-Schluss-Diagramm für „Kategorien einer Variablen" - die Aktienkursentwicklung in der Automobilbranche.

Nach der Befehlsfolge „Grafiken", „Hoch-Tief..." wird die Auswahlkombination „Einfach Hoch-Tief-Schluss" und „Auswertung über verschiedene Variablen" angeklickt. Abb. 26.19 zeigt links die durch Klicken von „Definieren" geöffnete Dialogbox mit dem Beispiel zur Grafikdefinition und rechts die resultierende Grafik. Die Variablen HOCH, TIEF und SCHLUSS wurden in die Eingabefelder „Hoch:", „Tief:" und „Schluss:" von „Bedeutung der Balken" übertragen. Standardmäßig wird die Auswertungsfunktion „MEAN" (arithmetisches Mittel) eingesetzt. Falls eine andere Berechnung dargestellt werden soll, werden die in die Eingabefelder übertragenen Variablen markiert und anschließend wird mit Klicken auf „Auswertungsfunktion" eine Dialogbox mit Wahlmöglichkeiten für andere Auswertungsfunktionen geöffnet (⇨ Abb. 26.4). Die Variable WOCHE wurde aus der Quellvariablenliste in das Eingabefeld „Kategorienachse:" übertragen.

In der Abb. 26.19 rechts ist das Diagramm dargestellt. Aus dem Diagramm kann man die durchschnittliche Aktienkursentwicklung der drei Unternehmen der Automobilbranche entnehmen.

Werte einzelner Fälle. Für die Auswahlkombination „Einfach Hoch-Tief-Schluss“ und „Grafikdaten repräsentieren Werte einzelner Fälle“ wird eine Dialogbox aufgerufen, die der in Abb. 26.19 ähnelt. Für die Aktienkurse einzelner Unternehmen können Hoch-, Tief-, und Schlusskurse z.B. einer Woche dargestellt werden.

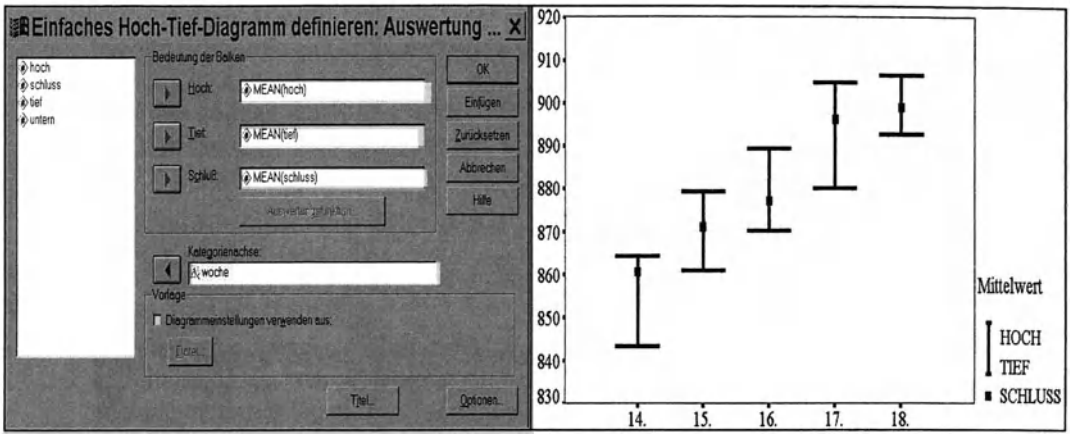


Abb. 26.19. Durchschnittliche Aktienkurse von Automobilunternehmen

26.6.2 Gruppiertes Hoch-Tief-Schluss-Diagramm

Auswertung über Kategorien einer Variablen. Das folgende Übungsbeispiel bezieht sich auf Daten, die ausschnittsweise in Abb. 26.20 dargestellt sind (Datei AKTIE.SAV). Für die 14. bis 18. Kalenderwoche (Variable WOCHE) sind mit den Variablen HOCH, TIEF und SCHLUSS die höchsten, tiefsten sowie Börsenschlusskurse von Unternehmen der Branchen (Variable BRANCHE) Automobilhersteller und Bierbrauereien erfasst (im in Abb. 26.20 dargestellten Dateneditorfenster ist im Menü Extras „Werte-Labels anzeigen“ aktiv geschaltet). Dargestellt werden soll die Entwicklung der Aktienkurse für die beiden Branchen.

Nach der Befehlsfolge „Grafiken“, „Hoch-Tief...“ wird die Auswahlkombination „Gruppiert Hoch-Tief-Schluss“ und „Auswertung über Kategorien einer Variablen“ angeklickt. Abb. 26.21 zeigt links die durch Klicken von „Definieren“ geöffnete Dialogbox mit dem Beispiel zur Grafikdefinition und rechts die resultierende Grafik. Die Variablen HOCH, TIEF und SCHLUSS wurden in die Eingabefelder „Hoch:“, „Tief:“ und „Schluss:“ von „Bedeutung der Balken“ übertragen. Standardmäßig wird die Funktion „MEAN“ eingesetzt. Falls eine andere Auswertungsfunktion dargestellt werden soll, werden die in das Eingabefeld übertragenen Variablen markiert und anschließend wird mit Klicken auf „Auswertungsfunktion“ eine Dialogbox mit Wahlmöglichkeiten für andere Auswertungsfunktionen geöffnet (⇨ Abb. 26.4). Die Variablen WOCHE und BRANCHE wurden aus der Quellvariablenliste in die Eingabefelder „Kategorienachse:“ und „Gruppen definieren durch:“ übertragen.

	woche	untern	branche	hoch	tief	schluss
1	14.	BMW	Auto	873,00	850,00	860,00
2	14.	Daimler Benz	Auto	880,00	860,00	869,00
3	14.	Porsche	Auto	840,00	820,00	853,00
4	14.	Haake Beck	Bier	640,00	640,00	640,00
5	14.	Henninger	Bier	600,00	590,00	597,00
6	14.	Holsten	Bier	545,00	540,00	543,00
7	15.	BMW	Auto	880,00	860,00	870,00
8	15.	Daimler Benz	Auto	878,00	870,00	873,00
9	15.	Porsche	Auto	880,00	853,00	870,00
10	15.	Haake Beck	Bier	642,00	636,00	640,00
11	15.	Henninger	Bier	625,00	615,00	620,00

Abb. 26.20. Ausschnitt aus der Datendatei AKTIE.SAV

In der Abb. 26.21 rechts ist das Diagramm dargestellt. Aus dem Diagramm kann man die durchschnittliche Aktienkursentwicklung der Unternehmen - jeweils für Branchen - entnehmen. Es handelt sich um die gleiche Darstellung wie in Abb. 26.19, mit dem Unterschied, dass die Kursentwicklung für zwei Unternehmensgruppen gleichzeitig dargestellt wird.

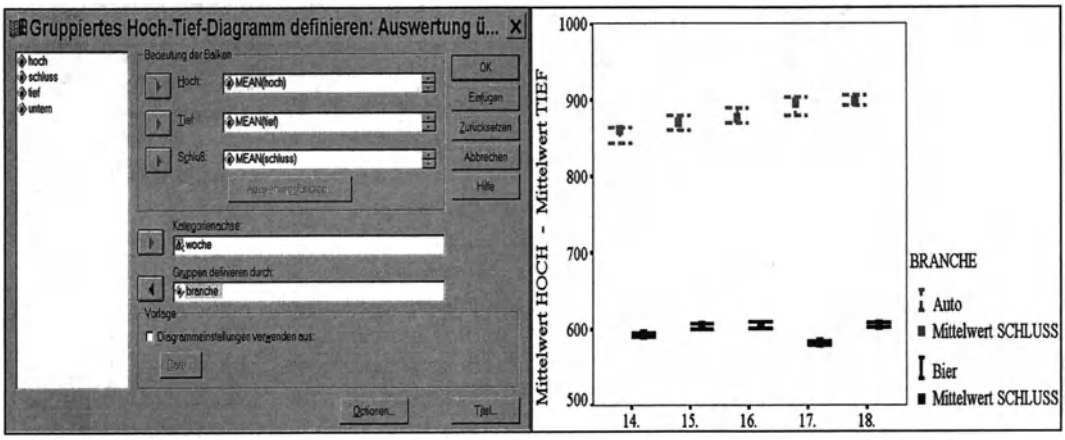


Abb. 26.21. Durchschnittliche Aktienkurse von Unternehmen der Automobil- und Bierbraubranche

Auswertung über verschiedene Variablen. Zur beispielhaften Darstellung sind die in Abb. 26.20 ausschnittsweise dargestellten Daten in anderer Weise im Dateneditorfenster erfasst. In Abb. 26.22 ist dieses dargestellt (Datei AKTIE1.SAV): Die Variablen HOCH_A, TIEF_A und SCHLUS_A erfassen die höchsten, tiefsten und Börsenschluss-Aktienkurse von Unternehmen der Automobilbranche, die Variablen HOCH_B, TIEF_B und SCHLUS_B die von Bierbrauern. Dargestellt werden soll die durchschnittliche Entwicklung der Aktienkurse für jede der beiden Unternehmensgruppen.

	woche	auto_unt	hoch_a	tief_a	schlus_a	bier_unt	hoch_b	tief_b	schlus_b
1	14.	BMW	873,00	850,00	860,00	Haake Beck	640,00	640,00	640,00
2	14.	Daimler Ben	880,00	860,00	869,00	Henninger	600,00	590,00	597,00
3	14.	Porsche	840,00	820,00	853,00	Holsten	545,00	540,00	543,00
4	15.	BMW	880,00	860,00	870,00	Haake Beck	642,00	636,00	640,00
5	15.	Daimler Ben	878,00	870,00	873,00	Henninger	625,00	615,00	620,00
6	15.	Porsche	880,00	853,00	870,00	Holsten	555,00	548,00	550,00
7	16.	BMW	890,00	873,00	880,00	Haake Beck	644,00	635,00	640,00
8	16.	Daimler Ben	890,00	870,00	880,00	Henninger	624,00	618,00	620,00
9	16.	Porsche	888,00	868,00	871,00	Holsten	558,00	550,00	555,00
10	17.	BMW	920,00	872,00	909,00	Haake Beck	640,00	640,00	640,00

Abb. 26.22. Ausschnitt aus der Datendatei AKTIE1.SAVE

Nach der Befehlsfolge „Grafiken", „Hoch-Tief..." wird die Auswahlkombination „Gruppiert Hoch-Tief-Schluss" und „Auswertung über verschiedene Variablen" angeklickt. Abb. 26.23 zeigt links die durch Klicken von „Definieren" geöffnete Dialogbox mit dem Beispiel zur Grafikdefinition und rechts die resultierende Grafik.

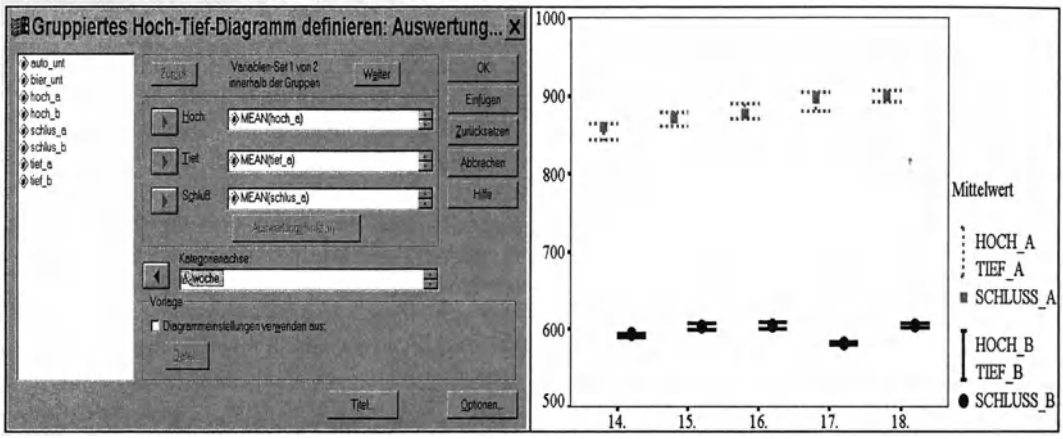


Abb. 26.23. Durchschnittliche Aktienkurse von Unternehmen der Automobil- und Bierbraubranche

Die Variablen HOCH_A, TIEF_A und SCHLUS_A wurden in die Eingabefelder „Hoch:", „Tief:" und „Schluss:" von „Variablen-Set 1 von 1 innerhalb der Gruppen" übertragen. Damit sind die Variablen für die erste Gruppe - die Automobilunternehmen - eingetragen. Nach Klicken auf „Weiter" wurden die entsprechenden Variablen für die Brauereien (HOCH_B, TIEF_B, SCHLUS_B) eingetragen. Durch Wiederholen des Vorgangs können weitere Gruppen dargestellt werden. Die Variable WOCHE wird in „Kategorienachse" übertragen.

In der Abb. 26.23 rechts ist das Diagramm dargestellt. Aus dem Diagramm kann man die durchschnittliche Aktienkursentwicklung der Unternehmen beider Branchen entnehmen. Es handelt sich um die gleiche Darstellung wie in Abb. 26.19,

mit dem Unterschied, dass die Kursentwicklung für zwei Unternehmensgruppen gleichzeitig dargestellt wird.

Werte einzelner Fälle. Die Dialogbox ähnelt der in Abb. 26.23. Im Unterschied dazu werden die einzelnen - nicht die durchschnittlichen - Hoch-, Tief- und Schlusskurse einer Woche von z.B. zwei Unternehmen dargestellt.

26.6.3 Einfaches Bereichsbalkendiagramm

Auswertung über Kategorien einer Variablen. Im folgenden Demonstrationsbeispiel aus dem Datensatz ALLBUS90.SAV werden die Häufigkeiten der Variablen SCHUL untergliedert nach dem Geschlecht ausgewertet und in einem einfachen Bereichsbalkendiagramm dargestellt.

Nach der Befehlsfolge „Grafiken", „Hoch-Tief..." wird die Auswahlkombination „Einfach Bereichsbalken" und „Auswertung über Kategorien einer Variablen" angeklickt. Abb. 26.24 zeigt links die durch Klicken von „Definieren" geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition und rechts die resultierende Grafik.

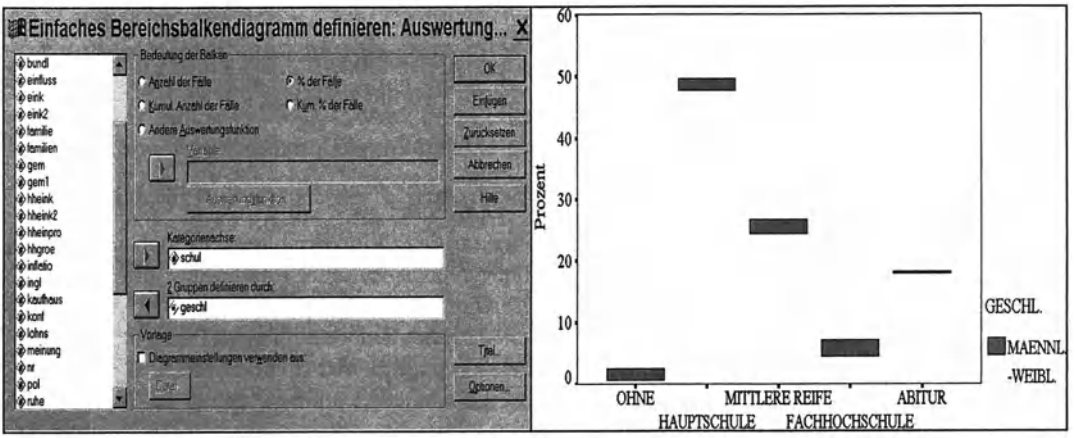


Abb. 26.24. Unterschiede der Häufigkeiten von Schulabschlüssen bei Männern und Frauen

Die Variable SCHUL wurde aus der Quellvariablenliste in das Eingabefeld „Kategorienachse:" und die Variable GESCHL in das Eingabefeld „2 Gruppen definieren durch" übertragen. Im Auswahlfeld „Bedeutung der Balken" ist „% Fälle" angeklickt. Damit basiert die Darstellung der Häufigkeiten auf Prozentwerte. Das Diagramm in Abb. 26.24 rechts entspricht in seinem Informationsgehalt dem gruppierten Balkendiagramm in Abb. 26.8. Dort sind die prozentualen Häufigkeiten der Schulabschlüsse von Männern und Frauen als Balken dargestellt. Im Unterschied dazu wird hier die Differenz dieser Häufigkeiten als Balken abgebildet.

Alternativ dazu kann sich die Darstellung auch auf absolute, kumulierte absolute oder kumulierte prozentuale Häufigkeiten stützen. Außerdem kann alternativ auch eine „Andere Auswertungsfunktion" gewählt werden. Dann muss danach eine

Variable in das Eingabefeld „Variable" übertragen werden. Analog zur Erzeugung von Balkendiagrammen wird dann standardmäßig das arithmetische Mittel „MEAN" dieser Variable ausgewertet. Wäre z.B. die übertragene Variable ALT (Alter), so würde im Balkendiagramm die Differenz im durchschnittlichen Alter von Männern und Frauen dargestellt. Man kann aber auch andere Funktionen wie z.B. den Median, den Modalwert, die Standardabweichung etc. auswerten lassen.

Auswertung über verschiedene Variablen. Im folgenden Demonstrationsbeispiel aus dem Datensatz ALLBUS90.SAV werden die Differenzen der durchschnittlichen HHEINPRO (Haushaltseinkommen pro Kopf des Haushalts = HHEINK/HHGROE) und durchschnittlichen EINKOM (Einkommen der Befragten) für die Schulabschlüsse der Befragten dargestellt.

Nach der Befehlsfolge „Grafiken", „Hoch-Tief..." wird die Auswahlkombination „Einfach Bereichsbalken" und „Auswertung über verschiedene Variablen" angeklickt. Abb. 26.25 zeigt links die durch Klicken von „Definieren" geöffnete Dialogbox mit dem Beispiel zur Grafikdefinition und rechts die resultierende Grafik. Die Variable SCHUL wurde aus der Quellvariablenliste in das Feld „Kategorienachse:" übertragen. In „Balkenpaar entspricht" wurden die Variablen HHEINPRO und EINK in die Felder „1" und „2" übertragen. Danach wird von SPSS - wie auch bei gruppierten Bereichsbalkendiagrammen für Kategorien einer Variablen - standardmäßig die Funktion „MEAN" eingetragen und ausgewertet. Man kann aber auch andere Funktionen (wie im Zusammenhang mit Abb. 26.4 beschrieben) auswerten lassen.

In der Abb. 26.25 rechts ist das Diagramm dargestellt. Für jeden Schulabschluss wird die Differenz des durchschnittlichen Pro-Kopf-Haushaltseinkommens und durchschnittlichen Einkommens der Befragten in Form von Balken dargestellt.

Werte einzelner Fälle. Wählt man die Kombination „Einfaches Bereichsbalkendiagramm" und „Werte einzelner Fälle" wird eine Dialogbox geöffnet, die der in Abb. 26.25 ähnelt. Im Unterschied zur entstehenden Grafik dort entsprechen die Balken nun den Differenzen von zwei Variablen für einzelne Fälle (und nicht den Differenzen von Mittelwerten).

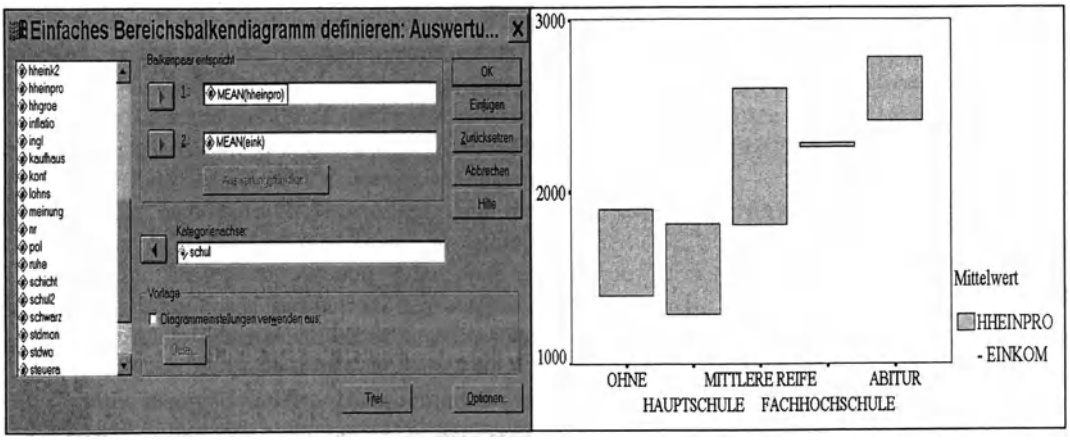


Abb. 26.25. Differenz des durchschnittlichen Pro-Kopf-Haushaltseinkommens und Einkommens der Befragten mit verschiedenen Schulabschlüssen

26.6.4 Gruppiertes Bereichsbalkendiagramm

Auswertung über Kategorien einer Variablen. In Erweiterung der Darstellung in Abb. 26.25 könnte man die Differenz von durchschnittlichem EINKOM (monatliches Nettoeinkommen des Befragten) und durchschnittlichem HHEINPRO (Haushaltseinkommen pro Kopf eines Haushaltsmitglieds) für unterschiedliche Schulabschlüsse untergliedert nach zwei Altersgruppen (z.B. < 40 und > 40 Jahre, erfasst in der Variable ALT4) darstellen.

Nach der Befehlsfolge „Grafiken", „Hoch-Tief..." wird die Auswahlkombination „Gruppiert Bereichsbalken" und „Auswertung über Kategorien einer Variable" angeklickt. In der nach Klicken von „Definieren" geöffneten Dialogbox wird ergänzend zu den Variablenübertragungen, die in Abb. 26.25 erfolgen, die Variable ALT4 in das Feld „Gruppen definieren durch:" übertragen. Im Diagramm wird für jeden Schulabschluß die Differenz des durchschnittlichen Pro-Kopf-Haushaltseinkommens und durchschnittlichen Einkommens der Befragten in Form von Balken dargestellt. Dabei wird eine Untergliederung nach den beiden Altersgruppen vorgenommen.

Auswertung über verschiedene Variablen. Zur Erläuterung mit Hilfe eines Demonstrationsbeispiels werden Variable aus dem zur Darstellung von Regelkarten-Diagrammen verwendeten Datensatzes ZIGARETT.SAV genutzt (⇨ Abb. 26.36 in Kap 26.8). In der Datei sind Einzelmesswerte von auf Anlage A und B produzierten Zigaretten erfasst. Dabei handelt es sich um zehn Stichproben mit einem Stichprobenumfang von je 24 Zigaretten. Die Stichprobennummer ist in der Variable PROBE erfasst. Die Variablen DM_A und DM_B erfassen die Durchmesser und die Variablen ZW_A und ZW_B die Zugwiderstände von auf Anlage A bzw. B produzierten Zigaretten.

Hier geht es um die Darstellung von Differenzen von Durchmessern (DM_A minus DM_B) sowie Differenzen in den Zugwiderständen (ZW_A minus ZW_B) der auf Anlage A und Anlage B produzierten Zigaretten. Dargestellt werden sollen

hier aber nicht die Differenzen von Einzelwerten, sondern von Durchschnittswerten für jede der zehn Stichproben. Aus Darstellungsgründen wurden die Messwerte von DM_A und DM_B mit zehn multipliziert (= Variable DM10_A und DM10_B).

Nach der Menüfolge „Grafiken", „Hoch-Tief..." wird die Auswahlkombination „Gruppiert Bereichsbalken" und „Auswertung über verschiedene Variablen" angeklickt. Abb. 26.26 zeigt links die durch Klicken von „Definieren" geöffnete Dialogbox mit dem Beispiel zur Grafikdefinition. Die Variablen DM10_A und DM10_B (die Durchmesser der auf den Produktionsanlagen A und B produzierten Zigaretten - jeweils mit 10 multipliziert zur besseren Darstellung im Diagramm) wurden in die Eingabefelder „1" und „2" durch Mausklick übertragen. Von SPSS wird standardmäßig die Funktion „MEAN" eingetragen. Damit ist das Variablenpaar für die erste darzustellende Gruppe bestimmt. Um das Variablenpaar der zweiten Gruppe festzulegen, wird auf „Weiter" geklickt. Nun kann man in die frei gewordenen Eingabefelder „1" und „2" das Variablenpaar der zweiten Gruppe übertragen. Für dieses Beispiel wurden für die zweite Gruppe die Variablen ZW_A und ZW_B (Zugwiderstand der auf den Anlagen A und B produzierten Zigaretten) gewählt. Möchte man weitere Variablenpaare darstellen, so können nach Klicken von „Weiter" diese Variablen in die beiden Eingabefelder übertragen werden. Durch Klicken von „Zurück" kann man wieder zu vorhergehenden Variablenpaaren zurückschalten. In das Auswahlfeld „Kategorienachse" wurde die Variable PROBE - sie gibt die Nummer der Stichprobe an - übertragen.

In der Abb. 26.26 rechts ist das Diagramm dargestellt. Für jede Stichprobe wird die Differenz des durchschnittlichen Durchmessers und durchschnittlichen Zugwiderstands der auf den Anlagen A und B produzierten Zigaretten durch Balken dargestellt. Aus der Grafik wird deutlich, dass sich die Durchmesser der auf den beiden Anlagen gefertigten Zigaretten nur wenig unterscheiden.

Werte einzelner Fälle. Man geht prinzipiell analog zum Fall der Auswertung über verschiedene Variablen vor (⇨ Abb. 26.26). Im Unterschied dazu werden die Differenzen von Variablen für einzelne Fälle (z.B. einzelne Zigaretten) dargestellt. Für die Grundachse („Kategorienachse") kann man wählen, ob als Achsenbeschriftung entweder die Fallnummer oder Werte einer anzugebenden Variablen verwendet werden sollen.

26.6.5 Differenzliniendiagramm

Auswertung über Kategorien einer Variablen. Ein Differenzliniendiagramm entspricht von der Informationsdarstellung im Prinzip einem einfachen Bereichsbalkendiagramm. Der Unterschied besteht darin, dass die Häufigkeiten von zwei Gruppen in Form von Linien und die Differenzen der Häufigkeiten durch Flächen dargestellt werden. Durch eine unterschiedliche Farbgebung kann man erkennen, welche Gruppe in der Häufigkeit überwiegt.

Beim Erstellen des Diagramms geht man wie bei der Erzeugung eines einfachen Bereichsbalkendiagramms vor, mit dem Unterschied, dass ein Differenzliniendiagramm gewählt wird. In Abb. 26.27 ist ein Beispiel dargestellt. Es werden die

Differenzen der prozentualen Häufigkeiten von Männern und Frauen für vier Altersgruppen dargestellt.

Auswertung über verschiedene Variablen bzw. Werte einzelner Fälle. Zur Erstellung des Diagramms geht man wie bei der Erzeugung eines einfachen Bereichsbalkendiagramms vor mit dem Unterschied, dass ein Differenzliniendiagramm gewählt wird.

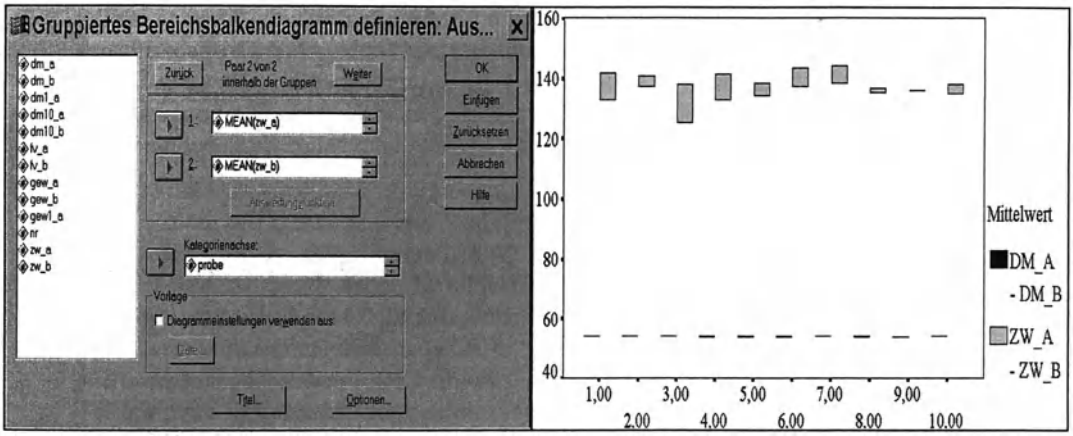


Abb. 26.26. Differenz der durchschnittlichen Durchmesser sowie der durchschnittlichen Zugwiderstände von auf Anlagen A und B produzierten Zigaretten für zehn Stichproben

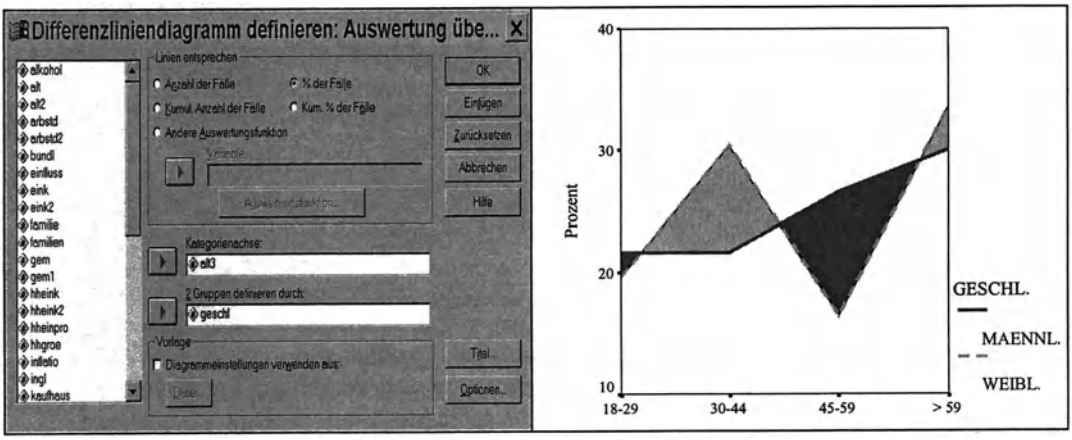


Abb. 26.27. Differenzen in den Altersgruppenhäufigkeiten von Männern und Frauen

26.6.6 Wahlmöglichkeiten

Für fast alle Hoch-Tief-Diagramme bestehen folgende Wahlmöglichkeiten:

- Versorgung mit Titel und Fußnoten („Titel“).
- Form der Behandlung fehlender Werte („Optionen“).

- ❑ Grafiklayout aus Vorlage übernehmen („Vorlage“).
- ❑ Missing-Werte als Kategorie anzeigen.

Außerdem: Überarbeitung im Diagramm-Editorfenster möglich (⇨ Kap. 25.4).

26.7 Pareto-Diagramme erzeugen

Pareto-Diagramme sind Balkendiagramme zur grafischen Darstellung von Häufigkeiten einer kategorialen Variablen, wobei in der Darstellung die Häufigkeiten der Kategorien der Größe nach geordnet werden: zuerst die Kategorie mit der größten Häufigkeit, dann die mit der zweitgrößten Häufigkeit usw. Optional kann eine Linie in dem Diagramm die kumulierten Häufigkeiten darstellen. Ein Pareto-Diagramm wird sinnvoll immer dann verwendet, wenn eine Variable viele Kategorien hat und man daran interessiert ist, welche Kategorien die größten Häufigkeiten haben.

Um ein Pareto-Diagramm zu erstellen, öffnet man durch Klicken der Befehlsfolge

▷ „Grafiken“, „Pareto...“

die in Abb. 26.28 dargestellte Dialogbox.

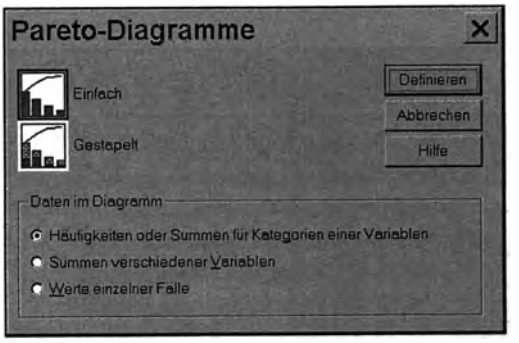


Abb. 26.28. Dialogbox zur Auswahl eines Pareto-Diagramms

Als Diagrammtypen sind ein *einfaches* und ein *gestapeltes* Diagramm wählbar wobei jeder Diagrammtyp unterschiedliche Daten darstellen kann: Häufigkeiten oder Summen für Kategorien einer Variablen, Summen verschiedener Variablen oder Werte einzelner Fälle.

Im folgenden werden einige dieser verschiedenen Diagrammformen anhand von Daten zu Qualitätsmerkmalen von Zigaretten dargestellt. Es handelt sich dabei um festgestellte Mängel von geprüften Zigaretten. Die Mängel bzw. Fehler sind in Form von Abweichungen des Durchschnitts, des Gewichts, des Zugwiderstandes und der Ventilation von Normgrenzwerten dieser Merkmale definiert. Die Normgrenzwerte wurden für diese Darstellungszwecke festgelegt. Basisdatei zur Herstellung der Variablen mit Fehlerinformationen (mittels Recodierung) ist die in

Abb. 26.36 (⇨ Kap. 26.8) ausschnittsweise dargestellte Datei ZIGARETT.SAV, die verschiedene metrische Messvariablen von Zigaretten enthält.

26.7.1 Einfaches Pareto-Diagramm

Häufigkeiten oder Summen für Kategorien einer Variablen. Das folgende Übungsbeispiel bezieht sich auf die Variable FEHL_A, die aus den Messdaten der Datei ZIGARETT.SAV (⇨ Abb. 26.36 in Kap. 26.8) durch Umkodierung entstanden ist (= ZIGARET1.SAV). Die Variable FEHL_A ist eine kategoriale Variable, die Mängelarten der auf der Anlage A produzierten Zigaretten erfasst: beispielsweise bedeutet der Variablenwert „1" ein zu kleiner Durchmesser (DM zu klein), „2" ein zu großer Durchmesser (DM zu groß), „3" eine zu geringe Filterventilation (FV zu klein), „5" zu kleiner Zugwiderstand (ZW zu klein), „7" zu kleines Gewicht (GEW zu klein) usw. Die Fälle mit dem Variablenwert „0" (ohne Fehler) werden mittels Menü „Daten, „Fälle auswählen" ausgeschlossen.

Nach der Befehlsfolge „Grafiken", „Pareto..." wird die Auswahlkombination „Einfach" und „Häufigkeiten oder Summen für Kategorien einer Variablen" angeklickt. Abb. 26.29 zeigt links die nach Klicken von „Definieren" geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition und rechts die resultierende Grafik. Die Variable FEHL_A, in der die verschiedenen Mängel der Zigaretten kodiert sind, wurde aus der Quellvariablenliste in das Eingabefeld „Kategorienachse:" übertragen. Im Auswahlfeld „Bedeutung der Balken" ist „Häufigkeiten" angeklickt. Damit wird eine Darstellung der Häufigkeiten der Fehlerarten in Form von Balken angefordert. Alternativ dazu kann auch „Summe der Variablen" gewählt und eine metrische Variable in das vorgesehene Feld übertragen werden. Für jeden Variablenwert der kategorialen Variablen werden dann die Summen der metrischen Variablen dargestellt. Diese Option wäre z.B. dann angebracht, wenn in der Datendatei die Daten anders erfasst sind: neben der kategorialen Variablen erfasst eine metrische Variable die Häufigkeiten von Fehlern.

In der Abb. 26.29 rechts ist das Pareto-Diagramm dargestellt. Auf der waagerechten Grundachse sind die Fehlerarten und auf der senkrechten die Häufigkeiten (links als absolute Anzahl und rechts als prozentualer Wert) der Fehler abgebildet. Im Unterschied zu Balkendiagrammen wird die Reihenfolge der dargestellten Kategorien nach der Größe der Häufigkeit geordnet: der erste Balken stellt die Häufigkeit der Kategorie „mehrere Fehler" dar, da diese die größte Häufigkeit hat. Auf den Balken werden die Häufigkeiten auch zahlenmäßig aufgeführt. Die Kurve zeigt die kumulierten Häufigkeiten. Sie erscheint nur dann, wenn die Option „Kumulative Linie anzeigen" gewählt worden ist.

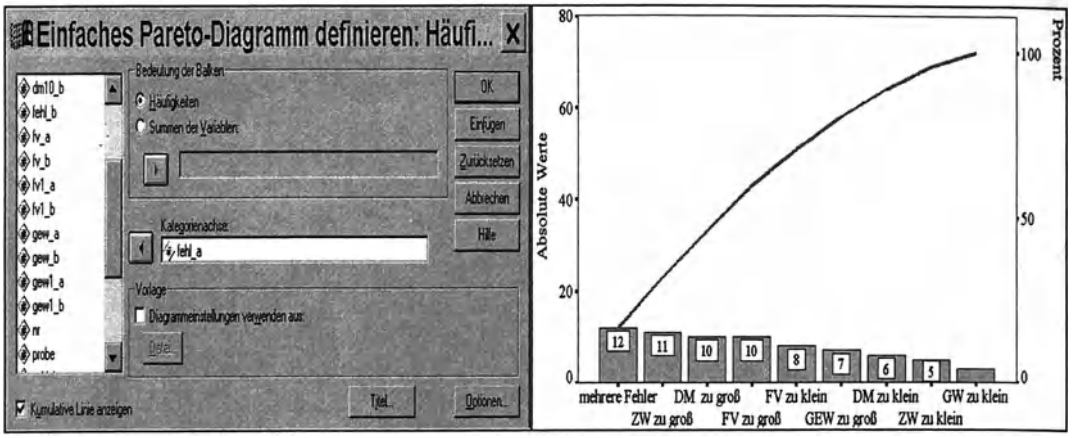


Abb. 26.29. Häufigkeit von Fehlern von Zigaretten im einfachen Pareto-Diagramm

Summen verschiedener Variablen. Für das folgende Demonstrationsbeispiel sind die Informationen zu Fehlern der Zigaretten in anderer Form aufbereitet. In der in Abb. 26.30 dargestellten Datei FEHLER.SAV stellt jeder SPSS-Fall eine Probe von aus der laufenden Produktion entnommenen Zigaretten mit je zwanzig Zigaretten dar. Mit den Variablen N_FDM und N_FGW werden die Häufigkeiten (n) eines fehlerhaften Durchmessers (FDM) bzw. fehlerhaften Gewichts (FGW) erfasst.

	probe	n_fdm	n_fgw
1	1	1	3
2	2	1	0
3	3	1	5
4	4	0	4
5	5	4	2
6	6	5	3

Abb. 26.30. Daten der Datei FEHLER.SAV

Nach der Befehlsfolge „Grafiken", „Pareto..." wird die Auswahlkombination „Einfach" und „Summen verschiedener Variablen" angeklickt. Abb. 26.31 zeigt links die geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition. Die Variablen N_FDM und N_FGW mit den Häufigkeiten eines fehlerhaften Durchmessers bzw. fehlerhaften Gewichts wurden aus der Quellvariablenliste in das Eingabefeld „Variablen:" übertragen.

In der Abb. 26.31 rechts ist das Pareto-Diagramm dargestellt. Auf der waagerechten Achse sind die beiden Variablen und auf der senkrechten die summierten Häufigkeiten abgebildet. Im Unterschied zu Balkendiagrammen wird die Reihenfolge der dargestellten Balken nach deren Höhe geordnet. Da die Option „Kumulative Linie anzeigen" nicht aktiv geschaltet war, wird das Diagramm ohne diese Linie dargestellt.

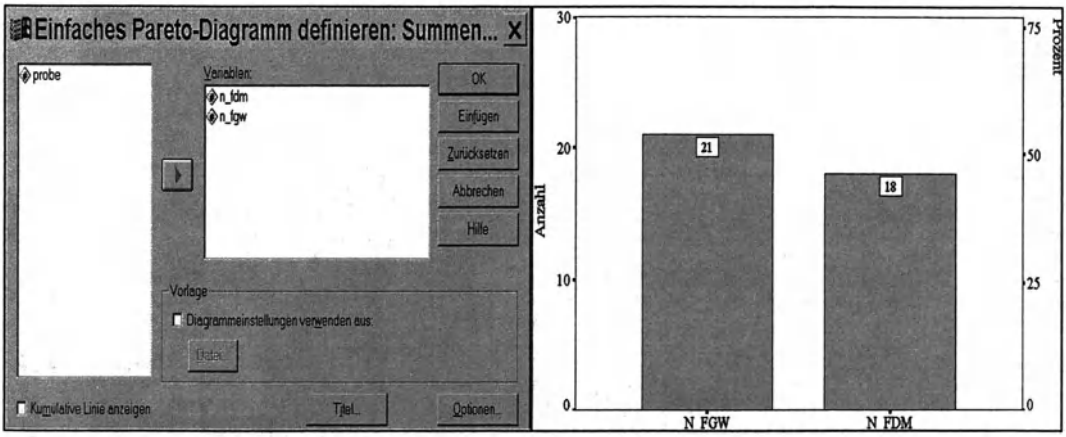


Abb. 26.31. Häufigkeiten fehlerhaften Gewichts bzw. Durchmessers im Pareto-Diagramm

Werte einzelner Fälle. In einem derartigen Diagramm werden für jeden Fall die Variablenwerte in der Reihenfolge ihrer Größe dargestellt: zuerst der Fall mit dem höchsten, dann mit zweithöchsten Wert usw.

Als „Achsenbeschriftung" kann „Fallnummer" oder eine Variable gewählt werden. Im zweiten Fall würde auf der waagerechten Achse für jeden dargestellten Fall der Wertelabel dieser Variablen erscheinen.

26.7.2 Gestapeltes Pareto-Diagramm

Häufigkeiten oder Summen für Kategorien einer Variablen. Zur beispielhaften Darstellung wird wieder die kategoriale Variable FEHL_A mit Werten für verschiedene Fehlerarten der Datei ZIGARET1.SAV (erstellt aus den Messdaten der Datei ZIGARETT.SAV ⇨ Abb. 26.36 in Kap. 26.8) verwendet. Stapelvariable ist die Variable SCHICHT, die mit den Werten „1" und „2" erfasst, ob die Zigaretten aus der Tages- oder Nachtschicht stammen.

Nach der Befehlsfolge „Grafiken". „Pareto..." wird die Auswahlkombination „Gestapelt" und „Häufigkeiten oder Summen für Kategorien einer Variablen" angeklickt. Abb. 26.32 zeigt links die nach Klicken von „Definieren" geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition. Die Variable FEHL_A wurde aus der Quellvariablenliste in das Eingabefeld „Kategorienachse" übertragen. Im Auswahlfeld „Bedeutung der Balken" ist wie im einfachen Pareto-Diagramm „Häufigkeiten" angeklickt. Wie dort beschrieben ist, kann auch die andere Option gewählt werden. Die Variable SCHICHT wurde in das Eingabefeld „Stapelvariable definieren durch" übertragen.

In der Abb. 26.32 rechts ist das Pareto-Diagramm dargestellt. Es gleicht dem für einfache Diagramme mit dem Unterschied, dass die Häufigkeit für jede Fehlerart nach den Werten der Stapelvariable untergliedert wird. Nun kann man erkennen, wie sich die Häufigkeiten eines jeden Fehlers auf die Produktionszeiten Tages- und Nachtschicht aufteilen. Da die Option „Kumulative Linie anzeigen" nicht aktiv geschaltet war, wird das Diagramm ohne diese Linie dargestellt.

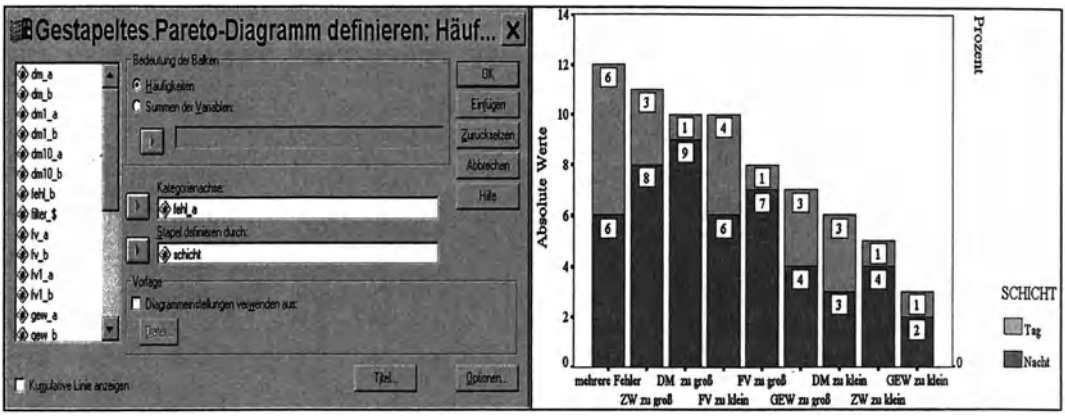


Abb. 26.32. Häufigkeit von Fehlern von Zigaretten untergliedert nach Tages- und Nachtschicht im gestapelten Pareto-Diagramm

Summen verschiedener Variablen. Das folgende Demonstrationsbeispiel bezieht sich auf die in Abb. 26.30 dargestellte Datendatei (FEHLER.SAV), in der die Häufigkeiten von zwei Fehlerarten von Zigaretten erfasst sind. Jeder SPSS-Fall ist eine Probe von aus der laufenden Produktion entnommenen Zigaretten mit je zwanzig Zigaretten. Mit den Variablen N_FDM und N_FGW werden die Häufigkeiten (n) eines fehlerhaften Durchmessers (FDM) bzw. fehlerhaften Gewichts (FGW) erfasst.

Nach der Befehlsfolge „Grafiken", „Pareto..." wird die Auswahlkombination „Gestapelt" und „Summen verschiedener Variablen" angeklickt. Abb. 26.33 zeigt links die geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition und rechts die resultierende Grafik.

Werte einzelner Fälle. Für die Auswahlkombination „Gestapelt" und „Werte einzelner Fälle" werden für jeden Fall die addierten Variablenwerte von zwei Variablen dargestellt. Die Reihenfolge bei der Darstellung der Balken orientiert sich wieder an der Höhe der Balken. Die Grafik entspricht der in Abb. 26.33 mit dem Unterschied, dass die Werte einzelner Fälle überlagert dargestellt werden.

26.7.3 Wahlmöglichkeiten

Für fast alle Pareto-Diagramme bestehen folgende Wahlmöglichkeiten

- Versorgung mit Titel und Fußnoten („Titel").
- Form der Behandlung fehlender Werte („Optionen").
- Grafiklayout aus Vorlage übernehmen („Vorlage").
- „Kumulative Linien anzeigen". Ist diese Option aktiv, so wird im Diagramm eine Kurve der kumulierten Häufigkeiten bzw. Werte angezeigt (⇨ Abb. 26.29).

Außerdem: Überarbeitung im Diagramm-Editorfenster möglich (⇨ Kap. 27.4).

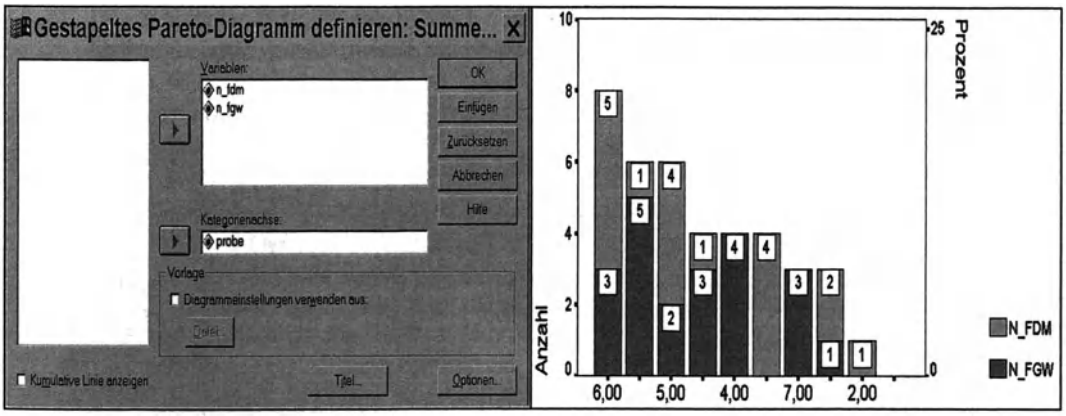


Abb. 26.33. Häufigkeiten fehlerhaften Gewichts bzw. Durchmessers im gestapelten Pareto-Diagramm

26.8 Regelkarten-Diagramme erzeugen

Regelkarten-Diagramme werden in der statistischen Qualitätskontrolle eingesetzt. Zur Überprüfung von laufenden Produktionsprozessen werden z.B. täglich produzierte Einheiten bzw. Stücke zufällig ausgewählt und auf ihre Qualität hin geprüft. Messwerte eines Qualitätsmerkmals - z.B. die Länge eines Werkstücks - können dann in Diagrammen abgebildet werden, um festzustellen, ob Abweichungen der Messwerte vom Normwert zufällig sind oder als bedeutsame Veränderung im Produktionsprozess interpretiert werden müssen.

