LERNEN EINFACH GEMACHT

2. Auflage

# Statistik mit SPSS

für dummies®

## Alles-in-einem-Band

Vom Fragebogen über erste Auswertungen zu Datendateien

Rundumschlag: deskriptive und induktive Statistik mit SPSS

Eigene Syntaxfiles freigeistig schreiben

Jürgen Faik

# Statistik mit SPSS Alles in einem Band für Dummies

# Schummelseite

## DATENANALYSE MIT SPSS

- *Daten einlesen*: manuell direkt über den SPSS-Dateneditor, Erfassung über andere Programme wie EXCEL, Erfassung mit einem Datenerfassungsprogramm oder Erfassung über eine Erfassungsmaske eines Datenbankprogramms
- *Daten formatieren/verarbeiten*: neue Variablen bilden, Variablen recodieren, Variablen und Werteetikette vergeben, Daten aggregieren, Daten filtern, Daten gewichten
- *Daten auswerten:* statistische Prozeduren von SPSS wie zum Beispiel FREQUENCIES oder DESCRIPTIVES verwenden oder beispielsweise Schnittstelle zu R nutzen
- *Daten interpretieren:* prüfen, ob statistische SPSS-Auswertung dem Skalenniveau der Variablen oder überhaupt der Fragestellung angemessen ist, sowie statistische Methodenkenntnisse anwenden und Ergebnissignifikanz beachten

## SCHNITTSTELLEN ZU ANDEREN PROGRAMMEN

- *EXCEL, WORD, …:* Importieren und Exportieren von Daten aus/nach andere(n) Programme(n)
- *R, Python, …:* Einbinden von statistischen Prozeduren aus anderen Programmen (auch umgekehrt SPSS-Einbezug zum Beispiel in R möglich)

## SPSS-DATEIARTEN

- *SAV-Datendateien:* stellen Rohdaten oder bereits erstellte Datensätze im SPSS-Datenformat zur Verfügung
- *SPS-Syntaxdateien:* für mehrere aufeinanderfolgende Anweisungen in der SPSS-»Sprache«, die dann auf eine *SAV*-Datendatei bezogen werden können
- *SPV-Outputdateien:* halten die Ergebnisse einer statistischen Analyse vor allem in Form von (Pivot-)Tabellen und Diagrammen fest, können in andere Programme (wie zum Beispiel EXCEL) eingebunden werden
- *JNL-Journaldateien:* protokollieren eine SPSS-Sitzung, sind nicht so wichtig wie *SAV*-, *SPS*- oder *SPV*-Dateien

## ÜBERSICHT ÜBER DIE STATISTISCHEN SPSS-PROZEDUREN

- *DESCRIPTIVES:* Summe, arithmetischer Mittelwert, Minimum, Maximum, Spannweite, Varianz, Standardabweichung, Schiefe, Wölbung, Standardfehler (in Bezug auf den arithmetischen Mittelwert)
- *FREQUENCIES:* Summe, arithmetischer Mittelwert, (gruppierter) Median, Modus, Quantile, Perzentile, Minimum, Maximum, Spannweite, Varianz, Standardabweichung, Schiefe, Wölbung, Standardfehler (in Bezug auf den arithmetischen Mittelwert)
- *EXAMINE:* arithmetischer Mittelwert, (gruppierter) Median, Perzentile, Interquartilbereich, Minimum, Maximum, Spannweite, Varianz, Standardabweichung, Schiefe, Wölbung, Standardfehler (in Bezug auf den arithmetischen Mittelwert)
- *SUMMARIZE und MEANS:* Summe, arithmetischer Mittelwert, (gruppierter) Median, harmonisches Mittel, geometrisches Mittel, Minimum, Maximum, Spannweite, Varianz, Standardabweichung, Schiefe, Wölbung, Standardfehler (in Bezug auf den arithmetischen Mittelwert, die Schiefe und die Wölbung)

## ÜBERSICHT ÜBER DIE MITTELWERTTESTS IN SPSS

| *Mindestskalenniveau der Daten* | *2 Gruppen: Unabhängige Stichproben* | *2 Gruppen: Abhängige Stichproben* | *Mehr als 2 Gruppen: Unabhängige Stichproben* | *Mehr als 2 Gruppen: Abhängige Stichproben* |
|---|---|---|---|---|
| Metrische Daten | Mittelwertdifferenzentest (t-Test) für unabhängige Stichproben beziehungsweise Welch-Test | Mittelwertdifferenzentest (t-Test) für abhängige Stichproben | Einfaktorielle ANOVA ohne Messwiederholungen | Einfaktorielle ANOVA mit Messwiederholungen |
| Ordinale Daten | Mann/Whitney-U-Test | Wilcoxon-Test beziehungsweise Vorzeichentest | Kruskal/Wallis-Test | Friedman-Test |

## ÜBERSICHT ÜBER ASSOZIATIONEN, KORRELATIONEN UND REGRESSIONEN

- ✔ *Zusammenhänge zwischen zwei Variablen (bivariate Zusammenhänge):*
  - Assoziation (erforderliches Skalenniveau: nominalskalierte Variablen)
  - Rangkorrelation (erforderliches Skalenniveau: ordinalskalierte Variablen)
  - Korrelation (erforderliches Skalenniveau in der Regel: metrisch skalierte Variablen; ungerichtet, das heißt, dass eine Ursache-Wirkungs-Beziehung keine Rolle spielt)
  - Regression (erforderliches Skalenniveau: metrisch skalierte Variablen; gerichtet im Sinne einer Ursache-Wirkungs-Beziehung)
- ✔ *Zusammenhänge zwischen mehr als zwei Variablen:*
  - Multiple Regression (erforderliches Skalenniveau: metrisch skalierte Variablen, aber auch Dummies als unabhängige Variablen möglich)
  - Logistische Regression (erforderliches Skalenniveau: dichotome Variablen als abhängige Variablen) und Ähnliches
  - Diskriminanz-, Cluster- und Faktorenanalyse

## ÜBERLEBENSANALYSEN

- ✔ *Sterbetafelanalyse:* Ermittlung einer Überlebens- und einer Hazardrate (und einer Sterbetafel)
- ✔ *Kaplan/Meier-Methode:* Schätzung einer Überlebens- und einer Hazardfunktion auf der Basis einer unabhängigen kategorialen Variablen
- ✔ *Cox-Regression:* regressionsanalytische Schätzung von Überlebenswahrscheinlichkeiten mit mehreren unabhängigen (auch metrisch skalierten) Variablen

## ÜBERSICHT ÜBER WICHTIGE REINE SCHLEIFEN UND »SCHLEIFEN-MAKROS« ZUM VEREINFACHEN DER PROGRAMMIERUNG IN SPSS

- ✔ *Schleifen:* DO-REPEAT-Schleifen, LOOP-Konstruktionen, DO-IF-Befehl
- ✔ *Makros:* DO-Loops, DO-IN-Loops

## ZU SPSS-BEFEHLEN

- ✔ *Allgemein:*

  BEFEHLSSCHLÜSSELWORT VARIABLE(N) OPERATION

  /UNTERBEFEHL(E)

- ✔ *Beispiel für einen speziellen SPSS-Befehl* (Aufruf der statistischen FREQUENCIES-Prozedur für die Variable *x1* ohne Häufigkeitstabelle, aber mit Balkendiagramm und dort mit Prozentwerten auf der senkrechten Achse sowie Ausweis des Medians und des arithmetischen Mittels für *x1*):

  FREQUENCIES VARIABLES = *x1*

  /FORMAT = NOTABLE

  /BARCHART PERCENT

  /STATISTICS = MEDIAN MEAN.

Jürgen Faik

# Statistik mit SPSS Alles in einem Band

für dummies®

2. Auflage

WILEY
WILEY-VCH GmbH

Телеграмм: lehrwerke

**Statistik mit SPSS Alles in einem Band für Dummies**

**Bibliografische Information der Deutschen Nationalbibliothek**

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

2. Auflage 2023

© 2023 Wiley-VCH GmbH, Boschstraße 12, 69469 Weinheim, Germany

All rights reserved including the right of reproduction in whole or in part in any form. This translation published by arrangement with John Wiley and Sons, Inc.

Alle Rechte vorbehalten inklusive des Rechtes auf Reproduktion im Ganzen oder in Teilen und in jeglicher Form. Diese Übersetzung wird mit Genehmigung von John Wiley and Sons, Inc. publiziert.

Wiley, the Wiley logo, Für Dummies, the Dummies Man logo, and related trademarks and trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries. Used by permission.