In der Regel werden die einer Qualitätsprüfung unterzogenen Einheiten bzw. Stücke in Bündeln - hier Untergruppen genannt - ausgewählt und es werden unterschiedliche Arten der Qualitätserfassung verwendet: entweder werden Daten mit metrischer Skala - z.B. die Länge des Werkstücks - erhoben oder es wird nur festgehalten, ob fehlerhafte Stücke vorliegen oder nicht. Davon abhängig und je nachdem, ob ein Bündel (bzw. eine Untergruppe) eine große oder kleine Stückzahl enthält, ob die Gruppen eine konstante oder unterschiedliche Anzahl von Einheiten umfasst, ob verschiedene Fehler von fehlerhaften Stücken erfasst werden oder nicht, ist ein spezifischer Diagrammtyp zur Darstellung der Daten geeignet.

In der folgenden Abbildung 26.34 wird eine Übersicht darüber gegeben.

Art der Qualitätsmessung	Einheiten je Untergruppe	Diagrammtyp
Metrische Skala	≥ 10	X-Quer und R
	< 10	X-Quer und s
	= 1	individuelle Werte und gleitende Spannweite□
Fehlerhafte Stücke	konstant	p oder np
	verschieden	p
Fehlerhafte Stücke mit je mehreren Fehlern	konstant	c oder u
	verschieden	u

Abb. 26.34. Arten von Regelkarten-Diagrammen

Um ein Regelkarten-Diagramm zu erstellen, öffnet man durch Klicken der Befehlsfolge

▷ „Grafiken", „Kontrollkarte..."

die in Abb. 26.35 dargestellte Dialogbox.

Die Daten in der Datendatei können unterschiedlich organisiert sein:

- *Fälle sind Einheiten.* Jeder Fall in der SPSS-Datendatei ist ein einzelnes kontrolliertes Stück.
- *Fälle sind Untergruppen.* Jeder Fall in der SPSS-Datendatei stellt eine Untergruppe dar. Eine Untergruppe ist in der Regel ein Zeitintervall (z.B. die Messergebnisse eines Tages).

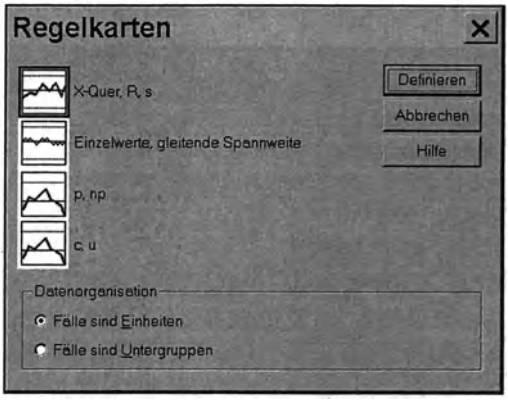


Abb. 26.35. Dialogbox zur Auswahl eines Regelkarten-Diagramms

Für Regelkarten-Diagramme sind folgende vier Formen wählbar:

- *X-Quer, R, s:* Es handelt sich hierbei um verschiedene Diagramme:
  - *X-Quer-Diagramm.* Wenn die Fälle Einheiten sind, wird für jede definierte Gruppe von Einheiten der Gruppen-Mittelwert einer metrischen Variablen (häufig mit dem Symbol $\bar{x}$ bezeichnet) dargestellt. Wenn die Fälle Untergruppen sind, so wird der Mittelwert von mehreren metrischen Variablen dargestellt.
  - *R-Diagramm.* Wenn die Fälle Einheiten sind, wird für jede definierte Gruppe von Einheiten die Gruppenspannweite R (= Range = die Differenz zwischen dem kleinsten und größten Wert in einer Gruppe) dargestellt. Wenn die Fälle Untergruppen sind, so wird die Spannweite einer Gruppe von Variablen dargestellt.
  - *s-Diagramm.* Wenn die Fälle Einheiten sind, wird für jede definierte Gruppe von Einheiten die Gruppenstandardabweichung s dargestellt. Wenn die Fälle Untergruppen sind, so wird die Standardabweichung von mehreren Variablen dargestellt.
- *Einzelwerte, gleitende Spannweite.* Es handelt sich um zwei Diagramme zur Darstellung der Messwerte einer metrischen Variablen:
  - *Einzelwerte.* Dargestellt werden die Messwerte von einzelnen Stücken.

- *Gleitende Spannweite.* Dargestellt wird jeweils die Differenz der Messwerte aufeinanderfolgender Stücke.

❑ *p, np.* Es handelt sich um zwei Diagramme zur Darstellung der Anzahl der fehlerhaften Stücke:
- *p-Diagramm.* Dargestellt wird für jede Untergruppe die Häufigkeit von fehlerhaften Stücken in Form des Anteils p an allen Stücken der Untergruppe.
- *np-Diagramm.* Dargestellt wird die absolute Anzahl der fehlerhaften Stücke in jeder Untergruppe (p*n = Fehleranteil*Stückzahl)

❑ *c, u.* Es handelt sich ebenfalls um zwei Diagramme zur Darstellung der Häufigkeit von fehlerhaften Stücken. Sie werden angewendet, wenn die Daten in anderer Form aufbereitet vorliegen.

Im folgenden werden einige dieser verschiedenen Diagrammformen anhand von Beispielen dargestellt.

26.8.1 Diagrammtyp: X-Quer, R, s

Datenorganisation: Fälle sind Einheiten. In der Datei ZIGARETT.SAV, die ausschnittsweise in Abb. 26.36 im Dateneditorfenster dargestellt ist, sind mehrere metrische Variablen mit Messwerten von Zigaretten erfasst. Jeder Fall ist eine Zigarette aus einem Produktionsprozess A bzw. B. Stündlich wurden Zufallsproben von je 20 Zigaretten aus den laufenden Produktionsprozessen entnommen und geprüft. Insgesamt sind in dem Datensatz je Produktionsprozess Messwerte von 10 Proben à 20 Zigaretten enthalten. Die Variable PROBE mit den Werten 1 bis 10 dient zur Identifikation der Proben. Es soll die Variable DM_A, die den Durchmesser der auf der Anlage A produzierten Zigaretten erfasst, in Kontrolldiagrammen dargestellt werden.

	nr	probe	gew_a	zw_a	fv_a	dm_a	gew_b	zw_b	fv_b	dm_b	dm10_	dm10_
1	1,00	1,00	614,00	123,00	57,9	5,42	622,00	140,00	59,3	5,42	54,20	54,20
2	2,00	1,00	572,00	123,00	50,4	5,37	607,00	145,00	56,2	5,42	53,70	54,20
3	3,00	1,00	634,00	126,00	54,2	5,42	638,00	150,00	59,9	5,42	54,20	54,20
4	4,00	1,00	605,00	132,00	52,0	5,38	600,00	145,00	55,6	5,41	53,80	54,10
5	5,00	1,00	612,00	136,00	52,2	5,39	584,00	130,00	54,7	5,42	53,90	54,20
6	6,00	1,00	576,00	130,00	46,2	5,40	647,00	157,00	57,8	5,42	54,00	54,20
7	7,00	1,00	618,00	141,00	53,3	5,41	605,00	152,00	55,7	5,42	54,10	54,20

Abb. 26.36. Ausschnitt aus dem Datensatz ZIGARETT.SAV

Nach der Befehlsfolge „Grafiken", „Regelkarten..." wird die Auswahlkombination „X-Quer, R, s" und „Fälle sind Einheiten" angeklickt. Abb. 26.37 zeigt links die nach Klicken von „Definieren" geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition. Als Diagramm ist „X-Quer und Spannweite" gewählt. Die metrische Variable DM_A wurde in das Feld „Prozessmessung" und die Variable PROBE in das Feld „Untergruppe definiert durch" übertragen. Nach Klicken auf „OK" werden zwei Diagramme erzeugt. In der Abb. 26.37 rechts ist das erste Diagramm - das X-Quer-Diagramm - zu sehen. Für jede der zehn Stichproben à 20 Zigaretten

wird der Mittelwert der Durchmesser dargestellt. Mit 5,3981 wird der Mittelwert aller Zigaretten als waagerechte Linie angezeigt. Um den Mittelwert werden als unterbrochene Linien der obere (UCL = upper control limit) und der untere (LCL = lower control limit) Kontrollwert in Form eines Drei-Sigma-Bereichs angezeigt. In die durch Mausklick auf „Optionen" geöffnete Unterdialogbox „X-Quer, R, s: Optionen" kann ein anderer Sigma-Bereich (beruhend auf der Normalverteilung) gewählt werden. Anstelle eines Sigma-Bereichs können auch feste Ober- bzw. Untergrenzen eingegeben werden. In „Mindest-Stichprobengröße der Untergruppen" kann eine Stichprobengröße vorgegeben werden. Ist die Stichprobengröße der Untergruppe kleiner, so wird diese Untergruppe nicht in die Grafik und die Berechnungen einbezogen.
Es zeigt sich, dass die Mittelwerte einiger Proben aus dem abgesteckten Kontrollintervall herausfallen.

In dem zweiten Diagramm erfolgt eine analoge Darstellung, mit dem Unterschied, dass die Spannweite (Differenz zwischen größtem und kleinstem Wert in jeder Probe) auf der senkrechten Achse abgebildet wird.

Wird „X-Quer und Standardabweichung" in Abb. 26.37 gewählt, so werden ebenfalls zwei Diagramme erstellt: Das erste ist das gleiche X-Quer-Diagramm, und das zweite bildet auf der senkrechten Achse die Standardabweichung s der Untergruppen ab.

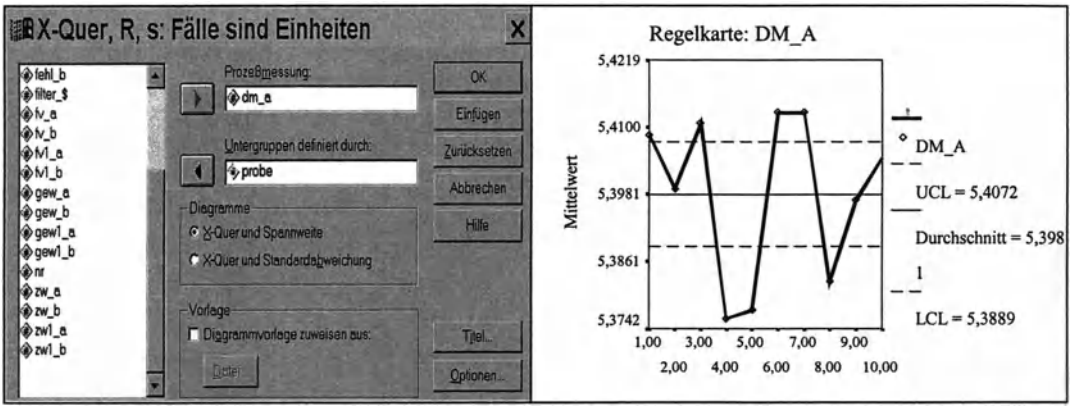


Abb. 26.37. Mittelwert des Durchmessers von Zigaretten in einem X-Quer-Diagramm

Datenorganisation: Fälle sind Untergruppen. Die Fälle der SPSS-Datei stellen Gruppen dar. Zur Erläuterung wird wieder die Datei ZIGARETT.SAV benutzt (⇨ Abb. 26.36). Es werden nun mehrere Variablen betrachtet: GEW_A und GEW_B sind die Gewichte von Zigaretten, die auf den Anlagen A und B produziert werden. Ein Fall erfasst nun eine Gruppe (hier nur zwei) von Messungen: GEW_A und GEW_B.

Nach der Befehlsfolge „Grafiken", „Regelkarten..." wird die Auswahlkombination „X-Quer, R, s" und „Fälle sind Untergruppen" geklickt. Die Variablen GEW_A und GEW_B werden in das Eingabefeld „Stichproben" übertragen. Es werden zwei Diagramme erzeugt: Im ersten wird für jeden Fall der Mittelwert der

Durchmesser und im zweiten die Spannweite der auf den Anlagen A und B produzierten Zigaretten dargestellt. Da sich die erzeugten Diagramme nicht von denen für „Fälle sind Einheiten" unterscheiden, sei zur Interpretation auf die Ausführungen oben verwiesen.

26.8.2 Diagrammtyp: Einzelwerte, gleitende Spannweite

Datenorganisation: Fälle sind Einheiten. Im Demonstrationsbeispiel wird als darzustellende Prozessvariable wieder DM_A des Datensatzes ZIGARETT.SAV verwendet (⇨ Abb. 26.36). Da auf der waagerechten Achse des Diagramms die einzelnen Zigaretten dargestellt werden, wurden mit der Befehlsfolge „Daten", „Fälle auswählen" nur die ersten acht Fälle (Zigaretten) selektiert.

Nach der Befehlsfolge „Grafiken", „Regelkarten..." wird die Auswahlkombination „Einzelwerte, gleitende Spannweite" und „Fälle sind Einheiten" angeklickt. Abb. 26.38 zeigt links die nach Klicken von „Definieren" geöffnete Dialogbox zur Grafikdefinition. Als Diagramme sind „Einzelwerte und gleitende Spannweite" angeklickt. Die metrische Variable DM_A wurde in das Feld „Prozessmessung" und die Variable PROBE in das Feld „Untergruppenbeschriftung:" übertragen. Nach Klicken auf „OK" werden zwei Diagramme erzeugt. In der Abb. 26.38 rechts ist das zweite Diagramm - es zeigt die gleitenden Spannweiten - dargestellt. Da für „Spanne" der Wert „2" gewählt wurde, werden die Durchmesserdifferenzen von im Datensatz aufeinander folgenden Zigaretten abgebildet. Da die Zigaretten aus der ersten Probe stammen, erhält jede die Beschriftung „1". Die Spannweite kann durch Eintragen einer anderen Zahl erhöht werden. Wie auch in X-Quer-Diagrammen wird der Durchschnitt sowie ein Drei-Sigma-Bereich durch Kontrollinien angezeigt. Die untere Kontrollinie wird aber nur dargestellt, wenn sie im positiven Bereich liegt.

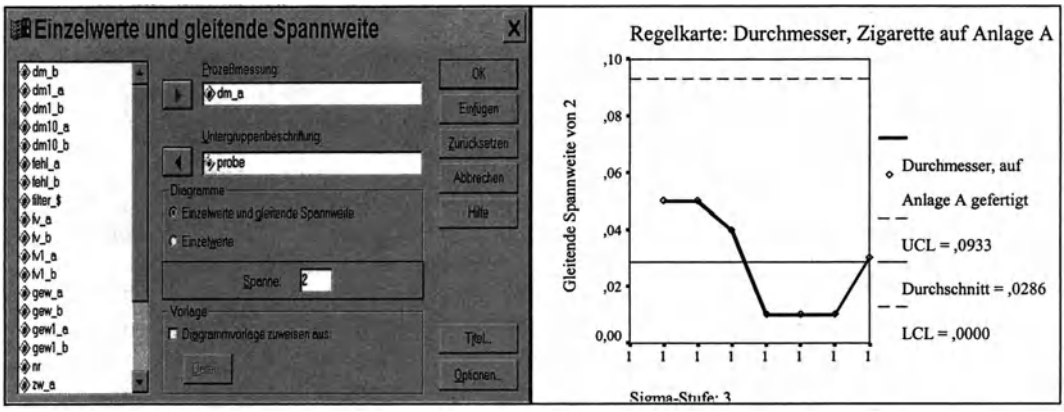


Abb. 26.38. Gleitende Spannweite von Zigarettendurchmessern

26.8.3 Diagrammtyp: p, np

Datenorganisation: Fälle sind Einheiten. Dieser Programmtyp wird gewählt, wenn die Qualitätsdaten in qualitativer Form vorliegen. Bei Wahl des p-Diagramms wird der Anteil und bei Wahl des np-Diagramms die absolute Anzahl von fehlerhaften Stücken in jeder Stichprobe grafisch dargestellt. Zur Anwendungsdemonstration wurde die Variable GEW_A im Datensatz ZIGARETT (⇨ Abb. 26.36), die das Gewicht der Zigaretten in mg misst, in eine Variable GEW1_A rekodiert: der Variablenwert „2" bildet ein (für dieses Beispiel) normgemäßes, „1" ein zu geringes und „3" ein zu großes Gewicht ab (Datei ZIGARET1.SAV).

Nach der Befehlsfolge „Grafiken", „Regelkarten..." wird die Auswahlkombination „p, np" und „Fälle sind Einheiten" angeklickt. Abb. 26.39 zeigt links die nach Klicken von „Definieren..." geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition. Es ist „p (Anteil der Abweichenden)" gewählt worden. Die Variable GEW1_A wurde in das Feld „Merkmal" und die Variable PROBE in das Feld „Untergruppen definiert durch" übertragen. Es wurde „Fehlerfreie" als „Zu zählender Wert" gewählt. Es wird der Anteil der Zigaretten mit fehlerhaftem Gewicht - gemessen an allen in einer Probe enthaltenen Zigaretten - grafisch dargestellt. Da normgemäße Zigaretten mit dem Wert „2" kodiert worden sind, wird in das Eingabefeld „Wert" eine 2 eingetragen. Nach Klicken auf „OK" wird die Regelkarte erzeugt. In der Abb. 26.39 rechts ist das Diagramm dargestellt. Für jede der 10 Proben à 20 Zigaretten wird die Quote der Zigaretten mit fehlerhaftem (zu geringes oder zu hohes) Gewicht dargestellt. Mit Zentrum = 0,09 wird in Form einer waagerechten Linie die mittlere Quote der Zigaretten mit fehlerhaftem Gewicht dargestellt. Um den Mittelwert werden als unterbrochene Linien der obere (UCL = upper control limit) und der untere (LCL = lower control limit) Kontrollwert in Form eines Drei-Sigma-Bereichs angezeigt. Da hier LCL im negativen Bereich liegen würde, wird die Linie nicht abgebildet. Im Menü „Optionen" kann ein anderer Sigma-Bereich gewählt werden.

Wird „np (Anzahl der abweichenden Einheiten)" in Abb. 26.39 gewählt, so werden die fehlerhaften Stücke in absoluter Anzahl grafisch dargestellt. Diese Auswahl macht nur dann Sinn, wenn die Anzahl der Stücke in jeder Stichprobe konstant ist.

Datenorganisation: Fälle sind Untergruppen. Dieser Diagrammtyp eignet sich für Qualitätsdaten, die in Form von Häufigkeiten vorliegen. Zur Demonstration werden die in Abb. 26.40 dargelegten Daten verwendet (Datei ZWISCHF.SAV). Es handelt sich bei ZWISCHF um die Häufigkeit von unvorhergesehenen Zwischenfällen bei in sechs Monaten durchgeführten Operationen. Diese Daten entsprechen denen in Abb. 26.42, mit dem Unterschied, dass sie anders aufbereitet sind: die wöchentlichen Operationen eines Monats sind hier zusammengefasst. Die Variable N gibt die Anzahl der Operationen pro Monat an:

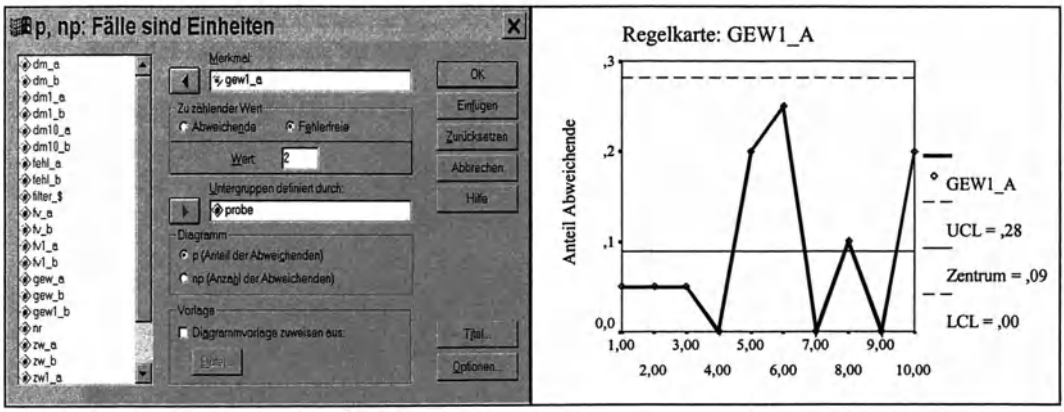


Abb. 26.39. Anteil nicht normgerechter Durchmesser von Zigaretten im p-Diagramm

	monat	zwischf	n
1	Januar	0	20
2	Februar	4	40
3	März	1	30
4	April	3	25
5	Mai	1	30
6	Juni	2	40

Abb. 26.40. Daten des Anwendungsbeispiels (ZWISCHF.SAV)

Nach der Befehlsfolge „Grafiken", „Regelkarten..." wird die Auswahlkombination „p, np" und „Fälle sind Untergruppen" angeklickt. Abb. 26.41 zeigt links die nach Klicken von „Definieren" geöffnete Dialogbox mit dem Beispiel zur Grafikdefinition. Es ist „p (Anteil der Abweichenden)" angeklickt. Die Variable ZWISCHF wurde in das Feld „Anzahl der Abweichenden:" und die Variable MONAT in das Feld „Untergruppenbeschriftung" übertragen. In „Größe der Stichprobe" wurde „Variable" angeklickt und die Variable n - sie enthält die Anzahl der Operationen je Monat - in das Eingabefeld übertragen. Diese Option ist zu wählen, wenn die Stichprobengröße je Untergruppe (hier ein Monat) variiert. Für den Fall gleicher Stichprobengröße je Untergruppe kann im Feld „Stichprobengröße" „konstant" gewählt werden und anschließend die Stichprobengröße in das dafür vorgesehene Eingabefeld eingetippt werden.

In Abb. 26.41 rechts ist das Kontrolldiagramm dargestellt. Für jede der sechs Untergruppen (= Monat) wird der Anteil der Zwischenfälle bei Operationen dargestellt. Das Diagramm entspricht dem in Abb. 26.39. Daher wird auf die dort gegebene Kommentierung verwiesen.

Wird „np (Anzahl der Abweichenden)" in Abb. 26.41 gewählt, so wird die Anzahl der Zwischenfälle in absoluter Anzahl grafisch dargestellt. Diese Auswahl macht nur dann Sinn, wenn die Anzahl der Operationen in jedem Monat (= Untergruppe) konstant ist.

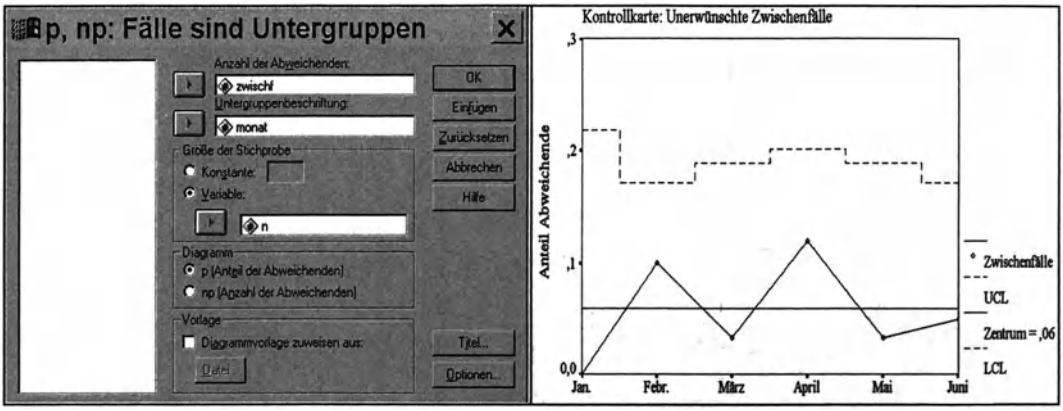


Abb. 26.41. Anteil der Zwischenfälle bei Operation im p-Diagramm

26.8.4 Diagrammtyp: c, u

Datenorganisation: Fälle sind Einheiten. Dieser Diagrammtyp ist bei einer anderen Datenlage geeignet. Die Daten liegen in Form von Häufigkeiten von Fehlern bzw. unerwünschter Ereignisse vor. In Abb. 26.42 wird ein Auszug aus der Datei ZWISCHF1.SAV gegeben: Die SPSS-Fälle sind operative Eingriffe in einer Woche. Mit der Variablen ZWISCHF wird die Anzahl unerwünschter Zwischenfälle bei den Operationen erfasst. Eine zweite Variable MONAT (= Untergruppe) erfasst, in welchem Monat eine Operation stattgefunden hat. Bei Wahl des u-Diagramms wird für jeden Monat die Anzahl von Zwischenfällen je Woche und bei Wahl des c-Diagramms die absolute Anzahl von Zwischenfällen pro Monat (= in jeder Untergruppe) grafisch dargestellt.

	monat	woche	zwischf	n
1	Januar	1	0	5
2	Januar	2	0	4
3	Januar	3	0	6
4	Januar	4	0	5
5	Februar	1	1	10
6	Februar	2	0	8
7	Februar	3	2	12
8	Februar	4	1	10

Abb. 26.42. Daten des Anwendungsbeispiels (ZWISCHF1.SAV)

Nach der Befehlsfolge „Grafiken", „Regelkarten..." wird die Auswahlkombination „c, u" und „Fälle sind Einheiten" angeklickt. Abb. 26.43 zeigt links die nach Klicken von „Definieren" geöffnete Dialogbox mit dem Beispiel zur Grafikdefinition. Als Diagramm ist „c (Anzahl der Abweichungen)" angeklickt. Die Variable ZWISCHF wurde in das Feld „Merkmal" und die Variable MONAT in das Feld „Untergruppen definiert durch" übertragen.

In der Abb. 26.43 rechts ist das Diagramm dargestellt. Für jeden in der Datei enthaltenen Monat, wird die Anzahl von Zwischenfällen dargestellt. Die Angaben

Zentrum, UCL und LCL sind oben erklärt. Wird „u (Abweichungen je Einheit)“ in Abb. 26.43 gewählt, so wird für jeden Monat die Anzahl der Zwischenfälle je Woche (= Einheit) grafisch dargestellt.

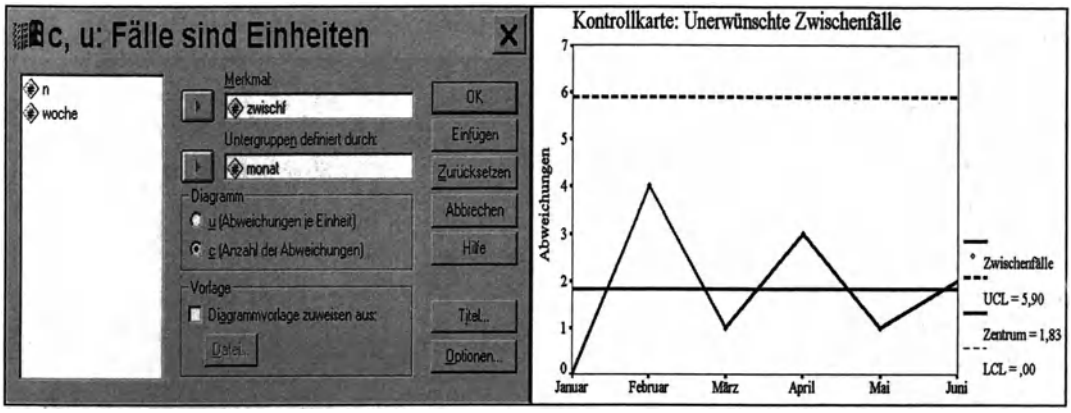


Abb. 26.43. Anzahl von Zwischenfällen im u-Kontrolldiagramm

26.8.5 Wahlmöglichkeiten

Für Kontrollinien-Diagramme bestehen folgende Wahlmöglichkeiten:

- „Optionen“:
  - Es kann der Sigma-Bereich, d.h. die Anzahl der Standardabweichungen oberhalb und unterhalb der Mittellinie, gewählt werden (voreingestellt ist ein Drei-Sigma-Bereich).
  - Für X-Quer-Diagramme können zusätzlich durch Eingabe von „Maximum“- und „Minimum“-Werten weitere Kontrollinien spezifiziert werden.
  - Untergruppen mit Missing-Werten können angezeigt werden.
  - In X-Quer-Diagrammen kann ein minimaler Umfang einer Untergruppe spezifiziert werden. Untergruppen mit kleinerem Umfang werden dann im Diagramm nicht dargestellt.
- Versorgung mit Titel und Fußnoten („Titel“, ⇨ Kap. 26.3).
- Grafiklayout aus einer Vorlage übernehmen („Vorlage“, ⇨ Kap. 26.3).

Außerdem: Überarbeitung im Diagramm-Editorfenster möglich (⇨ Kap. 27.4).

26.9 Boxplot-Diagramme erzeugen

In einem Boxplot-Diagramm wird für jede Kategorie einer kategorialen Variablen die Streuung einer anderen Variablen grafisch abgebildet.

Um ein Boxplot-Diagramm zu erstellen, öffnet man durch Klicken der Befehlsfolge

▷ „Grafiken“, „Boxplot...“

die in Abb. 26.44 dargestellte Dialogbox.

Als Boxplot-Diagrammtypen sind ein *einfaches* und ein *gruppiertes* Boxplot-Diagramm wählbar. Dabei können für beide Diagrammtypen die Grafikdaten auf der Grundachse des Boxplot-Diagramms entweder Kategorien einer Variablen oder verschiedene Variablen abbilden. Im folgenden werden einige dieser verschiedenen Boxplot-Diagrammformen anhand von Beispielen aus dem ALLBUS90-Datensatz erläutert.

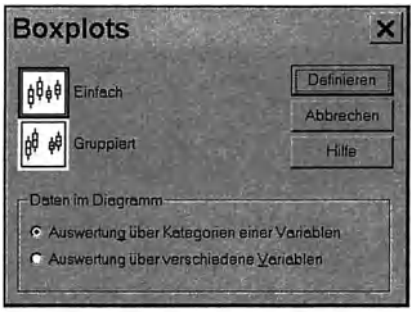


Abb. 26.44. Dialogbox zur Auswahl eines Boxplot-Diagramms

26.9.1 Einfaches Boxplot-Diagramm

Auswertung über Kategorien einer Variablen. Nach der Befehlsfolge „Grafiken", „Boxplot..." wird die Auswahlkombination „Einfach" und „Grafikdaten repräsentieren Kategorien einer Variable" angeklickt. Abb. 26.45 zeigt die nach Klicken von „Definieren" geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition und die resultierende Grafik (um den Zusammenhang Einkommen/Schulbildung möglichst eng abzubilden, wurden vorher mittels der Befehlsfolge „Daten", „Fälle auswählen" nur die Fälle mit positiven Arbeitsstunden des Befragten (ARBSTD > 0) einbezogen). Die Variable SCHUL mit den Schulabschlüssen als Kategorien wurde aus der Quellvariablenliste in das Eingabefeld „Kategorienachse:" und die Variable EINK in das Eingabefeld „Variable" übertragen. Ergebnis ist ein Boxplot-Diagramm, das zusammenfassende statistische Maßzahlen über die Verteilung der Einkommen der Befragten für die einzelnen Schulabschlüsse abbildet. Die untere Kante der Kästen zeigt den 25-Prozentwert (25. Perzentil = 1. Quartil), die waagerechte Linie innerhalb der Kästen den Median (auch Zentralwert bzw. 50-Prozentwert oder 50. Perzentil genannt) und die obere Kante den 75. Prozentwert (75. Perzentil = 3. Quartil): Daher liegen innerhalb der Kästen 50 % der Fälle. Aus einem Boxplot kann auch eine Erkenntnis über die Schiefe der Verteilung abgelesen werden. Aus Abb. 26.45 ist z.B. zu erkennen, dass die Verteilung der Nettoeinkommen der Befragten mit Fachhochschule und Abitur als höchstem Schulabschluss im mittleren Bereich schief ist: rechtssteil bei Fachhochschülern und linkssteil bei Abiturienten.

Des weiteren werden zwei Arten von entlegenen Fällen gezeigt. *Extremwerte* sind Fälle, die mehr als drei Kastenlängen vom oberen bzw. unteren Kastenrand entfernt liegen. Diese sind mit einem Stern (*) gekennzeichnet. *Ausreißer* sind

Fälle, die 1,5 bis 3 Kastenlängen vom oberen bzw. unteren Kastenrand entfernt liegen. Diese sind mit einem Kreis (∘) gekennzeichnet.

In der Dialogbox wurde die Variable GESCHL in das Eingabefeld „Fallbeschriftung" übertragen. Mit dieser optionalen Angabe wird veranlasst, dass die Extremwerte und Ausreißer mit dem Werte-Label dieser Variablen gekennzeichnet werden. In diesem Beispiel handelt es sich dabei um Männer. Verzichtet man auf diese optionale Angabe, so werden die Fallnummern zur Kennzeichnung genommen. Von der unteren und oberen Kastenkante sind senkrechte Linien mit Querbalken gezogen. Mit diesen Linien werden die größten und kleinsten Werte (ausgenommen Extremwerte und Ausreißer) eingegrenzt. Da diese Linien im angelsächsischen Sprachraum whiskers genannt werden, hat sich für das Diagramm auch der Ausdruck *Box-and-Whisker-Plot* eingebürgert.

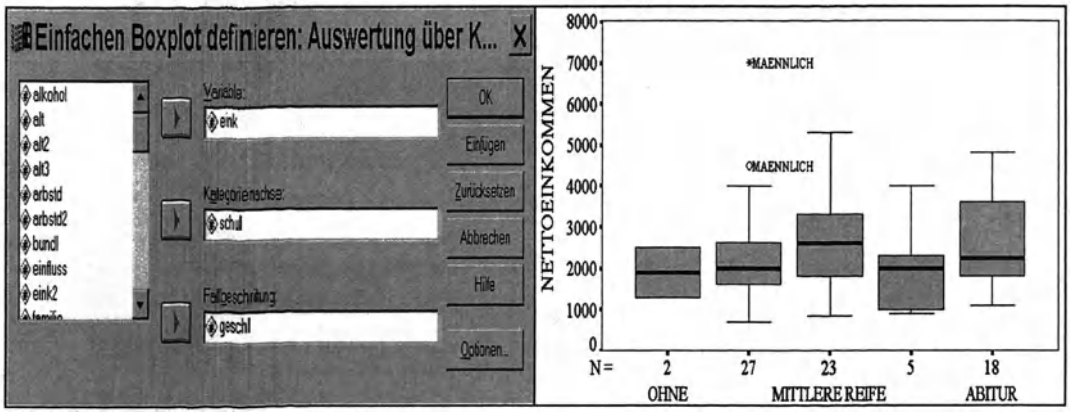


Abb. 26.45. Einkommensverteilung für Schulabschlüsse der Befragten

Auswertung über verschiedene Variablen. Nach der Befehlsfolge „Grafiken", „Boxplot..." wird die Auswahlkombination „Einfach" und „Auswertung über verschiedene Variablen" angeklickt. Abb. 26.46 zeigt die nach Klicken von „Definieren" geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition und die resultierende Grafik. Die Variablen LOHNS (= Nettoeinkommen/Arbeitsstunden) und ARBSTD (Arbeitsstunden) wurden in das Feld „Boxen entspricht" übertragen. Auf die Option „Fallbeschriftung" wurde verzichtet.

26.9.2 Gruppiertes Boxplot-Diagramm

Auswertung über Kategorien einer Variablen. Nach der Befehlsfolge „Grafiken", „Boxplot..." wird die Auswahlkombination „Gruppiert" und „Auswertung über Kategorien einer Variablen" angeklickt. Die nach Klicken von „Definieren" geöffnete Dialogbox ähnelt der in Abb. 26.45 für ein einfaches Boxplot-Diagramm. Ergänzend zu den Eingabefeldern wird in „Gruppen definieren durch:" (analog zu gruppierten Balkendiagrammen) eine Gruppierungsvariable (z.B. GESCHL) übertragen. Im Unterschied zur Abb. 26.45 wird im Diagramm dann

für jeden Schulabschluss die Verteilung der Einkommen untergliedert nach Männern und Frauen dargestellt.

Auswertung über verschiedene Variablen. Bei der Auswahlkombination „Gruppiert" und „Auswertung über verschiedene Variablen" wird ebenfalls eine Gruppierungsvariable (z.B. GESCHL) in ein Eingabefeld der Dialogbox übertragen. Der in Abb. 26.46 dargestellte Vergleich der Verteilung des Lohnsatzes (LOHNS = Einkommen/Arbeitsstunden) und der Arbeitstunden pro Woche wird nun jeweils für Frauen und für Männer dargestellt.

26.9.3 Wahlmöglichkeiten

Für alle Boxplot-Diagramme bestehen folgende Wahlmöglichkeiten:

- Fallbeschriftung (⇨ Abb. 26.45).
- Form der Behandlung fehlender Werte („Optionen") (⇨ Kap. 26.2.1).

Außerdem: Überarbeitung im Diagramm-Editorfenster möglich (⇨ Kap. 27.4).

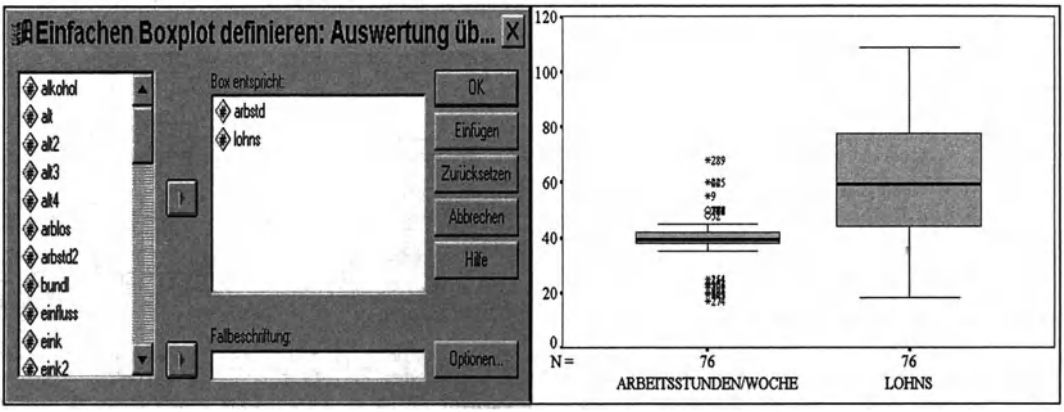


Abb. 26.46. Verteilung des Lohnsatzes und der Arbeitsstunden pro Woche

26.10 Fehlerbalkendiagramme erzeugen

Ein Fehlerbalkendiagramm hat ähnlich wie ein Boxplot-Diagramm die Aufgabe, für Kategorien von kategorialen Variablen die Streuung einer anderen metrischen Variablen zu visualisieren. Im Unterschied zu Boxplot-Diagrammen, in denen die Quartile und somit die Quartilsabstände der anderen Variablen als Streuungsmaß abgebildet werden, können in einem Fehlerbalkendiagramm Konfidenzbereiche für den unbekannten Mittelwert der Grundgesamtheit bzw. Streuungsbereiche der metrischen Variablen dargestellt werden.
Um ein Fehlerbalkendiagramm zu erstellen, öffnet man durch Klicken der Befehlsfolge

▷ „Grafiken", „Fehlerbalken..."

die in Abb. 26.47 dargestellte Dialogbox.

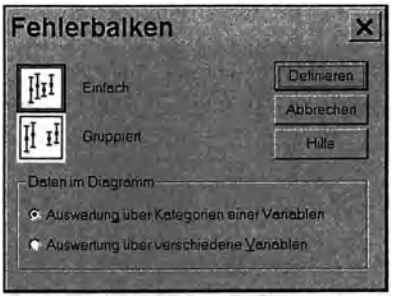


Abb. 26.47. Dialogbox zur Auswahl eines Fehlerbalkendiagramms

Aus Abb. 26.47 wird ersichtlich, dass ein *einfaches* und ein *gruppiertes* Fehlerbalkendiagramm zur Auswahl stehen. Dabei können wie bei Boxplot-Diagrammen für beide Diagrammtypen die Grafikdaten auf der Bodenachse des Boxplot-Diagramms entweder Kategorien einer Variablen oder verschiedene Variablen abbilden. Im folgenden wird das einfache Fehlerbalkendiagramm anhand des ALLBUS90-Datensatzes exemplarisch erläutert.

26.10.1 Einfaches Fehlerbalkendiagramm

Auswertung über Kategorien einer Variablen. Nach der Befehlsfolge „Grafiken", „Fehlerbalken..." wird die Auswahlkombination „Einfach" und „Auswertung über Kategorien einer Variablen" angeklickt. Abb. 26.48 zeigt die nach Klicken von „Definieren" geöffnete Dialogbox mit einem Beispiel zur Grafikdefinition und die resultierende Grafik. Die kategoriale Variable SCHUL wurde aus der Quellvariablenliste in das Eingabefeld „Kategorienachse:" und die metrische Variable ARBSTD (Arbeitsstunden/Woche) in das Eingabefeld „Variable" übertragen.

Zur Darstellung von Streuungsbereichen von ARBSTD in Form von Balken bestehen folgende Auswahlmöglichkeiten:

- *Konfidenzintervall für den Mittelwert.* Die auszuwertenden Fälle werden als eine Zufallsstichprobe aus einer Grundgesamtheit interpretiert. Ein Konfidenzbereich gibt an, in welchen Grenzen der unbekannte Mittelwert für die Arbeitsstunden der Grundgesamtheit bei einer vorzugebenden Wahrscheinlichkeit bzw. einem Sicherheitsgrad erwartet werden kann. Voreingestellt ist ein Sicherheitsgrad von 95 %. Es kann auch ein anderer Sicherheitsgrad gewählt werden (Eingabefeld „Niveau"). Das Konfidenzintervall ergibt sich als (zu Konfidenzintervalle ⇨ Kap. 8.4).

$$\bar{x} \pm t\frac{s}{\sqrt{n}} \qquad (26.1)$$

$\bar{x}$ = Mittelwert der metrischen Variablen der Stichprobe,
s = Standardabweichung der metrischen Variablen der Stichprobe,
t = Sicherheitsgrad (entspricht einer Wahrscheinlichkeit der t-Verteilung),

n = Stichprobenumfang (gültige Fallzahl).

- *Standardfehler Mittelwert* ($=s/\sqrt{n}$). Auch bei dieser Option wird ein Konfidenzintervall für den unbekannten Mittelwert dargestellt. Im Unterschied zu oben wird dieser durch t einer t-Verteilung, das einer Wahrscheinlichkeit entspricht, gewählt. Diese Variante sollte nur bei hohen Fallbesetzungen n für die einzelnen Kategorien gewählt werden, weil der in der Dialogbox einzugebende t-Wert (Eingabefeld „Multiplikator") für alle Kategorien angewendet wird. Nur bei hohen Fallbesetzungen kann man davon ausgehen, dass die in der Grafik abgesteckten Bereiche gleichen Wahrscheinlichkeiten entsprechen.
- *Standardabweichung.* Es wird ein Streuungsbereich um den Mittelwert gemäß Gleichung 26.2 durch Festlegen von t (das einer Wahrscheinlichkeit entspricht) dargestellt:

$$\bar{x} \pm ts \qquad (26.2)$$

Die Auswahl erfolgt aus einer Drop-Down-Liste, die man durch Klicken auf den Pfeil im Auswahlfeld „Bedeutung der Balken" öffnet.

In Abb. 26.48 wird links in der dargestellten Dialogbox der 95 %-Konfidenzbereich für die durchschnittlichen Arbeitsstunden für jeden Schulabschluss angefordert. Rechts ist das resultierende Diagramm zu sehen (CI = confidence interval). Bei der Grafikerstellung wurde durch Fallselektion SCHUL =1 ausgeschlossen.

Wird für „Bedeutung der Balken" „Standardfehler Mittelwert" bzw. „Standardabweichung" gewählt, so kann ein t-Wert für „Multiplikator" eingegeben werden. Auf der senkrechten Achse der Grafik erscheint sinngemäß die Beschriftung „Mean ± t*SE*Variablennamen" bzw. „Mean ± t*SD*Variablennamen" (Mean = Mittelwert, SE = Standarderror, SD = Standarddeviation).

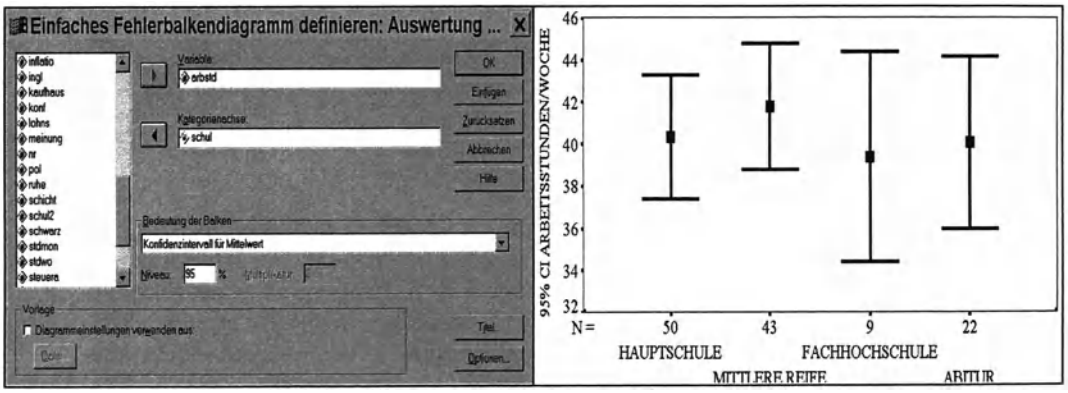


Abb. 26.48. 95 %-Konfidenzbereiche für die durchschnittlichen Arbeitsstunden/Woche von Befragten nach Schulabschluss

Auswertung über verschiedene Variablen. Nach der Befehlsfolge „Grafik", „Fehlerbalken..." wird die Auswahlkombination „Einfach" und „Auswertung über verschiedene Variablen" geklickt. Die Vorgehensweise und die Dialogbox entsprechen denen für Boxplots

26.10.2 Gruppiertes Fehlerbalkendiagramm

Auswertung über Kategorien einer Variablen bzw. über verschiedene Variablen. Die Vorgehensweise entspricht der für die Erstellung von Boxplots

Wahlmöglichkeiten. Folgende weitere Einstellungen sind möglich:

- Versorgung mit Titel („Titel").
- Form der Behandlung fehlender Werte („Optionen").
- Grafiklayout aus Vorlage entnehmen („Vorlage").

Außerdem: Überarbeitung im Diagramm-Editorfenster möglich (⇨ Kap. 27.4).

26.11 Streudiagramme erzeugen

Um ein Streudiagramm zu erstellen, öffnet man durch Klicken der Befehlsfolge

▷ „Grafiken", „ Streudiagramm..."

die in Abb. 26.49 dargestellte Dialogbox.

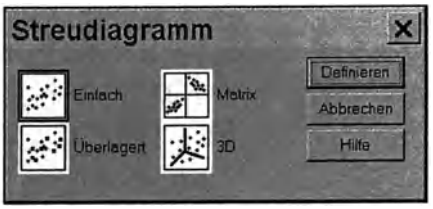


Abb. 26.49. Dialogbox zur Auswahl eines Streudiagramms

Als Streudiagrammtypen sind ein *einfaches,* eines in *Matrixform,* ein *überlagertes* sowie ein *dreidimensionales* (3D) wählbar. Im folgenden werden diese verschiedenen Diagrammformen anhand des Datensatzes MAKRO.SAV (⇨ Anhang B und C) kurz dargestellt.

26.11.1 Einfaches Streudiagramm

Nach der Befehlsfolge „Grafiken", „Streudiagramm..." wird in der in Abb. 26.49 dargestellten Dialogbox das gewünschte einfache Streudiagramm durch Mausklick auf „Einfach" gewählt und danach „Definieren" geklickt. Abb. 26.50 zeigt die geöffnete Dialogbox mit einem Beispiel zur Definition eines einfachen Streudiagramms und die resultierende Grafik. Die Variablen ZINS und INFLAT (Inflationsrate) wurden aus der Quellvariablenliste in die Eingabefelder „y-Achse:" und „x-Achse" übertragen. Außerdem wurde die Variable WBSP2, in der die Wachstumsrate des Bruttosozialprodukts [= (BSP - LAG(BSP)) / LAG(BSP) * 100] zur Bildung von drei Wachstumsklassen (< 1,5 %, 2,5 bis 3 %, > 3 %) rekodiert worden ist, in das Eingabefeld „Gruppenvariable" übertragen. Diese Angabe ist optional und bewirkt, dass die einzelnen Punkte des Streu-

diagramms je nach Größenklasse der Wachstumsrate mit unterschiedlichen Farben im Diagramm ausgewiesen werden.

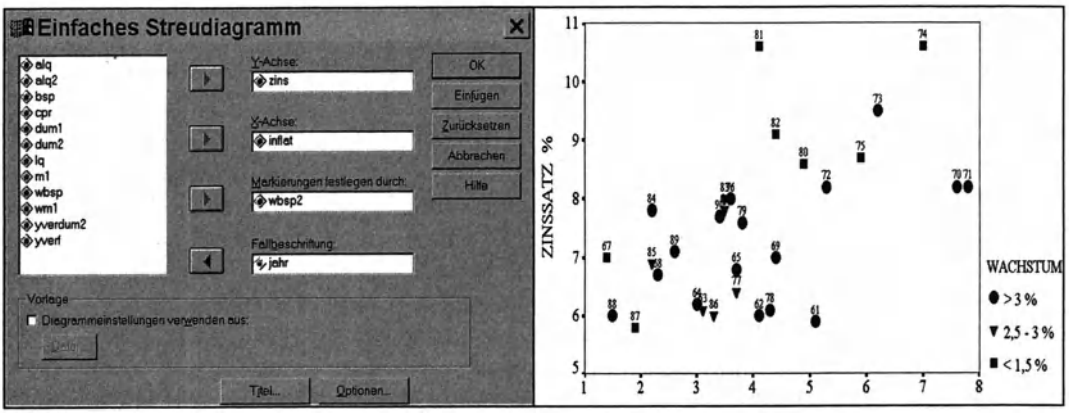


Abb. 26.50. Einfaches Streudiagramm: Zinssatz und Inflationsrate

In Abb. 26.50 wurde die Grafik mit SPSS derart überarbeitet, dass Punkte mit verschiedenen Wachstumsraten durch unterschiedliche Symbolformen dargestellt sind. Die Übertragung der Variable JAHR in das Eingabefeld „Fallbeschriftung" ist optional. Sie bewirkt, dass jeder Punkt des Streuungsdiagramms mit dem Variablenwert, der Jahreszahl, versehen wird.

26.11.2 Streudiagramm in Matrixform

Nach der Befehlsfolge „Grafiken", „Streudiagramm..." wird in der in Abb. 26.49 dargestellten Dialogbox „Matrix" gewählt und danach „Definieren" geklickt. Abb. 26.51 zeigt die geöffnete Dialogbox mit einem Beispiel zur Definition eines Matrix-Streudiagramms und die resultierende Grafik. Die Variablen INFLAT (Inflationsrate), ZINS und WM1 (Wachstumsrate der volkswirtschaftlichen Geldmenge M1 [= (M1-LAG(M1))/LAG(M1)∗100] wurden aus der Quellvariablenliste in das Eingabefeld „Matrix-Variablen" übertragen. In der entstandenen Grafik wird in Streuungsdiagrammen der Zusammenhang jeder Variablen mit jeder anderen dargestellt. Es lässt sich ein positiver Zusammenhang zwischen Inflationsrate und Zinssatz sowie ein negativer Zusammenhang zwischen Zinssatz und Wachstumsrate der Geldmenge M1 erkennen.

26.11.3 Überlagertes Streudiagramm

Nach der Befehlsfolge „Grafiken", „Streudiagramm..." wird in der in Abb. 26.49 dargestellten Dialogbox „Überlagert" gewählt und danach „Definieren" geklickt. Abb. 26.52 zeigt die geöffnete Dialogbox mit einem Beispiel zur Definition eines überlagerten Streudiagramms und die resultierende Grafik. Die Variablen ZINS und INFLAT (Inflationsrate) wurden durch Mausklick markiert und danach als

Variablenpaar in das Eingabefeld „y-x Paare:“ übertragen. Im nächsten Schritt wurden die Variablen ZINS und WM1 (Wachstumsrate der volkswirtschaftlichen Geldmenge M1 [= (M1-LAG(M1))/LAG(M1)∗100] markiert und als Paar in das Eingabefeld übertragen. In der entstandenen Grafik werden die durch die Variablenpaare definierten einfachen Streuungsdiagramme überlagert dargestellt.

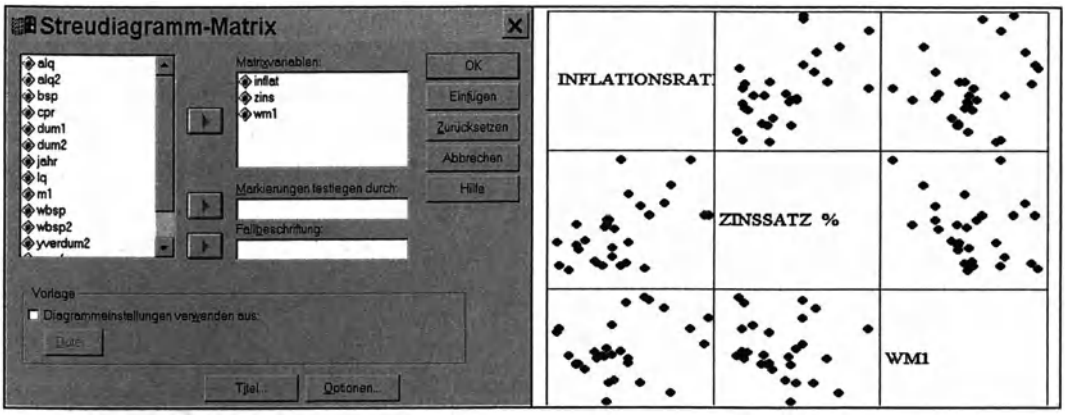


Abb. 26.51. Matrix-Streudiagramm: Inflationsrate, Zinssatz und Wachstumsrate der Geldmenge M1 (= WM1)

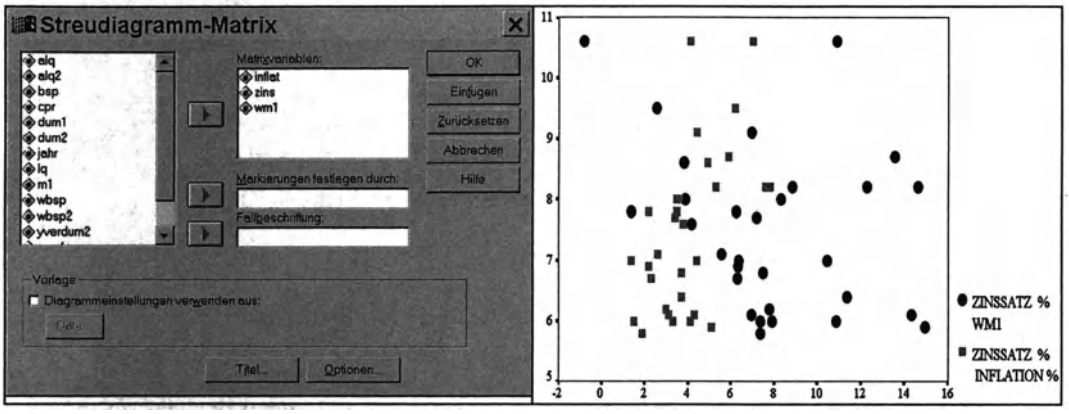


Abb. 26.52. Überlagertes Streudiagramm: Zinssatz-Inflationsrate und Zinssatz-Wachstumsrate der Geldmenge M1

26.11.4 Dreidimensionales Streudiagramm (3D)

Nach der Befehlsfolge „Grafiken“, „Streudiagramm...“ wird in der in Abb. 26.49 dargestellten Dialogbox „3D“ gewählt und dann „Definieren“ geklickt.

Abb. 26.53 zeigt die danach geöffnete Dialogbox mit einem Beispiel zur Definition eines 3D-Streudiagramms und die resultierende Grafik. Die Variablen INFLAT (Inflationsrate), WM1 (Wachstumsrate der volkswirtschaftlichen Geldmenge M1 [= (M1-LAG(M1))/LAG(M1)∗100] und ZINS wurden aus der Quell-

variablenliste in die Eingabefelder „y-Achse:“, „x -Achse:“ und „z-Achse:“ übertragen. Die Übertragung einer Gruppenvariablen sowie einer für die Fallbeschriftung ist optional.

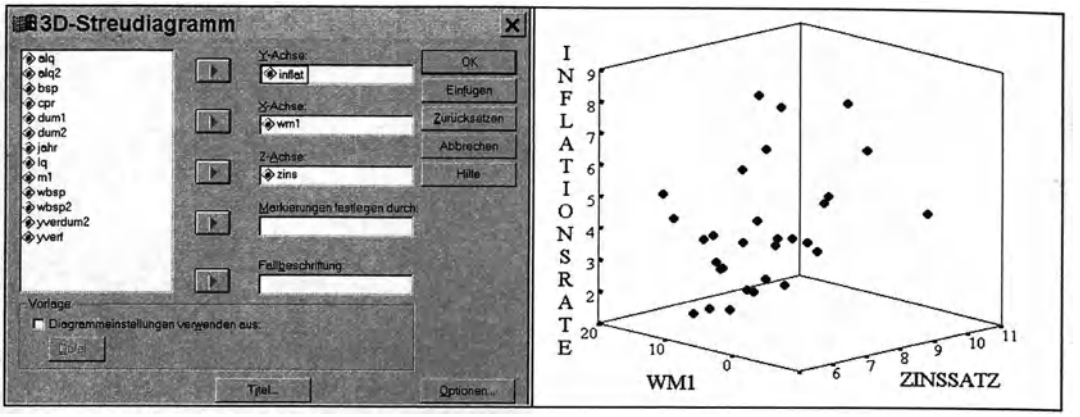


Abb. 26.53. 3D-Streudiagramm: Wachstumsrate der Geldmenge M1, Zinssatz und Inflationsrate

26.11.5 Wahlmöglichkeiten

Für fast alle Streudiagramme bestehen folgende Wahlmöglichkeiten:

- ❑ Ausweisen von Gruppen (⇨ Abb. 26.50).
- ❑ Fallbeschriftung (⇨ Abb. 26.50).
- ❑ Versorgung mit Titel und Fußnoten („Titel“) (⇨ Kap. 26.2.1).
- ❑ Form der Behandlung fehlender Werte („Optionen“) (⇨ Kap. 26.2.1).
- ❑ Grafiklayout aus Vorlage übernehmen („Vorlage“) (⇨ Kap. 26.2.1).

Außerdem: Überarbeitung im Diagramm-Editorfenster möglich (⇨ Kap. 27.4).

26.12 Histogramme erzeugen

Um ein Histogramm zu erstellen, öffnet man durch Klicken der Befehlsfolge

▷ „Grafiken“, „Histogramm...“

eine Dialogbox. Abb. 26.54 zeigt die geöffnete Dialogbox mit einem Beispiel aus dem ALLBUS90-Datensatz. Die Variable EINK wurde aus der Quellvariablenliste in das Eingabefeld „Variable:“ übertragen. Durch Wahl der Option „Normalverteilungskurve“ ist in das Histogramm eine Normalverteilungskurve gelegt worden. Es werden gemäß Voreinstellung die Standardabweichung, der Mittelwert sowie die Anzahl der gültigen Fälle angegeben.

In dem Histogramm wurde automatisch eine Klassenbreite von 500 gebildet. Die Klassenmitte dient zur Beschriftung der Achse. Die Klassenmitte der ersten Klasse ist 0. Damit wird sachlich eine falsche Darstellung erzeugt. Bei einer Klassenbreite von 500 mit 0 als Klassenmitte reicht die Klasse von -250 bis +250. Tatsächlich sollte die Klasse aber von 0 bis 500 gehen. Um diese automatisch gesetzten Klassengrenzen zu ändern, muss man im Menü „Grafik“ die Intervallachse der

Grafik zur Öffnung der Dialogbox „Intervallachse“ doppelklicken. In dieser kann man mit der Option „Anpassen“ und „Definieren“ die Klassengrenzen festlegen und z.B. die erste Klasse mit 0 beginnen lassen. (⇨ „Gestaltung von Intervallachsen“ in Kap. 27.4.3).

Wahlmöglichkeiten. Folgende Optionen bestehen:

- ❒ Überlagerung mit Normalverteilung (⇨ Abb. 26.54).
- ❒ Versorgung mit Titel und Fußnoten („Titel“)(⇨ Kap. 26.3.1).
- ❒ Grafiklayout aus Vorlage übernehmen („Vorlage“)(⇨ Kap. 26.3.1).

Außerdem: Überarbeitung im Diagramm-Editorfenster möglich (⇨ Kap. 27.4).

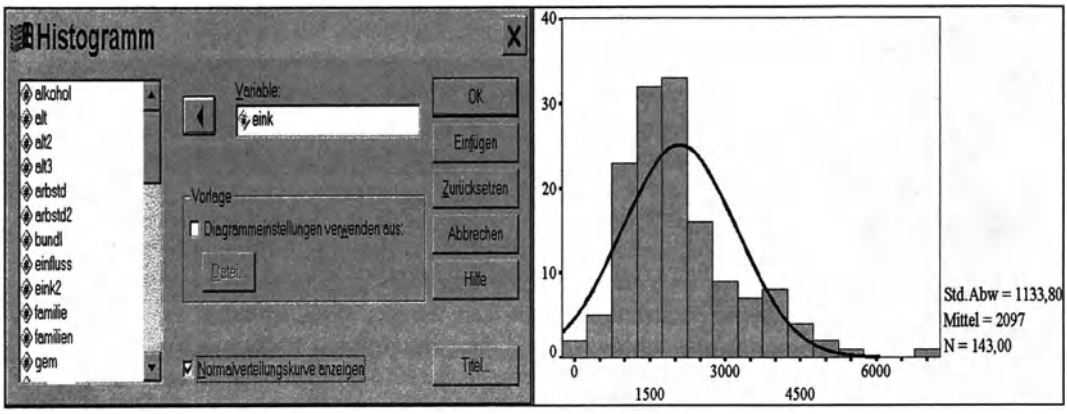


Abb. 26.54. Histogramm: Verteilung der Einkommen und Normalverteilung

26.13 P-P- und Q-Q-Diagramme erzeugen

In der statistischen Datenanalyse kommt es häufig vor, dass man überprüfen möchte, ob die untersuchten Daten als Stichprobe aus einer normalverteilten Grundgesamtheit anzusehen sind. Bei der Regressionsanalyse z.B. oder einem anderen statistischen Modell ist es von Bedeutung, ob die Residualwerte normalverteilt sind. Eine Darstellung als Histogramm bzw. ein statistischer Test wie der von Shapiro Wilks bzw. Kolmogorov-Smirnov (Lillifors) (⇨ Kap. 9.3.2) sind dafür hilfreiche Instrumente. Manchmal möchte man auch prüfen, ob Daten einer anderen theoretischen Verteilung entsprechen.

P-P bzw. Q-Q-Diagramme dienen dazu, in einem Streuungsdiagramm Daten mit einer Normalverteilung oder auch einer anderen theoretischen Verteilung zu vergleichen. In diesen Grafiken werden die empirischen Werte einer Variablen mit den gemäß einer Normalverteilung (oder einer anderen theoretischen Verteilung) zu erwartenden Werten gegenübergestellt. Bei Vorliegen einer Normalverteilung streuen die Datenpunkte eng und zufällig um eine Gerade.

Grundlage der Darstellung sind auf Rängen basierende Anteilswerte der Fälle, die nach unterschiedlichen Verfahren berechnet werden. Diese Anteilswerte werden gegen die Anteilswerte unter einer Normalverteilung (oder einer anderen

theoretischen Verteilung) geplottet. Bei der Ermittlung der Anteilswerte der Fälle kann man aus folgenden Verfahren wählen:

- *Blom.* Diese Berechnung geschieht nach der Formel (r - 3/8)/(n + 1/4) (Blom, 1958) (= Voreinstellung).
- *Rankit.* Die Berechnungsformel lautet (r - 1/2)/n (Chambers et. al., 1983).
- *Tukey.* Die Berechnungsformel lautet (r - 1/3)/(n + 1/3) (Tukey, 1962).
- *Van der Waerden.* Die Transformationsformel lautet r/(n + 1) (Lehmann, 1975).

Für alle Berechnungsansätze ist dabei
n = Anzahl der Beobachtungen
r = Rangziffer, r = 1,....,n

Für die vergleichende grafische Darstellung empirischer Daten und einer theoretischen Verteilung sind zwei Darstellungstypen möglich:

- *P-P-Diagramm* (Befehlsfolge: „Grafik", „P-P"). Es werden die (auf Rängen basierenden) kumulierten Anteile der Fälle denen einer theoretischen Verteilung (z.B. Normalverteilung) gegenübergestellt.
- *Q-Q-Diagramm* (Befehlsfolge: „Grafik", „Q-Q"). Bei dieser Grafik werden die Quantile der empirischen und der theoretischen Verteilung (z.B. Normalverteilung) einander gegenübergestellt.

Beim Erstellen dieser Diagramme kann aus in Tabelle 26.2 erfassten theoretischen Verteilungen gewählt werden:

Tabelle 26.2. Testverteilungen in P-P- und Q-Q-Diagrammen

Beta	Logistisch
Chi-Quadrat	Lognormal
Exponentiell	Pareto
Gamma	Student (t)
Halb-Normal	Weibull
Laplace	Gleich

Im folgenden Anwendungsbeispiel soll geprüft werden, ob die linkssteile Verteilung des Nettoeinkommens der Befragten (Datensatz ALLBUS90.SAV) annähernd einer logarithmierten Normalverteilung entspricht.

PP-Diagramm. Man öffnet durch Klicken der Befehlsfolge

▷ „Grafiken", „P-P..."

die in Abb. 26.55 dargestellte Dialogbox. Es wurde die Variable EINK in das Eingabefeld „Variablen" übertragen.

Im Auswahlfeld „Testverteilung" wird die theoretische Verteilung gewählt, mit der die Verteilung der empirischen Daten verglichen werden soll. Wir wählen Lognormal (alternativ hätte man auch „Normalverteilung" in Verbindung mit der Transformationsoption „Natürlicher Logarithmus" wählen können). Die Parameter der theoretischen Verteilung sollen aus den Daten geschätzt werden. Bei Wahl an-

derer theoretischer Verteilungen müssen eventuell die Anzahl der Freiheitsgrade bzw. andere Parameter angegeben werden.

Folgende Optionen für eine Transformation der Variablen sind möglich:

- *Natürlicher Logarithmus.* Bei Wahl dieser Option wird die untersuchte Variable logarithmiert (zur Basis e ≈ 2,7183).
- *Werte standardisieren.* Die untersuchte Variable x wird in Standardeinheiten transformiert gemäß der Transformation

$$\frac{x - \bar{x}}{s}$$

$\bar{x}$ = Mittelwert
s = Standardabweichung

Die resultierende standardisierte Variable hat einen Mittelwert von 0 und eine Standardabweichung von 1.

- *Differenz.* Diese Transformation ist für Zeitreihen von Bedeutung. Es wird die Differenz zu vorherigen Werten gebildet. Durch Angabe einer Zahl kann festgelegt werden, zu welchem vorhergehenden Wert die Differenz gebildet werden soll.
- *Saisonale Differenz.* Hat man Zeitreihen mit einer Saisonkomponente vorliegen und mit der Befehlsfolge „Daten", „Datum definieren" definiert, so können Differenzen von Werten gleicher Saisonzeitzugehörigkeit gebildet werden. Analog zu oben kann man angeben, zu welchem vorhergehenden Saisonzeitwert die Differenz gebildet werden soll.

In „Formel für Anteilsschätzungen" kann man eine Berechnungsmethode für die Anteilswerte der Fälle wählen.

Außerdem kann gewählt werden, wie bei Rangbindungen (= gleiche Variablenwerte bei mehreren beobachteten Fällen, englisch: ties) vorgegangen werden soll. Folgende Wahlmöglichkeiten bestehen:

- *Mittelwert* (= Voreinstellung ab Version 6.1). Es wird der Mittelwert der Rangzahlen den Fällen als Rang zugewiesen.
- *Maximum.* Die höchste Rangzahl wird den Fällen als Rang zugewiesen.
- *Minimum.* Die kleinste Rangzahl wird den Fällen als Rang zugewiesen.
- *Bindungen willkürlich lösen* (= Voreinstellung vor Version 6.1). Jeder gebundene Fall wird in die Grafik als Datenpunkt aufgenommen. Bei den oben genannten Wahlmöglichkeiten gilt, dass bei Bindung mehrere Fälle in einem Datenpunkt abgebildet werden.

In Abb. 26.55 (rechts) sind auf der senkrechten Achse die nach der Transformationsformel von Blom berechneten erwarteten kumulierten Häufigkeiten (gemäß einer Lognormalverteilung) und auf der waagerechten Achse die empirischen kumulierten Häufigkeiten für das logarithmierte Einkommen dargestellt.

Es zeigt sich, dass die Abweichungen von der Geraden und damit von einer Normalverteilung erheblich sind. Dieses wird auch durch eine zweite, gleichzeitig erzeugte Grafik (Abb. 26.56) unterstrichen. Dort werden auf der senkrechten

Achse die Abweichungen von der Geraden abgebildet, die eine Lognormalverteilung repräsentiert.

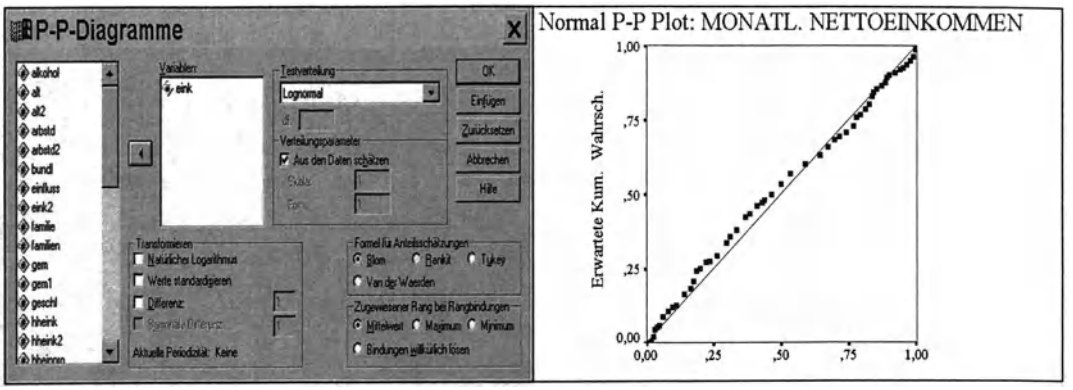


Abb. 26.55. P-P-Lognormalverteilungs-Diagramm für das Nettoeinkommen der Befragten

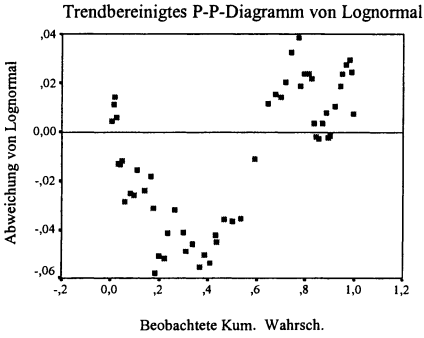


Abb. 26.56. Abweichungen vom P-P-Lognormalverteilungs-Diagramm für das Nettoeinkommen der Befragten

Q-Q-Diagramm. Man öffnet durch Klicken der Befehlsfolge

▷„Grafik“, „Q-Q...“

eine Dialogbox, die der in Abb. 26.55 ähnelt. In der Dialogbox „Q-Q-Diagramme“ bestehen die gleichen Wahlmöglichkeiten wie bei P-P- Diagrammen. Auf den Achsen werden die Quantile der empirischen und theoretischen Verteilung dargestellt.

Außerdem: Überarbeitung im Diagramm-Editorfenster möglich (⇨ Kap. 27.4).

Anmerkung. Ab Version 6.1 ist es möglich, bei der Erzeugung von P-P- bzw. Q-Q-Normalverteilungsdiagrammen Fälle zu gewichten.

26.14 Sequenzdiagramme erzeugen

Zur grafischen Darstellung von Zeitreihenwerten in der Zeit öffnet die Befehlsfolge

▷ „Grafiken", „Sequenz..."

die in Abb. 26.57 links dargestellte Dialogbox. Die darzustellenden Zeitreihen BSP (Bruttosozialprodukt) und M1 (Geldmenge M1) aus dem Datensatz MAKRO (⇨ Anhang B) wurden aus der Quellvariablenliste in das Eingabefeld „Variablen" übertragen. In das Eingabefeld „Zeitachsenbeschriftung" wurde die Variable JAHR übertragen.
Die Variablen können auch in transformierter Weise dargestellt werden. Dafür stehen im Feld „Transformieren" folgende Transformationen für die Variablen zur Auswahl:

- *Natürlicher Logarithmus.* Logarithmus zur Basis e (e ≅ 2,7183). Diese Option wurde zur Darstellung der Variablen BSP und M1 gewählt. Bei der Wahl einer logarithmischen Skala kann der Verlauf der beiden Variablen im Diagramm besser verglichen werden.
- *Differenz.* Man kann wählen, welche Differenz dargestellt werden soll. Voreingestellt ist die erste Differenz, d.h. die Differenz zum vorhergehenden Wert. Die zweite Differenz - die Differenz der ersten Differenz - erhält man durch Eintragen einer 2 in das Eingabefeld usw.
- *Saisonale Differenz.* Diese Option steht nur dann zur Verfügung, wenn zuvor mit Hilfe der Befehlsfolge „Daten", „Datum definieren" die Datenreihe als Zeitreihe definiert wurde. Es kann analog zur „Differenz" auch die erste, zweite usw. Differenz abgebildet werden. Voreingestellt ist die erste Differenz. Mit Periodizität wird die Häufigkeit von Zeitreihenwerten pro Periode angegeben. Bei Quartalsdaten z.B. ist die Periodizität gleich vier.

Wahlmöglichkeiten:

① *Ein Diagramm je Variable.* Hat man mehrere Variablen in das Eingabefeld „Variablen" eingetragen, so wird für jede Variable eine Grafik erstellt.

② *Zeitlinien.* Mit diesem Untermenü kann man Bezugslinien auf die Zeitachse des Diagramms projizieren. Nach Klicken auf „Zeitlinien..." öffnet sich die links in Abb. 26.58 dargestellte Dialogbox. Es ist „Linie bei jedem Wechsel von:" gewählt und als „Bezugsvariable" KONJTIEF aus dem Quellvariablenfenster übertragen worden. Diese Variable bildet die konjunkturellen Tiefpunkte in den Jahren 1967, 1975 und 1982 ab: für die Jahre 1960 bis 1966 hat die Variable den Wert 1, für 1967 bis 1974 den Wert 2, für 1975 bis 1981 den Wert 3 und für 1982 bis 1990 den Wert 4. Die Referenzlinie wird bei jedem Wertewechsel der Variablen KONJTIEF in das Diagramm eingefügt. In Abb. 26.58 rechts ist das erzeugte Diagramm zu sehen.

Es ist auch möglich, eine Bezugslinie für eine bestimmte Beobachtung bzw. Zeitperiode in das Diagramm einzufügen. Dann wird „Linie bei Zeitpunkt" angeklickt, und in das Eingabefeld von „Beobachtung" wird die Fallzahl eingegeben. Ist per „Datum definieren" der Datensatz als Zeitreihe definiert, so nennt

das Eingabefeld die definierte Zeiteinheit, und es ist hier die gewünschte Zeitperiode einzugeben.

③ *Format.* Klicken auf „Format..." öffnet eine Dialogbox, in der folgende Spezifizierungen möglich sind:

- *Zeit auf horizontaler Achse.* Mit dieser Option wird die senkrechte Achse der Grafik als Zeitachse genommen.
- *Diagramme einzelner Variablen.*
  - Mit „Flächendiagramm" kann zu diesem Diagrammtyp gewechselt werden.
  - „Bezugslinie für Mittelwert der Zeitreihe" fügt eine Linie in Höhe des Mittelwerts der Reihe ein.
- *Grafiken mit mehreren Variablen.* Es werden für jede Zeitperiode (Beobachtung) Verbindungslinien zwischen den Variablen gezogen.

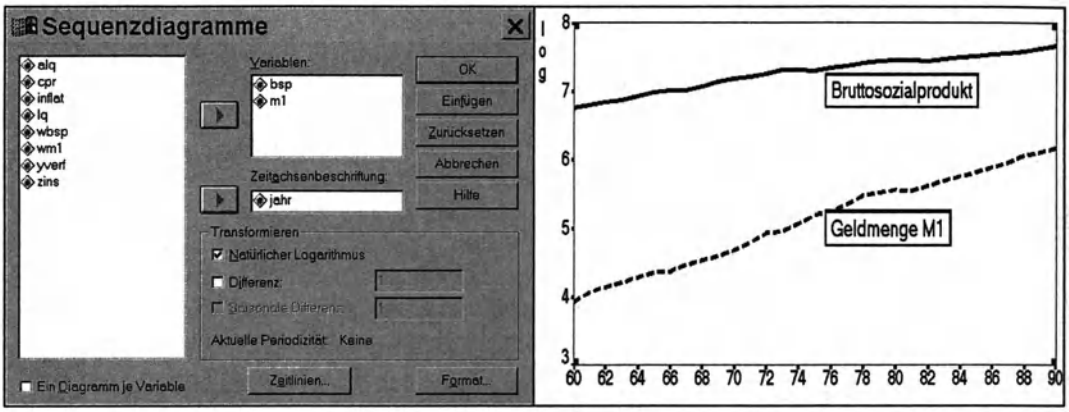


Abb. 26.57. Zeitreihendarstellung in logarithmischer Skala

26.15 ROC-Kurve erzeugen

Theoretische Grundlagen. Insbesondere in der Medizin werden diagnostische Tests eingesetzt, um zu prüfen, ob Patienten eine bestimmte Erkrankung haben oder nicht. Die ROC-Kurve[1] ist ein Instrument, derartige Tests zu bewerten. Aber auch in anderen Bereichen findet die ROC-Kurve Anwendung.

Das für die Diskriminanzanalyse verwendete Beispiel zur Diagnose von viraler Hepatitis soll zur näheren Erläuterung dienen. Messwerte von Enzymen werden für eine diagnostischen Test verwendet, ob Patienten eine virale Hepatitis (virH) haben oder nicht. In dem Beispiel werden für Patienten Messwerte von Enzymen in der Variablen (neben anderen) ALT erfasst. Zur diagnostischen Unterscheidung

[1] ROC = Receiver Operating Characteristic. Der Begriff hat seine historische Wurzeln im 2. Weltkrieg als Radargeräteoperatoren zu entscheiden hatten, ob ein Signal auf dem Bildschirm feindliche oder freundliche Schiffe bzw. Flugzeuge bedeuten und Messmethoden zur Fähigkeit des Operator dieses zu unterscheiden entwickelt worden sind.

von an virH erkrankten und nicht an virH erkrankten Patienten muss ein Trenn-

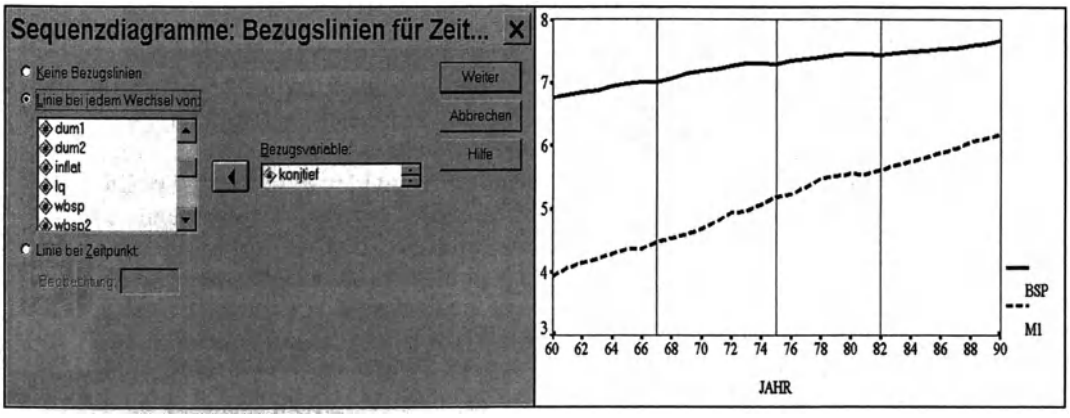


Abb. 26.58. Einfügen von Bezugslinien

messwert (⇨ $LALT_{krit}$ in Abb. 20.1)[2] von ALT festgelegt werden: Patienten mit ALT-Messwerten oberhalb dieses Trennwerts (Testergebnis positiv) werden als erkrankt und Patienten mit ALT-Messwerten unterhalb dieses Trennwerts (Testergebnis negativ) werden als nicht an virH erkrankt diagnostiziert. In Abb. 20.1 sowie 20.2 wird dargestellt, dass sich die Häufigkeitsverteilungen von ALT für beide Gruppen überlappen: es gibt Patienten mit über dem Trennwert liegenden ALT-Werten, die nicht an virH erkrankt sind und umgekehrt gibt es Patienten, die ALT-Messwerte haben, die unterhalb des Trennwert liegen und an virH erkrankt sind. Im Bereich der Überlappung versagt der Diagnosetest. Je kleiner der Überlappungsbereich, umso genauer kann der Test die Kranken von den nicht Kranken trennen.

In einer Vierfeldertabelle (Tabelle 26.3) kann man die Ergebnisse des diagnostischen Tests zusammenfassen.

Tabelle 26.3. Vierfeldertafel mit Ergebnissen eines Tests auf virale Hepatitis

An vir. Hepatitis erkrankt	Testergebnis		Summe
	positiv	negativ	
Ja	a	b	a+b
nein	c	d	c+d
Summe	a+c	b+d	n

Wird der Stichprobenumfang n = a+b+c+d sehr groß, so können die Anteile $a/(a+b)$ (Anteil positiv getesteter Patienten an Erkrankten) und $d/(c+d)$ (Anteil negativ getesteter Patienten an nicht Erkrankten) als Wahrscheinlichkeiten interpretiert werden. Diese werden Sensitivität und Spezifität genannt. Wird der Diag-

[2] In der Diskriminanzanalyse wurden die Variablen logarithmiert, um die Modellvoraussetzung einer Normalverteilung annähernd zu erreichen.

nosetrennwert verändert, so verändert sich auch die Sensitivität und Spezifität. Erhöht man den ALT-Trennwert für den Diagnosetest, so wird die Sensitivität größer und die Spezifität kleiner. Umgekehrtes gilt für eine Senkung des Trennwerts.

In der ROC-Kurvendarstellung werden auf der senkrechten Achse eines Koordinatensystems die Stichprobenschätzwerte für die Sensitivität (Anteil richtig positiv Getesteter) und auf der waagerechten Achse die für 1- Spezifität (Anteil falsch positiv Getesteter) abgetragen. Trägt man die Sensitivitätswerte und 1-Spezifitätswerte eines Tests für unterschiedliche Trennwerte des Tests als Punkte in das Koordinatensystem ein und verbindet die Punkte, so entsteht die ROC-Kurve eines Diagnosetests. Da mit wachsender Sensibilität die Differenz 1- Sensitivität größer wird, hat die ROC-Kurve eine positive Steigung. Für einen guten (möglichst genauen) Test sollte die Kurve auf der senkrechten Achse möglichst weit oben beginnen und dann nach rechts oben streben. Je näher die ROC-Kurve an der 45-Grad-Linie liegt, umso ungenauer wird der Test. Vergleicht man z.B. zwei Tests, so zeigt sich der bessere (genauere) Test durch eine oberhalb der anderen liegende ROC-Kurve. Der Flächenanteil unterhalb der ROC-Kurve ist ein Maß für die Testgenauigkeit. Flächenanteilsgrößen größer als 0,9 gelten als ausgezeichnet, zwischen 0,80 und 0,90 als gut und zwischen 0,70 und 0,80 noch als akzeptabel.

Praktische Anwendung. Die Daten aus der Datei LEBER.SAV wurden für die Diskriminanzanalyse genutzt, um eine Diskriminanzfunktion zur Trennung von an viraler Hepatitis und anderen Lebererkrankungen erkrankten Patienten zu gewinnen. Aus den standardisierten Koeffizienten der Diskriminanzfunktion (⇨ Tabelle 20.3) ergab sich, dass die (logarithmierte) Variable ALT einen höheren Beitrag zur Trennung der Gruppen leistet als die (logarithmierten) Variable AST.

Im folgenden sollen die ROC-Kurven der Enzym-Variablen ALT und AST ermittelt und die Trenngenauigkeit dieser Variablen für eine Diagnose von viraler Hepatitis verglichen werden. Nach Laden der Datei LEBER.SAV gehen Sie wie folgt vor:

- ▷ Wählen Sie per Mausklick die Befehlsfolge "Grafiken", "ROC-Kurve". Es öffnet sich die in Abb. 26.59 links dargestellte Dialogbox.
- ▷ Übertragen Sie die Variablen ALT und AST aus der Quellvariablenliste in das Eingabefeld "Testvariablen:".
- ▷ In das Eingabefeld „Zustandsvariable:" wird die Variable GRUP1 (mit den Variablenwerten 0 für virale Hepatitis und 1 für andere Lebererkrankungen) übertragen sowie in das Eingabefeld „Wert der Zustandsvariablen" eine 0 eingetragen. Im Feld „Anzeigen" werden alle Optionen angefordert. Mit „OK" wird die Grafikerstellung gestartet.

In Abb. 26.59 rechts sind die beiden ROC-Kurven zu sehen. Da die ROC-Kurve für die Diagnosetestvariable ALT oberhalb der ROC-Kurve von AST liegt, wird hier deutlich, dass sie besser für eine Trennung beider Patientengruppen geeignet ist.

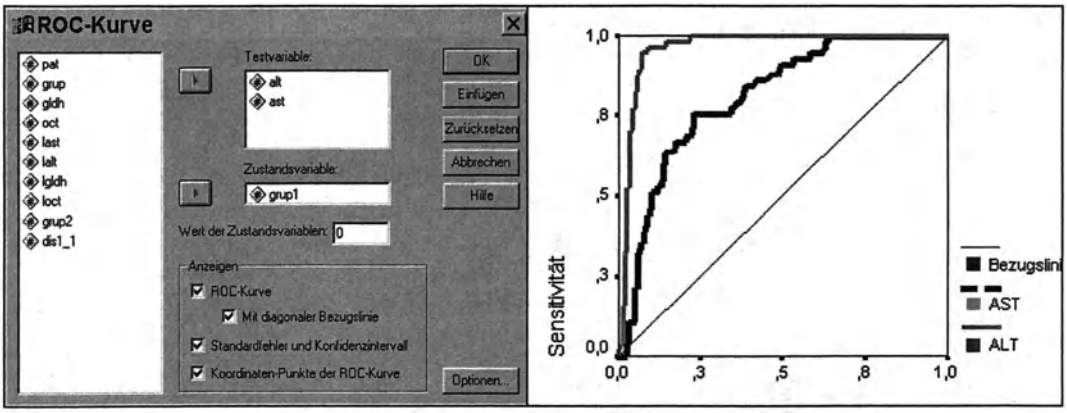


Abb. 26.59. ROC-Kurven für ALT und AST

Auch die in Tabelle 26.4 gezeigten Daten untermauern die obige Aussage. Der Flächenanteil für die Variable ALT ist mit 0,964 größer als der von AST und liegt nahe bei 1. Er weist damit ein exzellentes Ergebnis aus. In einem statistischen Test kann geprüft werden, ob der ausgewiesenen Flächenanteil der ROC-Kurve sich signifikant vom Wert 0,5 (Hypothese H_0) unterscheidet. Bei einem Test mit einem Signifikanzniveau von $\alpha = 0{,}05$ wird die H_0-Hypothese abgelehnt. Da auch das asymptotische 95-%-Konfidenzintervall den Flächenwert = 0,5 nicht einschließt, wird dies Schlussfolgerung auch hier deutlich.

In Tabelle 26.5 werden die Koordinatenpunkte der ROC-Kurve ausschnittsweise für verschiedene Trennwerte der Testgrößen ALT gezeigt. Der kleinste Trennwert ist der kleinste beobachtete Testwert minus 1, und der größte Trennwert ist der größte beobachtete Testwert plus 1. Alle anderen Trennwerte sind Mittelwerte von zwei aufeinanderfolgenden, geordneten beobachteten Testwerten.

Tabelle 26.4. Ausgabeergebnis zur Anzeige der Fläche unter der ROC-Kurve

Fläche unter der Kurve

Variable(n) für Testergebnis	Fläche	Standardfehler[a]	Asymptotische Signifikanz[b]	Asymptotisches 95% Konfidenzintervall	
				Untergrenze	Obergrenze
ALT	,964	,012	,000	,940	,988
AST	,810	,031	,000	,750	,871

Bei der bzw. den Variable(n) für das Testergebnis: ALT, AST liegt mindestens eine Bindung zwischen der positiven Ist-Zustandsgruppe und der negativen Ist-Zustandsgruppe vor. Die Statistiken sind möglicherweise verzerrt.

a. Unter der nichtparametrischen Annahme

b. Nullhypothese: Wahrheitsfläche = 0.5

Tabelle 26.5. Ausschnitt des Ausgabeergebnisses zur Anzeige der Koordinaten der ROC-Kurve für unterschiedliche Trennwerte der Testvariablen

Koordinaten der Kurve

Variable(n) für Testergebnis	Positiv, wenn größer oder gleich(a)	Sensitivität	1 - Spezifität
ALT	17,00	1,000	1,000
	18,50	1,000	,994
	20,50	1,000	,988
.....	.....	.....	.....
	1209,50	,000	,012
	1929,50	,000	,006
	2299,00	,000	,000

Bei der bzw. den Variable(n) für das Testergebnis: ALT liegt mindestens eine Bindung zwischen der positiven und der negativen Ist-Zustandsgruppe vor.
a Der kleinste Trennwert ist der kleinste beobachtete Testwert minus 1, und der größte Trennwert ist der größte beobachtete Testwert plus 1. Alle anderen Trennwerte sind Mittelwerte von zwei aufeinanderfolgenden, geordneten beobachteten Testwerten.

Wahlmöglichkeiten (Klicken der Schaltfläche Optionen):

- *Klassifikation.* Man kann wählen, ob der jeweilige Trennwert bei einer positiven Klassifikation ein – oder ausgeschlossen werden soll..
- *Test-Richtung.* Man kann die Darstellung der ROC-Kurvern um die Bezugslinie spiegeln.
- *Parameter für Standardfehler der Fläche.* Bei der Schätzung des Standardfehlers für die berechnete Fläche unter der ROC-Kurve kann aus zwei Methoden gewählt werden („Nichtparametrisch" und „Bi-negativ exponentiell"). Außerdem kann man das Niveau des Konfidenzintervalls festlegen (Werte zwischen 50,1% und 99,9%).
- *Fehlende* Werte. Es kann aus zwei Optionen gewählt werden.

26.16 Autokorrelations- und Kreuzkorrelationsdiagramme erzeugen

26.16.1 Autokorrelationsdiagramme

In modernen Modellen der Zeitreihenanalyse wird die Entstehung einer Zeitreihe als ein stochastischer Prozess interpretiert. Die Struktur eines solchen Prozesses kann in verschiedenen Modellen [z.B. moving average oder kurz MA(k), autoregressiver Prozess oder kurz AR(k) der Ordnung k bzw. Mischformen] erfasst und

spezifiziert werden. Für die Spezifizierung sowie Beurteilung eines derartigen Modells hat das *Autokorrelations-* sowie das *partielle Autokorrelationsdiagramm* eine wichtige Funktion, da das Muster dieser Diagramme Hinweise für die Modellierung gibt und zur Überprüfung eines gewählten Modells dient.

In einem Autokorrelationsdiagramm werden Autokorrelationskoeffizienten dargestellt. Autokorrelationskoeffizienten messen die Korrelation zwischen den Werten einer Zeitreihe und den um 1, 2, 3,...k... Perioden verschobenen Werten (lags) dieser Zeitreihe. Der Schätzwert für den *Autokorrelationskoeffizienten* r_k (der Korrelation einer Zeitreihe mit sich selbst bei um k Perioden verschobenen (lags) Zeitreihenwerten) wird in Gleichung 26.3 angeführt. Er ist analog zum Korrelationskoeffizienten nach Bravais-Pearson definiert:

$$r_k = \frac{\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \qquad (26.3)$$

y_t = Zeitreihenwert mit den Beobachtungen t = 1,2,....n
$\bar{y}$ = arithmetisches Mittel
k = lag von k Perioden

Der *partielle Autokorrelationskoeffizient* misst analog zum partiellen Korrelationskoeffizienten die Stärke des Zusammenhangs zwischen Werten und um k Perioden verschobenen Werten einer Zeitreihe, nachdem der Effekt der Korrelation der vorhergehenden lags statistisch eliminiert worden ist. Das Muster der partiellen Korrelationskoeffizienten bietet eine Hilfe zur Entscheidung über die Ordnung eines AR-Modells für die Zeitreihe.

Um ein Autokorrelationsdiagramm zu erzeugen, klickt man die Befehlsfolge

▷ „Grafiken", „Zeitreihen ▷", „Autokorrelationen..."

zur Öffnung der in Abb. 26.60 links dargestellten Dialogbox. Für ein Anwendungsbeispiel wurde die Variable ZINS (Zinssatz) aus dem Datensatz MAKRO.SAV (⇨ Anhang B) aus der Quellvariablenliste in das Feld „Variablen" übertragen. In der Abb. 26.60 rechts wird das Korrelationsdiagramm dargestellt. Auf der waagerechten Achse sind die Länge der lags und auf der senkrechten Achse die Autokorrelationskoeffizienten abgebildet. „ACF" ist die Abkürzung für Auto-Correlation-Function. So wird der Ausweis der Autokorrelationskoeffizienten als Funktion der lag-Länge bezeichnet. In der Grafik werden auch die 95 %-Konfidenzbereiche um den Wert Null ausgewiesen. Es zeigt sich, dass für lags in Höhe von 1 und 12 bis 14 die Koeffizienten außerhalb des Konfidenzbereiches liegen.

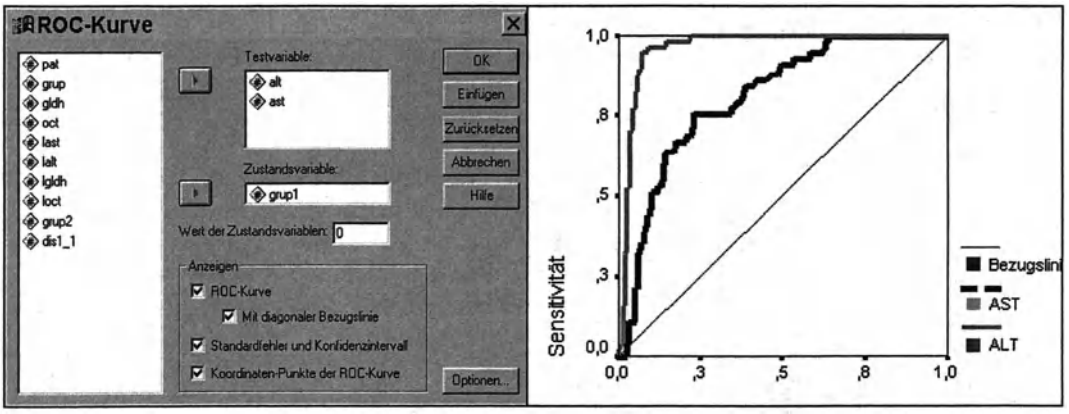


Abb. 26.60. Autokorrelationsdiagramm für die Zeitreihe Zinssatz

Folgende Spezifizierungen sind möglich:

- *Transformieren.* Es können für die Variablen die gleichen Transformationen wie in Sequenzdiagrammen gewählt werden (⇨ Kap. 26.14).
- *Anzeigen.*
  - *Autokorrelationen.* Es wird ein Autokorrelationsdiagramm erzeugt. In Abb. 26.60 ist das Autokorrelationsdiagramm für die Variable ZINS dargestellt. Bei Erzeugung der Grafik wurde per Schaltfläche „Optionen" die Dialogbox „Autokorrelationen: Optionen" geöffnet und es wurden 25 Zeitintervalle (lags) gewählt. Die Größe der Autokorrelationskoeffizienten entspricht der Höhe der Balken. Jeder Balken gehört zu einem bestimmten lag.
  - *Partielle Autokorrelationen.* Es wird ein partielles Autokorrelationsdiagramm erstellt.
- *Optionen.* In einer Dialogbox können folgende Vorgaben festgelegt werden:
  - *Maximale Anzahl an Zeitintervallen.* Im Beispiel wurden 25 gewählt.
  - *Methode für Standardfehler.* Es stehen zwei Modelle zur Berechnung des Standardfehlers der Autokorrelationskoeffizienten für den Ausweis von Konfidenzbereichen zur Auswahl:
    - *Unabhängigkeitsmodell.* In diesem Modell wird zur Berechnung eines Standardfehlers als H_0-Hypothese angenommen, dass der Prozess zur Entstehung der untersuchten Zeitreihe y durch unabhängige Ziehungen aus gleichen Populationen entstanden ist. Man nennt einen derartigen Verlauf auch white noise. Der Standardfehler von r_k berechnet sich für dieses Modell (wenn keine Variablenwerte fehlen) nach der Gleichung

$$S_{rk} \cong \sqrt{\frac{1}{n}\left(\frac{n-k}{n+2}\right)} \qquad (26.4)$$

    - *Bartletts Approximation.* Diese Approximation ist dann angemessen, wenn für die Reihe als Modell ein moving-average-Prozess unterstellt werden kann. Nach Bartlett beträgt (wenn keine Variablenwerte fehlen

und unter der Annahme eines MA-Prozesses der Ordnung k – 1) der Standardfehler von r_k approximativ

$$S_{rk} \cong \sqrt{\frac{1}{n}\left(1+2\sum_{j=1}^{k-1} r_j^2\right)} \tag{26.5}$$

- *Autokorrelation in periodischen Intervallen anzeigen.* Diese Möglichkeit zielt auf den Fall, dass die Daten z.B. saisonalen Schwankungen unterliegen und die Datenreihe mittels „Daten", „Datum definieren" als Zeitreihe mit z.B. Quartalen definiert worden ist.