Wiley, die Bezeichnung »Für Dummies«, das Dummies-Mann-Logo und darauf bezogene Gestaltungen sind Marken oder eingetragene Marken von John Wiley & Sons, Inc., USA, Deutschland und in anderen Ländern.

Das vorliegende Werk wurde sorgfältig erarbeitet. Dennoch übernehmen Autoren und Verlag für die Richtigkeit von Angaben, Hinweisen und Ratschlägen sowie eventuelle Druckfehler keine Haftung.

**Print ISBN:** 9783527720682
**ePub ISBN:** 9783527842384

**Coverfoto:** Molnia - stock.adobe.com
**Korrektur:** Petra Heubach-Erdmann

# Über den Autor

Jürgen Faik studierte Volkswirtschaftslehre und Soziologie in Frankfurt am Main und promovierte ebenda. Er war Lehrbeauftragter an verschiedenen Hochschulen (Universität Lüneburg und Hochschulen in Darmstadt, Frankfurt am Main, Wiesbaden, Bad Hersfeld sowie an der DIPLOMA) und zudem Privatdozent an der Universität Vechta. Aktuell ist er als Lehrkraft für besondere Aufgaben in den Feldern Wirtschaftsmathematik und Statistik an der Hochschule Mainz tätig. Er ist Autor des »Wiley-Schnellkurs Volkswirtschaftslehre«, des »Wiley-Schnellkurs Wirtschaftsmathematik« und des Lehrbuchs »Statistik für Wirtschafts- und Sozialwissenschaftler« sowie Mitherausgeber von »Mathematik für Wirtschaftswissenschaftler für Dummies«.

# Inhaltsverzeichnis


- Cover
- Titelblatt
- Impressum
- Über den Autor
- Einführung
  - Über dieses Buch
  - Törichte Annahmen über meine Leser
  - Wichtige Konventionen in diesem Buch
  - Wie dieses Buch aufgebaut ist
  - Wie geht es weiter im Buch?
- Teil I: Grundlagen der statistischen Datenverarbeitung und von SPSS
  - Kapitel 1: Vorbemerkungen zu SPSS
    - Bedeutung von Statistik
    - Bedeutung von Statistikprogrammen
    - Tabellenkalkulationsprogramme als Alternative
    - Verschiedene Statistikprogramme
    - Zur Relevanz von SPSS
    - Start von SPSS
    - Der Eingangsbildschirm
    - Die SPSS-Datenansicht
    - Was Sie so alles in SPSS öffnen können
  - Kapitel 2: Daten als Mütter statistischer Analysen
    - Bedeutung von Daten und Informationen
    - Vorgehen statistischer Untersuchungen
    - Regeln im Umgang mit Daten in SPSS
    - Datenskalierungen

SPSS-Variablentypen
Zum Nutzen quantitativer Ausdrücke in SPSS
Zum Problem fehlender Werte
Einlesen von (massenstatistischen) Daten in SPSS
Manuelle Dateneingabe
Vorhandene Datendateien
Einlesen von TXT-Files
Einlesen von EXCEL-, STATA- und SAS-Daten

## Kapitel 3: Hier wächst zusammen, was zusammengehört: Datendateien zusammenbringen

Fälle zusammenbringen
Variablen zusammenbringen

## Kapitel 4: Datendateien in Form bringen

Daten aggregieren
Daten (um-)formatieren
Zulässige Variablennamen
Variablentyp
Variablenbreite und Dezimalstellen
Weitere Einstellungen
Variablen umcodieren
Neue Variable berechnen
Daten bearbeiten
Doppelte Fälle ermitteln

## Kapitel 5: Was Sie an Syntax reinstecken, bekommen Sie an Output raus

Syntaxfiles
Vorteile von SPSS-Syntaxfiles
Erforderliche Fähigkeiten und »Fallstricke« im Zusammenhang von SPSS-Syntaxprogrammen
Zum Aufbau von SPSS-Syntaxprogrammen
Zugänge zur SPSS-Syntax
»SPSS-Grammatik«
SPSS-»Grundbefehle«

Ein kleines SPSS-Beispiel-Syntaxprogramm

Ausführen eines SPSS-Syntaxprogramms

Outputfiles

Viewer-Einstellungen

Navigationsbereich

Inhaltsbereich

## Teil II: Deskriptive SPSS-Statistiken

### Kapitel 6: Von der Schönheit des Anhäufens: Eindimensionale Häufigkeitsverteilungen

Zum Häufigkeitsbegriff

Das »Basketball-Beispiel« mit fehlenden Werten

Zur Fehler- und Ausreißeranalyse

Häufigkeitsdiagramme

### Kapitel 7: Viele Wege führen nach Rom: Eindimensionale deskriptive Statistiken

Unterschiede zwischen statistischen Prozeduren und statistischen Funktionen

Art der statistischen Kennziffern bei den einzelnen Statistikprozeduren

Häufigkeits-Prozedur (FREQUENCIES)

DESCRIPTIVES-Prozedur

Explorative Datenanalyse (EXAMINE)

Mittelwertvergleichs-Prozedur (MEANS)

Fallzusammenfassungs-Prozedur (SUMMARIZE)

## Teil III: Deskriptive und induktive SPSS-Statistiken gemischt

### Kapitel 8: Was alles mehr oder weniger miteinander zusammenhängt: Mehrdimensionale Häufigkeitsverteilungen

Vorbemerkungen zu den Zusammenhängen zwischen mehreren Variablen

Kreuztabelle

Assoziationen

Korrelationen

Rangkorrelationskoeffizienten

Korrelationskoeffizient nach Bravais/Pearson

Bootstrapping

## Kapitel 9: Über Abhängigkeiten: Simple Regressionen

Zum Unterschied zwischen Korrelation und Regression

Regressionsarten

Lineare Einfachregression

Annahmen

Die Schätzmethode der Kleinst-Quadrate-Methode

Eine lineare Regression in SPSS schätzen

Beeinflussung der linearen Regressionsgeraden durch Ausreißerwerte

Keine Variation der abhängigen Variablen

Bereichsspezifische Konstanz der abhängigen Variablen

Entgegengesetzte allgemeine Regression im Vergleich zu bereichsspezifischen Regressionen

Zur Aussagekraft einer Regressionsbeziehung

Quasilineare Regression

Quadratische Funktion

Potenzfunktion

Exponentialfunktion

Logistische Funktion

Lineare und quasilineare Regressionen gleichzeitig

Modifikationen/Erweiterung der linearen Regressionsanalyse

0/1-Dummies als unabhängige Variablen

Zeitreihenwerte als abhängige Variable

Residuenanalyse

Multiple lineare Regression

Schätzgleichung

Multikollinearität

## Kapitel 10: Komplexere Regressionen

Binär-logistische Regression (Logit-Ansatz)
Ordinale Regression
Multinomiale Regression

## Kapitel 11: Vom Gruppieren und Zusammenfassen von Daten: Diskriminanz-, Cluster- und Faktorenanalyse

Diskriminanzanalyse
Clusteranalyse
Faktorenanalyse

## Kapitel 12: Zeit spielt eine Rolle – und überhaupt: Überleben erwünscht!

Zeitreihenanalyse
Zeitreihenkomponenten und ihre Verknüpfung
Trendermittlung
Saisonkomponente
Sterbetafelanalyse
Kaplan/Meier-Methode
Cox-Regression

# Teil IV: Induktive SPSS-Statistiken

## Kapitel 13: Tests I: Welcher Mittelwert ist am größten?

Allgemeines
Null- und Alternativhypothese
Zweiseitiger Test
Einseitiger Test
Schema für einen statistischen Test
Eine Variable: t-Test
Zwei Gruppen: Unabhängigkeit
t-Test bei unabhängigen Stichproben
t-Test bei abhängigen Stichproben
Kruskal/Wallis-Test
Friedman-Test