Im Ausgabefenster von SPSS werden ergänzende Informationen zu den Autokorrelationsdiagrammen bereitgestellt. In Tabelle 26.6 ist ein Ausschnitt aus der Ausgabe für das Beispiel der Variablen ZINS wiedergeben. Für die lags werden mit „Auto-Corr." die Autokorrelationskoeffizienten und mit „Stand-Err." ihre Standardfehler aufgeführt. Mit „Box-Ljung" wird eine Test-Prüfgröße bereitgestellt, die einen Hypothesentest zur Prüfung auf Vorliegen von Autokorrelation ermöglicht. Die Test-Statistik ist für den lag k definiert als

$$Q_k = n(n+2)\sum_{j=1}^{k}\frac{r_j^2}{n-j} \tag{26.5}$$

Für große n hat Q_k eine Chi-Quadrat-Verteilung mit $k-p-q$ Freiheitsgraden, wobei p und q die Ordnungen des autoregressiven bzw. moving-average-Prozesses sind.

Hohe Werte der Test-Prüfgröße sind ein Zeichen dafür, dass Autokorrelation vorliegt. Mit „Prob." wird in der SPSS-Ausgabe die Wahrscheinlichkeit (ein Signifikanzniveau) angeführt, mit der man sich bei Ablehnung der Hypothese H_0 (es besteht keine Autokorrelation) irren kann. Für die Variable ZINS wird die H_0-Hypthese abgelehnt, da das angeführte Signifikanzniveau eine Irrtumswahrscheinlichkeit von 5 % ($\alpha = 0{,}05$) übersteigt.

Im partiellen Autokorrelationsdiagramm für die Variable ZINS stellen die Balken die Größe des partiellen Korrelationskoeffizienten für lags der Länge von 1 bis 25 Perioden dar. Um den Wert Null wird der Zwei-Sigma-Konfidenzbereich markiert. Nur die Koeffizienten mit lags von 1 und 2 sind größer als der angezeigte Bereich. Diese Informationen deuten zusammen mit der Darstellung der Autokorrelationskoeffizienten darauf hin, dass die Zeitreihe eventuell mit einem autoregressiven Modell der Ordnung zwei prognostiziert werden kann. Zu dem partiellen Autokorrelationsdiagramm wird im Ausgabefenster auch eine Ausgabe analog der Tabelle 26.6 angeboten.

Tabelle 26.6. Ergebnisausgabe von Autokorrelation

```
Autocorrelations:   ZINS   ZINSSATZ  %

     Auto- Stand.
Lag  Corr. Err.  -.5 -.25   0   .25  .5   .75   1   Box-Ljung
Prob.
                +----+----+----+----+----+----+----+
 1  ,729  ,171      .      |******.********         18,098  ,000
 2  ,328  ,168      .      |*******                 21,894  ,000
 3 -,004  ,165      .      *      .                 21,895  ,000
 4 -,196  ,162      . ****|      .                 23,346  ,000
 5 -,208  ,159      . ****|      .                 25,040  ,000
 6 -,115  ,156      .   **|      .                 25,581  ,000
 7  ,080  ,153      .     |**    .                 25,852  ,001
 8  ,193  ,150      .     |****  .                 27,501  ,001
.....................

Plot Symbols:   Autocorrelations *     Two Standard Error Limits
```

26.16.2 Kreuzkorrelationsdiagramme

In einem Kreuzkorrelationsdiagramm werden Korrelationskoeffizienten zur Messung der Stärke des Zusammenhangs zwischen zwei Zeitreihenvariablen dargestellt, wobei für die Korrelationen unterschiedliche lags (Zeitverzögerungen) zugrunde gelegt werden. Kreuzkorrelationsdiagramme bieten eine Entscheidungsgrundlage für die Frage, mit welchem lag eine Zeitreihe eine andere Zeitreihe am besten vorhersagen kann. Die Kreuzkorrelationskoeffizienten werden bei k lags wie folgt berechnet:

$$r_{xy}(k) = \frac{C_{xy}(k)}{S_x S_y} \tag{26.6}$$

$C_{xy}(k)$ = Kovarianz bei k lags

$$= \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(y_{t+k} - \bar{y}) \text{ für } k = 0,1,2...$$

$$= \frac{1}{n}\sum_{t=1}^{n+k}(y_t - \bar{y})(x_{t-k} - \bar{x}) \text{ für } k = -1,-2...$$

$$S_x = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2} \text{ und } S_y = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

Der Standardfehler von $r_{xy}(k)$ beträgt unter der Annahme, dass die Zeitreihen nicht kreuzkorreliert sind und eine der Reihen approximativ white noise ist [Box and Jenkins (1976)]

$$S_{rk} \cong \sqrt{\frac{1}{n-|k|}} \text{ für } k = 0,\pm1,\pm2... \tag{26.7}$$

Um ein Kreuzkorrelationsdiagramm zu erzeugen, klickt man die Befehlsfolge

▷ „Grafiken", „Zeitreihen ▷", „Kreuzkorrelationen..."

zur Öffnung der in Abb. 26.61 links dargestellten Dialogbox. Für ein Anwendungsbeispiel aus dem Datensatz MAKRO.SAV (⇨ Anhang B und C) sind die Variablen ZINS (Zinssatz) und WBSP (Wachstumsrate des Bruttosozialprodukts) aus der Quellvariablenliste in das Feld „Variablen" übertragen worden. In Abb. 26.61 rechts wird das Korrelationsdiagramm dargestellt. Auf der waagerechten Achse sind die lags und auf der senkrechten Achse die Kreuzkorrelationskoeffizienten abgebildet. „CCF" ist die Abkürzung für Cross-Correlation-Function. So wird der Ausweis der Kreuzkorrelationskoeffizienten als Funktion von lags bezeichnet. In der Grafik werden auch die 95 %-Konfidenzbereiche um den Wert Null ausgewiesen. Es zeigt sich, dass nur für den lag von eins der Koeffizient außerhalb des Konfidenzbereiches liegt.

Für die Erstellung eines Kreuzkorrelationsdiagramms sind folgende Spezifizierungen möglich:

- *Transformieren.* Es können für die Variablen die gleichen Transformationen wie in Sequenzdiagrammen gewählt werden (⇨ Kap. 26.14).
- *Optionen.* Klickt man auf „Optionen", so öffnet sich die in Abb. 26.62 dargestellte Dialogbox. Es können folgende Vorgaben festgelegt werden:
  - *Maximale Anzahl der Zeitintervalle.* Im Beispiel wurde 7 gewählt.
  - *Kreuzkorrelationen bei periodischen Intervallen.* Diese Möglichkeit zielt auf den Fall, dass die Daten z.B. saisonale Schwankungen aufweisen und die Datenreihe mittels „Daten", „Datum definieren" als Zeitreihe mit z.B. Quartalen definiert worden ist.

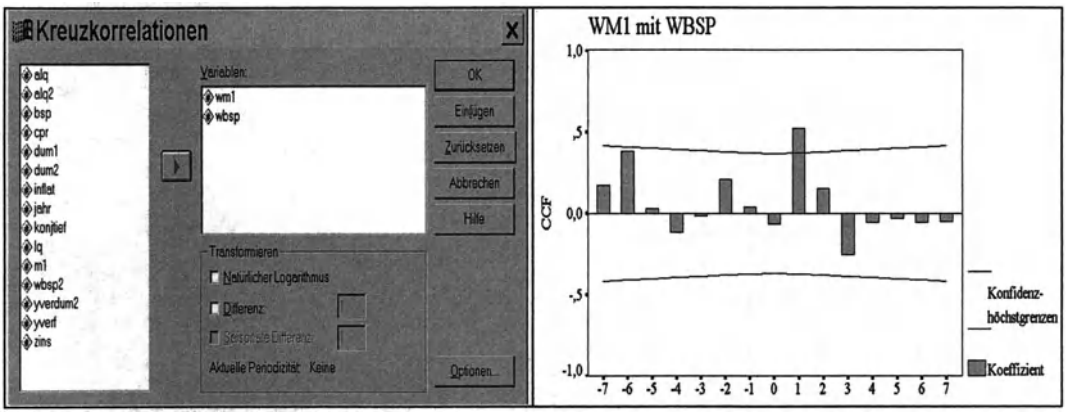


Abb. 26.61. Kreuzkorrelationsdiagramm für die Zeitreihen Zinssatz und Wachstumsrate des Sozialprodukts

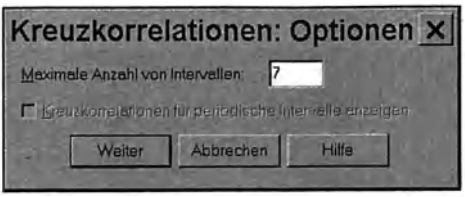


Abb. 26.62. Dialogbox „Kreuzkorrelationen: Optionen"

Analog zur Erstellung von Autokorrelationsdiagrammen werden im Ausgabefenster die Kreuzkorrelationskoeffizienten mit ihren Standardfehlern für die angeforderte Anzahl von lags aufgeführt und in einer einfachen Grafik dargestellt.

27 Herkömmliche Grafiken gestalten

27.1 Das Diagramm-Editorfenster

Nachdem man eine herkömmliche Grafik erzeugt hat (⇨ Kap. 26), möchte man die Grafik für Präsentationszwecke ansprechender gestalten. Die Überarbeitung und Layoutgestaltung einer Grafik geschieht im Diagramm-Editorfenster. Um dieses zu öffnen, doppelklickt man auf die im Ausgabefenster (Viewer) befindliche Grafik (alternativ über Menü: „Bearbeiten", „Objekt: SPSS-Diagramm", „Öffnen"). Die Grafik erscheint nun zur Bearbeitung im Diagramm-Editorfenster (⇨ Abb. 27.1). Im Ausgabefenster bleibt die Grafik erhalten, wird aber schraffiert angezeigt. Es können mehrere Grafikfenster mit je einer Grafik parallel geöffnet sein. Eine Begrenzung der Anzahl ist durch die Systemressourcen bedingt. Eventuell müssen Fenster geschlossen werden, um neue zu öffnen.

Sobald man in das Diagramm-Editorfenster wechselt, werden sowohl die Menüs als auch die Symbolleisten des Ausgabefensters durch die des Diagramm-Editorfensters ersetzt. Neben Standardsymbolen, die auch in Symbolleisten anderer Fenster enthalten sind, treten Formatierungssymbole zur Layoutgestaltung (⇨ Abb. 27.1). Beide Symbolleisten können verschoben und nach eigenen Wünschen zusammengestellt werden.

Im folgenden werden die Menüs sowie die Formatierungssymbole kurz erläutert. Anhand von Beispielen werden sie unten ausführlicher erklärt.

Die Menüs im Diagramm-Editorfenster.

① *Datei.* Mit dem Befehl „Diagrammvorlage speichern" öffnet sich eine Dialogbox zum Speichern der Grafik als Layoutvorlage für andere Grafiken. Man wählt einen Ordner (Looks ist das Standardverzeichnis für Grafikvorlagen) und vergibt einen Dateinamen (Dateiendung .sct). Mit „Diagramm exportieren" kann die Grafik in einem anderen Grafikformat gespeichert werden (⇨ unten: Grafik exportieren).

② *Bearbeiten.* Mit „Diagramm kopieren" wird die Grafik in die Zwischenablage kopiert. Von dort kann sie z.B. in eine Datei eines Textverarbeitungsprogramms eingefügt werden(⇨ Kap. 26.9). Mit „Optionen" können allgemeine Voreinstellungen für Diagramme, den Viewer, Pivot-Tabellen etc. festgelegt werden (⇨ Kap. 26.5).

③ *Ansicht.* „Statusleiste" erlaubt Anzeigen oder Ausblenden der Statusanzeigeleiste (⇨ Abb. 2.1). „Symbolleisten" öffnet die Dialogbox „Symbolleisten anzeigen". Es können Symbolleisten ein- bzw. ausgeblendet, angepasst sowie neue Symbolleisten erstellt werden. (⇨ Kap. 26.4).

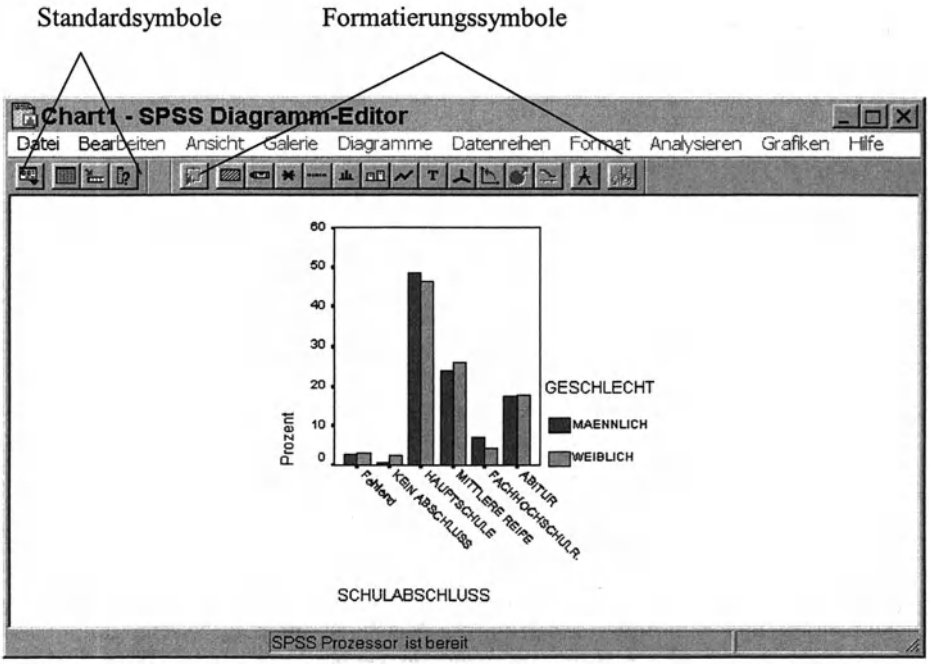


Abb. 27.1. Das Diagramm-Editorfenster

④ *Galerie.* Mit den Befehlen dieses Menüs kann man zu einem neuen Grafiktyp wechseln, der zu den Daten der aktuellen Grafik im Grafikfenster passt. So kann z.B. ein Balkendiagramm in ein entsprechendes Linien- oder Kreisdiagramm überführt werden und umgekehrt. Nach Auswahl eines Diagrammtyps aus einer Dialogbox wird mit „Ersetzen" die Grafik durch die neue ersetzt (⇨ Kap. 27.3).

⑤ *Diagramme.* Mit den Befehlen dieses Menüs können viele Layout- und Beschriftungsmerkmale wie Titel, Labels, Achsenbeschriftung, Rahmen etc. geändert werden. (⇨ Kap. 27.4).

⑥ *Datenreihen.* Mit „Angezeigt" können Datenreihen und Kategorien in einer Grafik weggelassen oder angezeigt werden. Für Balken-, Linien-, und Flächendiagramme kann man in der Darstellung bestimmen, ob eine Reihe als Linie, Fläche oder als Balkenreihe dargestellt werden soll. Auf diese Weise entstehen gemischte Balken-, Linien- bzw. Flächendiagramme. Des weiteren kann bei bestimmten Grafiktypen mit „Daten transponieren" die Art der Darstellung verändert werden (⇨ Kap. 27.5).

⑦ *Format* (vormals „Grafikattribute"). Mit den Befehlen in diesem Menü kann man eine Reihe von Grafikmerkmalen wie Füllmuster und Farben von Balken, Markierungssymbole für Punkte in Streudiagrammen, Stile der Linien in Liniengrafiken, Schattierungs- und 3D-Effekte für Balkendiagramme, Anordnung von Labels in Diagrammen, Schriftart- und größe verändern und gestalten sowie Rotationen von 3D-Streudiagrammen bewirken. Durch Mausklick auf das zu ändernde Element in der Grafik öffnet sich ein Palettenfenster, aus dem das

Gewünschte ausgewählt werden kann. Alternativ zur Menübedienung - aber schneller - können dazu die Formatierungssymbole im Grafikfenster (⇨ Abb. 27.1) verwendet werden (⇨ Kap. 27.6).

⑧ *Analysieren* (vormals Statistik) Enthält die Befehle zum Aufruf statistischer Prozeduren.

⑨ *Grafiken*. Enthält die Befehle zum Erzeugen von Grafiken.

⑩ *Hilfe*. Das Hilfe-Menü (⇨ Kap. 26).

Formatierungssymbole. Die Standardsymbole sind in Kapitel 2.2 erläutert worden. Die im folgenden kurz erläuterten Formatierungssymbole (⇨ Abb. 27.1) starten zum größten Teil die gleichen Funktionen wie die Befehle im Menü „Format" (vormals „Grafikattribute"). Der Vorteil liegt in der schnelleren Bedienung.


Punkte in Streudiagrammen und Boxplots identifizieren und mit Labels versehen.


Flächen mit Füllmuster versehen.


Grafikobjekten Farben zuweisen.


Markierungen von Datenpunkten in Größe und Stil verändern.


Datenlinien in Stil und Stärke verändern.


Schattierungen und 3D-Effekt für Balken wählen.


Balken mit Werte-Label versehen.


Interpolationsart für Datenlinien in Streu-, Linien- und Flächendiagrammen wählen.


Textelemente von Grafiken in Schriftart und -stil verändern.


3D-Scatterplot drehen.

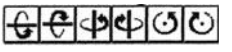

X- und Y-Achsen in 2D-Diagrammen vertauschen.


Segment in Kreisdiagrammen absetzen.


Darstellung fehlender Werte in Liniendiagrammen (Linien verbunden oder unterbrochen ).


Optionen für Grafiken aufrufen.


Drehmodus für 3D-Streudiagramme ein- bzw. ausschalten. Nach Einschalten erscheinen mehrere neue Schaltflächensymbole zum Drehen des Streudiagramms um seine Achsen:


Mit „Zurücksetzen" wird der Ursprungszustand wiederhergestellt.

Grafik exportieren. Eine im Ausgabe- oder Grafik-Editorfenster befindliche Grafik kann über die Zwischenablage von Windows in ein anderes parallel laufendes Anwendungsprogramm (z.B. Textverarbeitung) übernommen werden (⇨ 28.8).

Ab Version 6.1 von SPSS für Windows kann man eine im Ausgabe- bzw. Grafik-Editorfenster enthaltene Grafik in einem anderen Grafikformat speichern, um sie später in ein anderes Programm zu importieren. Mit der Befehlsfolge „Datei", „Diagramm exportieren..." öffnet sich die Dialogbox „Diagramm exportieren" zur Auswahl eines Laufwerks und Verzeichnisses sowie zur Vergabe eines Dateinamens. Befindet man sich im Ausgabefenster, wird mit „Datei", „Exportieren" die Dialogbox „Ausgabe exportieren" geöffnet, und in dieser wird in „Export:" „Nur Diagramme" gewählt. In beiden Dialogboxen kann man in „Dateityp" aus einer Drop-Down-Liste aus folgenden Grafikformaten auswählen:

- Enhanced Metafile (*.EMF).
- JPEG (*.JPG)
- Macintosch PICT (*.pct). Rasterstandard von Macintosh.
- PostScript (*.eps). Betriebssystemunabhängiger Druckerstandard.
- Tagged Image File (*.tif). Betriebssytemunabhängiger Rasterstandard, meistens komprimiert und damit kompakter als andere Rasterstandards.
- Windows Bitmap (*.bmp). Rasterstandard, von den meisten Windowsanwendungen unterstützt.
- Windows-Metadatei (*.wmf). Vektorstandard, von den meisten Windowsanwendungen unterstützt (z.B. von Word, Word Perfect, Pagemaker, Ami-Pro).

Klicken auf „Optionen" öffnet eine Unterdialogbox, in der man je nach Grafikformat eine Reihe von spezifischen Festlegungen vornehmen kann.

27.2 Ein Beispiel zum Gestalten einer Grafik

Beispielhaft soll nun die Überarbeitung einer Grafik anhand der in Abb. 27.1 dargestellten demonstriert werden.

① *Die Kategorie „Fehlend" und den Achsentitel „SCHULABSCHLUSS" entfernen sowie den Label „KEIN ABSCHLUSS" in „OHNE" ändern.*

Die Kategorie „Fehlend" in der Grafik hätte man bei der Erzeugung der Grafik durch Klicken auf „Optionen" in der Dialogbox „Gruppierte Balken: Auswertung über Kategorien einer Variablen" und Deaktivieren von „Fehlende Werte als Kategorie anzeigen" unterdrücken können (⇨ Abb. 20.8).

Es ist aber auch nachträglich möglich, diese Kategorie - wie auch jede andere - aus der Grafik zu entfernen. Die Befehlsfolge „Datenreihen", „Angezeigt..." öffnet die in Abb. 27.2 links dargestellte Dialogbox (alternativ: Doppelklicken auf die Balken). In der Gruppe „Kategorien" wird die im Feld „Anzeigen" markierte Kategorie „Fehlend" durch Klicken auf den Pfeilschalter in das Feld „Weglassen" übertragen. Mit „OK" wird die Dialogbox verlassen. Es entsteht die Grafik ohne die Kategorie.

Zur Entfernung des Achsentitels „SCHULABSCHLUSS" doppelklickt man auf die Achse, die Achsenbeschriftung oder den Achsentitel der Grafik. Es öffnet sich die in Abb. 27.2 rechts oben dargestellte Dialogbox „Kategorienachse". Nun löscht man den in „Achsentitel" angezeigten Text „SCHULABSCHLUSS". Zur Änderung des Labels „KEIN SCHULABSCHLUSS" in „OHNE" klickt man in

der Dialogbox „Kategorienachse“ auf die Schaltfläche „Beschriftungen...“. Es öffnet sich die in Abb. 27.2 rechts unten dargestellte Dialogbox „Kategorienachse: Beschriftungen“. Nach Markieren des Labels „KEIN SCHULABSCHLUSS“ wird dieser Text im Texteingabefeld „Label:“ angezeigt und kann dort durch Überschreiben zum neuen Label-Text (hier „OHNE“) verändert werden. Dieses muss durch Klicken auf „Ändern“ bestätigt werden. Mit „Weiter“ schaltet man eine Dialogbox-Ebene zurück und bestätigt mit „OK“.

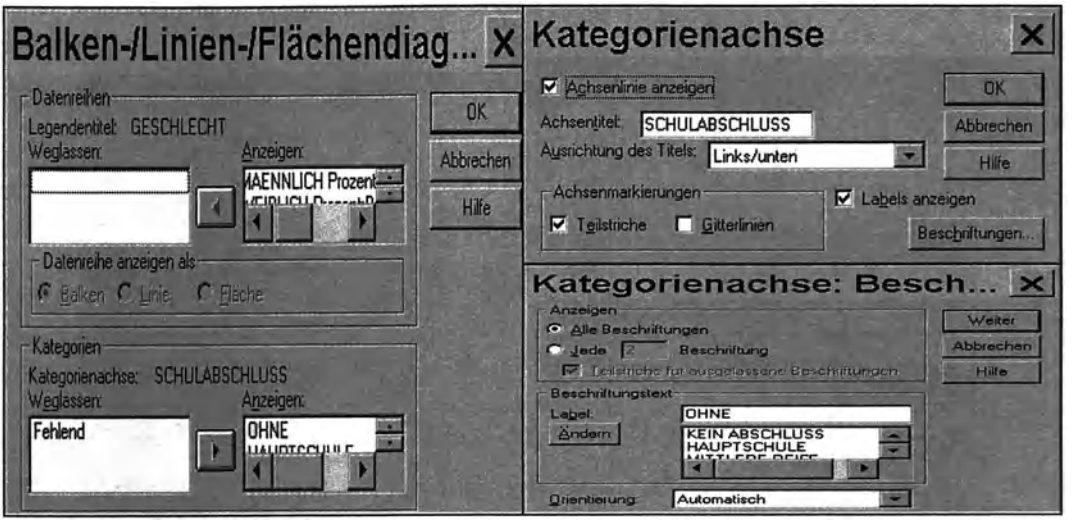


Abb. 27.2. Dialogboxen „Balken-/Linien-/Flächendiagramm: Daten anzeigen“, „Kategorienachse“ und „Kategorienachse: Beschriftungen“

② *Die Skalenachsenbeschriftung „Prozent“ in „%“ ändern und dann die Schriftart und -größe verändern.*

Durch Doppelklicken auf das Grafikelement „Prozent“ oder auf die Skalenachse der Grafik öffnet sich die in der Abb. 27.3 links dargestellte Dialogbox „Skalenachse“. Nun wird der Achsentitel „Prozent“ durch „%“ ersetzt. Für die „Ausrichtung des Titels“ wird „Mitte“ gewählt. „OK“ schließt den Vorgang ab.
Um anschließend die Schriftgröße von „%“ zu vergrößern, markiert man „%“ in der Grafik durch Einfachklick. Die Markierung wird durch einen Rahmen um „%“ angezeigt. Danach klickt man auf T in der Symbolleiste. Es öffnet sich dann die in der Abb. 27.3 rechts dargestellte Dialogbox. Es wird die Schriftart „Times New Roman“ und die Schriftgröße „14“ gewählt und damit die alte Schrift ersetzt. Mit Klicken auf „Zuweisen“ wird die neue Schrift angewendet. Auf gleiche Weise kann man auch andere Texte (Titel, Legenden, Label etc.) sowie die Skalenwerte der Grafik in Schriftart und -größe verändern.

Die Schriftartwahl kann aber auch durch Voreinstellen auf der Registerkarte „Diagramme“ der Dialogbox „Optionen“ erreicht werden.

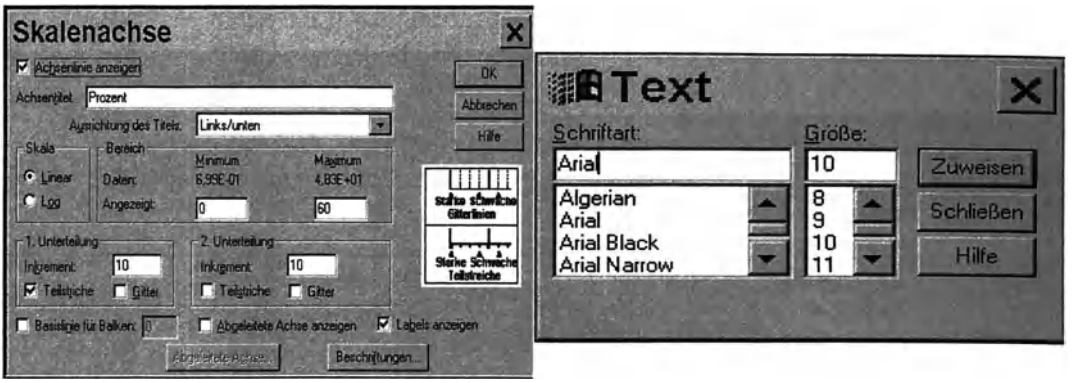


Abb. 27.3. Dialogboxen „Skalenachse“ und „Text“

③ *Titel, Untertitel sowie eine Fußnote einfügen.*

Titel, Untertitel und Fußnoten können schon bei Erzeugung der Grafik eingefügt werden (⇨ Kap. 26.2.1).

Nachträglich wird ein Titel bzw. ein Untertitel über die Befehlsfolge „Diagramme“, „Titel“ eingefügt. Es öffnet sich die in Abb. 27.4 links dargestellte Dialogbox „Titel“. Es wird ein Text für einen Titel und Untertitel eingegeben. Für die Ausrichtung des Titels in der Grafik wird „Mitte“ gewählt. „OK“ schließt die Titeleinfügung ab.

Durch Klicken von „Diagramme“, „Fußnote...“ öffnet sich die in Abb. 27.4 rechts dargestellte Dialogbox zur Eingabe eines Fußnotentextes. Es ist ein Fußnotentext eingetragen worden. Die voreingestellte Positionierung „Linksbündig“ wird beibehalten.

Abb. 27.4. Dialogboxen „Titel“ und „Fußnoten“

④ *Balken mit 3D-Effekt und Füllmuster versehen.*

Um die Balken mit 3D-Effekt abzubilden, klickt man auf den Symbolschalter . Es öffnet sich dann die in Abb. 27.5 links dargestellte Dialogbox „Balkenart“. Es wird ein 3D-Effekt gewählt und die „Tiefe“ von 24 auf 30 Prozent erhöht. Mit Klicken auf „Allen zuweis.“ wird der Effekt angewendet, und mit „Schließen“ wird der Vorgang beendet.

Zur besseren Unterscheidung der Balkenreihen für Männer und Frauen werden die Balken mit unterschiedlichen Füllmustern versehen. Zunächst wird durch Ein-

fachklicken auf einen Balken zur Darstellung der Schulabschlüsse für Männer diese Balkenreihe markiert (alternativ: Klicken auf „Maennlich“ in der Legende). Die Markierung wird durch Markierungspunkte gekennzeichnet (⇨ Abb. 27.5). Danach wird durch Klicken auf den Symbolschalter die in Abb. 27.5 rechts dargestellte Palette „Füllmuster“ geöffnet. Nach Auswahl des Füllmusters und Klicken auf „Zuweisen“ wird das gewählte Füllmuster in die Grafik eingefügt. Danach wird die Balkenreihe zur Darstellung der Schulabschlüsse für Frauen markiert, ein anderes Füllmuster ausgewählt und zugewiesen. Durch Klicken auf „Schließen“ wird die Palette geschlossen und der Vorgang beendet.

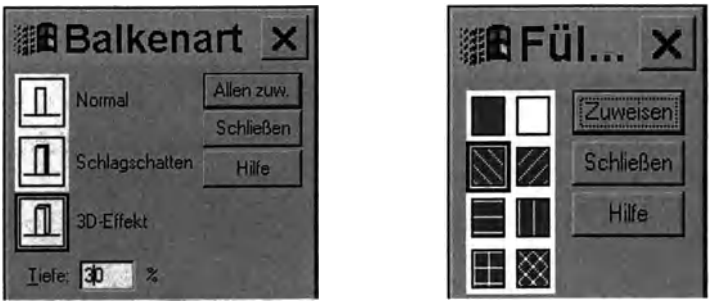


Abb. 27.5. Dialogbox „Balkenart“ und Palette „Füllmuster“

Balken mit einer Wertebeschriftung und anderen Farben versehen.

Um die Balken mit den Prozentwerten für die einzelnen Schulabschlüsse zu beschriften, wird auf den Symbolschalter geklickt. Es öffnet sich die in Abb. 27.6 links dargestellte Dialogbox zur Auswahl einer Wertebeschriftung. Nach Wahl von „Rahmen“ und Klicken auf „Allen zuweis.“ wird die Beschriftung durchgeführt.

Damit der 3D-Effekt noch besser zur Geltung kommt, werden die Balkenflächen mit unterschiedlichen Farben versehen. Für die Frontflächen wird jeweils eine helle und für die anderen Flächen eine dunklere Farbe gewählt. Um dieses zu erreichen, kann man die einzelnen Flächen der Balken durch Einfachklick markieren (wird durch Markierungspunkte angezeigt). Anschließend wird auf den Symbolschalter geklickt. Es öffnet sich die in Abb. 27.6 rechts dargestellte Palette „Farben“. Nach Auswahl einer Farbe und Klicken auf „Zuweisen“ wird die markierte Fläche mit der Farbe versehen. Danach wird die nächste Fläche auf gleiche Weise mit einer Farbe versehen. Ist die Farbgebung für die Grafik wunschgemäß, kann die Palette „Farben“ geschlossen werden.

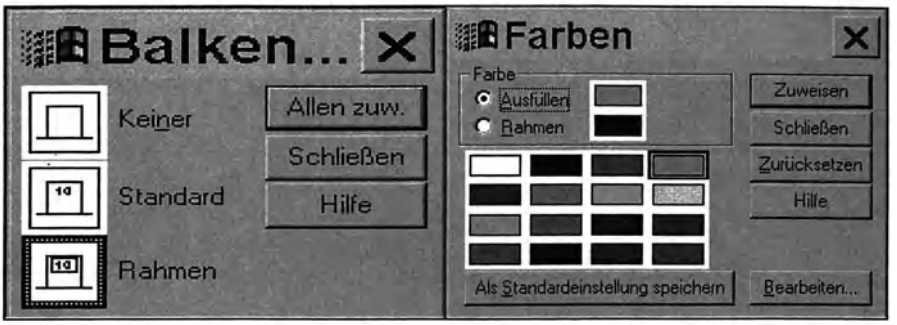


Abb. 27.6. Dialogbox „Balken“ und Palette „Farben“

© *Inneren Rahmen weglassen und äußeren Rahmen einfügen.*

Dazu wird im Menü „Diagramme“ „Äußerer Rahmen“ angeklickt und dieser damit aktiviert. Analog wird durch Klicken auf „Innerer Rahmen“ deaktiviert.

In Abb. 27.7 ist die in den beschriebenen sechs Schritten überarbeitete Grafik abgebildet.

Abb. 27.7. Zur Präsentation überarbeitete Grafik der Abb. 27.1

27.3 Wechseln zwischen Grafiktypen (Menü „Galerie“)

Im Diagramm-Editorfenster ist es möglich, zwischen Balken-, Linien-, Flächen-, Hoch-Tief- und Kreisdiagrammen hin und her zu wechseln. Soweit es die Datenlage zulässt, kann man auch innerhalb der Balken- und innerhalb der Liniengrafikarten etc. wechseln. Interessant ist insbesondere die Möglichkeit, gemischte Balken- und Liniengrafiken oder gemischte Flächen- und Liniengrafiken etc. zu erstellen. Ebenso kann man zwischen Streudiagrammarten - soweit es die Datenlage zulässt - und zu Histogrammen wechseln. Mittels einiger exemplarischer Beispiele soll dieses gezeigt werden.

Vom Balken- zum Kreisdiagramm. Ausgehend von dem in Abb. 27.1 dargestellten gruppierten Balkendiagramm kann eine der zwei Datenreihen (Männer bzw. Frauen) in ein Kreisdiagramm überführt werden. Mit der Befehlsfolge

▷ „Galerie“, „Kreis...“

wird die in Abb. 27.8 dargestellte Dialogbox geöffnet.

Abb. 27.8. Dialogbox „Kreisdiagramm“

Bei Wahl von „Einfach“ und Klicken auf „Ersetzen“ öffnet sich die in Abb. 27.9 links dargestellte Dialogbox. Im Feld „Anzeigen“ wird die Datenreihe „MAENNLICH Prozent“ aufgeführt und demgemäß für die Darstellung des Kreisdiagramms genutzt. Soll die Häufigkeitsverteilung der Schulabschlüsse für die Frauen im Kreisdiagramm dargestellt werden, so muss die entsprechende Datenreihe per Markieren und Pfeilschalter in das Feld „Anzeigen“ übertragen werden. In Abb. 27.9 rechts ist das Kreisdiagramm (ohne Titel und Fußnote) zu sehen.

Vom gruppierten Balken- zum gemischten Balken- /Liniendiagramm. Es ist möglich, in einer Grafik die Darstellung der Datenreihen in Form von Balken, Linien oder Flächen zu mischen. Im folgenden Beispiel wird die Mischform aus Balken und einer Linie gezeigt. Ausgehend von dem in Abb. 27.1 dargestellten gruppierten Balkendiagramm wird mit der Befehlsfolge

▷ „Galerie“, „Gemischt...“

die in Abb. 27.10 dargestellte Dialogbox geöffnet.

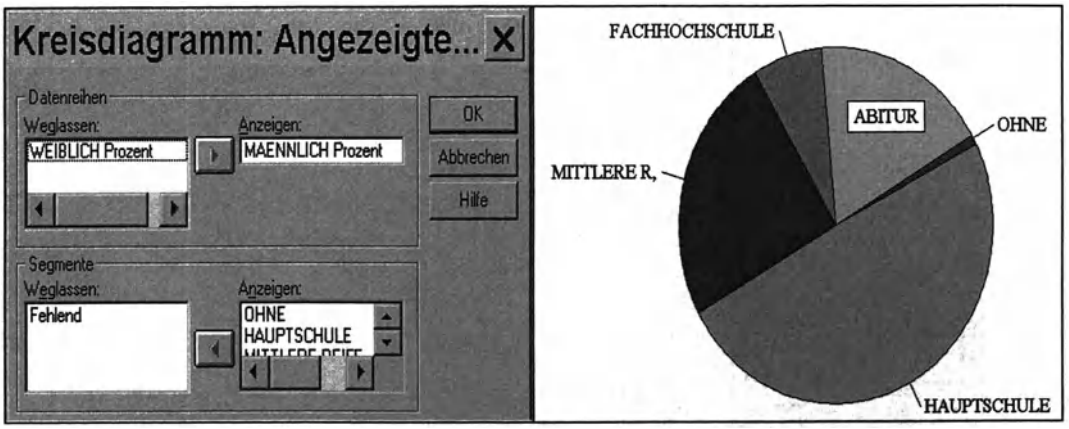


Abb. 27.9. Wechseln vom gruppierten Balkendiagramm zum Kreisdiagramm

Abb. 27.10. Dialogbox „Gemischtes Diagramm"

Nach Klicken auf „Ersetzen" öffnet sich die in Abb. 27.11 links dargestellte Dialogbox. Im Feld „Anzeigen" wird zunächst sowohl „WEIBLICH Prozent: Balken" als auch „MAENNLICH Prozent: Balken" angezeigt. Wird nun z.B. „MAENNLICH Prozent: Balken" markiert und anschließend in „Datenreihe anzeigen", „Linie" geklickt, so wandelt sich die Anzeige in „MAENNLICH Prozent: Linie". Nach Klicken von „OK" wird das gruppierte Balkendiagramm in ein gemischtes Balken-/Liniendiagramm überführt. Auf diese Weise kann jede Datenreihe eines Diagramms zwischen den Darstellungsarten Balken-, Linien- bzw. Flächendiagramm wechseln.

Vom Streudiagramm zum Histogramm. Man kann zwischen Streudiagrammtypen - soweit es die Datenlage zulässt - sowie zu Histogrammen wechseln. Am Beispiel eines Matrix-Streudiagramms sei ein Wechsel zum Histogramm demonstriert. Ausgehend von dem in Abb. 26.50 dargestellten Matrix-Streudiagramm wird durch Klicken der Befehlsfolge

▷ „Galerie", „Histogramm..."

die in Abb. 27.12 dargestellte Dialogbox geöffnet.

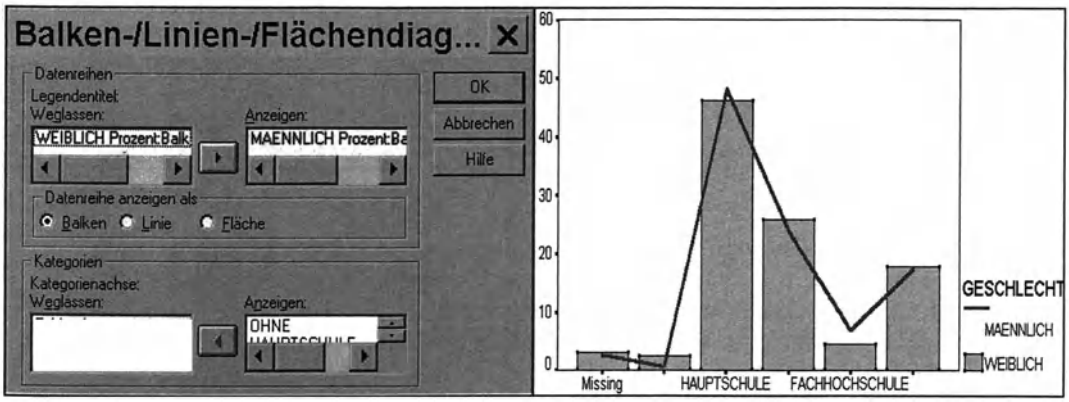


Abb. 27.11. Wechseln vom gruppierten zum gemischten Balken-/Liniendiagramm

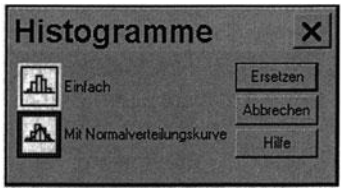


Abb. 27.12. Dialogbox „Histogramm“

Es kann ein einfaches Histogramm oder eines ergänzt um die Kurve einer Normalverteilung gewählt werden. Durch Wahl von „Einfach“ und Klicken von „Ersetzen“ öffnet sich die in Abb. 27.13 links dargestellte Dialogbox. Durch Markieren einer Variablen und Klicken auf den Pfeil-Schalter kann man die Variable zwischen den Feldern „Weglassen“ und „Anzeigen“ übertragen. In Abb. 27.13 rechts ist das Ergebnis der gewählten Einstellungen zu sehen.

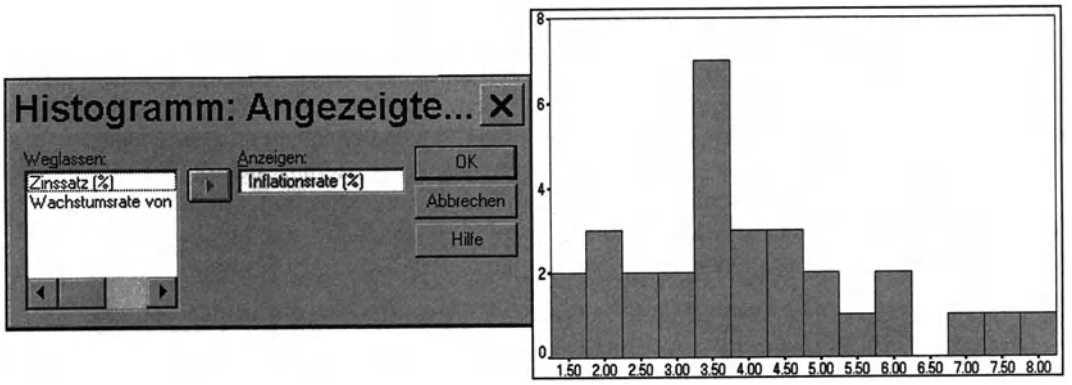


Abb. 27.13. Wechsel von einem Matrix-Streudiagramm zum Histogramm

27.4 Überarbeiten von Objekten einer Grafik (Menü „Diagramme")

27.4.1 Objekte einer Grafik

Eine Grafik besteht aus verschiedenen Bestandteilen (Objekten), die man allgemein in zwei Gruppen einteilen kann:

- *Datenreihen-Objekte.* Dies sind die Balken, Linien, Markierungspunkte der Grafiken. Sie repräsentieren die Daten. Sie können als Datenreihen gewählt und in ihrer grafischen Darstellung verändert werden.
- *Grafik-Objekte.* Dies sind die Elemente einer Grafik, die nicht Datenreihen-Objekte sind. Es handelt sich z.B. um Überschriften, Fußnoten, Skalen- und Kategorienachsen, Achsenbeschriftungen, Legenden, Rahmen etc. Sofern diese Objekte in der Grafik enthalten sind, können sie ausgewählt und überarbeitet werden.

Die Objekte einer Grafik haben Attribute, die auch verändert werden können: z.B. können die Farbe und das Füllmuster von Balken, die Schriftart bzw. -größe von Überschriften, die Datenpunkte von Streudiagrammen in Art, Farbe und Stärke verändert werden.

Eine Grafik wird im Diagramm-Editorfenster (⇨ Kap. 27.1) überarbeitet. Die Überarbeitung eines Objekts einer Grafik vollzieht sich in zwei Schritten: Im ersten Schritt wird das Objekt ausgewählt (z.B. eine Legende oder die Balken eines Balkendiagramms). Im zweiten Schritt wird das Objekt überarbeitet. Die Auswahl eines Objekts kann in alternativer Weise geschehen:

- *Durch Auswahl über die Menüs „Diagramme" oder „Datenreihen" im Grafikfenster.* Ein Grafik-Objekt wird über das Menü „Diagramme" (vormals „Grafik") und ein Datenreihen-Objekt über das Menü „Datenreihen" ausgewählt. Aus den Drop-Down-Listen dieser Menüs können Untermenüs ausgewählt werden. Damit werden Dialogboxen geöffnet, die zur Spezifikation der Darstellung dienen.
- *Durch Doppelklick mit der Maus auf das Objekt in der Grafik.* Je nachdem ob man auf ein Datenreihen-Objekt (z.B. eine Linie oder einen Balken) oder ein Grafik-Objekt (z.B. eine Überschrift, Legende oder eine Achse) geklickt hat, öffnet sich eine Dialogbox aus dem Menü „Datenreihen" oder dem Menü „Diagramme". In den Dialogboxen wird die Spezifizierung der grafischen Darstellung vorgenommen. Die Auswahl per Doppelklick mit der Maus ist komfortabler und schneller, und man wird diese gegenüber der Menüauswahl im allgemeinen bevorzugen.

Folgende Grafikbearbeitungen kann man aber nur über eines der beiden Menüs vornehmen:

- Nur über das Menü „Diagramme":
  - Veränderung des Abstandes der Balken in einem Balkendiagramm.
  - Einfügen von bislang nicht in der Grafik enthaltenen Titeln, Fußnoten und Anmerkungen.

❒ Nur über das Menü „Datenreihen“:
- Transponieren der Daten.

Des weiteren gibt es spezielle Änderungsmöglichkeiten von Grafiken, die nicht durch Doppelmausklick auf ein Datenreihen- oder Grafik-Objekt einer Grafik im Grafikfenster ausgewählt werden können. Diese Änderungsmöglichkeiten sind spezifisch für einen Grafiktyp. Die Befehle zur Änderung eines jeweiligen Grafiktyps sind im Untermenü „Optionen“ des Menüs „Diagramme“ zusammengefasst.

Auf dreifache Weise kann die Dialogbox „Optionen“ geöffnet werden, um damit die optionalen Gestaltungsmöglichkeiten für den aktuell im Diagramm-Editorfenster befindlichen Grafiktyp wahrzunehmen:

❒ Durch Klicken auf das in der Symbolleiste des Diagramm-Editorfensters.
❒ Durch die Befehlsfolge „Diagramme“, „Optionen“ im Diagramm-Editorfensters.
❒ Durch Doppelmausklick auf eine „freie“ Stelle der Grafik (also nicht auf eine Legende, eine Achse, einen Balken etc.). Auch wenn man versehentlich ein erwünschtes Objekt beim Doppelklicken nicht erfasst, öffnet sich die Dialogbox „Optionen“.

In der Dialogbox „Optionen“ können Spezifizierungen für die Grafik festgelegt werden.

Damit sind die Möglichkeiten von SPSS, Grafiken zu ändern, aber noch nicht ausgeschöpft. Sowohl Datenreihen-Objekte als auch Grafik-Objekte haben Merkmale (Attribute) verschiedenster Art wie Farbe und Muster, Schriftart und -größen etc., die man zur Layoutgestaltung nutzen kann. Zur Veränderung des Attributs eines Objekts wird das Objekt im ersten Schritt durch Einfachklick ausgewählt. Im zweiten Schritt wird über das Menü „Format“ (vormals „Attribute“) oder schneller über die Symbole des Diagramm-Editorfensters eine Box geöffnet, die es erlaubt, Spezifizierungen zuzuweisen. Dabei kann man einem ausgewählten Objekt nacheinander mehrere Attribute zuweisen.

In diesem Kapitel werden nur die Befehle des Menüs „Diagramme“ behandelt. Die Menüs „Datenreihen“ und „Format“ (vormals „Attribute“) werden danach erläutert. Die folgende Tabelle 27.1 gibt einen Überblick.

Tabelle 27.1. Überblick über den Inhalt von Kapiteln

Menü DIAGRAMME ⇨ Kap. 27.4	Menü DATENREIHEN ⇨ Kap. 27.5	Menü FORMAT ⇨ Kap. 27.6

27.4.2 Optionen zum Gestalten von Diagrammen (Menü „Optionen“)

Im (Unter-)Menü „Optionen“ des Menüs „Diagramme“ sind spezielle Änderungsmöglichkeiten von Grafiken zusammengefasst. Diese Dialogboxen sind für Grafiktypen spezifisch: die Dialogbox „Optionen“ eines Balkendiagramms z.B. unterscheidet sich von der für ein Kreisdiagramm. Zur Öffnung dieser Dialogboxen stehen drei Wege bereit:

- Klicken auf den Symbolschalter.
- Durch die Befehlsfolge „Diagramme", „*Optionen*".
- Durch Doppelmausklick auf eine „freie" Stelle der Grafik (also nicht auf eine Achse, Legende, Achsenbeschriftung etc.).

Im folgenden wird auf diese Optionen zur Grafikgestaltung eingegangen.

Optionen zum Gestalten von Balken-, Linien-, und Flächendiagrammen. Befindet sich eines dieser Diagramme im Diagramm-Editorfenster und klickt man auf (alternativ: Doppelklicken auf eine „freie" Stelle in der Grafik oder die Befehlsfolge „Diagramme", „Optionen"), so öffnet sich die in Abb. 27.14 dargestellte Dialogbox.

Abb. 27.14. Dialogbox „Optionen für Balken-/Linien-/Flächendiagramme"

Skala auf 100 % setzen. Für Balken- und Flächendiagramme kann die Skala der senkrechten Achse auf 100 % festgelegt werden. Abb. 27.15 zeigt dieses für die Schulabschlüsse von Männern und Frauen.

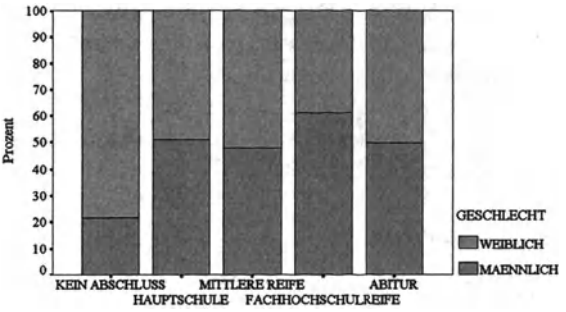


Abb. 27.15. Balkendiagramm mit Option „Skala auf 100 % setzen"

Markierungen innerhalb Kategorien verbinden. Zwischen Markierungspunkten der Linien eines Mehrfachliniendiagramms werden senkrechte Verbindungslinien eingefügt.

Projektion anzeigen. Wird diese Option gewählt (nur für Liniendiagramme möglich), so wird die Schaltfläche „Kategorie" aktiviert. Klicken auf „Kategorie" öffnet die in Abb. 27.16 links dargestellte Dialogbox. In der Abbildung wird diese Option am Beispiel der Häufigkeitsverteilung von Arbeitstunden für Männer und

Frauen (Variable ARBSTD2 für Fallauswahl ARBSTD2 > 0) dargelegt. In der Dialogbox ist der Trennwert „45 BIS 49,5“ markiert. Außerdem ist die Option „Bezugslinie bei Kategorie anzeigen“ gewählt. Auf der rechten Seite der Abb. 27.16 ist das Ergebnis dieser Spezifizierungen zu sehen. Diese Option ist z.B. interessant, wenn man für Zeitreihendarstellungen Prognosewerte deutlich herausstellen möchte.

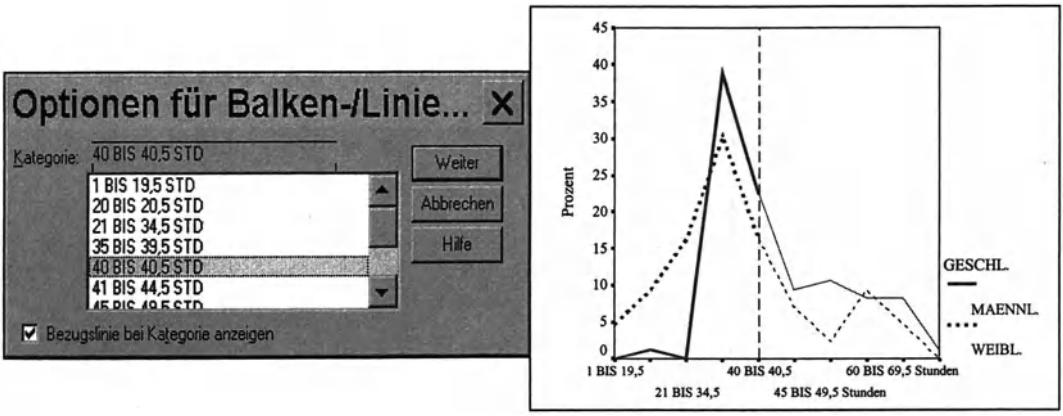


Abb. 27.16. Liniendiagramm durch Trennwert teilen

Optionen zum Gestalten von Kreisdiagrammen. Befindet sich ein Kreisdiagramm im Diagramm-Editorfenster und klickt man auf (bzw. Doppelklicken auf eine „freie“ Stelle in der Grafik oder die Befehlsfolge „Diagramme“, „Optionen“), so öffnet sich die in Abb. 27.17 dargestellte Dialogbox.

Abb. 27.17. Dialogbox „Optionen für Kreisdiagramm“

Am Beispiel der Häufigkeitsverteilung von Schulabschlüssen seien die folgenden wählbaren Optionen erläutert:

- *Erstes Segment bei: Uhr.* Voreingestellt ist 12 Uhr. Durch Eingabe einer anderen Uhrzeit zwischen 1 und 12 kann das Kreisdiagramm gedreht werden (hier: 3 Uhr).
- *Zusammenfassung aller Segmente kleiner als %.* Voreingestellt sind 5 %. Andere Eingaben zur Zusammenfassung von Segmenten sind möglich. Das

zusammengefasste Segment erhält automatisch das Label „Andere". Mit der Option „Text" für Labels kann es anders bezeichnet werden.

❒ *Beschriftung.* Mit den Optionen „Text bearbeiten" und „Format" können der Label-Text sowie das Label-Format überarbeitet werden. Dabei sind „Text", „Werte" und/oder „Prozente" wählbar.

- *Text bearbeiten.* Mit dieser Option kann der Text der Labels verändert werden. Wird „Text" mit „Text bearbeiten" gewählt, so öffnet sich die in Abb. 27.18 dargestellte Dialogbox. Das markierte Label „MITTLERE REIFE" erscheint im Überarbeitungsfeld und ist dort in „MITTLERE R." geändert worden. Mit „Ändern" muss dieses bestätigt werden, ehe man mit „Weiter" zur vorhergehenden Dialogbox zurückgeht. Für den Fall von Zusammenfassungen könnte man das Label „Andere" für das Restsegment mit einem neuen Label versehen.

Abb. 27.18. Dialogbox „Kreisdiagramm: Beschriftung"

- *Format.* Das Format des Labels kann sich auf den Label-Text, die Angabe von Werten oder von Prozenten beziehen. Klickt man auf „Format", so öffnet sich die in Abb. 27.19 links dargestellte Dialogbox zur Gestaltung des Label-Formats.

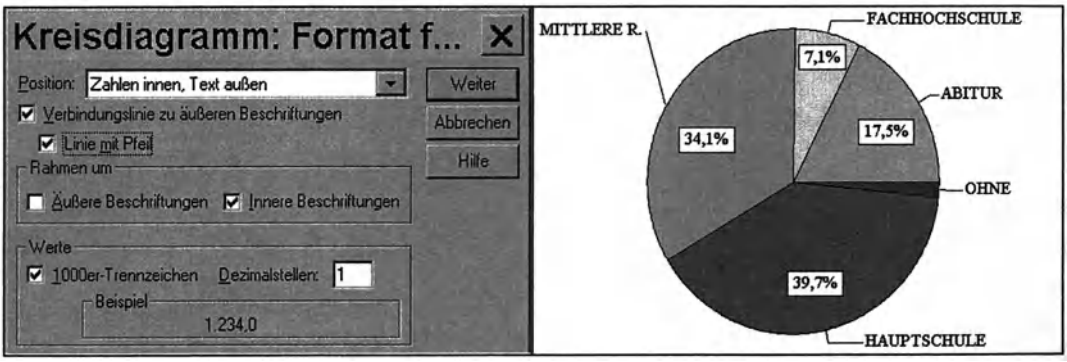


Abb. 27.19. Labelgestaltung im Kreisdiagramm

Für das Format des Labels können folgende Optionen gewählt werden:

- *Position.* Für die Positionierung des Labels gibt es Wahlmöglichkeiten. Je nach Wahl wird das Label inner- oder außerhalb des Kreisdiagramms positioniert (hier: „Zahlen innen, Text außen").

- *Verbindungslinie zu äußeren Beschriftungen.* Wahlweise kann angefordert werden, dass die Verbindungslinien als Pfeile dargestellt werden.
- *Rahmen um.* Diese Option erlaubt es, Labels bzw. Werte zu umrahmen (hier: Rahmen um „Innere Beschriftungen“).
- *Werte.* Mit „Werte“ ist es möglich, die Anzahl der Dezimalstellen für die Zahlenwerte (0 bis 19 Stellen) festzulegen. Voreingestellt sind zwei Dezimalstellen (hier: 1). Werte größer als 1000 werden mit einem Trennzeichen für Tausender (je nach Voreinstellung im Windows-System ein Punkt oder ein Komma) versehen.

Optionen zum Gestalten von Boxplot-Diagrammen. Befindet sich ein Boxplot-Diagramm im Grafikfenster und klickt auf [A] (bzw. Doppelklicken auf eine „freie“ Stelle in der Grafik oder die Befehlsfolge „Diagramme“, „Optionen“), so öffnet sich die in Abb. 27.20 dargestellte Dialogbox.

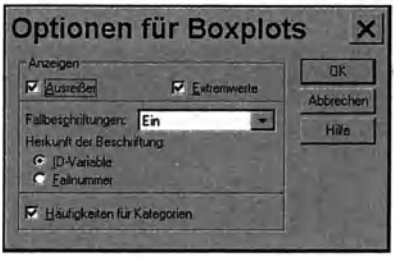


Abb. 27.20. Dialogbox „Boxplot-Optionen“

Folgende Möglichkeiten zur Gestaltung bestehen:

- *Anzeigen.* Man kann festlegen, ob Ausreißer und/oder Extremwerte im Diagramm angezeigt werden sollen. Zusätzlich kann gewählt werden, ob dafür Labels verwendet werden sollen (Fall-Labels: „Ein“, „Aus“ bzw. „Wie gegeben“). Bei einer Label-Anzeige kann man wählen, ob die Labels der Fallbeschriftungsvariablen (natürlich nur, wenn bei der Erzeugung der Grafik eine derartige Variable gewählt worden ist, ⇨ z.B. Abb. 26.44) oder die „Fallnummer“ als Label verwendet werden soll. Haben zwei Ausreißer oder Extremwerte den gleichen Variablenwert mit unterschiedlichen Labels, so wird die Labelanzeige unterdrückt. Für einzelne Ausreißer/Extremwerte kann man individuell festlegen, ob eine Labelanzeige erfolgen soll oder nicht. Dazu schaltet man am besten zunächst über die in Abb. 27.20 dargelegte Dialogbox die Labelanzeige auf „AUS“. Danach klickt man auf die Symbolschaltfläche [🔍]. Führt man anschließend den Mauscursor auf die Grafik, so erscheint der Cursor als [⊕]. Klickt man nun mit dem Cursor auf einen Ausreißer/Extremwert, so wird die Labelanzeige eingeschaltet bzw. ausgeschaltet. Klickt man jedoch auf einen Ausreißer-/Extremwert, bei dem die Labelanzeige wegen gleichen Variablenwertes bei unterschiedlichen Labels unterdrückt wird, so öffnet sich ein kleines Fenster, in dem beide Labels angezeigt werden. Die Symbolschaltfläche [🔍] dient auch zur Fallauffindung im

Dateneditor. Hat man für einen Ausreißer/Extremwert eine Labelanzeige eingeschaltet und klickt dann auf die Symbolschaltfläche, so wird zum Dateneditor umgeschaltet und dort der Fall aufgefunden und angezeigt. Mit kann man anschließend zur Grafik zurückschalten.

- *Häufigkeiten für Kategorien*. Es kann festgelegt werden, ob bei der Erzeugung von Boxplots unter jeder Kategorie in der Zeile „N =„ die Anzahl der Fälle angezeigt werden sollen.

Optionen zum Gestalten von einfachen und Matrix-Streudiagrammen. Die Möglichkeiten zur Gestaltung von Streudiagrammen sind je nach Diagrammart unterschiedlich. Auf sie wird im folgenden eingegangen.

Befindet sich ein einfaches oder ein Matrix-Streudiagramm im Diagramm-Editorfenster, so öffnet ein Klicken auf die in Abb. 27.21 dargestellte Dialogbox zur Festlegung von Gestaltungsmöglichkeiten. Am Beispiel des einfachen Streudiagramms zur Darstellung des Zusammenhangs zwischen ZINS (Zinssatz) und INFLAT (Inflationsrate) sollen die Möglichkeiten gezeigt werden.

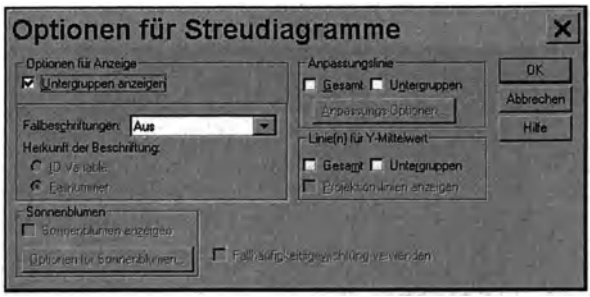


Abb. 27.21. Dialogbox „Optionen für Streudiagramme“

Folgende Gestaltungsmöglichkeiten bestehen:

① *Optionen für Anzeige*. Die Anzeige von Untergruppen bzw. die Anzeige von Fall-Labels als „ID-Variable“ ist nur anwendbar, wenn bei der Erzeugung des Diagramms Untergruppen definiert wurden und/oder eine Fallbeschriftung angefordert worden ist. Mit „Untergruppen“ kann gewählt werden, ob Untergruppen durch unterschiedliche markierte Datenpunkte im Streuungsdiagramm angezeigt werden sollen. Für die Anzeige von „Fall-Labels“ kann man zwischen „Ein“, „Aus“ bzw. „Wie gegeben“ wählen. Für die Art der Labelanzeige kann man wählen zwischen: „ID-Nummer“ (natürlich nur, wenn bei der Erzeugung eine Variable als „Fallbeschriftung“ gewählt worden ist) oder „Fallnummer“. Ähnlich wie bei Boxplot-Diagrammen kann man eine Labelanzeige für einzelne Punkte im Streudiagramm mit dem Symbolschalter ein- bzw. ausschalten. Man geht dabei wie oben beschrieben vor (⇨ „Optionen zur Gestaltung von Boxplot-Diagrammen“). Bei Streudiagrammen ist dies noch unkomplizierter, weil durch Klicken mit dem -Cursor auf einen Punkt die Labelanzeige sowohl ein- als auch ausgeschaltet werden kann. Analog zu der bei Boxplots beschriebenen Vorgehensweise, kann auch eine Fallauffindung im Dateneditor vorgenommen werden.

② *Sonnenblumen.* Diese Option macht Sinn, wenn sich viele Datenpunkte in einem Streuungsdiagramm überlappen bzw. aufeinanderfallen. Anstelle einer Darstellung der einzelnen Datenpunkte werden in Zellfeldern des Streudiagramms liegende Datenpunkte zusammenfasst und in Form einer „Sonnenblume“ dargestellt. Ein Datenpunkt wird nach wie vor als kleiner Kreis, mehrere zusammenliegende Datenpunkte eines Zellenfeldes werden durch kleine vom Kreis ausgehende Linien (Blütenblätter der Sonnenblume) dargestellt. Dabei wird die Anzahl der Datenpunkte in dem Zellenfeld durch die Anzahl der Blütenblätter der Sonnenblume dargestellt. Die Sonnenblumen können durch Klicken auf „Optionen für Sonnenblumen“ gestaltet werden. Es öffnet sich dann die in Abb. 27.22 dargestellte Dialogbox.

Abb. 27.22. Dialogbox „Optionen für Streudiagramme“

Es gibt folgende Gestaltungsmöglichkeiten für die Sonnenblumen:

- *Ein Blatt entspricht.* Voreingestellt ist die Darstellung eines Datenpunktes in einem Blatt der Blume. Soll ein Blütenblatt mehrere Datenpunkte repräsentieren, kann man dieses durch Überschreiben der 1 in dem Eingabefeld verändern. Alternativ kann man die Anzahl der Fälle je Blütenblatt automatisch bestimmen lassen. Ist die Anzahl der Blütenblätter kleiner 1,5, so wird nur das Zentrum der Sonnenblume angezeigt. Ist sie größer als 1,5, so wird die gerundete Anzahl der Blütenblätter angezeigt.
- *Auflösung.* Die Größe der Zellenfelder („Grob“ oder „Fein“) kann festgelegt werden.
- *Position.* Die Lage der Sonnenblume im Zellenfeld kann bestimmt werden. Mit „Zentriert“ wird die Sonnenblume in der Mitte plaziert. Wird „Mittelwert“ gewählt, so liegt die Sonnenblume im Schnittpunkt der Mittelwerte der Datenpunkte in dem Zellenfeld.

Das einfache Streudiagramm mit Sonnenblumen in Abb. 27.23 zeigt die Beziehung zwischen LOHNS (Lohnsatz = EINK/ARBSTD) und ALTER (ALLBUS90.SAV). Für „Ein Blatt entspricht“ wurde „Automatisch“, für die Auflösung „Grob“ und für die „Position“ „Zentrum“ gewählt.

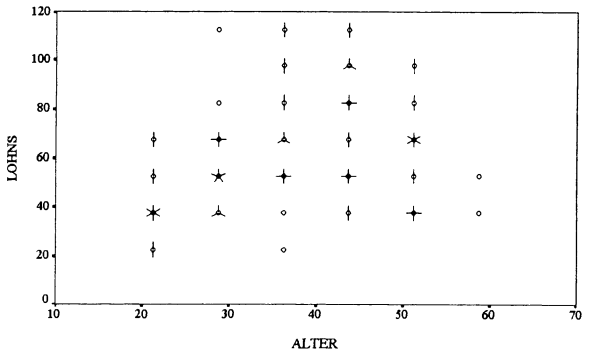


Abb. 27.23. Sonnenblumen-Streudiagramm: Lohnsatz und Alter

③ *Anpassungslinie.* In Streudiagramme können Anpassungskurven gelegt werden. Klicken auf „Anpassungs-Optionen" öffnet die in Abb. 27.24 dargestellte Dialogbox zur Auswahl von Anpassungslinien.

Abb. 27.24. Dialogbox „Scatterplot-Optionen: Anpassungslinie"

Folgende Auswahlmöglichkeiten bestehen für eine Kurvenanpassung:

- *Anpassungsmethode.* Man kann zwischen einer „Linearen", „Quadratischen" und Kubischen Regression auswählen. Die Kurven dieser Regressionsgleichungen (eine Gerade, die Kurve einer quadratischen oder kubischen Regression) werden mit Hilfe der Methode der kleinsten Quadrate berechnet und in das Diagramm plaziert. Alternativ dazu kann auch die Anpassungsmethode „Lowess" (= locally weighted regression scatterplot smoothing method, siehe dazu Cleveland, 1979, und Chambers et. al., 1983) gewählt werden. Bei diesem Verfahren wird eine iterativ gewichtete Methode der kleinsten Quadrate zur Anpassung an die Datenpunkte angewendet. Dadurch nimmt der Einfluss eines Beobachtungspunktes auf die Glättung an einem Punkt mit der Entfernung von diesem Punkt ab. Dabei sind mindestens 13 Datenpunkte erforderlich. Aufgrund der Rechenintensität des Verfahrens wird die Kurve für einen festgelegten Prozentsatz von Punkten angepasst. Außerdem wird eine feste Anzahl von Iterationen angewendet. Voreingestellt sind jeweils 50 % und drei Iterationen. Diese Voreinstellungen können durch Überschreibungen

verändert werden. In Abb. 27.25 ist für den Zusammenhang zwischen LOHNS und ALTER die Anpassungslinie „Lineare Regression" einerseits (links) und „Lowess" andererseits (rechts) eingepasst. Es wird eine Tendenz deutlich, dass mit zunehmendem Alter der Lohnsatz sich zunächst erhöht, aber dann wieder absinkt.

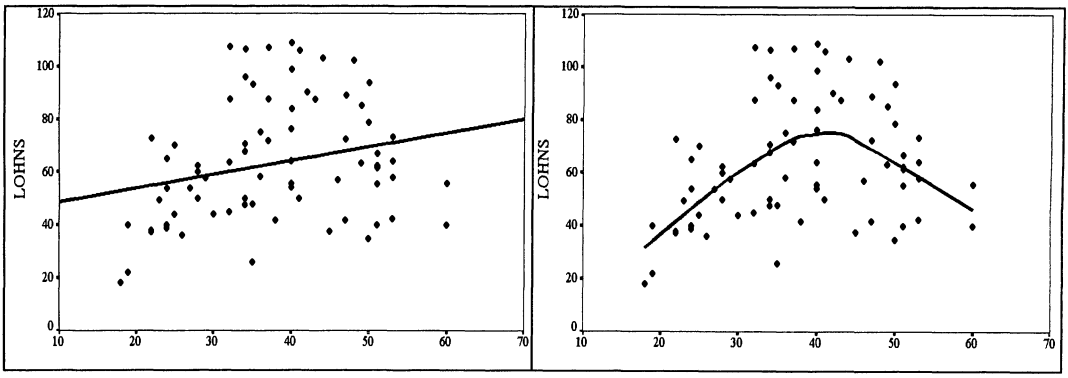


Abb. 27.25. Anpassungslinien „Lineare Regression „ und „Lowess"

- *Linien für Regressionsschätzer.* Es werden Konfidenzintervalle für die Vorhersagewerte der Regressionsgleichung angezeigt. Bei Wahl von „Mittelwert" beziehen sich die Intervalle auf die mittleren, bei „Einzelwert" auf individuelle Vorhersagewerte (⇨ ③ in Kap. 17.2.4). Das voreingestellte Konfidenzintervall beträgt dabei 95 %. Es kann verändert werden.
- *Optionen für Regression.* Mit „Konstante in Gleichung einschließen" kann bestimmt werden, ob die Regressionslinie durch den Ursprung des Koordinatensystems gehen soll. Mit „R-Quadrat in Legende anzeigen" kann das Bestimmtheitsmaß in der Legende anzeigt werden. Für Matrix-Streudiagramme ist diese Option nicht vorhanden.

④ *Linie(n) für Y-Mittelwert.* Es kann der Mittelwert der Y-Werte als Referenzlinie in das Streuungsdiagramm gelegt werden. Wenn Untergruppen bestehen, wird bei der Wahl von „Untergruppen" die Referenzlinie für jede Untergruppe eingefügt. Mit „Projektionslinien anzeigen" werden senkrechte Verbindungslinien zwischen Datenpunkten und der Referenzlinie eingefügt.

⑤ *Fallhäufigkeitsgewichtung verwenden.* Wurde im Menü „Daten" mit „Fälle gewichten" eine Gewichtungsvariable eingetragen, so werden die Gewichte für die Streudiagramme verwendet.

Optionen zum Gestalten von überlagerten Streudiagrammen. Befindet sich ein überlagertes Streudiagramm im Diagramm-Editorfenster und klickt man auf [A] (alternativ: Doppelklicken auf eine „freie" Stelle im Diagramm oder die Befehlsfolge „Diagramme", „Optionen"), so öffnet sich die in Abb. 27.26 dargestellte Dialogbox.

Wie bei einfachen und Matrix-Streudiagrammen kann man Kurvenanpassungen vornehmen. Im Unterschied dazu wird die ausgewählte Kurvenanpassung aber für

jedes Variablenpaar separat angewendet. Analoges gilt für die Einfügung von Y-Mittelwertlinien. Falls eine Fall-Beschriftung vorgenommen wurde, kann diese unterdrückt werden („Fallbeschriftungen" auf „Aus" schalten). Wahlweise kann für die Fallbeschriftung die „ID-Variable" (nur wenn bei Erzeugung der Grafik eine Variable für die Fallbeschriftung gewählt worden ist ⇨ Abb. 26.49) oder die „Fallnummer" herangezogen werden. Ähnlich wie bei Boxplot-Diagrammen und anderen Streudiagrammarten, kann man eine Labelanzeige für einzelne Punkte sowie eine Fallauffindung im Dateneditor mittels des Symbolschalters [🔍] erzielen (⇨ „Optionen zur Gestaltung von Boxplot-Diagrammen" und „Optionen zur Gestaltung von einfachen und Matrix-Streudiagrammen"). Außerdem kann eine Gewichtung angewendet werden.

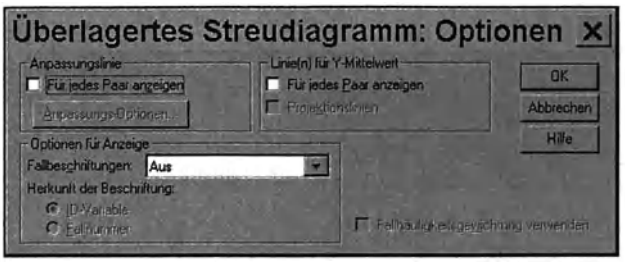


Abb. 27.26. Dialogbox „Überlagertes Streudiagramm: Optionen"

Optionen zum Gestalten von 3D-Streudiagrammen. Befindet sich solches Diagramm im Diagramm-Editorfenster, so öffnet ein Klicken auf [🔍] (oder Doppelklick auf eine freie Stelle) die in Abb. 27.27 dargestellte Dialogbox zur Festlegung von Gestaltungsmöglichkeiten.

Falls bei der Erstellung des Streudiagramms Untergruppen und/oder Fall-Beschriftungen definiert wurden, kann dieses unterdrückt werden. Für die Fall-Labels hat man die Auswahl: „ID-Variable" (nur wenn bei Erzeugung der Grafik eine Variable für die Fallbeschriftung gewählt worden ist) oder „Fallnummer". Ähnlich wie bei Boxplot-Diagrammen und anderen Streudiagrammarten kann man eine Labelanzeige für einzelne Punkte sowie eine Fallauffindung im Dateneditor mittels des Symbolschalters [🔍] erzielen (⇨ „Optionen zur Gestaltung von Boxplot-Diagrammen" und „Optionen zur Gestaltung von einfachen und Matrix-Streudiagrammen"). Des weiteren können Projektionslinien verschiedenster Art in das Streudiagramm eingefügt werden. Zudem kann man zwischen zwei Rahmen für das Diagramm wählen oder auf einen Rahmen verzichten.

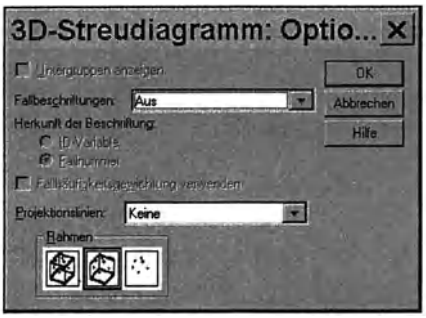


Abb. 27.27. Dialogbox „3 rechts D-Streudiagramm: Optionen“

Optionen zum Gestalten von Histogrammen. Befindet sich ein Histogramm im Diagramm-Editorfenster und klickt man auf [A] (alternativ: Doppelklicken auf eine „freie“ Stelle im Diagramm oder die Befehlsfolge „Diagramme“, „Optionen“), so öffnet sich die in Abb. 27.28 dargestellte Dialogbox. Man kann bestimmen, ob eine Normalverteilungskurve, die statistischen Werte Mittelwert und Standardabweichung sowie die Anzahl der Fälle N in der Legende angezeigt werden sollen oder nicht. Außerdem kann eine Gewichtung verwendet werden.

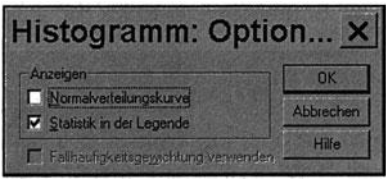


Abb. 27.28. Dialogbox „Histogramm: Optionen“

27.4.3 Gestalten der Achsen von Diagrammen (Menü „Achse“)

Die Veränderung der Achsen eines Diagramms geschieht mit achsenspezifischen Dialogboxen. Der Zugriff auf diese Dialogboxen im Diagramm-Editorfenster kann im allgemeinen in dreierlei Weise erfolgen:

- Doppelklicken auf die Achse, die modifiziert werden soll.
- Markieren der Achse oder des Achsen-Labels und anschließende Befehlsfolge „Diagramme“, „Achse...“.
- Befehlsfolge „Diagramme“, „Achse...“. Es öffnet sich eine Dialogbox zur Auswahl der Achsenart. Man wähle eine der Achsen.

Je nach Art der Achse (Skalenachse, Kategorienachse, Intervallachse) öffnet sich eine bestimmte Dialogbox.

Gestalten der Skalenachse von Balken-, Linien-, Flächen- und Boxplotdiagrammen. In Abb. 27.29 ist links die Dialogbox zur Modifizierung der Skalenachse (im allgemeinen die senkrechte Achse) für das Beispiel eines Balkendiagramms zur Darstellung der Inflationsraten von 1961 bis 1990 zu sehen. Folgende Gestaltungsmöglichkeiten bestehen:

- *Achsenlinie anzeigen.* Man kann die Achsenlinie anzeigen lassen oder ausblenden.
- *Achsentitel.* Man kann einen Achsentitel in das Eingabefeld einfügen oder einen vorhandenen durch einen anderen austauschen (hier: INFLATIONS-RATE).
- *Ausrichtung des Titels.* Zur Auswahl stehen verschiedene Lagen der Titel zur Achse (hier: „Rechts/oben").
- *Skala.* Wählen kann man die voreingestellte lineare Skalierung der Achse oder eine logarithmische zur Basis 10 umstellen. Für Boxplotdiagramme gibt es nicht die logarithmische Achse.
- *Bereich.* Der auf der Skala angezeigte Wertebereich kann festgelegt werden. Zur Information wird der kleinste (1,4) und größte (7,8) Datenwert angegeben. Bei Umstellung auf eine logarithmische Skala werden die Bereichswerte in der gleichen Einheit wie die Datenwerte angezeigt. Das Minimum muss positiv sein.
- *Unterteilung.* Man kann die Achsenunterteilung bestimmen. Der angezeigte Bereich (Differenz zwischen Maximum und Minimum = 10) muss ein Vielfaches des „Inkrementes", des Teilungsabstandes, sein. Des weiteren muss die Unterteilung des 1. Inkrements (hier = 2) ein mehrfaches des 2. (hier = 1) sein. Falls man die Teilungsstriche auf der Achse unterdrücken möchte, kann man „Teilstriche" deaktivieren. „Gitter" erlaubt das Einfügen von senkrecht zur Skalenachse verlaufenden Gitternetzlinien.
- *Basislinie für Balken.* Man kann eine Basislinie festlegen, von der sich die Balken absenken oder erheben (hier: 3 %).
- *Abgeleitete Achse anzeigen.* Man kann die Grafik um eine weitere Skalenachse ergänzen. Durch Aktivschaltung von „Abgeleitete Achse anzeigen" und Klicken auf „Abgeleitete Achse..." öffnet sich eine Dialogbox. In dieser kann man die Gestaltung dieser Achse (ähnlich wie bei der ursprünglichen Skalenachse) festlegen.
- *Labels anzeigen.* Ist diese Option aktiv geschaltet, so kann durch Klicken auf die Schaltfläche „Beschriftungen..." die Dialogbox zur Gestaltung der Labels auf der Skalenachse geöffnet werden. (⇨ Abb. 27.29 rechts). Folgende Gestaltungsmöglichkeiten bestehen:
  - *Dezimalstellen.* Die Anzahl der darzustellenden Dezimalstellen ist hier mit 1 bestimmt worden.
  - *Führendes Zeichen, Abschlusszeichen.* Es ist möglich, ein Zeichen vor oder nach (z.B. ein Währungszeichen) dem Skalenwert einzufügen (hier: % für die Inflationsrate).
  - *1000er-Trennzeichen.* Werte auf der Skalenachse größer 1000 werden mit einem Tausendertrennzeichen (ein Komma oder ein Punkt je nach Einstellung des Windows-Systems) ausgewiesen.
  - *Skalierungsfaktor.* Man kann z.B. durch Angabe eines Skalierungsfaktors 1000 die Skalenwerte 1500, 2600 etc. als 1,5 und 2,6 darstellen und dann im Achsentitel den Zusatz „in Tsd." aufnehmen. Voreingestellt ist der Faktor 1.

- *Orientierung*. Verschiedene Auswahlmöglichkeiten für die Lage der Achsenbeschriftung.

Abb. 27.30 zeigt das Balkendiagramm mit den in Abb. 27.29 gezeigten Gestaltungsfestlegungen.

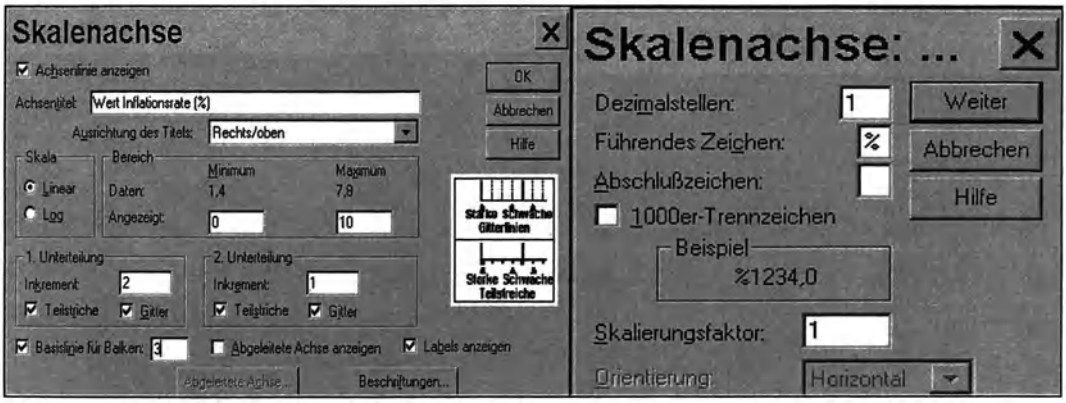


Abb. 27.29. Dialogboxen „Skalenachse“ und „Skalenachse: Labels“

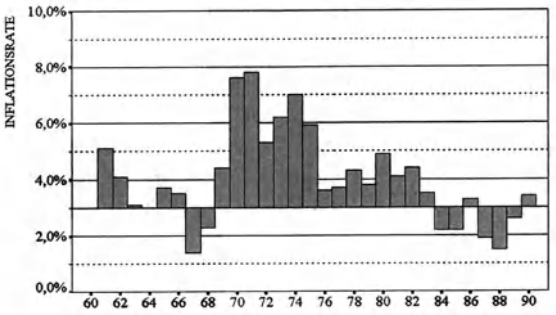


Abb. 27.30. Darstellung der Inflationsraten als hängendes Balkendiagramm

Gestalten der Achsen von Streudiagrammen. Die Dialogbox zur Modifizierung einer Skalenachse kann durch Doppelklicken auf eine Achse oder die Befehlsfolge „Diagramm“, „Achse...“ geöffnet werden. Bei einfachen und überlagerten Streudiagrammen öffnen sich Dialogboxen, die den oben erläuterten Dialogboxen zur Gestaltung von Skalenachsen in Balkendiagrammen etc. gleichen.

Im folgenden wird auf die Gestaltung der Achsen von Matrix-Streudiagrammen eingegangen. Die Öffnung der Dialogbox zur Gestaltung der Achsen kann auch durch Doppelklicken auf eines der Variablen-Labels in den Diagonalfeldern erfolgen. In Abb. 27.31 links ist die Dialogbox zur Modifizierung der Skalenachse abgebildet.

Abb. 27.31. Dialogboxen für eine Scatterplot-Matrix: „Skalenachsen"

Folgende Gestaltungsmöglichkeiten bestehen:

- *Variablen in der Diagonalen anzeigen.* Durch Deaktivierung können die Variablen-Labels in den Diagonalfeldern der Matrix unterdrückt werden.
- *Achsentitel anzeigen.* Bei Aktivierung werden die Variablen-Labels außerhalb der Achsen angezeigt.
- *Horizontal* und *Vertikal.* Man kann separat für die horizontale bzw. vertikale Achse festlegen, ob Achsenlinien, Achsenbeschriftungen (Achsenlabels), Teilstriche sowie Gitternetze erscheinen sollen oder nicht.
- *Bearbeiten einzelner Achsen.* Wird in Abb. 27.31 links eine der Matrixvariablen markiert (z.B. WM1 = Wachstumsrate der Geldmenge M1) und dann auf „Bearbeiten" geklickt, so öffnet sich die in Abb. 27.31 rechts dargestellte Dialogbox zum Modifizieren der gewählten Achse. Folgende Gestaltungsmöglichkeiten bestehen:
  - *Titel.* Die Variablen-Labels in den diagonalen Feldern und für die Beschriftung außerhalb der Achsen können gelöscht und es können dann neue eingegeben werden. Des weiteren kann zwischen alternativen Lagen der Variablen-Labels zu den Achsen gewählt werden.
  - *Skala.* Die Skalierung der Achse kann auf eine logarithmische zur Basis 10 umgestellt werden.
  - *Bereich.* Der auf der Skala angezeigte Wertebereich kann festgelegt werden. Zur Information wird der minimale (-2) und der maximale (14,95) Datenwert der Variablen angezeigt. Wichtig ist, dass der angezeigte Bereich (Minimum bis Maximum) ein ganzzahliges Vielfaches des Inkrementes, des Unterteilungsabstandes auf der Achse, beträgt. Bei logarithmischer Darstellung sind negative Werte nicht zulässig.
  - *Beschriftungen.* Die Optionen in dieser Gruppe können nur dann eingesetzt werden, wenn man für „Horizontal" oder „Vertikal" die Option „Achsenbeschriftung" gewählt hat. Analog zur Wertedarstellung für Skalenachsen kann festgelegt werden mit wieviel Dezimalstellen die Werteunterteilung angezeigt werden sollen, ob ein Zeichen davor (z.B. ein

Währungszeichen) oder danach (z.B. %), ob ein Tausendertrennzeichen oder ein anderer Skalierungsfaktor für die Achsenunterteilung verwendet werden soll. Des weiteren kann zwischen Alternativen zur Lage der Werte auf der horizontalen Achse gewählt werden.

Gestalten von Kategorienachsen. Der Zugang zur Modifizierung der Kategorienachse (im allgemeinen die horizontale Achse) von Balken-, Linien-, Flächen- sowie Boxplotdiagrammen vollzieht sich über die Öffnung der Dialogbox „Kategorienachse“. Diese Dialogbox kann durch Doppelklicken auf die Achse bzw. auf die Labels der Achse, markieren der Achse durch Einfachklick und anschließender Befehlsfolge „Diagramm“, „Achse...“ oder durch die Befehlsfolge „Diagramm“, „Achse...“ mit anschließender Auswahl der Kategorienachse geöffnet werden. In Abb. 27.32 links ist die Dialogbox zur Modifizierung der Skalenachse abgebildet.

Abb. 27.32. Dialogboxen „Kategorienachse“ und „Kategorienachse: Beschriftung“

Folgende Gestaltungsmöglichkeiten gibt es:

- *Achsenlinie anzeigen.* Bei Deaktivierung wird die Achsenlinie unterdrückt. Da die Achsenlinie und der innere Rahmen sich überlagern, muss zur Unterdrückung auch gleichzeitig der innere Rahmen ausgeschaltet sein.
- *Achsentitel.* Der Achsentitel kann verändert werden.
- *Ausrichtung des Titels.* Die Lage des Variablen-Labels auf der Achse kann gewählt werden („Links/unten“, „Mitte“, „Rechts/oben“).
- *Achsenmarkierungen.* Möglich sind „Teilstriche“ und „Gitternetzlinien“.
- *Labels anzeigen bzw. verändern.* Ist „Labels anzeigen“ aktiv, so werden die Labels der Kategorien in der Grafik angezeigt. Nun ist es möglich, die Wertelabels (die Achsenbeschriftung) zu verändern. Durch Doppelklicken auf die Schaltfläche „Beschriftungen...“ öffnet sich die in der Abb. 21.32 rechts dargestellte Dialogbox zur Veränderung der Kategorienlabels.
  Folgende Spezifizierungen sind möglich:
  - *Anzeigen.* Man kann wählen, ob alle oder jedes n-te Kategorienlabel (voreingestellt ist n = 2) in der Grafik angezeigt werden soll. Nicht verwechselt werden darf dieses mit dem Ausschluss von Kategorien aus der Grafik (⇨ Kap. 27.5.1). Für nicht angezeigte Labels kann man die Markierungspunkte (Teilstriche) durch Deaktivieren unterdrücken.

- *Beschriftungstext.* Der Label-Text kann verändert werden. Dazu markiert man dieses Label, z.B. „MITTLERE REIFE", in der in Abb. 27.32 rechts dargestellten Dialogbox. Es erscheint im Feld „Label:". Hier kann man es verändern und muss dieses anschließend durch Klicken auf „Ändern" bestätigen.
- *Orientierung.* Die Ausrichtung der Labels zur Achse kann festgelegt werden.

Gestalten von Intervallachsen. Handelt es sich bei der Grafik um ein Histogramm, so nennt man die Achse auf der die Balken fußen eine Intervallachse. Analog zu anderen Achsen wird durch Mauseinsatz eine Dialogbox zur Gestaltung der Intervallachse geöffnet. In Abb. 27.33 links ist die Dialogbox für das Beispiel eines Histogramms der Variable ALTER aus dem ALLBUS90-Datensatz dargestellt. Folgende Möglichkeiten zur Gestaltung bestehen:

- *Achsenlinie anzeigen.* Durch Deaktivierung kann die Achsenanzeige unterdrückt werden. Wie bei Kategorienachsen muss zur Unterdrückung der Anzeige der innere Rahmen ausgeschaltet sein.
- *Achsentitel.* Das Variablen-Label (z.B. ALTER in Abb. 27.33) kann verändert werden. Aus mehreren Möglichkeiten zur Positionierung des Labels an der Achse kann gewählt werden.
- *Achsenmarkierungen.* Man kann festlegen, ob Achsenunterteilungspunkte und/oder Gitterlinien eingefügt werden sollen oder nicht.
- *Intervalle.* Standardmäßig werden die Breite und Anzahl der Balken (die Intervalle) automatisch festgelegt. Durch Wahl von „Anpassen" und „Definieren" kann die Breite der Intervalle selbst bestimmt werden. Dabei kann man entweder die Anzahl oder die Breite der Intervalle wählen. In „Bereich" kann man durch die Festlegung eines Datenbereichs bei Angabe eines kleinsten und größten Wertes einen Bereichsausschnitt für das Histogramm erzeugen. Es könnte z.B. ein Histogramm nur für die Altersgruppe zwischen 30 und 70 Jahren erstellt werden.
- *Labels anzeigen und verändern.* Ist „Labels anzeigen" aktiv, so wird die Intervallbeschriftung (die Altersklassen im Beispiel) angezeigt. Es ist möglich, diese zu verändern. Durch Klicken auf die Schaltfläche „Beschriftungen..." öffnet sich die in der Abb. 27.33 rechts dargestellte Dialogbox zur Veränderung der Beschriftung. Folgende Spezifizierungen sind möglich:
  - *Anzeigen.* Man kann wählen, ob alle oder jedes n-te Intervall (voreingestellt ist n = 2) eine Beschriftung erhalten soll. Bei n = 2 wird jedes 2. Intervall beschriftet. Für unterdrückte Intervallbeschriftungen können die Unterteilungsmarkierungspunkte (Marker) auf der Achse unterdrückt werden.
  - *Typ.* Man kann bestimmen, ob die Intervalle mit ihrem Mittelpunkt (z.B. 35,5 für das Altersklassenintervall 30-40) oder mit den Grenzen der Intervalle (30-40) beschriftet werden sollen. Analog zu „Werte anzeigen" bei der Gestaltung von Skalenachsen kann auch hier die Anzahl der Dezimalstellen bestimmt werden, ob ein 1000er-Trennzeichen bzw. ein anderer Skalierungsfaktor verwendet werden soll.

- *Orientierung.* Die Ausrichtung der Labels zur Achse kann gewählt werden (nur bei horizontaler Intervallachse).

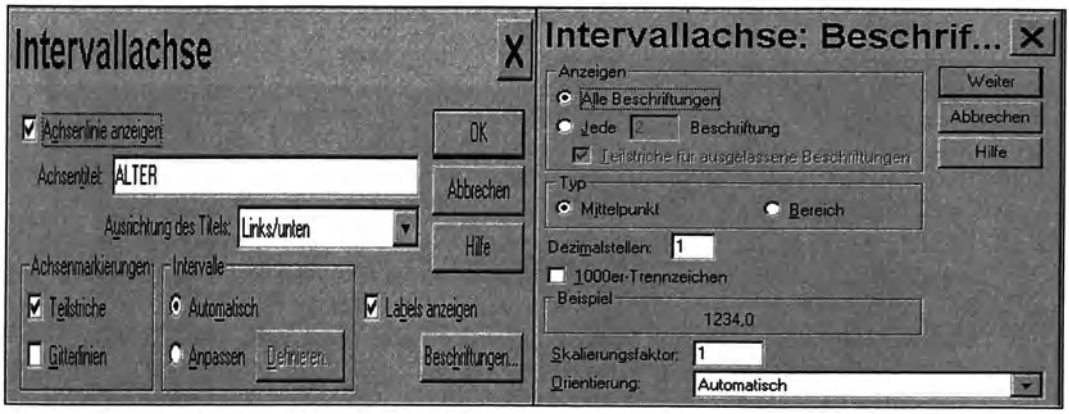


Abb. 27.33. Dialogboxen „Intervallachse“ und „Intervallachse: Labels“

27.4.4 Balkenabstände festlegen (Menü „Balkenabstand“)

Um die Balkenabstände von Balkendiagrammen bzw. Histogrammen zu verändern, wird die Befehlsfolge

▷ „Diagramm“, „Balkenabstand...“

abgesetzt. Es öffnet sich bei einem Balkendiagramm die in Abb. 27.34 dargestellte Dialogbox. Der Balkenrand, der Balkenabstand und der Abstand von Balkengruppen können festgelegt werden.

Bei einem Histogramm öffnet sich eine ähnliche Dialogbox. Es kann aber nur der Abstand der Balken zum Rand verändert werden.

Abb. 27.34. Dialogbox „Balkenabstände“

27.4.5 Titel, Fußnoten, Legenden und Anmerkungen einfügen bzw. verändern

In Diagramme können einerseits erklärende Texte in Form von Titel, Fußnoten, Legenden und/oder Anmerkungen eingefügt werden und andererseits können

schon bei Erzeugung der Grafik definierte erklärende Texte nachträglich verändert werden.

Ist die Grafik schon bei der Erzeugung mit Titel, Untertitel und Fußnoten versorgt worden (⇨ Kap. 26.2.1) oder werden standardmäßig Legenden eingefügt, so kann man durch Doppelklicken auf diese Textstellen Dialogboxen öffnen und in diesen die Texte verändern.

Hat die Grafik noch keine erklärenden Texte, so können diese über das Menü „Diagramme" mit den Untermenüs „Titel...", „Legende...", „Fußnote..." oder „Anmerkung..." nachträglich eingefügt werden (⇨ Abb. 26.4).

Die Dialogbox zur Veränderung von Legenden ermöglicht es, die Anzeige einer Legende auszublenden, die Legendenüberschrift (der Titel) zu verändern sowie ihre Ausrichtung zum inneren Rahmen festzulegen. Zur Veränderung eines Labels der Legende wird dieses in „Beschriftungen:" markiert (z.B. MAENNLICH) und im darunterliegenden Feld „Ausgewählte Beschriftung" verändert. Mit Klicken auf „Ändern" muss dieses bestätigt werden.

Mit dem Untermenü „Anmerkung" können an bestimmten Punkten innerhalb des Diagramms erläuternde Texte eingefügt werden bzw. schon eingefügte Anmerkungen überarbeitet werden. Dafür klickt man die Befehlsfolge

▷ „Diagramme", „Anmerkung..."

Es öffnet sich die in Abb. 27.35 links dargestellte Dialogbox. Ist die Grafik schon mit Anmerkungen versorgt, so führt Doppelklicken auf eine Anmerkung zur Öffnung der Dialogbox.

Die Plazierung von Anmerkungen sei am Beispiel eines Balkendiagramms zur Darstellung der prozentualen Häufigkeiten von Schulabschlüssen demonstriert (Datensatz ALLBUS90). In das Eingabefeld „Text" wurde der erste Anmerkungstext „Schulabschlüsse" eingetippt. Zur Festlegung der Positionierung der Anmerkung innerhalb der Grafik müssen die Koordinaten der Achsen angegeben werden. Im Beispiel wurde die Position auf der Skalenachse mit 40 und auf der Kategorienachse mit FACHHOCHSCHULREIFE angegeben bzw. gewählt. Zur Positionierung auf der Kategorienachse mit der gewählten Kategorie „FACHHOCHSCHULREIFE" wurde „Mitte" gewählt. Die Option „Textrahmen anzeigen" wurde angeklickt. Durch Klicken auf „Hinzufügen." ist der eingegebene Text in das oberhalb liegende Anzeigefeld „Anmerkung(en):" übertragen worden. Anschließend wurde der 2. Anmerkungstext (- alte Bundesländer -) in das Eingabefeld eingetragen, die Positionierung mit 35 auf der Skalen- und FACHHOCHSCHULREIFE auf der Kategorienachse eingegeben bzw. gewählt. Des weiteren wurde wieder die Ausrichtung „Mitte" von FACHHOCHSCHULREIFE gewählt. Auf einen Rahmen wurde verzichtet. Auch diese Angaben mussten durch Klicken auf „Hinzufügen" übertragen werden. In Abb. 27.35 rechts sind die eingefügten Anmerkungen im Diagramm zu sehen.