## Kapitel 14: Tests II: Glockenkurve – oder was?

Vorbemerkung

Chi-Quadrat-Verteilungstest

Kolmogorov/Smirnov-Verteilungstest

Shapiro/Wilk-Normalverteilungstest

## Kapitel 15: Tests III: Was hängt mit wem zusammen?

Vorbemerkung

Test für den Chi-Quadrat-Kontingenzkoeffizienten

Test für den Rangkorrelationskoeffizienten nach Spearman

Test für den Korrelationskoeffizienten nach Bravais/Pearson

# Teil V: SPSS-Syntax für Fortgeschrittene

## Kapitel 16: Koppelung mit R und Python

Zu den R- und den Python-Essentials

Statistische Einbindung von R in SPSS

R-Anwendungsbeispiel: Robuste Regression

## Kapitel 17: Schleifen zum schnelleren Programmieren

Allgemeines zu Schleifen

DO-REPEAT-Schleifen

LOOP-Konstruktion

DO-IF-Befehl

## Kapitel 18: Auch Makros machen das Leben leichter

Grundlagen der Makroprogrammierung: Makros ohne Parameter

Makros mit Parametern

Loop-Makros

Einbeziehen einzelner Syntaxfiles in größere Syntaxfiles

## Kapitel 19: Beispielhafte Syntaxprogramme

Armutsquoten

Gini-Koeffizient und andere Ungleichheitsmaße

Armut und Ungleichheit gleichzeitig

# Teil VI: Top-10-Teil

## Kapitel 20: Die 10 wichtigsten SPSS-Befehle

AGGREGATE

COMPUTE

DESCRIPTIVES

DO … IF

EXECUTE

FILTER BY

FREQUENCIES

RECODE

TEMPORARY

WEIGHT BY

## Kapitel 21: Die 10 wichtigsten SPSS-Tipps

Nutze auch mal das PDF-Benutzerhandbuch!

Nutze Copy-and-Paste beim Einlesen von Daten!

Beachte grundsätzlich die gleiche Sortierung (und natürlich auch gleiche Variableninhalte) beim Daten-»Matchen«!

Erstelle »Spiel-Datenfiles« für Syntaxprogramme!

Verwende »sprechende« Variablen- und Dateibezeichnungen!

Kommentiere ausreichend viel in Syntaxprogrammen!

Verändere Pivottabellen und Grafiken vor ihrer Präsentation!

Achte auf fehlende Werte!

Verwende Schleifen und Makros!

Sei dir bewusst, dass kein »Unsinn« berechnet wird! Gebrauche dein Hirn!

## Abbildungsverzeichnis

## Stichwortverzeichnis

## End User License Agreement


# Tabellenverzeichnis


### Kapitel 2

Tabelle 2.1: Unterschiedliche statistische Auswertungsmöglichkeiten bei untersch...

## Kapitel 4

Tabelle 4.1: Beispielhafter Codierplan für die Datei *PERSONEN2.SAV*

## Kapitel 6

Tabelle 6.1: »Basketball-Beispiel« zu Häufigkeiten

## Kapitel 7

Tabelle 7.1: Statistische Funktionen: erweitertes »Basketball-Beispiel«

Tabelle 7.2: Statistische SPSS-Prozeduren und verschiedene statistische Kennziff...

Tabelle 7.3: »Basketball-Beispiel«

Tabelle 7.4: Absolute Häufigkeitsverteilung im »Basketball-Beispiel«

## Kapitel 8

Tabelle 8.1: Beispiel 1 für Rangkorrelationen

Tabelle 8.2: Beispiel 2 für Rangkorrelationen

Tabelle 8.3: Beispiel 1 (keine Ties): Paarbeziehungen (*SPEERWERFEN.SAV*)

Tabelle 8.4: Beispiel 2 (mit Ties): Paarbeziehungen (*SPEERWERFEN2.SAV*)

Tabelle 8.5: Arbeitstabelle zur Berechnung der Kovarianz (in Bezug auf *SPEERWERF*...

## Kapitel 9

Tabelle 9.1: X/Y-Punktepaare als Basis einer Regressionsrechnung

Tabelle 9.2: Arbeitstabelle für eine lineare Beispielregression

Tabelle 9.3: Arbeitstabelle für die Berechnung des Bestimmtheitsmaßes

Tabelle 9.4: Übersicht über die Signifikanz der geschätzten linearen Regressione...

## Kapitel 13

Tabelle 13.1: Tests zur Analyse von Unterschieden

## Kapitel 19

Tabelle 19.1: (Personenbezogene) Ungleichheitsergebnisse (Gesamtvermögen) für di...

Tabelle 19.2: Personen- und haushaltsbezogene Gini-Koeffizienten für die Ungleic...

Tabelle 19.3: (Personenbezogene) Ungleichheitsergebnisse (Gesamtvermögen) für di...

# Illustrationsverzeichnis

## Kapitel 1

Abbildung 1.1: Mitteilung nach erfolgreicher Lizenzierung (Auszug; hier bezogen ...