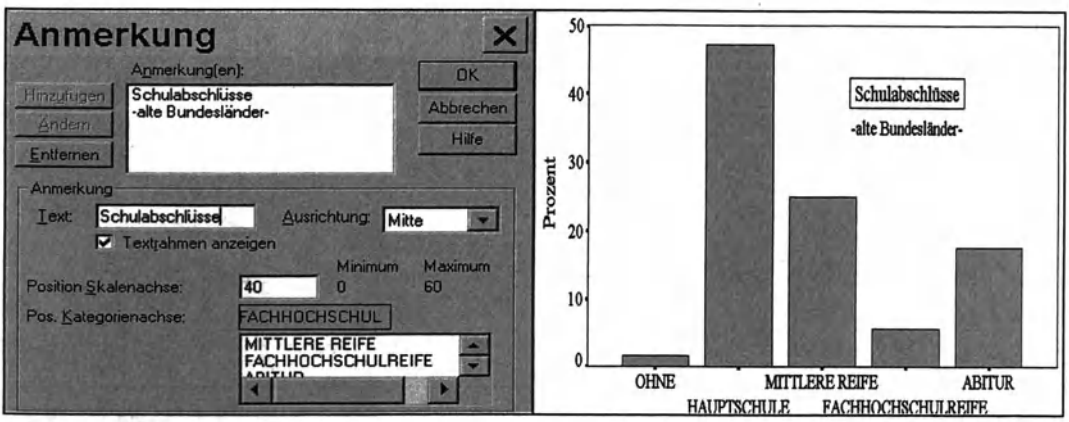


Abb. 27.35. Einfügen von Anmerkungen

27.4.6 Bezugslinien einfügen bzw. verändern (Menü „Bezugslinie“)

In eine Grafik können senkrechte oder waagerechte Bezugslinien eingefügt werden. Die Bezugslinie schneidet also entweder die Kategorien- oder die Skalenachse. Im folgenden Beispiel soll für ein Liniendiagramm zur Darstellung der Wachstumsraten des Bruttosozialprodukts im Zeitraum 1961 bis 1990 (Datensatz MAKRO) die durchschnittliche Wachstumsrate in Höhe von 2,84 % als Bezugslinie eingefügt werden. Dafür wählt man nach Übergabe der erzeugten Grafik in den Diagramm-Editor die Befehlsfolge

▷ „Diagramme“, „Bezugslinie...“

zur Öffnung der Dialogbox „Achse auswählen“. Aus den zur Auswahl stehenden Achsen „Skala“ und „Kategorie“ wird die Skalenachse gewählt. Es öffnet sich dann die in Abb. 27.36 links dargestellte Dialogbox. In das Eingabefeld „Postion der Linie(n):“ wurde 2,84 eingetippt und mit „Hinzufügen“ in das Anzeigefeld übertragen (aus unklaren Gründen verändert sich der Wert auf 2,83999). In Abb. 27.36 rechts ist das Ergebnis zu sehen. Auf diese Weise können weitere Bezugslinien eingefügt werden.

Das Einfügen einer Bezugslinie für die Kategorienachse vollzieht sich in analoger Weise.

27.4.7 Innerer und äußerer Rahmen für Grafiken

Durch Aktivieren bzw. Deaktivieren von „Rahmen innen“ bzw. „Rahmen außen“ kann eine wunschgemäße Rahmengestaltung für die Grafik vorgenommen werden (⇨ Kap. 27.2).

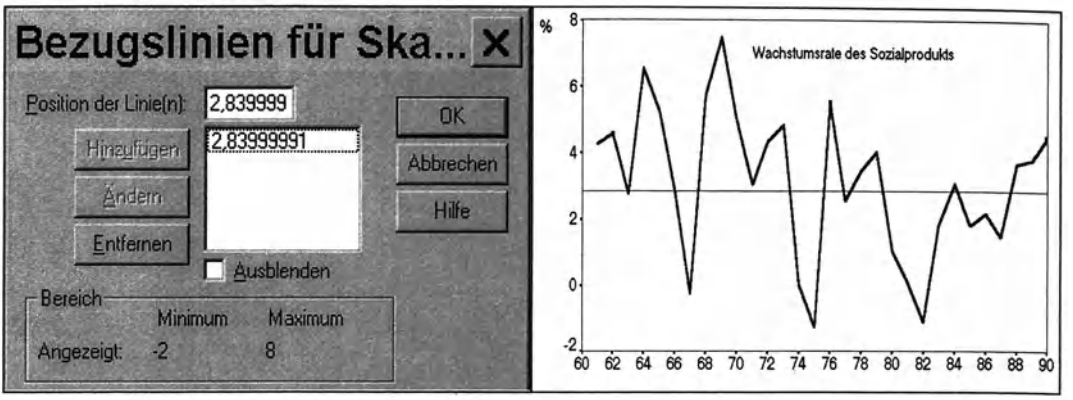


Abb. 27.36. Einfügen einer Bezugslinie für die Skalenachse

27.5 Daten anzeigen und transponieren (Menü „Datenreihen")

27.5.1 Datenreihen anzeigen

In diesem Abschnitt wird auf das Menü „Datenreihen" mit der Option „Angezeigt..." eingegangen. Damit kann man die Auswahl und Zuordnung von Daten für die Grafik steuern. Die Steuerung bezieht sich auf die bei der Erzeugung der Grafik enthaltenen Daten, d.h. es können keine Daten hinzugefügt, sondern nur weggelassen werden. Die im Menü verfügbaren Optionen hängen vom Typ der Grafik ab:

- In Balken-, Linien- und Flächendiagrammen kann man sowohl Datenreihen als auch Kategorien weglassen. Bei der Entfernung von Kategorien in Grafiken mit kumulativen Verteilungen ist zu beachten, dass die Werte verbleibender Kategorien nicht neu berechnet werden, so dass Interpretationsprobleme entstehen.
  Auch die Anordnung der Datenreihen und der Kategorien kann verändert werden. Des weiteren kann man für jede individuelle Datenreihe einer Grafik entscheiden, ob sie als Balken, Linien oder Flächen dargestellt werden soll, so dass gemischte Diagramme entstehen.
- In Kreisdiagrammen können Segmente (Kategorien) weggelassen werden. Sind bei Erzeugung der Grafik mehrere Datenreihen verwendet worden, so kann ausgewählt werden, welche dargestellt werden sollen.
- Für Boxplotdiagramme sind keine Optionen verfügbar.
- In Streudiagrammen kann man die Zuordnung der Datenreihen zu den Achsen verändern. Für die unterschiedlichen Streudiagramme gibt es dafür unterschiedliche Dialogboxen.
- Durch Weglassen von Datenreihen können einzelne Datenreihen aus einem Scatterplot als Histogramm dargestellt werden.

Angezeigte Daten in Balken-, Linien und Flächendiagrammen. Zur Veränderung der Darstellung von Datenreihen und Kategorien eines dieser Diagramme wird die Befehlsfolge

▷ „Datenreihen", „Angezeigt..."

geklickt. Für das Beispiel eines gruppierten Balkendiagramms zur Darstellung der prozentualen Häufigkeitsverteilung der Schulabschlüsse von Männern und Frauen öffnet sich die in Abb. 27.2 dargestellte Dialogbox.

Die Gestaltungsmöglichkeiten beziehen sich auf Datenreihen und auf Kategorien:

- *Datenreihen.* Eine angezeigte Datenreihe (z.B. „MAENNLICH Prozent: Balken") kann mit dem Pfeilschalter in das Listenfeld „Weglassen" verschoben werden. Auch kann man die Reihenfolge der angezeigten Datenreihen verändern, indem man zunächst alle weglässt und dann in der gewünschten Reihenfolge in das Feld „Anzeigen" schiebt.
- *Datenreihen anzeigen als.* Für jede der angezeigten Datenreihen besteht die Möglichkeit, die Form der Darstellung zu verändern: so kann z.B. eine als Balken dargestellte Datenreihe in Linien oder Flächen überführt werden und umgekehrt. Dazu markiert man die Datenreihe und wählt anschließend aus einer der Darstellungsformen „Balken", „Linie" oder „Fläche" aus.
- *Kategorien.* Kategorien können durch Verschieben einer Kategorie aus dem Feld „Anzeigen" in das Feld „Weglassen" aus der Grafik entfernt werden (⇨ Beispiel in Kap. **27**.2). Wie bei den Datenreihen kann auch die Reihenfolge der Kategorien auf der Achse neu geordnet werden.

Angezeigte Daten in Kreisdiagrammen. Zur Modifikation der Darstellung von Datenreihen bzw. Kategorien eines Kreisdiagramms (⇨ Abb. 26.14) wird mittels der Befehlsfolge

▷ „Datenreihen", „Angezeigt..."

eine Dialogbox geöffnet, in der man analog zu der in Abb. 27.2 Datenreihen und/oder Kategorien weglassen oder anzeigen (d.h. darstellen) lassen kann. Wurde das Diagramm als einfaches Kreisdiagramm erzeugt, so wird natürlich nur eine Datenreihe angezeigt. Nur wenn man über das Menü „Galerie" von einem Mehrfachreihendiagramm zu einem Kreisdiagramm für einzelne Datenreihen übergehen möchte, wird ein Weglassen von Datenreihen relevant.

Angezeigte Daten in Streudiagrammen. Auch für Streudiagramme kann man die angezeigten und damit dargestellten Datenreihen über die Menüfolge „Datenreihen", „Angezeigt..." verändern. Je nach Art des Streudiagramms öffnet sich eine dazugehörige Dialogbox. Es können einerseits die Zuordnung der Variablen zu den Achsen vertauscht werden und andererseits bei Übergängen zu anderen Streudiagrammarten bzw. bei Matrix-Streudiagrammen Variablen weggelassen werden.

Angezeigte Daten in Histogrammen. Über die Menüfolge „Datenreihen", „Angezeigt..." kann man ausgehend von Streudiagrammen, einzelne Datenreihen als Histogramme darstellen.

27.5.2 Daten transponieren

Daten zu transponieren bedeutet für Balken-, Linien- oder Flächendiagrammen mit mehreren Datenreihen, die Rollen von Kategorien und Datenreihen zu vertauschen. Anhand des Beispiels zur Darstellung der Schulabschlüsse von Männern und Frauen in Form eines gruppierten Balkendiagramms sei dieses erklärt. In Abb. 27.37 wird links das Ausgangsdiagramm und rechts die nach der Datentransponierung entstandene Grafik dargestellt. Die Grafik wurde durch die Befehlsfolge

▷ „Datenreihen", „Daten transponieren..."

transponiert. Die Kategorien des Schulabschlusses sind zu Datenreihen und die Datenreihen Männer und Frauen sind zu Kategorien geworden.

27.6 Layoutmerkmale von Grafikobjekten modifizieren

Die Elemente einer Grafik (Grafikobjekte) können in Art und Stil verändert und so für Präsentationszwecke aufbereitet werden:

- ❑ Balken, Linien und Flächen können eine andere Farbe erhalten.
- ❑ Linien - seien es Datenlinien, Achsen oder Rahmen - können in Stil und Dicke verändert werden.
- ❑ Flächen können verschiedene Füllmuster erhalten.
- ❑ Balken können schattiert oder mit 3D-Effekt dargestellt sowie mit Werte-Labels versehen werden.
- ❑ Datenlinien können auf verschiedene Art interpoliert werden.
- ❑ Kreissegmente in Kreisdiagrammen können abgesetzt werden.
- ❑ Die Achsen von zweidimensionalen Diagrammen können vertauscht, und die von dreidimensionalen gedreht werden.

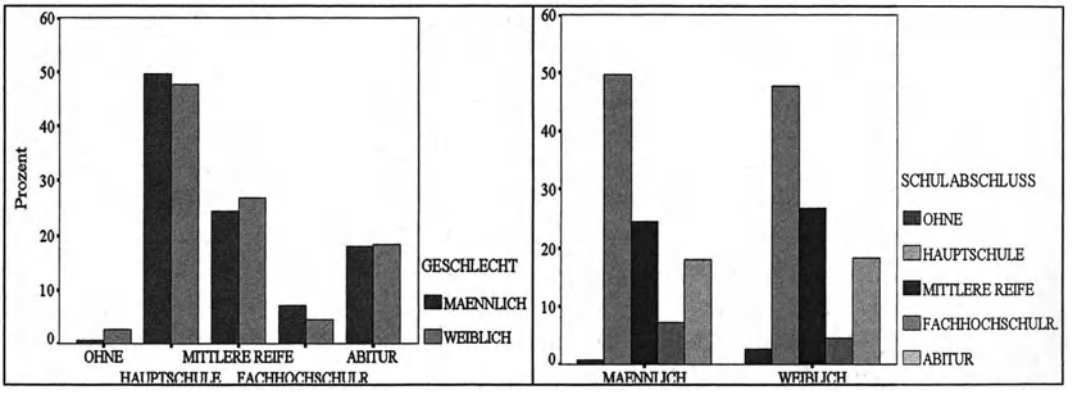


Abb. 27.37. Ausgangs- und transponiertes gruppiertes Balkendiagramm

Die Vergabe von Layoutmerkmalen eines Grafikobjekts geschieht dabei wie folgt:

- ❑ Das Grafikobjekt im Grafikfenster, z.B. eine Balkenreihe, wird durch Einfachklick markiert.

- Durch anschließendes Klicken auf eines der Formatierungs-Symbole des Diagramm-Editorfensters (z.B.  = Füllmuster) wird eine Palette bzw. Dialogbox geöffnet. Alternativ kann die Palette oder Dialogbox auch über das Menü „Format“ (vormals „Grafikattribute“) geöffnet werden.
- Aus den in diesen enthaltenen Mustern oder Stilen wird eines(r) gewählt und anschließend mit Klicken auf „Zuweisen“ oder „Allen zuweisen“ entweder auf die markierte Datenreihe bzw. das Objekt oder auf alle Reihen übertragen. Mit „Schließen“ wird die Palette bzw. Dialogbox geschlossen. Es ist möglich, mehrere Paletten gleichzeitig zu öffnen und diese auf dem Bildschirm zu verschieben.

Flächen mit Füllmuster versehen. Um z.B. die in Abb. 27.1 dargestellten Balkengruppen in der Grafik zur Abbildung der prozentualen Häufigkeiten von Schulabschlüssen von Männern und Frauen mit Füllmustern zu versehen, wird zunächst die linke Balkenreihe (Datenreihe Schulabschlüsse der Männer) durch einfaches Klicken auf einen Balken markiert. Die Markierung wird durch schwarze Markierungspunkte an den Ecken der Balken angezeigt (⇨ Abb. 27.38 links). Durch anschließendes Klicken auf das Symbol für Füllmuster

 (oder über Menü: „Format“, „Füllmuster...“)

öffnet sich die in Abb. 27.38 rechts dargestellte Palette mit Füllmusterarten. Durch Auswahl eines Musters und Klicken auf die Schaltfläche „Zuweisen“ wird das Füllmuster auf die markierten Balken übertragen. Danach kann die Balkenreihe der Schulabschlüsse von Frauen markiert und ein Füllmuster zugewiesen werden.

Es ist auch möglich, dem Bereich zwischen innerem und äußerem Rahmen einer Grafik ein Füllmuster zuzuweisen. Dafür muss diesem Bereich aber vorher eine Farbe zugewiesen worden sein. Der Bereich wird durch Klicken auf eine Stelle des Bereichs markiert. Die Markierung wird durch Markierungspunkte an den Ec??ken des äußeren Rahmens angezeigt.

Werden die Balken schattiert oder im 3D-Effekt dargestellt, so kann jede Fläche des Diagramms mit Füllmustern versorgt werden. In Abb. 27.39 wird das gruppierte Balkendiagramm mit Füllmustern für die Balken und für den Bereich zwischen innerem und äußerem Rahmen dargestellt.

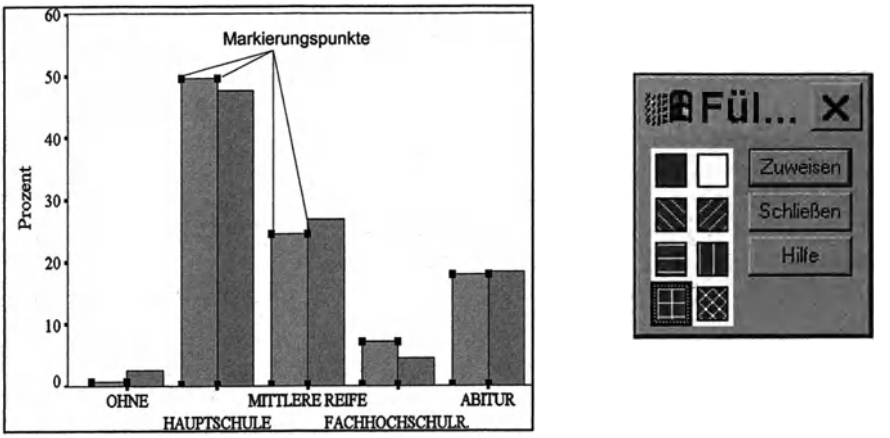


Abb. 27.38. Balken mit Füllmustern versehen

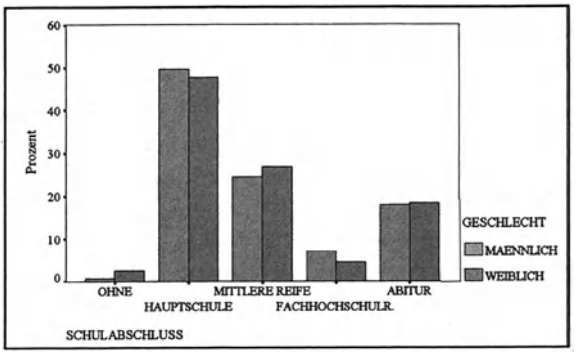


Abb. 27.39. Füllmuster für Balken und zwischen innerem und äußerem Rahmen

Grafikobjekten Farben zuweisen. Die Farben von Grafikobjekten wie Flächen, Daten- oder Achsenlinien, Markierungen in Streudiagramme und Texte können verändert werden. Dieses geschieht, indem das Objekt durch Anklicken markiert und anschließend eine Palette mit Auswahlfarben geöffnet wird. Die Öffnung der Palette erfolgt durch Klicken auf das Symbol für Farben

 (oder über Menü: „Format", „Farbe...").

In Abb. 27.40 ist die Palette „Farben" dargestellt. Man kann je nach Wahl die Farbe für eine Fläche (oder von Linien bzw. Text) oder für den Rahmen (die Umrandung) einer Fläche durch Farbauswahl verändern. Nach Klicken von „Zuweisen" erhält das markierte Objekt die gewählte Farbe.

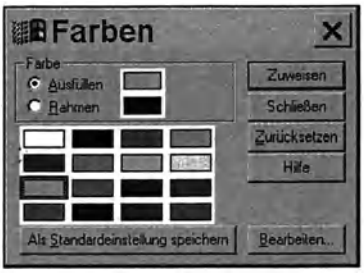


Abb. 27.40. Palette „Farben"

Wenn eine oder mehrere Farben der Farbpalette per „Bearbeiten..." verändert aber noch nicht gespeichert worden ist, dann führt ein Klicken auf „Zurücksetzen" zur Wiederherstellung der voreingestellten Farbpalette sowie zur voreingestellten Farbgebung bei Grafikerzeugung.

Folgende weitere Wahlmöglichkeiten bestehen:

- *Als Standardeinstellung speichern.* Die Farben der Palette können per „Bearbeiten" verändert und dann gespeichert werden.
- *Bearbeiten.* Zum Verändern der Farben der Farbpalette wählt man eine Farbe der Farbpalette und klickt auf „Bearbeiten...". Es öffnet sich die Dialogbox „Farben bearbeiten". Die in der Farbpalette gewählte Farbe ist hervorgehoben. Wählt man nun eine andere Farbe, so wird die alte Farbe durch die neue ersetzt. Ist in der Grafik die alte Farbe für ein Objekt verwendet worden, so verändert sich auch diese Farbe. Mit Klicken von „Farben definieren" können auch voreingestellte Farben der Farbpalette durch selbstdefinierte ersetzt werden.

Größe und Stil von Markierungen verändern. Markierungen dienen zur Kennzeichnung von Datenpunkten in Linien-, Flächen- und Streudiagrammen. In Abb. 27.41 wird dieses am Beispiel eines einfachen Streudiagramms gezeigt: Variable EINK (Nettoeinkommen) und ARBSTD (Arbeitsstunde/Woche) mit GESCHL als Markierungsvariable). Zur Veränderung der Markierung der Datenpunkte für Männer werden diese durch Anklicken ausgewählt (sichtbar durch schwarze Kästchen) und anschließend die Palette „Marker" geöffnet. Die Öffnung der Palette erfolgt durch Klicken auf das Symbol für Marker

 (oder über Menü: „Format", „Marker...").

Auf der linken Seite der Abb. 27.41 ist die Palette „Marker" zu sehen. Zur Veränderung des Markierungsstils und der Markierungsgröße für die gewählten Datenpunkte wurde ein Kästchen von kleiner Größe bestimmt. Mit „Zuweisen" wird diese Markierung in die Grafik übertragen. Anschließend wurde die zweite Datenpunktwolke angeklickt und ein Kreis von kleiner Größe als Markierungsstil gewählt und zugewiesen.

Wird „Allen zuw." angeklickt, so wird die gewählte Markierungsart auf alle Datenpunkte angewendet.

Erscheinen in einem Liniendiagramm die Markierungen der Datenlinien nicht, so können diese in der Palette „Linien-Interpolation“ durch Wahl von „Gerade“ und „Markierungen anzeigen“ angefordert werden.

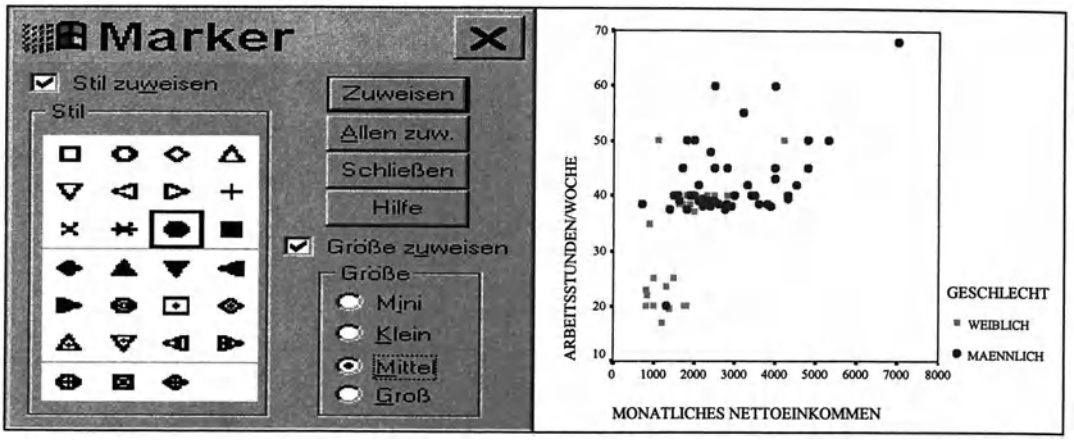


Abb. 27.41. Unterschiedliche Markierungsstile in einem Streudiagramm

Linienarten verändern. In Abb. 27. 42 ist rechts ein Mehrfachliniendiagramm zur Darstellung des Zinssatzes und der Inflationsrate von 1960 bis 1990 dargestellt. Nach Markierung einer Linie durch Anklicken wurde die Linienart der Datenreihe durch Auswahl von „Stil“ und „Dicke“ in der Palette „Linienstil“ verändert. Die Palette öffnet sich durch Klicken auf das Symbol für Linienstil (oder über Menü: „Format“, „Linienstil...“).

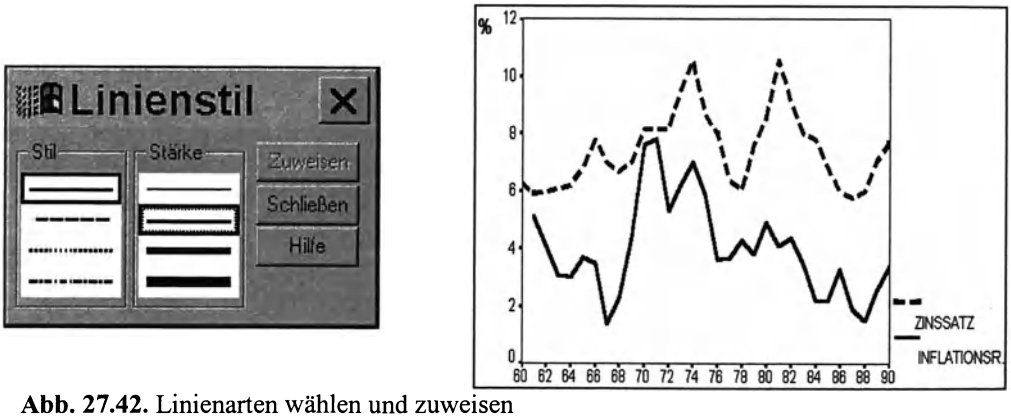


Abb. 27.42. Linienarten wählen und zuweisen

Schattierungen und 3D-Effekt für Balkendiagramme. In Abb. 27.43 rechts ist ein gruppiertes Balkendiagramm zur Darstellung der Schulabschlüsse von Männern und Frauen (Datensatz ALLBUS90) mit 3D-Effekt dargestellt. Die Palette zur Erzielung von Balkenschattierungen bzw. eines 3D-Effektes (⇨ Abb. 27.43

links) öffnet sich durch Klicken auf das Symbol für Balkenarten (oder über Menü: „Format", „Balkenart..").

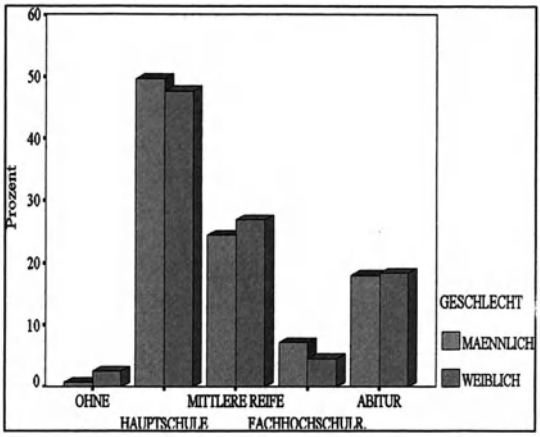


Abb. 27.43. Balkenarten wählen und allen zuweisen

Balken mit Werten beschriften. Das in Abb. 27.43 dargestellte Balkendiagramm ist in Abb. 27.44 mit Werten beschriftet. Die Palette zur Balkenbeschriftung (⇨ Abb. 27.44 links) öffnet sich durch Klicken auf das Symbol für die Balkenbeschriftung (oder über Menü: „Format", „Balkenbeschriftung..."). Außer „Rahmen" kann auch „Standard" gewählt werden.

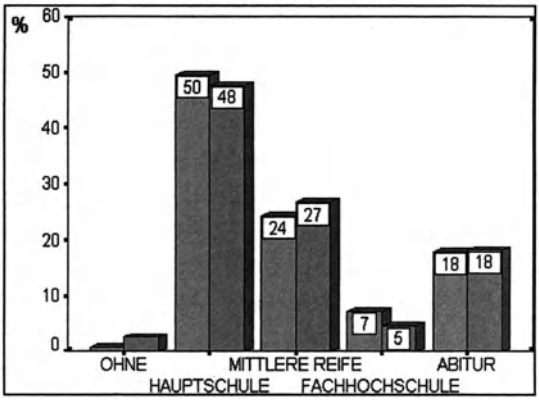


Abb. 27.44. Balken mit Werten beschriften

Datenpunkte in Linien- bzw. Streudiagrammen mit Linien verbinden. In Abb. 27.45 rechts ist in einem Liniendiagramm die Entwicklung der Inflationsrate und des Zinssatzes von 1961 bis 1990 abgebildet. Zur Verbindung der Datenpunkte sind unterschiedliche Interpolationsarten angewendet worden: für den Zinssatz eine Treppenkurve und für die Inflationsrate eine Lagrange-Interpolation dritter Ordnung bei der ein Polynom dritten Grades an die jeweils vier am nächsten

gelegenen Punkte angepasst wird. Um eine Interpolationsart auf eine Datenlinie anzuwenden, wird diese durch Anklicken markiert und anschließend wird durch Klicken auf das Symbol [~] (oder über Menü „Format", „Interpolation...") die in Abb. 27.45 links dargestellte Palette „Geradeninterpolation" geöffnet. Aus dieser kann eine Interpolationsart zur Linienverbindung ausgewählt und mit Klicken auf „Zuweisen" angewendet werden. Dabei ist es auch möglich, eine gewählte Interpolationsart durch Klicken auf „Allen zuw." auf alle Datenlinien anzuwenden.
Optional kann mit Aktivieren bzw. Deaktivieren von „Markierungen anzeigen" die Anzeige von Datenpunkten durch Markierungszeichen angefordert bzw. unterlassen werden.

Für Streudiagramme stehen neben diesen Typen im Menü „Optionen" weitere Formen von Verbindungslinien zur Verfügung.

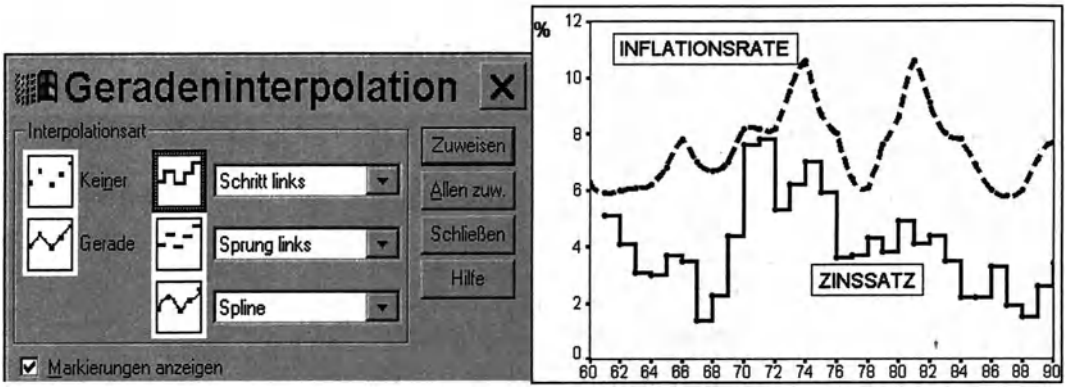


Abb. 27.45. Verbindungslinien für Datenpunkte in Liniendiagrammen

Textelemente einer Grafik in Schriftart und Schriftgröße verändern. Zur Veränderung der Schriftart und/oder Schriftgröße eines Textelementes in der Grafik (z.B. Achsenbeschriftung, Titel, Legende etc.) wird der Text durch Anklicken markiert. Durch anschließendes Klicken auf das Symbol [T] (oder über Menü: „Format", „Text...") öffnet sich die in Abb. 27.46 dargestellte Palette „Text". Es kann dann eine Schriftart sowie die Schriftgröße ausgewählt und mit „Zuweisen" auf den ausgewählten Text übertragen werden.

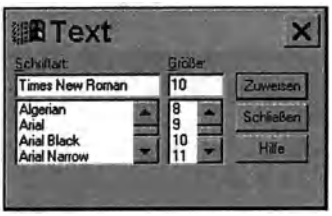


Abb. 27.46. Palette „Text"

Rotation eines 3D-Streudiagramms. Ein 3D-Streudiagramm kann in sechs Richtungen gedreht werden. Hat man ein 3D-Streudiagramm im Diagramm-Editor, so öffnet Klicken auf das Symbol [ ] (oder über Menü: „Format", „3D-Rota-tion...") die in Abb. 27.47 dargestellte Dialogbox „3D-Rotation". Durch Klicken auf einen der sechs Schalter mit der durch Pfeilrichtung gekennzeichneten Drehrichtung um eine Achse wird die Drehung in der Mitte der Dialogbox angezeigt. Ist die gewünschte Drehrichtung erreicht, so wird sie mit Klicken auf „Zuweisen" auf das Diagramm angewendet.

Mit Wahl der Option „Dreifuß einblenden" wird ein Dreifuß eingeblendet, dessen Linien parallel zu den drei Achsen verlaufen und sich im Zentrum des Rahmens schneiden.

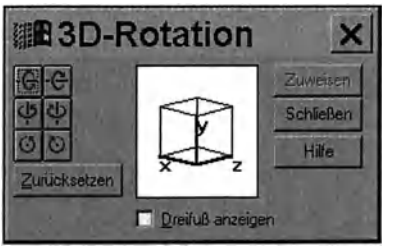


Abb. 27.47. Dialogbox „3D-Rotation"

Eine zweite Möglichkeit zur Drehung des 3D-Diagramms bietet das Klicken auf das Symbol [ ] (oder über Menü: „Format", „Dreh-Modus").

Es erscheint eine Symbolleiste mit den gleichen Rotationsschaltern wie in der Dialogbox „3D-Rotation": [ ]. In diesem Drehmodus wird die Grafik stark vereinfacht: es werden zur Zeit der Drehung nicht die Achsen, sondern nur der Dreifuß und die Punktwolke angezeigt. Der Vorteil ist, dass sich im Drehmodus die Punktwolke mitdreht. Hat man die gewünschte Drehung erreicht, so führt erneutes Klicken auf [ ] („Ende" in Vers. 6.0) zum Schließen des Drehmodus. Nun gewinnt die Grafik wieder die übliche Gestalt, aber in gedrehter Richtung. Mit [ ] wird die gedrehte Grafik wieder in die Ausgangslage zurückgesetzt.

Achsen in zweidimensionalen Diagrammen vertauschen. In Histogrammen, Balken-, Linien-, Flächen-, gemischten oder Boxplot-Diagrammen können durch Klicken auf [ ] (oder über Menü: „Format", „Achsen vertauschen") die Achsen vertauscht werden. In Abb. 27.48 wird das Vertauschen der Achsen am Beispiel eines gruppierten Balkendiagramms gezeigt.

Das Vertauschen von Achsen ist nicht mit dem Transponieren von Datenreihen zu verwechseln. Beim Transponieren werden die in der Legende angezeigten Daten zu Kategorien auf der Kategorienachse und die Kategorien zu Daten in der Legende (⇨ Kap. 27.5.2).

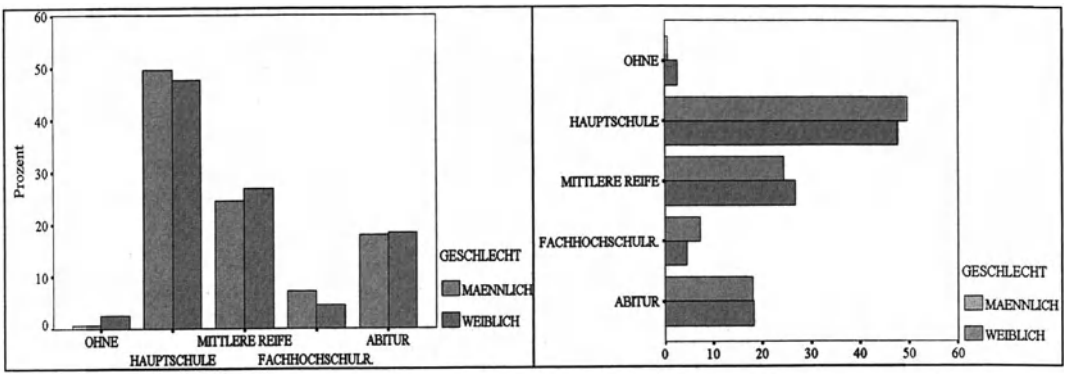


Abb. 27.48. Vertauschen der Achsen eines Balkendiagramms

Soll in einem Streudiagramm die Zuordnung der Variablen zu den Achsen vertauscht werden, so kann über das Menü „Datenreihen", „Angezeigt" die Zuordnung der Variablen zu den Achsen bestimmt werden.

Segmente in Kreisdiagrammen absetzen. Ein Segment eines Kreisdiagramms kann zur Hervorhebung abgesetzt werden. Dazu wird erst das abzusetzende Segment angeklickt und damit markiert. Dann klickt man auf den Schalter . In Abb. 27.49 ist ein Kreisdiagramm mit abgesetztem Segment dargestellt.

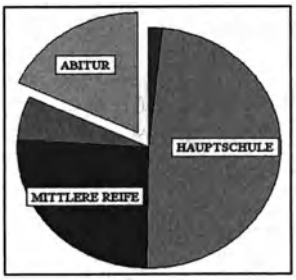


Abb. 27.49. Kreisdiagramm mit abgesetztem Segment

Darstellen fehlender Werte in Liniendiagrammen. In Liniendiagrammen können bei Einschluss fehlender Werte die Linien zwischen fehlenden Werten verbunden oder unterbrochen dargestellt werden.

Klicken auf den Schalter verbindet und auf den Schalter unterbricht die Linien bei fehlenden Werten. Alternativ kann über das Menü „Format" mit „Linie bei fehlendem Wert unterbrechen" eine Linienverbindung ein- oder ausgeschaltet werden.

28 Verschiedenes

28.1 Drucken

Aus SPSS heraus ist es möglich, Inhalte von Ausgabefenstern, Syntaxfenstern, des Datenfensters, des Skriptfensters und von Grafikfenstern direkt auszudrucken. Auch der Inhalt von Hilfefenstern kann gedruckt werden. Gedruckt wird immer die Datei des aktiven Fensters.

SPSS für Windows bedient sich dabei der Druckerinstallationen von Windows. Deshalb muss zunächst unter Windows mindestens ein Drucker installiert sein. (Informieren Sie sich hierüber gegebenenfalls im Windows-Handbuch.) In den meisten Fällen wird man einen Drucker als Standarddrucker und einige weitere in Windows installieren und einrichten. Der Druckvorgang wird jeweils durch die Befehlsfolge „Datei", „Drucken" oder Anklicken des Drucksymbols gestartet. Danach vollzieht sich der Ablauf in den verschiedenen Fenstern etwas unterschiedlich. Im Skriptfenster wird der Druckbefehl ohne weitere Einstellungsmöglichkeiten direkt ausgeführt. Beim Drucken aus den anderen Fenstern erscheint eine Dialogbox, in der Sie den Drucker auswählen. Weiter kann die Zahl der ausgedruckten Exemplare bestimmt werden. Außerdem kann man festlegen, welcher Teil des Fensters ausgedruckt werden sollen. Im „Ausgabefenster stehen dazu die Optionsschalter „Alle angezeigten Ausgaben" (zum Drucken der gesamten Datei, sofern Teile davon nicht ausgeblendet sind) und „Auswahl" (nur markierter Output wird gedruckt) zur Verfügung. Im Daten-Editor und Skriptfenster kann man dagegen zwischen „Alle" (der gesamte Inhalt wird gedruckt), „Seiten" (nur der durch die Anfangs- und Endseite bestimmte Bereich wird ausgedruckt) und „Markierung" wählen. Schließlich ist es möglich, die Ausgabe in eine Datei umzuleiten. Auch die Sortierung kann für mehrseitige Ausdrucke bestimmt werden. Die Dialogbox zum Drucken im Hilfefenster unterscheidet sich noch etwas von den anderen. Insbesondere kann die aktuelle Seite zum Druck gewählt werden. Abb. 28.1 zeigt die Dialogbox „Drucken" des Ausgabefensters. Dort können Sie, wie auch im Syntaxfenster oder Daten-Editor, über eine Dialogbox, die sich beim Anklicken von „Eigenschaften" öffnet den Drucker einrichten. Falls Sie mit dem Standarddrucker und dessen Standardeinstellung arbeiten wollen, klicken Sie auf „OK". Ansonsten nehmen Sie erst die Druckereinrichtung vor.

Druckereinrichtung. Um einen Drucker einzurichten, gehen Sie wie folgt vor:

- Klicken sie auf den Pfeil neben dem Auswahlfeld „Name:". Es öffnet sich eine Liste der installierten Drucker. Dort wählen Sie zunächst den gewünschten Drucker aus.

▷ Falls Sie nicht mit dessen Standardeinstellung arbeiten wollen, klicken Sie auf „Eigenschaften". Es öffnet sich eine Dialogbox, die je nach Drucker unterschiedlich aussieht.

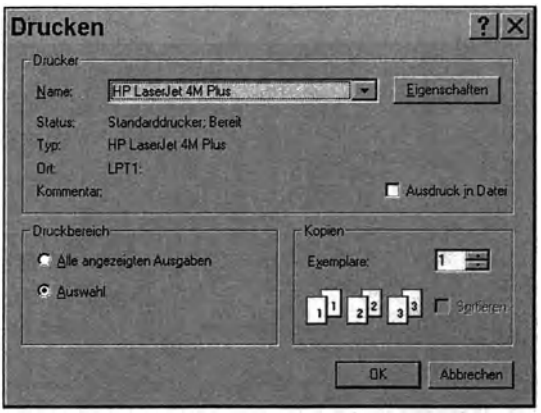


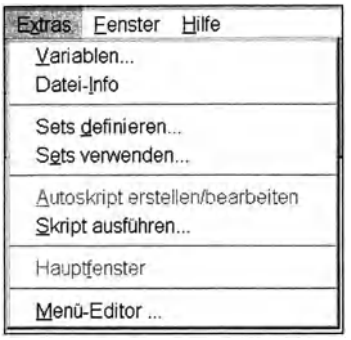
Abb. 28.1. Dialogbox „Drucken" im Ausgabefenster

Je nach Drucker können Sie z.B. im Register „Papier" Papiergröße und Format des zu bedruckenden Papiers festlegen, evtl. auch die Papierzufuhr, im Register „Grafik" Eigenschaften wie Grafikauflösung, Farbmischung. Das Register „Schriftarten" ermöglicht es u.U., weitere Schriftarten zu laden, und im Register „Geräteoptionen" werden Eigenschaften wie Druckdichte, Druckqualität, Speicherbelegung oder Bildsteuerung geregelt. (Ziehen Sie hier das Handbuch Ihres Druckers zu Rate.)

28.2 Das Menü „Extras"

Im Menü „Extras" bietet SPSS eine Reihe (in den verschiedenen Fenstern leicht divergierende) Arbeitshilfen an. Der unten dargestellte Bildschirmausschnitt zeigt die Optionen, wie sie im Menü „Extras" des „Viewers erscheinen. Die Optionen „Autoskript erstellen/bearbeiten" und „Hauptfenster" sind in andren Fenstern nicht verfügbar.

Die Befehle sind, durch Querstriche getrennt, in vier bzw. fünf Gruppen unterteilt. Die zwei Optionen der ersten Gruppe dienen beide dem Aufruf von Informationen über die Variablenstruktur der Datendatei.



Variablen. Öffnet eine Dialogbox. (Dasselbe bewirkt das Anklicken von in der Symbolleiste.) In dieser ist links die Liste aller Variablen des Datensatzes enthalten. Diese kann man in der üblichen Weise durchblättern. In der Gruppe „Variablenbeschreibung:“ werden Namen, Variablen-Label, Variablentyp, Werte, Werte-Label und Missing-Werte der jeweils markierten Variablen angezeigt. Hat man die gewünschte Variable markiert, gelangt man durch Anklicken der Schaltfläche „Gehe zu“ im Datenfenster direkt mit dem Cursor zu der gewünschten Variablen. Durch Anklicken von „Einfügen“ übertragen Sie den Variablennamen der markierten Variablen in das Syntaxfenster. Beides ist insbesondere bei der Arbeit mit langen Variablenlisten nützlich.

Abb. 28.2. Dialogbox „Variablen“

Datei-Info. Beim Anklicken dieser Option wird im „Ausgabefenster“ unter der Überschrift „File Information“ eine vollständige Liste der Variablen mit den dazugehörigen Informationen wie Name, Variablen-Label, Werte, Werte-Labels, Druck- und Schreibformat und fehlende Werte ausgegeben. Tabelle 28.1 zeigt einen Auszug aus der Liste für die Datei ALLBUS90.SAV. Häufig wird es nützlich sein, diese Informationen auszudrucken.

Tabelle 28.1. Datei-Information

```
List of variables on the working file

Name                                                        Position

POL       POLITISCHES INTERESSE, BEFR. <ORDINAL>                   3
          Print Format: F1
          Write Format: F1
          Missing Values: 9

          Value    Label

              1    SEHR STARK
              2    STARK
              3    MITTEL
              4    WENIG
              5    UEBERHAUPT NICHT
              7    VERWEIGERT
              8    WEISS NICHT
              9 M  KEINE ANGABE
```

Die Datei-Info wird bei größeren Datensätzen zu einer großen Textdatei führen, die in der normalen Ausgabe nicht vollständig zu sehen ist. Vollständig zugänglich wird diese erst durch Doppelklicken auf das Ausgabeobjekt. Je nachdem, was Sie im Menü „Optionen" im Register „Pivot-Tabellen" im Feld „Standardbearbeitungsmodus" eingestellt haben (⇨ unten), passiert dann etwas Unterschiedliches. Ist eine der Varianten eingestellt, nach der die Tabelle im „Viewer" bearbeitet werden soll, ändert sich auf den ersten Blick nicht viel. Aber die Tabelle wird um eine Bildlaufleiste ergänzt (evtl. müssen Sie diese durch Verschieben im Viewer erst sichtbar machen), mit der Sie die Tabelle durchscrollen können. Ist eingestellt, dass solche Tabellen in einem „eigenen Fenster" bearbeitet werden, öffnet sich ein Fenster, das die ausgewählte Tabelle samt Bildlaufleiste enthält.[1]

(Variablen-)Sets definieren und verwenden. Die beiden Optionen „Sets definieren" und „Sets verwenden" des Menüs „Extras" erleichtert den Umgang mit langen Variablenlisten. Man kann damit erreichen, dass in der Liste der Quellvariablen nur eine durch den Set definierte Auswahl aller Variablen angezeigt wird. Man wird damit übersichtliche Variablenlisten mit den Variablen zusammenstellen, die man für die jeweils anstehenden Analysen benötigt.

Beispiel: Der ALLBUS von 1990 weist im Original 559 Variablen auf. Sie wollen aber nur eine Untersuchung über die Einkommensverteilung vornehmen. Dazu benötigen Sie neben dem Einkommen noch einige Sozialdaten wie Alter, Geschlecht, Schulabschluss. Um diese immer im Auswahlfeld schnell parat zu haben, stellen Sie sie zu einem Set EINKOMMEN zusammen. (Vorteil dieses Verfahrens ist es, dass alle Variablen verfügbar bleiben. Das wäre nicht der Fall,

1 In älteren Versionen bis 7.5 erscheint dagegen eine Dialogbox „Großes Objekt: Pivot-Tabelle oder Text", in der zwischen diesen Bearbeitungsmodi gewählt werden muss.

wenn Sie eine neue Datei erstellen würden, in der nur die interessierenden Variablen vorhanden sind. Nachteil ist allerdings, dass zusätzlicher Speicherplatz benötigt wird.)

Set definieren. Um einen Set verwenden zu können, müssen Sie ihn zunächst definieren. Dazu gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge „Extras“ und „Sets definieren...“. Die Dialogbox „Variablen-Sets definieren“ öffnet sich (⇨ Abb. 28.3).
- ▷ Tragen Sie in das Feld „Name des Sets:“ einen selbst gewählten Namen ein.
- ▷ Übertragen Sie die gewünschten Variablen aus der Auswahlliste in die Liste „Variablen im Set:“.
- ▷ Klicken Sie auf die Schaltfläche „Set hinzufügen“. Der Name des Sets wird in das Feld unter dem Eingabefeld verschoben, der Set ist definiert.

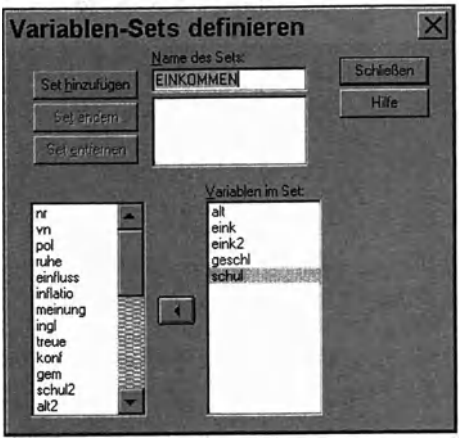


Abb. 28.3. Dialogbox „Variablen-Sets definieren“

Sie können anschließend weitere Sets definieren. Sind alle Sets definiert, schließen Sie die Dialogbox.

Sie können Sets entfernen, indem Sie in der Dialogbox „Variablen-Sets definieren“ den Namen dieses Sets markieren und auf die Schaltfläche „Set entfernen“ klicken. Sie können Namen und Variablenliste eines Sets ändern, nachdem Sie den Namen markiert, mindestens eine Variable hinzugefügt oder entfernt und evtl. den Namen im Eingabefeld geändert haben. Klicken Sie dann auf die Schaltfläche „Set ändern“.

Sets verwenden. Sie können nun die Sets verwenden.

- ▷ Wählen Sie die Befehlsfolge „Extras“ und „Sets verwenden...“, oder klicken Sie auf Es öffnet sich die Dialogbox „Sets verwenden“ (⇨ Abb. 28.4).

Dieses hat zwei Felder. Im linken stehen die Name der Sets, die nicht in Verwendung sind, im rechten diejenigen der Sets, die in Verwendung sind. Übertragen Sie jeweils die Sets, die Sie verwenden wollen, in das Feld „Verwendete Sets:“, alle

anderen in das linke Feld. Bestätigen Sie mit „OK". (Zwei spezielle Sets sind außer den nutzerdefinierten bereits vorhanden und zunächst in Verwendung, ALLVARIABLES und NEWVARIABLES. Bei ALLVARIABLES handelt es sich um einen speziellen Set, der sämtliche Variablen Ihrer aktiven Datendatei enthält. In NEWVARIABLES sind dagegen sämtliche Variablen enthalten, die Sie nach dem Öffnen ihrer aktiven Datendatei hinzugefügt haben.) Nachdem Sie bestimmt haben, welche Sets in Verwendung sind, werden im weiteren nur noch die in diesen Sets definierten Variablen in der Quellvariablenliste angezeigt.