Abbildung 1.2: Der Start-Bildschirm von SPSS 29

Abbildung 1.3: Das SPSS-Menü

Abbildung 1.4: Datei in SPSS öffnen

Abbildung 1.5: OPTIONEN bei SPSS

## Kapitel 2

Abbildung 2.1 Beispiel für eine rechteckige Datentabelle (*123.SAV*)

Abbildung 2.2: Verschiedene Datentypen (Quelle: Faik, Jürgen (2015): Statistik f...

Abbildung 2.3: Zum Problem fehlender Werte bei mathematischen Variablenoperation...

Abbildung 2.4: Manuelle Werteeingabe im Dateneditor

Abbildung 2.5: Variablenumbenennung im Variableneditor

Abbildung 2.6: Kontextmenü für FÄLLE EINFÜGEN

Abbildung 2.7: Eingefügter Fall

Abbildung 2.8: Kontextmenü für das Löschen von Fällen

Abbildung 2.9: Datendatei nach Löschen eines Falls

Abbildung 2.10: Abspeichern einer manuell erstellten Datendatei als SAV-File (*FA*...

Abbildung 2.11 In SPSS einlesbare Datenfile-Formate

Abbildung 2.12 Semikolon-Trennung bei ASCII-Daten (DATEN_ALLE_SEM.TXT)

Abbildung 2.13: Tabulator-Trennung bei ASCII-Daten (DATEN_ALLE_TABS.TXT)

Abbildung 2.14: Öffnen einer ASCII-Datei (mit Trennzeichen) in SPSS (DATEN_ALLE_...

Abbildung 2.15: Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Sc...

Abbildung 2.16: Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Sc...

Abbildung 2.17 Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Sch...

Abbildung 2.18 Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Sch...

Abbildung 2.19: Eingelesene Stichprobe (1)

Abbildung 2.20: Eingelesene Stichprobe (2)

Abbildung 2.21: Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Sc...

Abbildung 2.22: Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Sc...

Abbildung 2.23: Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Sc...

Abbildung 2.24: In SPSS mittels des Textassistenten eingelesene ASCII-Daten mit ...

Abbildung 2.25: In SPSS mittels Copy-and-Paste eingelesene ASCII-Daten (Tabulato...

Abbildung 2.26: In SPSS mittels Copy-and-Paste eingelesene ASCII-Daten (Semikolo...

Abbildung 2.27: ASCII-Daten mit fester Satzlänge (DATEN_ALLE_NICHTS.TXT)

Abbildung 2.28 SPSS-Warnung beim Versuch des Einlesens von ASCII-Daten mit feste...

Abbildung 2.29 Lokale Codierung von ASCII-Daten

Abbildung 2.30: Textassistent zum Einlesen von ASCII-Daten ohne Trennzeichen, mi...

Abbildung 2.31: Textassistent zum Einlesen von ASCII-Daten ohne Trennzeichen, mi...

Abbildung 2.32: Textassistent zum Einlesen von ASCII-Daten ohne Trennzeichen, mi...

Abbildung 2.33: Textassistent zum Einlesen von ASCII-Daten ohne Trennzeichen, mi...

Abbildung 2.34 In SPSS mittels Textassistenten eingelesene ASCII-Daten ohne Tren...

Abbildung 2.35: EXCEL-Beispieldaten (DATEN_ALLE.XLSX)

Abbildung 2.36: Bildschirm zum Einlesen von EXCEL-Daten in SPSS (DATEN_ALLE.XLSX...

Abbildung 2.37: EXCEL-Datei mit Fällen in den Spalten und Variablen in den Zeile...

Abbildung 2.38: SPSS-Anzeige für EXCEL-Datei DATEN_ALLE_INVERTIERT.XLSX mit Fäll...

Abbildung 2.39: Zwischenschritt beim Transponieren von SPSS-Daten

Abbildung 2.40: Transponierte SPSS-Daten

Abbildung 2.41 Transponierte und um die Vorspalte bereinigte SPSS-Daten

## Kapitel 3

Abbildung 3.1: SPSS-Datei mit den Fällen 1–5 (siehe *var00001*; *DATEN_1-5.SAV?...*

Abbildung 3.2: SPSS-Datei mit den Fällen 6–10 (siehe *var00001*; *DATEN_6-10.SAV?...*

Abbildung 3.3: Fälle hinzufügen (1)

Abbildung 3.4: Fälle hinzufügen (2)