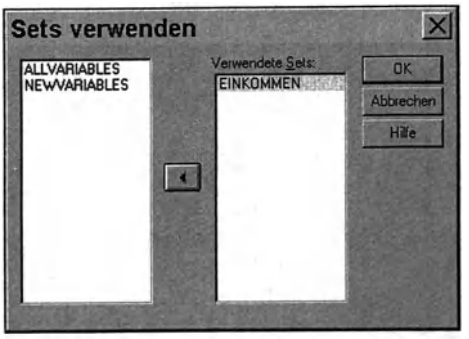


Abb. 28.4. Dialogbox „Sets verwenden"

Hauptfenster. Nur in Syntax- und Ausgabefenstern vorhanden. Macht das „aktive" Syntax- oder Ausgabefenster zum Hauptfenster. (Ist nur aktiv, wenn das aktive Fenster nicht das Hauptfenster ist, die Ausgabe also normalerweise in ein anderes Fenster geleitet wird. Dieselbe Wirkung erreichen Sie durch Anklicken der Schaltfläche .)

Autoskript erstellen/bearbeiten (Nur im Ausgabefenster). Wenn Sie im Ausgabefenster ein Objekt markieren und im Menü „Extras" diese Option anwählen, öffnet sich der Skript-Editor mit dem zu diesem Objekt gehörigen Skript. Sie können dieses Skript bearbeiten oder ein neues erstellen (⇨ Kap. 28.3)

Skript ausführen. Wenn Sie im Ausgabefenster ein Objekt markieren und in irgendeinem Fenster im Menü „Extras" diese Option anwählen, öffnet sich die Dialogbox „Skript ausführen", in der sie ein vorgefertigtes Skript auswählen und auf das ausgewählte Objekt anwenden können (⇨ Kap. 28.3.1).

Menü-Editor. Mit dieser Option können Sie das Menü anpassen, d.h. neue Menüs oder Optionen einbauen (⇨ Kap. 28.4).

28.3 Verwenden von Skripts und Autoskripts

Die SPSS-Ausgabe wird normalerweise in einer von SPSS vorgegebenen Weise formatiert. Man kann diese Gestaltung im Ausgabefenster überarbeiten. Es steht aber zu diesem Zwecke auch eine Programmiersprache (Skriptsprache) zur Verfügung, mit deren Hilfe man die Ausgabe automatisch anpassen kann (das Gegenstück zur Syntaxsprache für die Programmierung der Statistikprozeduren). SPSS liefert eine Reihe vorgefertigter Skripts. Man kann aber auch solche Skripts selbst programmieren bzw. vorgegebene Skripts nach eigenen Wünschen überarbeiten. SPSS stellt dazu im „Skript-Editor" eine eigene Programmierumgebung und ein umfangreiches Hilfesystem zur Verfügung. Das Schreiben von Skripts setzt einige Programmierkenntnisse voraus. Eine Einführung in das Programmieren von Skripts würde den Rahmen dieses Buches sprengen. Dagegen können die mitgelieferten Skripts einfach eingesetzt werden. Dies soll hier dargestellt werden.

Dazu sind zwei Arten von Skripts zu unterscheiden.

- *Skripts.* Sie werden zur Formatierung eines ausgewählten Ausgabeobjektes verwendet.
- *Autoskripts.* Ein Autoskript ist eine Sammlung von Skripts, die bestimmten Ausgabeobjekten zugeordnet sind. Wird ein solches Objekt erzeugt, wird es automatisch mit dem dazugehörigen Skript formatiert.

28.3.1 Verwenden eines vorgefertigten Beispielskripts

Beispiel. SPSS liefert das Beispielskript „Gesamt fett". Dieses Skript bewirkt, dass in einer Tabelle die Werte in Zeilen und/oder Spalten, die mit „Gesamt" überschrieben sind fett und blau ausgegeben werden. Dies soll auf eine fertige Tabelle, etwa die in Kap. 2 erzeugte Häufigkeitstabelle „Politisches Interesse" (⇨ Tab. 2.2) angewandt werden.

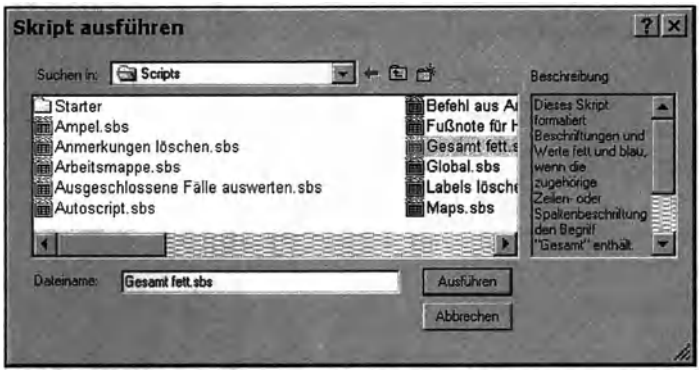


Abb. 28.5. Dialogbox „Skript ausführen"

Dazu gehen Sie wie folgt vor:

- ▷ Erstellen Sie die Häufigkeitstabelle. Markieren Sie diese im Ausgabefenster.
- ▷ Wählen Sie „Extras“ und „Skript ausführen...“. Es öffnet sich die Dialogbox „Skript ausführen“.
- ▷ Wählen Sie im Fenster „Suchen in:“ das Verzeichnis, in dem sich die Skriptdateien befinden (hier: c:\SPSS\Skripts).
- ▷ Markieren Sie den Namen des gewünschten Skripts (hier: „Gesamt fett“). Im Fenster „Beschreibung:“ erscheint eine Beschreibung dessen, was das Skript bewirkt, im Feld „Dateiname:“ der Name der Skriptdatei (hier: „Gesamt fett“).
- ▷ Klicken Sie auf die Schaltfläche „Ausführen“. Das Skript wird ausgeführt. In der Tabelle erscheinen die Werte in den Zeilen/Spalten mit der Überschrift „Gesamt“ fett und blau.

Anpassen oder neu Erstellen von Skripts ist möglich im Skript-Editor. Diesen erreichen Sie mit der Befehlsfolge „Datei“, „Neu“, „Skript“. In der dann erscheinenden Box „Starterskript verwenden“ wählen Sie ein zu bearbeitendes Skript aus und öffnen es im Skript-Editor. Oder Sie wählen „Abbrechen“. Dann öffnet sich ein, bis auf die Anfangs- und Schlussbefehle „Sub Main” und “End Sub”, leeres Skript Fenster.

Abb. 28.6 zeigt den Skript-Editor mit einem Ausschnitt aus dem Skript “Gesamt fett”, in dem u.a. die Farbe des Texte für die “Gesamt”-Werte auf blau (“Blue”) festgelegt wird.

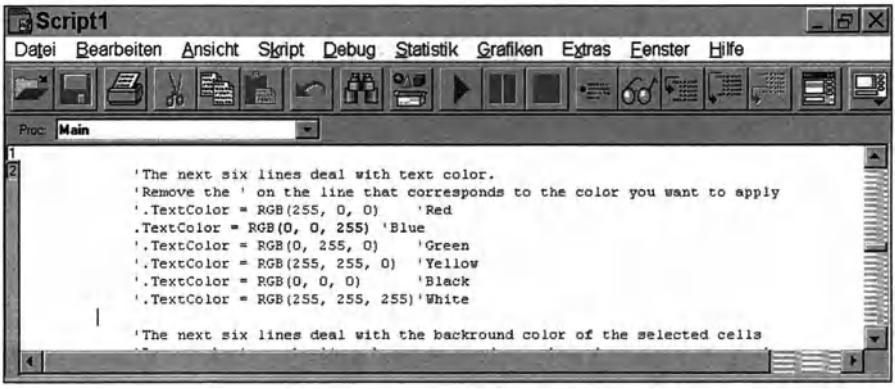


Abb. 28.6. SPSS Skript-Editor mit Auszug aus der Datei „Gesamt fett“

28.3.2 Verwenden eines vorgefertigten Autoskripts

SPSS liefert auch eine Autoskriptdatei mit, die bereits Skripts für verschiedene Elemente enthält. Diese Skripts sind aber nicht aktiviert.

Wählen Sie “Bearbeiten”, “Optionen” und das Register “Skripts”. Es erscheint die Registerkarte „Skripts“.

In dieser lässt sich zunächst die Datei auswählen, in der sich “globale Prozeduren” befinden. Dies sind Prozeduren, die in unterschiedlichen Skripts verwendet

werden. Die von SPSS mitgelieferten globalen Prozeduren befinden sich in "global.sbs". Sie werden in den mitgelieferten Skripts verwendet. Deshalb sollte man hier nur eine Änderung der Einstellung vornehmen, wenn man sich mit der Anwendung von Skripts gut auskennt.

Außerdem kann man diejenige "Autoskript-Datei" auswählen, die Verwendung finden soll. Die von SPSS mitgelieferte heißt "Autoskript.sbs" und wird von uns verwendet. Soll die Autoskriptdatei verwendet werden, muss das Auswahlkästchen "Autoskript Ausführung aktivieren" angewählt sein (Voreinstellung). Damit wird aber noch keine der Subroutinen wirklich ausgeführt. Durch Anklicken des Auswahlkästchens vor der gewünschten Subroutine wird diese aktiviert. *Beispiel:* Es wird mit „Deskriptive Statistiken" eine Tabelle für die Variable EINK erstellt. Mit den voreingestellten Autoskriptroutinen erscheint der Variablennamen in der ersten Spalte, die Statistiken in den folgenden. Im „Autoskript.sbs" steht eine per Voreinstellung nicht aktivierte Routine zur Verfügung, die Zeilen und Spalten gegenüber der Standardeinstellung von "Deskriptive Statistiken" vertauscht, so dass sich die Statistiken in den Zeilen und die Variablen in den Spalten befinden. Diese heißt *Descriptives_Table_DescriptiveStatistics_Create.*

Aktivieren Sie diese und erstellen Sie die Tabelle zum zweiten Mal. Der Vergleich zeigt die entsprechende Veränderung der Ausgabe.

28.4 Anpassen von Menüs und Symbolleisten

In SPSS für Windows ist es möglich, Menüs und Symbolleisten nach eigenen Wünschen umzugestalten. Bei den Menüs heißt dies, neue Menüs oder Optionen einfügen. Den Symbolleisten können neue Symbole hinzugefügt werden. Es ist auch möglich zu bestimmen, in welchen Fenstern die Leisten angezeigt werden sollen. Schließlich können gänzlich neue Symbolleisten erstellt werden. *Beispiel:* Es soll ein neues Menü "Export" mit nur einer Option "Exportieren nach Excel" erstellt werden. Dieses Menü soll es ermöglichen, die Daten des Dateneditors unmittelbar in eine Excel-Datei zu exportieren. Zu demselben Zweck soll eine neue Symbolleiste mit nur einem Symbol "Exportieren nach Excel" kreiert werden. Diese Symbolleiste soll nur im Daten-Editor erscheinen.

28.4.1 Anpassen von Menüs

Die Menüs können um folgende Typen von Optionen ergänzt werden:

- ❒ Optionen, mit denen angepasste SPSS-Skripts ausgeführt werden.
- ❒ Optionen, mit denen SPSS-Befehlssyntax-Dateien ausgeführt werden.
- ❒ Optionen, mit denen andere Anwendungen gestartet und Daten aus SPSS automatisch an andere Anwendungen übergeben werden.

Übergaben von Daten sind an folgende Anwendungen möglich: SPSS, Excel, Lotus 1-2-3 Version 3, SYLK, Tabulatorzeichen als Trennzeichen und dBASE IV.

In unserem Beispiel geht es um den letzten Typ von Optionen und zwar die Übergabe von Daten an eine Excel-Datei. Menüeinträge für die anderen Zwecke werden aber analog erstellt.

Um ein neues Menü "Export" mit der Option "Exportieren nach Excel" zu erstellen, gehen Sie wie folgt vor:

- ▷ Wählen Sie die Befehlsfolge "Extras" und "Menü-Editor...". Es erscheint die Dialogbox "Menü-Editor"(⇨ Abb. 28.7).
- ▷ Markieren Sie im Fenster "Menü:" den Namen des Menüs, vor dem das neue Menü "Export" eingefügt werden soll (hier: "&Hilfe").
- ▷ Im Fenster "Anwenden auf:" geben Sie an, für welches Fenster das neue Menü gelten soll (hier: „Daten-Editor").
- ▷ Klicken Sie auf die Schaltfläche „Menü einfügen". Es erscheint der Eintrag „Neues Menü". Überschreiben Sie diesen mit dem gewünschten Namen (hier: „Export"). Wenn Sie auf den Namen doppelklicken, sehen Sie auf der nächsten Ebene den Eintrag „Ende des Menüs Export". (Wenn Sie ein schon bestehendes Menü durch eine Option ergänzen, sehen Sie, wenn Sie auf den Menünamen doppelklicken, auf der unteren Ebene die Namen aller Optionen.)
- ▷ Markieren Sie die Option, vor der Sie die neue Option einsetzen wollen (hier: „Ende des Menüs Export").
- ▷ Klicken Sie auf die Schaltfläche „Eintrag einfügen". Es erscheint eine neue Option mit der vorläufigen Bezeichnung „Neuer Menüeintrag".
- ▷ Ersetzen Sie nun noch den vorläufigen Namen der Option durch „Exportieren nach Excel" .
- ▷ Wählen Sie den Dateityp aus (hier: „Anwendung").
- ▷ Wenn es sich um eine Anwendung handelt, muss jetzt hier angegeben werden, um welche Anwendung es sich handeln soll. Wählen Sie diese aus der Liste aus, die sich beim Anklicken des Pfeils neben dem Feld „Daten übergeben als" öffnet. Im Beispiel wählen wir „XLS-Excel-Dateien".
- ▷ Klicken Sie auf die Schaltfläche „Durchsuchen", und wählen Sie in der Dialogbox „Öffnen" auf die übliche Weise zunächst das Laufwerk und das Verzeichnis aus, in dem sich die Anwendung befindet. Aus der Liste der Dateien wählen Sie die exe-Datei der Anwendung aus (hier: „Excel.exe") und übertragen sie in das Feld „Dateiname:". Klicken Sie auf „Öffnen". Pfad und Dateiname erscheinen jetzt im Feld „Dateiname" der Dialogbox „Menü-Editor".
- ▷ Bestätigen Sie das Ganze mit "OK".

Die Menüleiste des Dateneditors enthält nun ein weiteres Menü "Export" mit der Option "Exportieren nach Excel". Wenn Sie diese anklicken, wird automatisch Excel geöffnet und der Inhalt des Dateneditors in eine Excel-Datei exportiert.

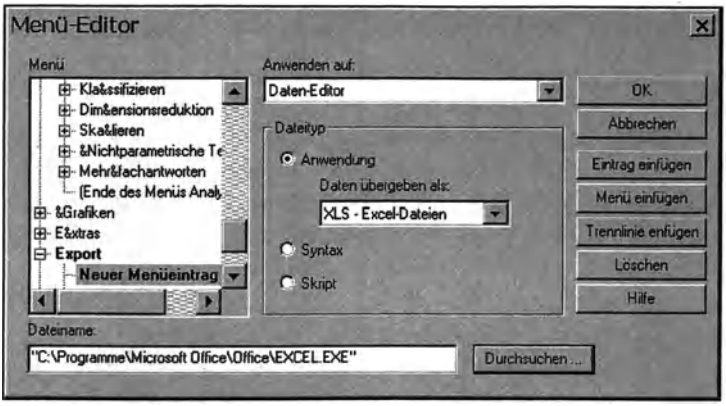


Abb. 28.7. Dialogbox „Menü-Editor"

28.4.2 Anpassen von Symbolleisten

Neue Symbole können folgende Zwecke erfüllen:

- Aufrufen von in SPSS verfügbaren Funktionen (d.h. auch alle über Menüs verfügbaren Aktionen). Bei weitem nicht alle sind in den vordefinierten Symbolleisten verfügbar. Nicht gewünschte Symbole können aus diesen entfernt werden, neue für andere Funktionen eingefügt.
- Starten anderer Anwendungen sowie von Befehlssyntax-Dateien und Skriptdateien.

In unserem Beispiel wird ein Symbol zum Starten einer anderen Anwendung eingefügt. Das Einfügen von Symbolen zum Aufrufen von SPSS-Funktionen folgt im Prinzip demselben Weg. Für alle verfügbaren Funktionen von SPSS stehen in Listen vordefinierte Symbole zur Verfügung, die in die Symbolleisten übertragen werden können.

Um eine neue Symbolleiste "Exportieren" mit nur dem einem Symbol "Exportieren nach Excel" zu erstellen, gehen Sie wie folgt vor:

- Wählen Sie "Ansicht", "Symbolleisten". Es öffnet sich die Dialogbox "Symbolleisten anzeigen".
- Klicken Sie auf die Schaltfläche "Neue Symbolleiste". Die Dialogbox "Symbolleiste: Eigenschaften" erscheint.
- Wählen Sie in der Gruppe "In den folgenden Fenstern anzeigen" die Fenster aus, in denen die neue Symbolleiste erscheinen soll. In unserem Beispiel ist es nur das Fenster "Daten-Editor". Alle anderen müssen ausgeschaltet werden.
- Geben Sie in das Feld "Name der Symbolleiste" den Namen der neue Leiste (hier: "Export") ein.
- Klicken Sie auf „Anpassen". Die Dialogbox „Symbolleiste anpassen" erscheint (⇨ Abb. 28.8).

Hier sehen Sie zwei Fenster. Im linken Fenster „Kategorien" sind Kategorien von Funktionen vorhanden, für die jeweils eine Liste von Symbolen zur Verfügung

steht. Auf der rechten Seite sind im Fenster „Symbole“ die Symbole für alle Funktionen der gerade angewählten Kategorie angezeigt. So etwa in der Kategorie „Datei“ Symbole für die Funktionen „Neue Daten“, „Neue Ausgabe“, „Neue Syntax“ etc. Die Dialogbox enthält außerdem im unteren Teil eine zunächst noch leere Symbolleiste. In unserem Beispiel beschriftet mit „Symbolleiste anpassen: Export“. (Wird eine schon existierende Symbolleiste angepasst, enthält sie bereits Symbole.) Würden wir jetzt eines der bereits vorhandenen Symbole verwenden, würde es einfach aus der Liste auf die Symbolleiste gezogen. Wir müssen aber ein eigenes Symbol definieren, d.h. schaffen ein nutzerdefiniertes Symbol.

Abb. 28.8. Dialogbox „Symbolleiste anpassen“

- Um ein nutzerdefiniertes Symbol zu erstellen, klicken Sie auf die Schaltfläche „neues Symbol“. Es erscheint die Dialogbox „Neues Symbol erstellen“. Geben Sie in der Gruppe „Beschreibung“ im Feld „Beschriftung“ einen Namen für das Symbol ein (hier: „Export nach Excel“). Da über das Symbol eine Anwendung gestartet werden soll, muss der Optionsschalter „Anwendung“ angewählt sein. In der Liste zum Feld „Daten übergeben als:“ muss die Art der Anwendung ausgewählt werden (im Beispiel: „XLS-Excel-Dateien“). Wiederum öffnen Sie über die Schaltfläche „Durchsuchen“ die Dialogbox „Öffnen“. Wählen Sie dort auf die übliche Weise zunächst das Laufwerk und das Verzeichnis aus, in dem sich die Anwendung befindet. Aus der Liste der Dateien wählen Sie die exe-Datei der Anwendung aus und übertragen sie in das Feld „Dateiname:“. Klicken Sie auf „Öffnen“. Pfad und Dateiname erscheinen jetzt im Feld „Dateiname“ der Dialogbox „Neues Symbol erstellen“.
- Beenden Sie mit „OK“. Die Liste der Kategorie „Benutzerdefiniert“ enthält jetzt ein neues Symbol mit der Bezeichnung „Export nach Excel“.
- Um dieses auf der Symbolleiste zu plazieren, klicken Sie auf das Symbol im Fenster „Symbole“ und ziehen Sie es auf die Symbolleiste „Export“ im unteren Drittel der „Box“. (Wenn sie mehrere nutzerdefinierte Symbole erstellen, sehen sie zunächst alle gleich aus. Sie sollten diese deshalb vielleicht noch im über die

Schaltfläche „Symbol Bearbeiten“ zu erreichenden Bitmap-Editor etwas umgestalten).

▷ Mit zweimal „OK“ beenden Sie die Definition.

Im Daten-Editor finden Sie nun eine neue Symbolleiste mit dem eben erstellten Symbol. Wenn Sie dieses anklicken, wird automatisch Excel geöffnet und der Inhalt des Dateneditors in eine Excel-Datei exportiert.

28.5 Ändern der Arbeitsumgebung im Menü „Optionen“

Mit SPSS arbeiten Sie in einer bestimmten Arbeitsumgebung, die Sie teilweise gestalten können. Das betrifft zunächst die allgemeine Arbeitsumgebung, z.B. die Reihenfolge der Variablen in den Quellvariablenlisten, die Führung der Protokolldatei, die Anordnung der Fenster nach der Ausführung eines Befehls. Vor allem aber wird die Gestalt der verschiedenen Ausgaben beeinflusst, die Gestaltung der Ausgabe der Pivot-Tabellen, der Diagramme etc.

Diese Einstellungen können geändert werden. Wählen Sie dazu „Bearbeiten“, „Optionen...“. Es öffnet sich die Dialogbox „Optionen“(⇨ Abb. 28.9). Sie enthält verschiedene Register. Auf jeder der Registerkarte können für einen speziellen Bereich Einstellungen verändert werden.

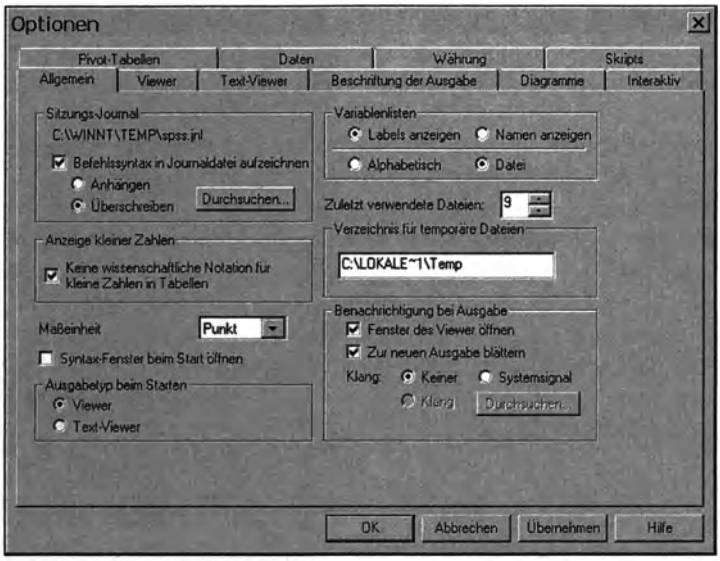


Abb. 28.9. Dialogbox „Optionen“ mit geöffnetem Register „Allgemein“

Register „Allgemein“. Hier können Sie jetzt die Arbeitsumgebung nach Ihren Wünschen gestalten.

- *Sitzungs-Journal.* Die abgearbeiteten Befehle einer SPSS-Sitzung werden in einer Protokolldatei (Voreinstellung: C:\WINNT\TEMP\SPSS.JNL) protokolliert. Dabei ist es gleichgültig, ob sie aus einem Dialog- oder einem Syntaxfenster gestartet wurden. Die Gruppe „Sitzungs-Journal" zeigt Verzeichnis und Namen der Protokolldatei an. Wollen Sie diese ändern, klicken Sie auf die Schaltfläche „Durchsuchen". Es öffnet sich die Dialogbox „Speichern unter". Hier können Sie auf die übliche Weise ein Verzeichnis auswählen und im Feld „Name:" den Dateinamen ändern. Per Voreinstellung wird jede SPSS-Sitzung protokolliert. Sie können aber nach Belieben das Protokoll während der Sitzung durch Anklicken des Kontrollkästchens „Befehlssyntax in Journaldatei aufzeichnen" an- und ausschalten. Wählen Sie den Optionsschalter „Anhängen", wird das Protokoll der laufenden Sitzung den Protokollen früherer angefügt, wählen Sie dagegen „Überschreiben", wird mit jeder neuen Sitzung das alte Protokoll gelöscht und durch das der neuen Sitzung ersetzt.
- *Variablenlisten.* In dieser Gruppe bestimmen Sie zweierlei:
  - *Anzeigeform in der Quellvariablenliste.* Entweder werden dort die je nach Auswahl des Optionsschalters Namen der Variablen oder die Variablenlabels (Voreinstellung) angezeigt. Ersteres ist übersichtlicher, letzteres bei nichtssagenden Variablennamen informativer.
  - *Variablensortierung in der Quellvariablenliste.* Variablen können in den Qellvariablenlisten der Dialogboxen entweder alphabetisch (Optionsschalter „Alphabetisch", Voreinstellung) oder in der Reihenfolge, in der die Variablen in die Datei eingegeben wurden („Wie in Datei") sortiert sein. Ersteres wird man bei langen unübersichtlichen Listen bevorzugen, letzteres, wenn in kürzeren Dateien die Eingabesortierung eine sinnvolle Orientierung ermöglicht.
- *Anzeigen kleiner Zahlen.* Hier kann man verfügen, dass kleine Zahlen in Ausgabetabellen nicht in wissenschaftlicher Notation angezeigt werden. Dies verbessert für mathematisch wenig Geübte in der Regel die Lesbarkeit der Tabelle,
- *Syntaxfenster bei Start öffnen.* Beim Starten von SPSS wird normalerweise kein Syntaxfenster geöffnet. Das geschieht erst, wenn man ein solches neu erstellt oder eine Syntaxdatei lädt oder aber einen Befehl mittels der Schaltfläche "Einfügen" aus einer Dialogbox überträgt. Durch Ankreuzen des Kontrollkästchens „Ja" bewirken Sie dagegen, dass mit dem Start ein Syntaxfenster geöffnet wird. Das ist vor allem dann interessant, wenn Sie ausschließlich oder überwiegend mit der Befehlssyntax arbeiten wollen.
- *Maßeinheit.* Betrifft Zeilenränder, Zeilenabstände usw. in Pivot-Tabellen. Wird normalerweise in Punkt ausgedrückt. Alternativen sind Zoll und cm.
- *Zuletzt verwendete Dateien.* Öffnet man das Menü „Datei", findet sich an vorletzter Stelle eine Liste der zuletzt geöffneten Dateien und davor eine ebensolche der zuletzt geöffneten Daten, so dass man leicht durch Anklicken des Dateinamens eine dieser Dateien zur Analyse auswählen kann. Man kann im Feld „Zuletzt verwendete Dateien" bestimmen, wie viele der zuletzt verwendeten Dateien in diesen Listen angeführt werden. Die Voreinstellung ist 4. Man verändert dies durch Anklicken eines der Pfeile neben dem Kästchen mit der Zahlenangabe.

- *Ausgabetyp bei Starten.* Man kann die Ergebnisse einer Prozedur im „Viewer“ ausgeben lassen (Voreinstellung). Dann stehen alle Ergebnisse in einem graphischen Format. Das hat für die Bearbeitung der Objekte in SPSS selbst viele Vorteile, benötigt aber auch viel Speicher und kann für die Übernahme in andere Anwendungen manchmal hinderlich sein. Deshalb steht jetzt wieder ein „Text-Viewer" als Alternative zur Verfügung. Hier werden die Ergebnisse – mit Ausnahme der Grafiken – in reinem ASCII-Format ausgegeben.
- *Benachrichtigung bei der Ausgabe.* Führt man in SPSS einen Befehl aus, der zu einer Ausgabe führt, wird per Voreinstellung automatisch der Viewer in den Vordergrund gebracht, so dass man sofort das Ergebnis sehen kann. Wünscht man dies nicht, sondern möchte in dem Fenster bleiben, in dem man sich beim Starten des Befehls befand, so muss man das durch Abwahl des Auswahlkästchens „Fenster des Viewers öffnen“ ändern. Ebenso springt SPSS per Voreinstellung nach Abarbeitung eines Befehls an den Anfang der neuen Ausgabe. Möchte man dagegen lieber, dass der Viewer an der Stelle stehen bleibt, an der er sich vor Abschicken des Befehls befand, wählt man „Zur neuen Ausgabe blättern“ ab. Schließlich kann man durch Auswahl des entsprechenden Optionsschalters bestimmen, ob die Beendigung einer Ausgabe durch ein Klangsignal angezeigt werden soll oder nicht. Falls der Computer über die entsprechende Hard- und Softwareausstattung verfügt, kann man auch die Art des Klangsignales selbst bestimmen. Dazu wählt man den Optionsschalter „Klang“ und mit „Durchsuchen“ die Datei, die den gewünschten Klang erzeugt.

Register „Daten“.

- *Optionen für Transformieren und zusammenfügen.* In dieser Gruppe bestimmen Sie, ob Datentransformationen sofort ausgeführt werden („Werte sofort berechnen“) oder erst dann, wenn eine Operation gestartet wird, die diese benötigt („Werte vor Verwendung berechnen“). Letzteres wird man dann verwenden, wenn bei aufwendigen Transformationen Rechnerzeit gespart werden soll.
- *Anzeigeformat für neue numerische Variablen.* In dieser Gruppe bestimmen Sie die Voreinstellung für die Anzeige neuer numerischer Variablen. Im Feld „Breite:“ wird die Gesamtanzeigenlänge (inklusive Dezimaltrennzeichen und Vorzeichen) der Anzeige eingestellt. Das Feld „Dezimalstellen:“ bestimmt die Zahl der angezeigten Stellen nach dem Dezimaltrennzeichen. Die Einstellung hat keinen Einfluss auf die Genauigkeit, mit der die Werte intern gespeichert werden.
- Jahrhundertbereich für 2-stellige Jahreszahlen. Mit diesem Bereich reagiert SPSS auf das bekannte Problem mit der Jahrtausendwende. Wenn zweistellige Jahreszahlen im Datumsbereich eingegeben wurden, wurden sie bisher automatisch um 1900 ergänzt. Jetzt kann man das beeinflussen.
  - *Automatisch.* Das bedeutet, dass eine Spanne von 69 Jahren vor dem aktuellen Jahr bis 30 Jahre nach dem aktuellen angenommen wird. So wird aus „98“ „1998“, aus „1“ dagegen „2001“.
  - Benutzerdefiniert. Hier kann man eine Zeitspanne von 100 Jahren durch Eingabe eines Wertes in „Erstes Jahr“ festlegen (voreingestellt ist diese Option mit der bisherigen Zeitspanne 1900 bis 1999). Statt dessen könnte man z.B.

1950 bis 2049 einstellen. Aus „49" würde dann „2049", aus „51" dagegen „1951".

Register „Währung" (⇨Abb. 3.2). In diesem Register kann man bis zu fünf Währungsformate definieren. Diese werden unter den in der Box links oben angezeigten Formatbezeichnungen „CCA" (bedeutet Custom Currency A), „CCB" usw. gespeichert. Per Voreinstellung entsprechen zunächst alle Formate dem numerischen Standardanzeigeformat mit zwei Nachkommastellen.

Um ein Format zu definieren, ändern Sie diese Voreinstellung:

- ▷ Wählen Sie eine der Formatbezeichnungen aus. Die Voreinstellung wird in der Gruppe „Beispiel" angezeigt.
- ▷ Geben Sie dann die gewünschten Definitionen ein. Im unteren Teil der Dialogbox befinden sich die Gruppen zur Änderung eines Formats. Die in der Gruppe „Alle Werte" festgelegten Definitionen werden jedem Wert zugeordnet, die in der Gruppe „Negative Werte" definierten, nur negativen Werten. Man kann ein „Präfix bestimmen", d.h. ein Zeichen, das vor dem Wert angezeigt wird, oder ein „Suffix", d.h. ein Zeichen, das nach dem Wert angezeigt wird. (Es kann sich auch um eine kurze Zeichenfolge handeln.) In der Gruppe „Dezimalzeichen" legt man durch Auswahl der entsprechenden Optionsschalter fest, ob der Punkt oder das Komma als Dezimaltrennzeichen verwendet wird (beachten Sie, dass letzteres sich nicht bei jeder statistischen Routine auswirkt). *Beispiel:* Sie definieren ein Währungsformat mit nachgestelltem „DM", Dezimaltrennzeichen sei das Komma, negative Zahlen werden durch vorangestelltes Minus gekennzeichnet.
- ▷ Klicken Sie auf die Schaltfläche „Übernehmen". Die Gruppe „Beispiel" zeigt nun das Beispiel mit der veränderten Einstellung. Bestätigen Sie mit „OK".

Die so definierten Währungsformate stehen nun für die Definition von Variablen im Register „Variablenansicht" des "Daten-Editors", Spalte „Typ" zur Verfügung. Sie können dann durch Anklicken der Schaltfläche in der Spalte „Typ" die Dialogbox „Variablentyp definieren" öffnen und dort „Spezielle Währungen" anwählen. Darauf öffnet sich eine Auswahlliste, aus der sie das erstellte Format zuweisen können.

Register „Viewer". Hier werden grundlegende Formatierungen des Ausgabenfensters festgelegt.

- ❒ *Anfänglicher Ausgabestatus.* In dieser Gruppe finden sich links 10 Symbole, die alle bestimmte Elemente der Ausgabe bezeichnen. Diese können durch Anklicken des jeweiligen Symbols ausgewählt werden. Das ausgewählte Element wird in Feld „Objekt" angezeigt. Die Auswahl kann auch aus einer Liste erfolgen, die sich beim Anklicken des Pfeils neben dem Feld „Objekt" öffnet. Objekte sind: Log, Warnungen, Anmerkungen, Titel, Seitentitel, Pivot-Tabelle, Diagramm, Textausgabe, Grafik und Karte. Für jeden dieser Objekttypen kann durch Anklicken des entsprechenden Optionsschalters festgelegt werden, ob Objekte dieses Typs nach Beendigung eines Laufs im Viewer angezeigt werden („Eingeblendet") oder nicht („Ausgeblendet"). Voreingestellt ist – mit Ausnahme von „Anmerkung" – eingeblendet. Dass ein Objekt ausgeblendet ist,

heißt jedoch nicht, dass es im Lauf nicht erstellt wurde. Es wird nur nicht angezeigt. Im Viewer selbst kann es jederzeit eingeblendet werden. Außer für Log- und Textausgaben kann auch die Ausrichtung („Linksbündig“, „Zentriert“ oder „Rechtsbündig“) festgelegt werden (Voreinstellung: „Linksbündig“).

❒ *Befehle im Log anzeigen.* Klickt man dieses Kontrollkästchen an, bewirkt das, dass vor dem Ergebnis einer Operation die Befehlssyntax dieser Operation angezeigt wird. Man kann diese z.B. dann verwenden, um eine Syntaxdatei zu erstellen.

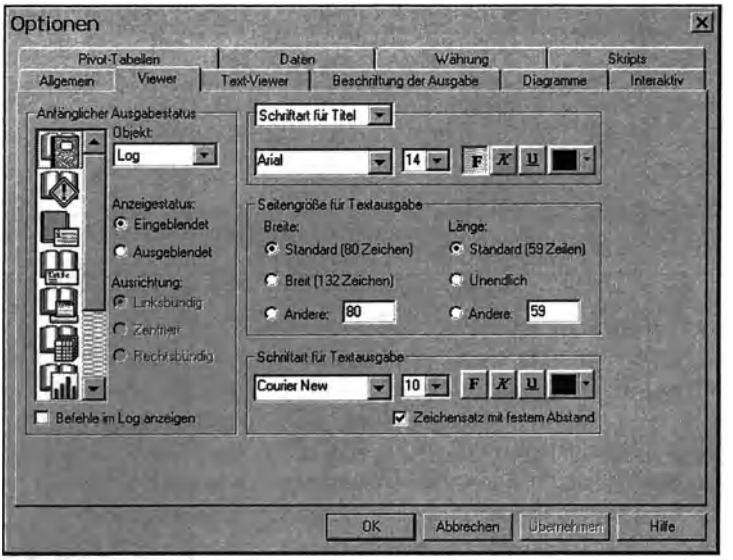


Abb. 28.10. Dialogbox „Optionen“, Register „Viewer“

❒ *Schriftart für Titel.* In dieser Gruppe bestimmt man die Formatierung der Zeichen von Überschriften (Titel) in der Ausgabe. Geändert werden können Schrifttyp, Schriftgröße, die Farbe der Schrift und die Auszeichnung (fett „F“, kursiv „K“ oder unterstrichen „U“). Die ersten drei Eigenschaften wählt man jeweils aus einer Liste aus, die sich öffnet, wenn man auf den Pfeil neben dem Anzeigefeld für dieses Merkmal klickt. Zur Auswahl der Auszeichnung klickt man auf das entsprechende Symbol.

❒ *Schriftart für Textausgabe.* Hier gilt dasselbe wie für Schriftart Titel. Allerdings existiert ergänzend die Auswahlbox „Zeichensatz mit festem Abstand“. Ist es angewählt, stehen nur solche Zeichensätze zur Verfügung. Dies ist insbesondere dann sinnvoll, wenn man Tabellenergebnisse als Texte in andere Anwendungen übertragen will. Ist das Kästchen nicht markiert, stehen auch proportionale Schriften zur Verfügung.

❒ *Seitengröße für Textausgabe.* Hier werden Länge (in Zeilen) und Breite (in Zeichen) einer Seite für Textausgaben im Viewer festgelegt. In dieser Gruppe kön-

nen Sie zwischen zwei Standardeinstellungen und einem selbstdefinierten Wert wählen. Die Seitenbreite ist mit 80 Zeichen als „Standard" voreingestellt. Das ist die übliche Bildschirmbreite. Standardlänge ist 59 Zeilen. Beides ist auf das amerikanische Papierformat bei Verwendung der Standardschriftgröße 10 ausgerichtet. Alternativ ist eine Breite von 132 Zeichen vorgeschlagen. Das ist abgestellt auf die Druckbreite des mit Seitendruckern häufig verwendeten Endlospapiers, die Länge wird dann sinnvollerweise unendlich. Sie können andere Seitenbreiten und -längen festlegen. Sie ändern die Einstellung, indem Sie den Optionsschalter „Andere:" anklicken und in die Eingabefelder die gewünschten Werte eintragen. Der selbstdefinierte Wert für die Breite muss aber zwischen 80 und 255 Zeichen liegen, derjenige für die Länge zwischen 24 und 9999 Zeilen. Wegen der Mindestzeichenbreite von 80 ist eine geeignete Anpassung auf deutsche Papiermaße und anderer Schriftgrößen (verbreitet ist die Größe 12) nur bedingt möglich. Die Alternative „Unendlich" unterdrückt alle Seitenvorschubzeichen in der Kopfzeile.

Register „Beschriftung der Ausgabe". Hier wird festgelegt, wie Variablen bzw. Variablenwerte bei der Beschriftung der Ausgabe verwendet werden.

- ❒ *Variablen.* Für die Variablen kann man sich entweder den „Namen" oder den „Variablenlabel" oder aber beides („Namen und Label") ausgeben lassen.
- ❒ *Werte.* Für die Werte kann man entweder die „Wert" oder die „Labels" oder beides („Werte und Labels") anzeigen lassen.
  Damit kann man die Lesbarkeit und den äußeren Eindruck der Tabellen und Überschriften nach Wunsch gestalten.
  - *Gliederungsbeschriftung.* In der oberen Gruppe legt man das für Objekte, insbesondere die Gliederungsansicht im linken Fenster des Viewers fest.
  - *Beschriftung für Pivot-Tabellen.* Die untere Gruppe dagegen bestimmt, wie Variablen und Werte bei der Beschriftung der Tabellen selbst verwendet werden.

Register „Pivot-Tabellen". In diesem Register werden weitere Eigenschaften der Pivot-Tabellen festgelegt.

- ❒ *Tabellenvorlage.* SPSS gibt per Voreinstellung den Pivot-Tabellen eine bestimmte Form. Diese kann später durch Bearbeitung verändert werden. U.a. ist das dadurch möglich, dass man eine der zahlreichen von SPSS mitgelieferten „Tabellenvorlagen" auswählt. Im Register „Pivot-Tabellen" können sie eine dieser mitgelieferten Tabellenansichten zur Standardansicht erklären. In der Gruppe „Tabellenvorlage" werden die verfügbaren Ansichten angezeigt. Ein Beispiel für die jeweils markierte Ansicht sehen Sie im rechten Feld „Beispiel". Wählen Sie die Tabellenansicht, die Ihnen am meisten zusagt aus, indem Sie deren Namen markieren und mit „OK" bestätigen.

  (Im Viewer kann man solche Standardtabellenvorlagen nach eigenen Wünschen überarbeiten und unter neuem Namen der Liste hinzufügen, evtl. auch in einem anderen Verzeichnis speichern (⇨ Kap. 4.1.5). Das Verzeichnis, in dem sich die gewünschte Tabellenvorlage befindet, stellt man dann in einem Dialogfenster, das sich nach Anklicken „Durchsuchen..." öffnet, in der üblichen Weise ein. Sollen in Zukunft immer die Tabellenvorlagen aus diesem gerade einge-

stellten Verzeichnis angeboten werden, klicken Sie auf „Verzeichnis für Tabellenvorlagen".

- ❒ *Spaltenbreite einstellen für.* Per Voreinstellung richtet sich SPSS bei der Gestaltung der Spaltenbreiten einer Tabelle nach der Beschriftung der Spalte „Beschriftungen", nicht aber nach der Größe der Zahl in der Spalte. Bei großen Zahlen kann das dazu führen, dass sie nicht ganz in die Spalte passen. Sie werden dann in wissenschaftlicher Notation angezeigt. Will man das verhindern, wählt man besser die Option „Beschriftungen und Daten".
- ❒ *Standardbearbeitungsmodus.* Man kann in SPSS Tabellen entweder im Viewer oder in einem eigenen Fenster bearbeiten. Ob nach Doppelklicken auf eine Tabelle diese im Viewer oder in einem eigenen Fenster bearbeitet werden soll, legt man in dieser Gruppe fest. Durch Klicken auf den Pfeil neben dem Eingabefenster öffnet sich eine Auswahlliste, in der man dies auch größenabhängig regeln kann. Besonders bei großen Tabellen empfiehlt sich evtl. die Bearbeitung in einem einfachen Fenster, weil sie sich dann ohne die störende Gliederungsleiste und überflüssige Menüs und Symbolleisten des Viewers etwas leichter handeln lassen.

Register „Diagramme".

- ❒ *Diagrammvorlage.* Generell kann man in diesem Register einige Merkmale der durch SPSS-Prozeduren erzeugten Diagramme festlegen. Man kann aber auch bestimmen, ob diese tatsächlich Verwendung finden oder durch eine andere, mitgelieferte Vorlage ersetzt werden.
  - *Aktuelle Einstellungen verwenden.* Es werden die in diesem Register festgelegten Einstellungen verwendet.
  - *Diagrammvorlagendatei verwenden.* Man kann im Viewer Diagramme bearbeiten und dann die Formatierung als Diagrammvorlage für spätere Diagramme der gleichen Art speichern (⇨ Kap. 4). Existieren solche Vorlagen, dann können Sie eine davon als Standardvorlage verwendet. In diesem Falle werden per Voreinstellung alle neuen Diagramme des gleichen Typs mit dieser Vorlage formatiert. Um dies zu erreichen, klicken Sie zunächst auf den Optionsschalter „Diagrammvorlage-Datei verwenden." Mit „Durchsuchen" öffnen Sie eine Dialogbox, in der auf die übliche Weise die gewünschte Vorlage ausgewählt wird.
- ❒ *Aktuelle Einstellungen.* In dieser Gruppe, und verschiedenen darin enthaltenen Untergruppen, werden einige Voreinstellungen für die Ausgabe von Diagrammen festgelegt.
  - *Schriftart.* In diesem Eingabefeld geben Sie die zur Beschriftung der Grafiken gewünschte Schriftart ein. Dazu öffnen Sie durch Klicken auf den Pfeil neben dem Eingabefeld „Schriftart" eine Auswahlliste. Die Auswahlliste zeigt nur die dem installierten Drucker verfügbaren Schriftarten an. Durch Klicken auf einen der Namen in der Liste bestimmen Sie die gewünschte Schriftart.
  - *Füllmuster und Linienstil.* Diese Gruppe bestimmt die Anfangszuweisung von Farben und Mustern zu neuen Grafiken. „Erst Farbpalette, dann Muster durchlaufen" (Voreinstellung) verwendet 14 Farben für die Gestaltung der

Diagramme. Reichen diese nicht aus, werden sie durch Muster ergänzt. Z.B. werden Balken in einem gruppierten Balkendiagramm zur Unterscheidung der Gruppen mit verschiedenen Farben gefüllt. „Muster durchlaufen" verwendet nur Muster. Dies ist vorzuziehen, wenn Schwarzweiß-Drucker verwendet werden. Dann entspricht das Bild der Grafik weitgehend der Druckausgabe. Z.B. werden Balken in einem einfachen gruppierten Balkendiagramm zur Unterscheidung der Gruppen mit unterschiedlichen schwarz-weiß Mustern (bzw. der Vollfarbe schwarz) gefüllt. (Eine Änderung ist aber im Grafikfenster jederzeit möglich.)

- *Rahmen.* In dieser Gruppe bestimmen Sie, ob ein Rahmen um die ganze Grafik („Äußerer") und/oder innen entlang den Achsen („Innerer") gezogen werden soll (Voreinstellung „Innerer"). Auch beides ist gleichzeitig möglich.
- *Gitterlinien.* In dieser Gruppe kann man durch Anklicken der Auswahlkästchen „Skalen-Achse" und/oder „Kategorien-Achse" bestimmen, dass per Voreinstellung Grafiken mit Gitterlinien auf der senkrechten (Skalen-Achse) und/oder auf der waagerechten (Kategorien-Achse) als Hilfslinien versehen werden. Per Voreinstellung sind keine Gitterlinien vorgesehen. Die Einstellung kann bei jeder Grafik geändert werden.

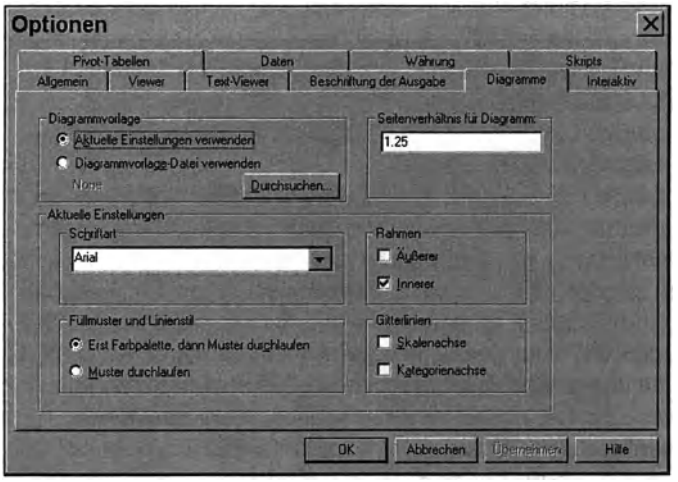


Abb. 28.11. Dialogbox „Optionen", Register „Diagramme"

- *Seitenverhältnis.* In dieser Gruppe legen Sie durch Anklicken der entsprechenden Optionsschalter das Seitenverhältnis (Breite zu Höhe, gemessen am Außenrahmen) der Grafik fest. Diese Einstellung hat Auswirkungen auf die Darstellung am Bildschirm und beim Druck der Grafik. Das „Bildschirmformat (1.67)" entspricht den Seitenverhältnissen eines Bildschirms im VGA-Modus. „Druckerformat (1.25)" entspricht dem Seitenverhältnis des amerikanischen Papierformats im Querformat (Voreinstellung). Durch Eingabe einer Dezimalzahl (Punkt als Dezimalzeichen) im Eingabefeld können Sie ein eigenes Seitenverhältnis bestimmen. Der Wert muss zwischen 0.1

und 10.0 liegen. (Ansonsten kommt eine Fehlermeldung.) Werte unter 1 erzeugen Grafiken mit größerer Höhe als Breite. 1 ergibt ein quadratisches Bild, Werte über 1 ergeben ein Bild mit größerer Breite als Höhe.

Register „Text-Viewer“. Hier legt man zunächst durch Ankreuzen von Kontrollkästchen fest, welche Ausgabeelemente im Viewer gezeigt werden sollen. Zur Auswahl stehen dieselben wie im „Viewer“, allerdings kann die Ausrichtung nicht verändert werden. Auch Seitenbreite und -länge werden auf dieselbe Weise bestimmt. Schriftart- und Größe können verändert werden, allerdings keine Auszeichnungen. Anders als für den Viewer kann in der Gruppe „Seitenumbruch zwischen“ der Seitenumbruch gesteuert werden. Das Markieren von „Prozeduren“ bewirkt, dass nach Beendigung jeder Prozedur (z.B. Kreuztabellen) ein Seitenvorschub durchgeführt wird. Bei Anwahl von „Elementen“ wird sogar nach jedem Objekt (Tabelle, Diagramm etc.) ein Seitenvorschub durchgeführt. Weiter sind einige Formatierungen der Tabellen in der Gruppe „Tabellenausgabe“ steuerbar. Mit der Option „Spalte trennen durch“ kann man einen Spaltentrenner festlegen. „Leerzeichen“ hat Vorteile, wenn man die Tabelle in ein anderes Programm übernimmt und nicht weiter bearbeiten möchte. Allerdings muss man dann nicht-proportionale Schriften (wie Courier) verwenden. Bei Verwendung von „Tabulatoren“ kann man evtl. im anderen Programm über die Formatierung der Tabulatoren wieder dessen Tabellenfunktion nutzen. Schließlich kann man die Spaltenbreite steuern. Entweder wird diese automatisch angepasst oder man legt eine maximale Zeichenzahl fest. Die für den Rahmen festgelegten ASCII-Zeichen können ebenfalls geändert werden (allerdings macht dies wenig Sinn, da kaum geeignete Zeichen zur Verfügung stehen).

Register „Interaktiv“. Hier können Voreinstellungen für die interaktiven Diagramme verändert werden. Es ist zunächst möglich, zwischen verschiedenen Diagrammvorlagen in einer Auswahlliste zu wählen. Außerdem kann man bestimmen, ob die Diagramme mitsamt den Daten gespeichert werden sollen (sinnvoll, wenn spätere Bearbeitung beabsichtigt) oder ohne diese Daten. Schließlich können Druckerauflösung (Bitmap „Hoch“. „Mittel“ oder „Niedrig“ bzw. „Vektor-Metafile“) und Maßeinheiten („Punkt“, „Zoll“ oder „Zentimeter“) bestimmt werden. In einem weiteren Feld kann für Daten aus früheren (!) SPSS-Versionen bestimmt werden, wie viele Werte eine Variable mindestens umfassen muss, um als metrisch eingestuft zu werden (Voreinstellung: 24)

Register „Skripts“. Hier wird eingestellt, welche Datei die „globalen Prozeduren enthält“, welche die „Autoskripts“ enthält und welche davon aktiviert werden sollen (⇨Kap. 28.3).

Weitere Möglichkeiten bei Verwenden des „Set“-Befehls. Mit dem SET-Kommando im Syntaxfenster können zusätzlich weitere Einstellungen vorgenommen werden, die ansonsten in anderen Dialogboxen eingestellt werden, wie die Basiszahl für den Zufallsgenerator setzen, das Zeichen für die Kommandobegrenzung festlegen, die Interpretation von Leerstellen in Datendateien bestimmen.

Dem SET-Kommando korrespondieren die Kommandos SHOW, PRESERVE und RESTORE.

- Mit dem Kommando SHOW und seinen Optionen können Sie sich die gegenwärtig gültigen Einstellungen ausgeben lassen.
- Die Einstellungen des SET-Befehls gelten für eine Arbeitssitzung, können aber jederzeit während der Sitzung geändert werden. Häufig wird es aber von Interesse sein, eine Grundeinstellung aufzubewahren, um sie wiederverwenden zu können. Die Sicherung geschieht über das Kommando PRESERVE. Mit RESTORE kann man dann während der Sitzung zu dieser Einstellung zurückkehren. Das ist besonders bei Verwendung von Makros interessant.

Alle diese Befehle sind nur über die Befehlssyntax verfügbar, können also nicht aus Dialogboxen in das Syntaxfenster übertragen werden.

28.6 Verwenden des Produktionsmodus

Für größere oder sich wiederholende Analysen kann es interessant sein, nicht mit Hilfe der Menüs oder des Syntaxfensters im sogenannten *Managermodus* zu arbeiten, sondern eine Stapeldatei von SPSS-Syntax-Befehlen aufzurufen und ablaufen zu lassen. Dazu arbeitet man im *Produktionsmodus*. Dort wird der Befehlsstapel ohne Kontrolle im Hintergrund abgearbeitet und das Ergebnis ausgegeben.

Für das Arbeiten im Produktionsmodus müssen Sie zunächst mindestens eine Befehlsdatei erstellen. Es ist gleichgültig, ob Sie das im Syntaxfenster von SPSS für Windows oder in einem Textverarbeitungsprogramm als ASCII-Datei durchführen. Die Befehle müssen in einer oder mehreren Syntaxdateien mit der Extension „SPS“ gespeichert sein. Die Datei muss selbstverständlich in der ersten Befehlszeile eine Datendatei aufrufen.

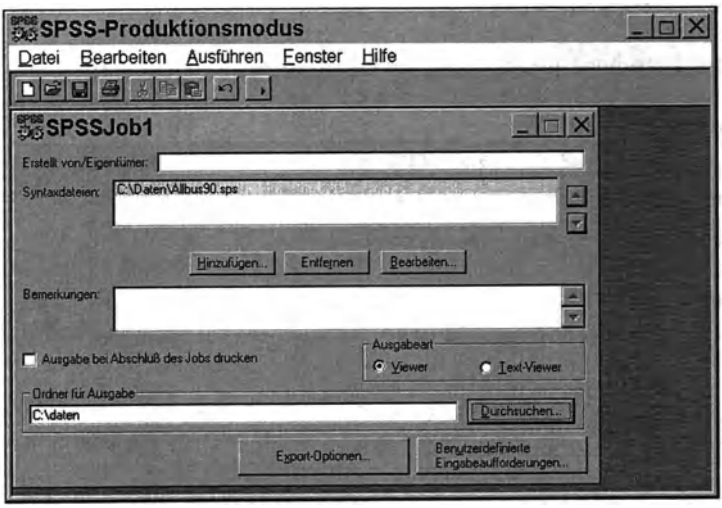


Abb. 28.12 Dialogbox „SPSS-Produktionsmodus“

Beispiel. Es existiert eine Datei ALLBUS90.SPS mit folgendem Inhalt:
GET File = "c:\Daten\ALLBUS90.SAV".
FREQUENCIES VARIABLES=POL,SCHUL.
Diese soll als Produktionsjob im Hintergrund ausgeführt werden. Dazu gehen Sie wie folgt vor:

- ▷ Wählen Sie im Menü „Start" „Programme" und dann aus der Liste der Programme „SPSS Produktionsmodus". Oder Doppelklicken Sie in der Programmgruppe von SPSS auf das Symbol „SPSS Produktionsmodus". Es öffnet sich die Dialogbox „SPSS Produktionsmodus".
- ▷ Legen Sie zunächst fest, welche Syntaxdateie(n) mit diesem Job abgearbeitet werden soll(en). (hier nur: „ALLBUS90.SPS"). Dazu klicken Sie auf die Schaltfläche „Hinzufügen..." unter dem Feld „Syntaxdateien:". Es öffnet sich die Dialogbox „SPSS Syntaxdatei anhängen". Dort wählen Sie in der üblichen Weise Laufwerk und Verzeichnis, in der sich die Syntaxdatei befindet, übertragen den Namen der Syntaxdatei aus der Liste in das Feld „Dateiname" und beenden die Auswahl mit „Öffnen".
- ▷ Wählen Sie den Ordner aus, in dem die Ausgabe des Jobs abgelegt werden soll (hier: „c:\daten"). Dazu klicken Sie auf die Schaltfläche „Durchsuchen..." neben dem Feld „Ordner für Ausgabe". Es öffnet sich die Dialogbox „Ausgabeordner angeben". Hier wählen Sie auf die übliche Weise den gewünschten Ordner aus und bestätigen mit „OK".
- ▷ Jetzt speichern Sie den Job. Sie wählen dazu „Datei", „Speichern unter". Es öffnet sich die Dialogbox „Als Produktionsjob speichern". Hier wählen Sie auf die übliche Weise Laufwerk und Verzeichnis aus und bestimmen einen Namen für die Datei des Produktionsjobs (Voreinstellung: SPSSJob mit angehängter laufender Nummer). Die Extension einer Produktionsjob-Datei ist „SPP". Bestätigen Sie mit „Speichern". So erstellte Jobs können später wieder geladen, evtl. überarbeitet, ergänzt und neu abgearbeitet werden.
- ▷ Falls Sie wünschen, dass das Ergebnis des Jobs ausgedruckt wird, markieren Sie schon in der Dialogbox zuvor das Auswahlkästchen „Ausgabe bei Abschluss des Jobs drucken". Außerdem können Sie durch Anklicken des entsprechenden Optionsschalters bestimmen, ob das Ergebnis im „Viewer" oder im „Text-Viewer" ausgegeben wird.
- ▷ Starten Sie den Job mit „Ausführen", „Produktionsjob" oder durch Anklicken von  . Die Stapeldatei wird abgearbeitet. Der Output wird im angegebenen Verzeichnis in einer Datei mit dem festgelegte Namen und der Extension „SPO" gespeichert, falls nicht „Ausgabe bei Abschluss des Jobs drucken" ausgewählt wurde.

Die so gespeicherte Datei können Sie in SPSS als Ausgabedatei öffnen.

Für das Arbeiten im Produktionsmodus stehen noch einige weitere Optionen zur Verfügung, die hier nur knapp erläutert werden können:

- ❒ **Export Optionen**. Über diese Schaltfläche öffnen Sie die Dialogbox „Export Optionen". Hier kann bestimmt werden, welche Teile der Ausgabe in ein fremdes Format exportiert werden (möglich sind „Viewer" mit oder ohne Diagramme und „Diagramme" alleine, Voreinstellung: „nichts"). Texte und Pivot-

Tabellen können in HTML- oder Textformat exportiert werden, Diagramme in verschiedenen Grafikformaten. Es sind Enhanced Metafile (EMF), Windows-Metafile (WMF), Windows-Bitmap (BMP), PostScript (EPS), JPEG File (JPG), Tagged Image File (TIF), PNG File (PNG) oder MacIntosh PICT (PCT).

- **Benutzerdefinierte Eingabeaufforderung**. Klicken Sie auf diese Schaltfläche, öffnet sich die gleichnamige Dialogbox. In dieser können Makrosymbole definiert werden, die in der Syntaxdatei enthalten sein können. Stößt SPSS auf ein solches Symbol in einer Syntaxdatei, öffnet sich ein Eingabefenster und eine Eingabe wird angefordert. Auf diese Weise könnte man z.B. in der Syntaxdatei ALLBUS90.SPS auf den Namen der Datendatei verzichten und statt dessen das Makrosymbol „@Ddatei" einsetzen. Wird dieses im Produktionsjob definiert, evtl. mit zusätzlicher Bestimmung einer „Eingabeanforderung", sie sei „Dateiname eingeben", so hält der Produktionsjob an, sobald er in der Syntaxdatei auf dieses Symbol stößt. Es öffnet sich ein Fenster mit der Aufforderung „Dateiname eingeben". Nach Eingabe fährt der Job fort. Dadurch wäre es möglich, dieselbe Befehlssyntax für mehrere Dateien (mit unterschiedlichen Dateinamen, aber gleichen Variablen) zu benutzen.
- **Bearbeiten**. Bei Anklicken dieser Schaltfläche öffnet sich ein einfacher Texteditor mit der Syntax der gerade markierten Syntaxdatei. Diese kann hierin bearbeitet und verändert abgespeichert werden.

28.7 Arbeiten mit großen Dateien

Arbeiten mit großen Dateien bringt mehrere Probleme mit sich, für die hier einige Lösungsmöglichkeiten vorgestellt werden.

Auswählen von Variablen aus langen Variablenlisten. Enthält eine Datei sehr viele Variablen, ist es oft schwierig, in der Quellvariablenliste eine Variable aufzufinden. Erleichtert wird das durch die Möglichkeit, mit Eingabe des ersten Buchstabens des Variablennamens (bzw. Labels) zur ersten Variablen in der Liste zu springen, die mit diesem Buchstaben beginnt. In alphabetisch geordneten Listen kann man so schnell die gesuchte Variable finden. Die Bildung kleiner Variablensets im Menü „Extras", mit den Optionen „Sets definieren" und „Sets verwenden" (⇨ Kap. 28.2) erleichtert die Auswahl aus der Quellvariablenliste. Sie können natürlich auch nicht benötigte Variablen löschen und die verkleinerte Datei – möglichst unter neuem Namen – abspeichern. Sinnvoll ist es, Dateien auf diesem Weg in Teildateien zu zerlegen und die Teildateien abzuspeichern. Das Anspringen einer Variablen im Dateneditor ist über die Befehlsfolge „Extras", „Variablen" möglich. Man markiert die gewünschte Variable in dieser Liste und klickt auf die Schaltfläche „Gehe zu".

Umgehen mit zu großen Dateien. Ein Problem besteht, wenn Dateien so viele Variablen umfassen, dass sie nicht in *einer* Quellvariablenliste angezeigt werden können. Maximal kann diese 4500 Variablen enthalten. (SPSS kann wesentlich mehr verarbeiten.) Enthält die Datei mehr Variablen, muss man entweder auf das Arbeiten mit den Windows-Dialogboxen verzichten und die Befehlssyntax benut-

zen oder aber die Datei auf weniger Variablen verkleinern. In diesem Falle muss das schon während des Einlesens geschehen. Zum Einlesen verwendet man die Befehlssyntax (zum Arbeiten mit der Befehlssyntax ⇨ Kap. 4.2). In Frage kommen in erster Linie die Befehle: GET (öffnet eine SPSS-Datei), IMPORT (importiert eine Datei im SPSS-Portable-Format) und GET TRANSLATE (öffnet eine Datei in einem der Formate der unterstützten Tabellenkalkulationsprogramme sowie im dBase oder Tab-delimited Format).

Liest man mit diesen Befehlen eine Datei ein, kann man gleichzeitig Variablen auswählen. Entweder man wählt positiv aus mit dem Befehl KEEP oder negativ mit dem Befehl DROP.

Beispiel.

```
IMPORT FILE=
 'c:\allbus\allbus90\s1800.exp'/DROP v200 to v500.
EXECUTE .
```

Es werden die Daten aus einer SPSS-Portable-Datei namens S1800.EXP eingelesen, die in dem Verzeichnis C:\ALLBUS\ALLBUS90 steht. Dabei werden die Variablen V200 bis V500 ausgeschlossen. Das Schlüsselwort „to" ermöglicht es, eine Reihe aufeinanderfolgender Variablen auf einfache Weise auszuschließen.

DROP kann auch als Unterkommando von SAVE (erzeugt eine Datendatei im SPSS-Format) oder EXPORT (erzeugt eine SPSS-Portable-Datei) verwendet werden, um die Datei zu verkleinern.

Mit dem Befehl MATCH FILES (kombiniert die Variablen zweier Dateien zu einer neuen) wird man solche Dateien wieder kombinieren, aus denen Variablen gemeinsam benötigt werden. Jeweils sind der KEEP und der DROP-Befehl sowie ein RENAME Befehl zum Umbenennen der Variablen verfügbar. Die Dialogboxen enthalten dieselben Optionen, aber (mit Ausnahme der Kombination von Dateien) ohne die Möglichkeit, Variablen auszuwählen. Beim Import von Datenbankdateien besteht generell die Möglichkeit, Variablen auszuwählen.

Sparen von Rechenzeit. Rechenzeit kann man auf verschiedene Weise sparen:

- ❒ Abbrechen eines als falsch erkannten Rechenlaufs durch Stoppen des SPSS-Prozessors mit der Befehlsfolge „Datei", „Prozessor anhalten".
- ❒ Datentransformationen und Berechnungen erst durchführen, wenn ein Rechenlauf benötigt wird. Dazu „Bearbeiten", „Optionen" und das Register „Daten" wählen. Dort ist in der Gruppe „Optionen für Transformieren und Zusammenfügen" der Optionsschalter „Werte vor Verwendung berechnen" auszuwählen.
- ❒ Auswahl von wenigen Fällen für Probeläufe. Dies geschieht über das Menü „Daten" und die Option „Fälle auswählen"(⇨ Kap. 7.4).

28.8 Zum Scrollen und Markieren in den Auswahllisten

Sie können in Auswahllisten (für Variablen, Dateien, Funktionen usw.) mit den Rollbalken oder mit den Richtungstasten scrollen. Beschleunigt wird das, wenn die Richtungstasten <Bild oben> bzw. <Bild unten> (scrollt um ein Bildschirmfenster weiter) oder <Anfang> bzw. <Ende> (scrollt an den Anfang bzw. das Ende der

Liste) benutzt werden. Eine schnelle Möglichkeit, eine gewünschte Variable oder Datei in einer langen Liste anzuwählen, besteht in der Eingabe des Anfangsbuchstabens des Variablen- bzw. Dateinamens. Der Cursor springt auf den ersten Namen in der Liste mit diesem Anfangsbuchstaben. Mehrere nebeneinander stehende Namen können mit der Technik Klicken und Ziehen markiert werden. Mehrere nicht nebeneinanderliegende Namen markieren Sie, indem Sie die <Ctrl>-Taste drücken und die gewünschten Namen anklicken. Außerdem können Sie in der Quellvariablenliste die Ordnung entweder nach Eingabe oder nach dem Alphabet bestimmen. Dies geschieht über „Bearbeiten" und „Optionen" i m Register „Allgemein"(⇨ Kap 28.5). In der Liste bereits ausgewählter Variablen stehen die Variablen in der Reihenfolge ihrer Auswahl und werden auch in dieser Reihenfolge abgearbeitet. Wollen Sie das ändern, ohne die Auswahl rückgängig zu machen, markieren Sie den Namen der zu verschiebenden Variablen. Mit <Alt>+<-> verschieben Sie diese jeweils eine Stelle nach oben, mit <Alt>+<+> eine nach unten. Sie können auch mehrere nebeneinander liegende Variablen gleichzeitig markieren und zusammen verschieben. Statt dessen können Sie auch durch Anklicken des Kästchens in der linken oberen Ecke einer Dialogbox das „Systemmenü" öffnen. Es enthält die Optionen „Auswahl nach oben" und „Auswahl nach unten". Diese bewirken dasselbe wie die Tastenkombinationen. (Da sich das Menü bei jedem Anklicken schließt, ist das bei größeren Verschiebungen sehr umständlich.)

28.9. SPSS-Ausgaben in andere Anwendungen übernehmen

28.9.1 Übernehmen in ein Textprogramm (z.B. Word für Windows)

Tabellenoutput können Sie über die Zwischenablage von Windows entweder als Grafik oder als Tab-delimited Text in eine Textverarbeitungsprogramm übertragen. Je nach Textverarbeitungsprogramm stehen zum Einfügen evtl. verschiedene Formate zur Verfügung. Wir beschreiben hier die Formate von Word für Windows.

- **Als Grafik übertragen**. Sie markieren eine Tabelle oder einen Text und übertragen ihn mit "Bearbeiten", "Kopieren" in die Windows-Zwischenablage. Dann wechseln Sie in das Textprogramm und fügen die Tabelle mit "Bearbeiten", "Inhalte einfügen" und "Grafik" in das Textprogramm ein. (Falls in Ihrer Version vorhanden sollten Sie das Auswahlkästchen „Über den Text legen" ausschalten.) Die Tabelle wird hier als Grafik eingefügt. Der Vorteil besteht darin, dass Umrandungslinien korrekt erhalten bleiben. Allerdings ist keine Textbearbeitung möglich.
- **Als Text übertragen**. Sie verfahren wie beschrieben. Beim Einfügen wählen Sie aber statt „Grafik" die Option „Unformatierten Text". Die Tabelle wird als durch Tabulatoren formatierter Text übertragen. Im Textverarbeitungsprogramm müssen die Tabulatoren erst angepaßt werden, damit die Tabelle wieder richtig formatiert erscheint.
- **Text im RTF-Format einfügen**. Mit dem RTF-Format steht eine weitere Möglichkeit zum Einfügen des Outputs zur Verfügung. Um dies zu benutzen, gehen

Sie wie beschrieben vor, wählen aber beim Einfügen die Option „Formatierter Text (RTF)".

- **Übernehmen mehrerer Objekte zur gleichen Zeit.** Es können auch mehrere Objekte auf einmal übertragen werden. Dazu markieren Sie alle gewünschten Objekte und wählen „Bearbeiten", „Objekte kopieren". Damit übertragen Sie alle markierten Objekte in die Windows-Zwischenablage. Im Textverarbeitungsprogramm setzen Sie den Cursor auf die Einfügestelle und wählen „Einfügen". Die Objekte werden als Grafik übertragen.

28.9.2 Übernehmen von Grafiken

Grafiken werden auf dieselbe Weise übertragen. Sie markieren den Ergebnisoutput im Ausgabefenster und übertragen ihn mit der Befehlsfolge "Bearbeiten", "Kopieren" in die Windows-Zwischenablage. Dann wechseln Sie in das Textprogramm und setzen den Cursor auf die Einfügestelle. Sie wählen "Bearbeiten", "Inhalte Einfügen". In Word für Windows stehen ihnen dann „Grafik", „Bitmap" „Bild (Enhanced Metafile)-Object" oder „Bild (Erweiterte Metadatei)" zur Verfügung. Wählen Sie eines der Formate aus. Eine mit "Grafik" übernommenes Bild kann z.B. im Programm Draw (Doppelklicken auf die Grafik in Word für Windows) überarbeitet werden. Durch Wahl von "Bitmap" erzeugen Sie ebenfalls ein Bild. Die Bilder können in der Regel in Größe und Format geändert werden. Wenn in Ihrer Version vorhanden, sollten Sie in Word für Windows die Option „Über den Text legen" ausschalten, damit die Grafik nicht an eine Position fixiert ist.

28.9.3 Übernehmen von Daten in ein Tabellenkalkulationsprogramm

Das Vorgehen beim Kopieren von Tabellen und Grafiken ist dasselbe, wie beim Textverarbeitungsprogramm beschrieben. Die zum Einfügen verfügbaren Formate hängen z.T. vom benutzten Tabellenkalkulationsprogramm ab. Als Beispiel wird hier lediglich Excel dargestellt. Das Übertragen von Grafiken, Text und Tabellen erfolgt wie in WORD. Allerdings steht für das Einfügen von Text nur die Formate „Text" und „Unicode Text" zur Verfügung. Grafiken können in der gleichen Weise eingefügt werden wie in WORD. Für das Einfügen von Tabellenoutputs steht dagegen ein zusätzliches Format „Biff" zur Verfügung.

Fügt man eine Tabelle mit „Bild (Erweiterte Metadatei)" ein, erhält man eine Grafik, deren Zahlenwerte in Excel nicht weiter verarbeitbar sind. Durch Einfügen mit „Text" „Unicode Text" oder „Biff" erhält man dagegen eine echte Excel-Tabelle. Die Zahlen können in Excel zu weiteren Berechnungen verwendet werden. Während die mit der Option „Text" aber die Zahlen nur mit der im SPSS-Output angezeigten Genauigkeit übernommen werden, bleibt bei Übernahme mit „Biff" die numerische Genauigkeit vollständig erhalten, d.h. es werden sämtliche in SPSS intern verwendeten Nachkommastellen mit übertragen.

Außerdem kann man in der Pivot-Tabelle auch vor der Übertragung erst die Teile auswählen, die übernommen werden sollen (⇨Kap 4).

28.9.4 Einbetten einer Pivot-Tabelle in eine andere Anwendung

Wenn die andere Anwendung ActiveX-Objekte unterstützt, können Sie auch eine Pivot-Tabelle einbetten. D.h., Sie können diese durch Doppelklicken aktivieren und in der anderen Anwendung wie in SPSS pivotieren. Dazu müssen Sie zuerst außerhalb von SPSS eine Datei „objs-on.bat" starten, die sich in dem Verzeichnis befindet, in dem Sie SPSS installiert haben. Beim Kopieren und Einfügen gehen Sie wie beschrieben vor. In der Dialogbox „Inhalte einfügen" finden Sie aber als weitere Option „SPSS-Pivot-Tabelle Steuerungselement" oder „"SPSS Interactive Graph Steuerungselemente". Diese wählen Sie aus und bestätigen mit „OK". Um eine Pivot-Tabelle einzubetten, benötigen Sie aber sehr viel Arbeitsspeicher. Für den Normalanwender ist eher davon abzuraten. Wenn Sie keine Einbettung von Pivot-Tabellen mehr wünschen, deaktivieren Sie die Einbettungsfunktion, indem Sie außerhalb von SPSS das Programm objs-off.bat starten. Sie befindet sich in dem Verzeichnis, in dem Sie SPSS installiert haben.

29 Exakte Tests

Einführung. Die in Kap. 13.3 dargestellte Vorgehensweise beim Testen von Hypothesen verwendet Testverteilungen, d.h. geht davon aus, dass die berechnete Prüfgröße einer bekannten und in Tabellenform vorliegenden theoretischen Verteilung (z.B. t-Verteilung, Standardnormalverteilung, Chi-Quadrat-Verteilung) folgt. Einschränkend muss man präzisieren, dass es sich um eine Approximation handelt: die Prüfgröße entspricht annähernd einer theoretischen Verteilung. Dabei gilt: je größer der Stichprobenumfang n ist, um so besser ist die Approximation. Man spricht daher auch von asymptotischen Tests.

Für die Chi-Quadrat-Tests, z.B den Unabhängigkeitstest (⇨ Kap. 10.2), der auf Kreuztabellen beruht, muss für die Approximation gewährleistet sein, dass der Stichprobenumfang nicht zu klein ist. Weiterhin muss eine „ausgewogene" Stichprobe vorliegen, d.h. die Zellenbesetzungen dürfen in allen Zellen der Kreuztabelle nicht zu klein und auch nicht konzentriert verteilt sein. Da diese Voraussetzungen in der empirischen Praxis nicht immer erfüllt sind, führt eine Anwendung asymptotischer Tests unter Umständen zu falschen Ergebnissen, d. h. zur falschen Hypothese.

Auch bei nichtparametrischen Tests stützt man sich auf Testverteilungen verschiedenster Art für die Prüfgrößen und führt insofern dann asymptotische Tests durch. Daher besteht das Risiko, dass bei kleinen Stichprobenumfängen fehlerhaft entschieden wird. Auch zu viele Bindungen (ties) sind problematisch für asymptotische Tests.

Will man Fehlermöglichkeiten hinsichtlich der Hypothesenentscheidung vermeiden, so muss man bei kleinen und unausgewogenen Stichproben exakte Tests durchführen. Auch bei exakten Tests werden die in Kapitel 13.3 dargestellten Schritte durchgeführt. Aber im Unterschied zu oben stützt man sich bei den Testverteilungen nicht auf bekannte theoretische Verteilungen, sondern es werden die Wahrscheinlichkeitsverteilungen der Prüfgrößen eigens für die Daten einer vorliegenden Stichprobe berechnet. Am Beispiel des auf Kreuztabellen basierenden Chi-Quadrat-Unabhängigkeitstest mit der Prüfgröße χ^2 (⇨ Gleichung 10.2) soll dieses näher erläutert werden. Im ersten Schritt wird die Prüfgröße χ^2 für alle denkbar möglichen Kreuztabellen berechnet, die die gleiche Zeilen- und Spaltenzahl und die gleichen Randsummenhäufigkeiten haben wie die als Stichprobe vorliegende empirische Kreuztabelle. Im nächsten Schritt werden alle Tabellen identifiziert, deren Prüfgröße χ^2 gleich bzw. größer ist als die der vorliegenden empirischen Tabelle. Die Häufigkeiten dieser Tabellen reflektieren noch stärkere Abweichungen von der H_0-Hypothese als die der empirischen Tabelle. Für jede dieser so be-

stimmten Tabellen wird dann die (hypergeometrische) Wahrscheinlichkeit ihres Auftretens berechnet. Die exakte Wahrscheinlichkeit P ergibt sich als Summe der Einzelwahrscheinlichkeiten. P ist also die Wahrscheinlichkeit, dass bei Geltung von H_0 der empirisch berechnete bzw. ein höherer Prüfgrößenwert zustande kommt. P wird - wie auch bei den asymptotischen Tests - mit dem Signifikanzniveau α verglichen. Bei $P > \alpha$ entscheidet man sich für die Hypothese H_0 und bei $P < \alpha$ für H_1.

Die Berechnung der P-Werte ist rechenaufwendig. Bei einer z.B. 5*6-Tabelle handelt es sich dabei um ca. 1,6 Millionen verschiedenen Tabellen mit gleichen Randverteilungen.

Ab der Vers. 6.1.2 von SPSS für Windows erlaubt das auf das Basismodul aufsetzende Ergänzungsmodul „Exact Tests" die exakte Berechnung von P durchzuführen. Dazu muss natürlich „Exact Tests" installiert sein. Da bei sehr großen Kreuztabellen (viele Spalten und Zeilen) und bei hohen Stichprobenumfängen für nichtparametrische Tests die Berechnung der Prüfgrößenverteilung sowie der Wahrscheinlichkeit P für die Prüfgröße sowohl aus Speicherplatz- als auch Rechenzeitgründen nicht möglich ist, bietet SPSS neben den asymptotischen Tests und der exakten Berechnung von P auch eine Schätzung des exakten Wertes von P mit Hilfe des Monte-Carlo-Verfahrens an. Bei diesem zweiten Verfahren werden aus der Verteilung der Prüfgröße zufällig z.B. 10000 ausgewählt und die dadurch entstehende Wahrscheinlichkeitsverteilung der Prüfgröße zur Grundlage für die Berechnung von Signifikanztest genommen. Für die empirisch berechnete Prüfgröße wird für das vorzugebene Signifikanzniveau (z.B. $\alpha = 0{,}05$) die Wahrscheinlichkeit P für das Auftreten der Prüfgröße berechnet. Außerdem wird für die berechnete Wahrscheinlichkeitswert P ein Konfidenzintervall ermittelt.

Unter bestimmten Bedingungen werden bei Anfordern des Monte Carlo-Verfahrens tatsächlich exakte P-Werte ausgegeben. In Tabelle 29.1 wird dafür eine Übersicht gegeben.

Tabelle 29.1. Bedingungen für die Ausgabe von exakten Tests

Test	SPSS-Prozedur	Bedingung
Binomial	Nichtparametr. Tests	stets exakt
Fisher's exakt	Kreuztabellen	2*2-Tabelle
Likelihood-ratio	Kreuztabellen	2*2-Tabelle
Linear-by-Linear A.	Kreuztabellen	2*2-Tabelle
McNemar	Nichtparametr. Tests	stets exakt
Median	Nichtparametr. Tests	$k = 2, n \leq 30$
Pearson Chi-Quadrat	Kreuztabellen	2*2-Tabelle
Sign	Nichtparametr. Tests	$n \leq 25$
Wald-Wolfowitz	Nichtparametr. Tests	$n \leq 30$

Für Stichprobenumfänge ≤ 30 und 3*3-Kreuztabellen bzw. kleiner ist eine exakte Berechnung von P einigermaßen schnell möglich. Bei 2*2-Tabellen darf der Stichprobenumfang sogar bis zu 100000 groß sein. Falls SPSS aus Gründen mangeln-

den Speicherplatzes das exakte P nicht berechnen kann, bricht die Prozedur ab. Dann sollte man das Monte Carlo-Verfahren einsetzen. Unter Umständen kann der Zeitbedarf zur Berechnung von P sehr hoch sein. Mit der Befehlsfolge „Datei", „Prozessor anhalten" kann man einen Berechnungsprozess abbrechen, um dann das Monte Carlo-Verfahren einzusetzen.

Ein Anwendungsbeispiel. Anhand eines Beispiels soll die Vorgehensweise näher erläutert werden. Mit Hilfe des Chi-Quadrat-Unabhängigkeitstests (⇨ Kap. 10.2) soll geprüft werden, ob der für alle Altersgruppen signifikante Zusammenhang zwischen dem politischen Interesse (POL) und dem Geschlecht eines Befragten (GESCHL) auch für die Altersgruppe der 18-29jährigen besteht (Datei ALLBUS90.SAV). Die Beschränkung der Auswertung auf die Altersgruppe geschieht über die Befehlsfolge „Daten", „Fälle auswählen" (ALT2 = 1). Gemäß der in Kapitel 10.1 und 10.2 beschriebenen Vorgehensweise wird dann die Dialogbox „Kreuztabellen" aufgerufen und es werden die Variablen GESCHL und POL als Zeilen- bzw. Spaltenvariable übertragen (⇨ Abb. 29.1). Danach wird nach Klicken der Schaltfläche „Statistik..." in der Dialogbox „Kreuztabellen: Statistik" „Chi-Quadrat" gewählt. Um neben den asymptotischen Test auch einen exakten Test anzufordern, wird jetzt die Schaltfläche „Exakt..." geklickt. Es öffnet sich die in Abb. 29.2 dargestellte Dialogbox „Exakte Tests". Man kann nun zwischen „Nur asymptotisch", „Exakt" und „Monte Carlo" wählen. „Nur asymptotisch" entspricht den Ergebnissen, die man bei einem Verzicht auf Durchführung von exakten Tests erhält. Bei Wahl von „Exakt" kann eine obere Zeitgrenze für den Test angegeben werden. Die Zeitgrenze ist standardmäßig auf 5 Minuten festgelegt und kann erhöht oder verringert werden. Bei der Wahl von „Monte Carlo" ist standardmäßig eine Zufallsauswahl von 10000 Stichproben aus der Verteilung der Prüfgröße festgelegt. Man kann die Anzahl der Stichproben verkleinern oder bis auf 1 Millionen erhöhen. Eine höhere Anzahl von Stichproben erhöht die Güte des Schätzwertes von P, verkleinert die Breite des ausgegebenen Konfidenzintervalls, benötigt aber mehr Rechenzeit. Mit der Ausgabe eines unverzerrten Schätzwertes für den exakten P-Wertes wird auch ein Konfidenzintervall für diesen P-Wert angegeben. Standardmäßig wird ein 99 %-Konfidenzintervall berechnet. Durch Überschreiben kann dieses wunschgemäß zwischen 0,01 und 99,9 verändert werden. Das Monte Carlo-Verfahren stützt sich auf den Zufallsgenerator von SPSS. Wenn man das Ergebnis der Monte Carlo-Schätzung wiederholen möchte, so muss man jeweils vorher einen Startwert des Zufallsgenerators mit der Befehlsfolge „Transformieren", „Startwert für Zufallszahlen" festlegen bzw. bestätigen (⇨ Kap. 7.4.2).

In Tabelle 29.2 ist das Ergebnis der Kreuztabellierung mit den Chi-Quadrat-Test-Ergebnissen dargestellt. In einer Warnungsmeldung wird angezeigt, dass 60 % der Zellen der Kreuztabelle eine erwartete Häufigkeit kleiner 5 haben. Damit wird eine Bedingung für die Zuverlässigkeit des asymptotischen Chi-Quadrat-Test verletzt (⇨ Kap. 10.2 und 22.2.1). Ein exakter Test ist daher angebracht. Für den asymptotischen Chi-Quadrat-Test wird ein (zweiseitiger) P-Wert von 0,133 ausgewiesen. Legt man das Signifikanzniveau für den Test auf $\alpha = 0{,}05$ (= 5 %) fest, so wird wegen $0{,}133 > 0{,}05$ die Hypothese H_0 (kein Zusammenhang) angenommen. Auch der exakte Test kommt mit P = 0,123 (zweiseitig) zum gleichen Testergebnis. In diesem Beispiel kommt der asymptotische Test trotz Verletzung der Anwendungs-

bedingungen zum gleichen Ergebnis wie der exakte Test. In Kapitel 22.3.3 wird in einem Beispiel für den Kolmogorov-Smirnov Z-Test deutlich, dass sich Ergebnisse der exakten Tests von denen der asymptotischen unterscheiden können.

Bei den exakten Tests wird für den Test „Zusammenhang linear-mit-linear" neben dem Wert von „Exakte Signifikanz" (dem exakten P-Wert) auch ein Wert für „Punkt-Wahrscheinlichkeit" für das Eintreffen der empirischen Prüfgröße ausgegeben. Dieser Wert ist ein Maß für die Diskretheit der exakten Verteilung der Prüfgröße. Von manchen Statistikern wird empfohlen, die Hälfte des Wertes von dem exakten P-Wert abzuziehen und für die Hypothesenentscheidung diese Differenz mit dem α-Wert zu vergleichen.

In Tabelle 29.3 wird das Testergebnis mit Hilfe des Monte Carlo-Verfahrens für ein angefordertes Konfidenzniveau von 99 % ausgewiesen. Das zweiseitige Signifikanzniveau wird mit P = 0,131 ausgewiesen (basierend auf 10000 Stichprobentabellen mit dem Startwert 1993510611). Das 99 %-Konfidenzintervall weist die Grenzen 0,122 und 0,139 aus. Auch das mit Hilfe der Monte Carlo Methode gewonnene Testergebnis führt zur Annahme der H_0-Hypothese.

Abb. 29.1. Dialogbox „Kreuztabellen"

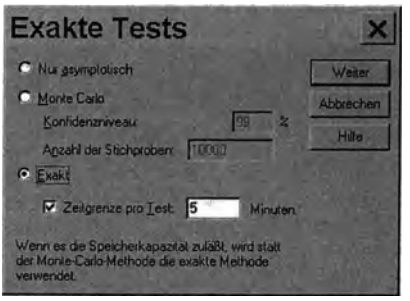


Abb. 29.2. Dialogbox „Exakte Tests"

Tabelle 29.2. Chi-Quadrat-Test für die Kreuztabelle Politisches Interesse nach Geschlecht: Exakter Test

GESCHL * POL Kreuztabelle

			POL					
			SEHR STARK	STARK	MITTEL	WENIG	UEBERHAUPT NICHT	Gesamt
GESCHL	MAENNLICH	Anzahl	6	9	11	1	1	28
		Erwartete Anzahl	3,2	8,6	12,6	2,7	,9	28,0
	WEIBLICH	Anzahl	1	10	17	5	1	34
		Erwartete Anzahl	3,8	10,4	15,4	3,3	1,1	34,0
Gesamt		Anzahl	7	19	28	6	2	62
		Erwartete Anzahl	7,0	19,0	28,0	6,0	2,0	62,0

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)	Punkt-Wahrscheinlichkeit
Chi-Quadrat nach Pearson	7,062[a]	4	,133	,123		
Likelihood-Quotient	7,640	4	,106	,145		
Exakter Test nach Fisher	6,949			,115		
Zusammenhang linear-mit-linear	4,388[b]	1	,036	,038	,024	,012
Anzahl der gültigen Fälle	62					

a. 6 Zellen (60,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist ,90.

b. Die standardisierte Statistik ist 2,095.

Tabelle 29.3. Chi-Quadrat-Test für die Kreuztabelle Politisches Interesse nach Geschlecht: Monte Carlo-Verfahren

Chi-Quadrat-Tests

				Monte-Carlo-Signifikanz (2-seitig)			Monte-Carlo-Signifikanz (1-seitig)		
					99%-Konfidenzintervall			99%-Konfidenzintervall	
	Wert	df	Asymptotische Signifikanz (2-seitig)	Signifikanz	Untergrenze	Obergrenze	Signifikanz	Untergrenze	Obergrenze
Chi-Quadrat nach Pearson	7,062[a]	4	,133	,131[b]	,122	,139			
Likelihood-Quotient	7,640	4	,106	,150[b]	,140	,159			
Exakter Test nach Fisher	6,949			,122[b]	,113	,130			
Zusammenhang linear-mit-linear	4,388[c]	1	,036	,039[b]	,034	,044	,026[b]	,022	,030
Anzahl der gültigen Fälle	62								

a. 6 Zellen (60,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist ,90.

b. Basierend auf 10000 Stichprobentabellen mit dem Startwert 1993510611.

c. Die standardisierte Statistik ist 2,095.

Anhang A

Datei ALLBUS: Variablen zu Kapitel 2 (Variablendefinitionen in Kap. 2)

LFDNR	NR	VN	GESCHL	SCHUL	EINK	POL	RUHE	EINFLUSS	INFLATIO	MEINUNG	TREUE
31	1	1	2	3	4000	3	1	2	4	3	1
126	2	1	2	1	250	4	2	3	4	1	4
690	3	1	1	3	99997	1	4	1	3	2	1
701	4	1	2	5	99997	3	2	3	4	1	0
897	5	1	1	4	3200	1	4	1	3	2	4
1144	6	1	3	4	4000	1	2	3	4	1	3
1186	7	1	1	2	2300	3	3	1	2	4	2
1459	8	2	1	3	99997	2	3	1	4	2	0
1776	9	1	2	3	0	4	1	2	4	3	2
2104	10	1	1	2	2000	3	1	3	4	2	3
2127	11	2	1	2	1500	4	2	3	4	1	0
2205	12	2	1	2	2500	2	1	4	3	2	0
2278	13	2	1	2	2600	2	1	4	3	2	0
2316	14	2	2	4	1000	1	3	2	4	1	0
2372	15	1	2	2	0	3	4	1	2	3	2
2568	16	2	2	2	0	3	2	3	4	1	0
2599	17	1	1	5	445	2	4	1	3	2	4
2610	18	2	2	5	600	1	4	1	3	2	0
2714	19	2	1	2	1400	3	1	4	2	3	0
2724	20	1	1	2	4500	3	3	2	4	1	1
2790	21	1	1	2	2400	1	2	1	4	3	4
2811	22	2	2	3	99997	4	3	2	4	1	0
3175	23	1	2	2	0	4	4	1	3	2	1
3537	24	1	1	3	2000	2	4	3	2	1	4
3831	25	1	2	3	0	4	1	4	3	2	2
3848	26	2	1	2	99997	2	4	1	3	2	0
4905	27	1	1	5	1000	2	4	1	3	2	4
4943	28	1	1	2	99997	2	1	3	2	4	2
4970	29	1	2	2	0	3	2	1	4	3	1
4124	30	1	2	3	0	3	3	2	4	1	3
5156	31	2	1	3	2640	3	1	4	3	2	0
5167	32	1	2	5	1000	3	3	2	4	1	4

Hinweise: LFDNR ist die originale Fallnummer der ALLBUS-Datei. NR ist die von den Autoren vergebenen Fallnummer. Sie ist in dieser Datei identisch mit der hier nicht angeführten automatisch von SPSS vergebenen Fallnummer. Um die Datenbereinigung demonstrieren zu können, sind bei den Fällen 6 für GESCHLECHT und 4 für TREUE zunächst falsche Werte eingegeben. Diese müssen korrigiert werden (Fall 6: GESCHL = 1 und Fall 4: TREUE = 3). Die Datei ALLBUS90.SAV (⇨ Anhang B) enthält die bereits korrigierten Daten.

Quelle: ALLBUS 1990

Anhang B

Dateienservice

Falls Sie die im Buch verwendeten Datendateien bzw. Ergänzungstexte (s.u.) haben möchten, so können Sie diese auf folgenden Wegen erhalten:

1. **Per Post.**
   Senden Sie eine formatierte MS-DOS-Diskette (3,5 Zoll) und einen ausreichend frankierten und an Sie selbst adressierten Rückumschlag an

   Jürgen Janssen
   Hochschule für Wirtschaft und Politik
   Von Melle-Park 9
   20146 Hamburg

   Die Datendateien und Ergänzungstexte (gepackt) werden Ihnen dann auf Ihrer Diskette in Ihrem Rückumschlag zugeschickt.

2. **Per Internet.**
   Unter folgender Internet-Adresse finden Sie den Zugang zu den (gepackten) Dateien:

   http://www.hwp-hamburg.de/JanssenJ/spss.html

E-mail: JanssenJ@hwp-hamburg.de

Ergänzungstexte (in MS Word)

1. Zum Hilfesystem von SPSS für Windows.
2. Zur Mehrweg-Varianzanalyse der Vorgängerversionen.
3. Zum Übernehmen von Daten aus Datenbanken über die ODBC-Schnittstelle in früheren Versionen.
4. Zur Erklärung der Varianz durch Polynome (Kap 14.5).
5. Tukeys Additivitätstest (Kap. 23).

Literaturverzeichnis

Backhaus, K., Erichson, B., Plinke, W., Weber, R., Multivariate Analysemethoden. Eine anwendungsoriente Einführung. Berlin u.a. 2000.

Bleymüller, J., G. Gehlert, H. Gülicher, Statistik für Wirtschaftswissenschaftler, München 2000.

Blom, G., Statistical estimates and transformed beta variables, New York 1958.

Böltken, F., Auswahlverfahren. Eine Einführung für Sozialwissenschaftler, Stuttgart 1976.

Bortz, J., G.A. Lienert, K. Boenke, Verteilungsfreie Methoden in der Biostatistik, Berlin et. al. 1990.

Büning, H., G. Trenkler, Nichtparametrische Methoden. Berlin, New York 1978.

Chambers, J.M., W.S. Cleveland, B. Kleiner, P.A. Tukey, Graphical methods for data analysis, Belmont 1983.

Claus, G., H. Ebner, Grundlagen der Statistik, Thun und Frankfurt a.M. 1977.

Cleveland, W.S., Robust locally weighted regression and smoothing scatterplots, in: Journal of the American Statistical Association, 74/ 1979, S. 829-836.

Cochran, W., Stichprobenverfahren, Berlin, New York 1973.

Cohen, J., Statistical Power Analysis for the Behavioral Sciences, Hillsdale, New Jersey 1988.

Freyhold, M. v., Autoritarismus und politische Apathie. Analyse einer Skala zur Ermittlung autoritätsgebundenen Verhaltens, Frankfurt a.M. 1971.

Inglehart, R., The silent revolution in Europe: Intergenerational change in post-industrial societies, in: American Political Science Review 65/1971, S. 991-1017.

Krüger, B., H. Ritter, C. Züll, SPSS Einsatz auf unterschiedlichen Plattformen in einem Netzwerk: Daten- und Ergebnisaustausch (ZUMA-Arbeitsbericht Nr. 93/17), ZUMA, Mannheim.

Kuritz, S.J., J.R. Landis, G.G. Koch, A general overview of Mantel-Haenszel methods: Applications and recent developments. In: Annual of Public Health, 9 (1988), S. 123-160.

Laatz, W., Empirische Methoden. Ein Lehrbuch für Sozialwissenschaftler, Frankfurt a.M. 1993.

Lehman, E.L., Nonparametrics: Statistical methods based on ranks, San Francisco 1975.

Siegel, S., Nonparametric statistics for the behavioral sciences, New York et. al. 1956.
Steinhausen, D., Langer, K., Clusteranalyse. Einführung in Methoden und Verfahren der automatischen Klassifikation. Berlin, New York 1977.
SPSS Inc., Statistical Algorithms, 2nd Edition, Chicago, Illinois 1991.
SPSS Inc., SPSS Base 8.0. Application Guide, Chicago 1998.
SPSS Inc., SPSS Base 11.0. Benutzerhandbuch, München 2001.
SPSS Inc., SPSS Exact Tests 6.1 for Windows, Chicago 1995.
Stenger, H., Stichproben, Heidelberg, Wien 1986.
Steinhausen, D.,Langer, K., Clusteranalyse. Einführung in Methoden und Verfahren der automatischen Klassifikation. Berlin, New York 1977.
Tukey, J.W., The future of data analysis, in: Annals of Mathematical Statistics, 33/1962.
Wolf, W., Statistik. Eine Einführung für Sozialwissenschaftler, Band 1, Weinheim und Basel 1974.

Bezugsquelle für den ALLBUS 1990:
Zentralarchiv für Empirische Sozialforschung. Universität zu Köln, Bachemer Str. 40, 50931 Köln

Sachverzeichnis

L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz

Statistik

Der Weg zur Datenanalyse

Das Buch bietet eine integrierte Darstellung der deskriptiven Statistik, moderner Methoden der explorativen Datenanalyse und der induktiven Statistik.

4., verb. Aufl. 2002. XVI, 608 S. 162 Abb., 25 Tab. (Springer-Lehrbuch) Brosch. **€ 29,95**; sFr 48,- ISBN 3-540-44000-3

Ergänzung zum Buch „Statistik"

L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz, A. Caputo, S. Lang

Arbeitsbuch Statistik

Es enthält die Lösungen zu den dort gestellten Aufgaben. Es dient damit der Vertiefung und der Einübung des im Lehrbuch vermittelten Stoffes zur Wahrscheinlichkeitsrechnung, deskriptiven und induktiven Statistik.

3., überarb. u. erw. Aufl. 2002. VIII, 284 S. 161 Abb. (Springer-Lehrbuch) Brosch. **€ 14,95**; sFr 24,- ISBN 3-540-44030-5

W. Stier

Methoden der Zeitreihenanalyse

Dieses Lehrbuch vermittelt einen umfassenden Überblick über die wichtigsten Methoden der Zeitreihenanalyse.
Das vollständige Inhaltsververzeichnis finden Sie unter: http://www.springer.de/books/toc-ascii/3540417001-c.txt

2001. XI, 400 S. 237 Abb., 6 Tab. (Springer-Lehrbuch) Brosch. **€ 34,95**; sFr 56,- ISBN 3-540-41700-1

H.-P. Nissen, Universität Paderborn

Das Europäische System Volkswirtschaftlicher Gesamtrechnungen

Das Buch informiert über die neuen Begrifflichkeiten und definitorischen Abgrenzungen.
Das Inhaltsverzeichnis finden Sie unter: http://www.springer.de/books/toc/379081444x-c.pdf

4., vollst. überarb. Aufl. 2002. XVII, 360 S. 51 Abb., 7 Tab. (Physica-Lehrbuch) Brosch. **€ 24,95**; sFr 40,- ISBN 3-7908-1444-X

J. Janssen, W. Laatz

Statistische Datenanalyse mit SPSS für Windows

Eine anwendungsorientierte Einführung in das Basissystem und das Modul Exakte Texts

Es werden das Basissystem von SPSS für Windows sowie das Ergänzungsmodul Exact Tests behandelt. Grundlage ist die Programmversion 11.

4., neubearb. u. erw. Aufl. 2002. XVI, 722 S. 550 Abb. Brosch. **€ 36,95**; sFr 59,50 ISBN 3-540-44002-X

Besuchen Sie uns unter:
http://www.springer.de/economics
www.springer.de/math-de

Springer · Kundenservice
Haberstr. 7 · 69126 Heidelberg
Tel.: (0 62 21) 345 - 217/-218 · Fax: (0 62 21) 345 - 229
e-mail: orders@springer.de

Die €-Preise für Bücher sind gültig in Deutschland und enthalten 7% MwSt.
Preisänderungen und Irrtümer vorbehalten. d&p · BA 44002/1

Springer