Abbildung 3.5: Nach Fällen zusammengeführte Datei (Variante 1)

Abbildung 3.6: Nach Fällen zusammengeführte Datei (Variante 2)

Abbildung 3.7: Sortieren von Fällen in einem Datenset

Abbildung 3.8: SPSS-Datei mit den Variablen 1 und 2 (*DATEN_VAR1_2.SAV*)

Abbildung 3.9 SPSS-Datei mit den Variablen 1 und 3 (*DATEN_VAR1_3.SAV* )

Abbildung 3.10: Variablen hinzufügen (1)

Abbildung 3.11: Variablen hinzufügen (2)

Abbildung 3.12: Nach Variablen zusammengeführte Datei

Abbildung 3.13: Datei mit den Variablen 1 und 3, absteigend nach Variable 1 sorti...

Abbildung 3.14: Fehlerhafte, nach den Variablen zusammengeführte Gesamtdatei

Abbildung 3.15: Umsortieren einer Datei

Abbildung 3.16: Korrekte, nach den Variablen zusammengeführte Gesamtdatei – nach...

Abbildung 3.17: Änderung der in SPSS abgebildeten Dateikennzeichnung nach Änderun...

## Kapitel 4

Abbildung 4.1: Beispiels-SPSS-Personendatendatei (*PERSONEN.SAV* )

Abbildung 4.2: Daten-Aggregation in SPSS starten

Abbildung 4.3: Mögliche Funktionen im Zusammenhang mit dem Aggregieren von Daten

Abbildung 4.4 Durchführen des Aggregierens in SPSS

Abbildung 4.5: Aggregierter Datensatz

Abbildung 4.6: Daten-Aggregation ohne Breakvariable

Abbildung 4.7: Ergebnis der Daten-Aggregation ohne Breakvariable

Abbildung 4.8: Ein zu formatierendes SPSS-Datenfile (Datenansicht) von *PERSONEN2*...

Abbildung 4.9: Die SPSS-Variablenansicht von *PERSONEN2.SAV*

Abbildung 4.10: SPSS-Fehlermeldung bei falschem Namensanfang einer Variablen bez...

Abbildung 4.11: SPSS-Fehlermeldung bei unzulässigem Sonderzeiche...

Abbildung 4.12: SPSS-Fehlermeldung bei unzulässiger Verwendung ein...

Abbildung 4.13: Verändern eines Variablennamens im Variableneditor

Abbildung 4.14: Verändern des Variablentyps im Variableneditor (*PERSONEN2.SAV*)

Abbildung 4.15: KOMMA-Formatierung der Variablen beitrag (*PERSONEN2.SAV*)

Abbildung 4.16 PUNKT-Formatierung der Variablen *beitrag* (*PERSONEN2.SAV*)

Abbildung 4.17: Formatierung der Variablen *beitrag* mit WISSENSCHAFTLICHE NOTATIO...

Abbildung 4.18: DOLLAR-Formatierung der Variablen *beitrag* (*PERSONEN2.SAV*)

Abbildung 4.19: Zahlendarstellung im Dateneditor bei zu geringer Spaltenbreite (

Abbildung 4.20: Skalierungsmöglichkeiten in SPSS: NOMINAL, ORDINAL, METRISCH (*PER*...

Abbildung 4.21: Zur ROLLE-Formatierung (*PERSONEN4A.SAV* )

Abbildung 4.22: Umcodieren in derselben Variablen am Beispiel der nominalskalier...

Abbildung 4.23: Umcodieren in derselben Variablen am Beispiel der nominalskalier...

Abbildung 4.24: Ergebnis des Umcodierens in derselben Variablen am Beispiel der ...

Abbildung 4.25: Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV* ) (1...

Abbildung 4.26: Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (2)

Abbildung 4.27: Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (3)

Abbildung 4.28: Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (4)

Abbildung 4.29: Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV* ) (5)

Abbildung 4.30: Ergebnis des Umcodierens in einer anderen Variablen (*PERSONEN4A*....

Abbildung 4.31: Gleichzeitiges Umcodieren von zwei Variablen in zwei anderen Var...

Abbildung 4.32: Gleichzeitiges Umcodieren von zwei Variablen in zwei anderen Vari...

Abbildung 4.33: Ergebnis des gleichzeitigen Umcodierens von zwei Variablen in zw...

Abbildung 4.34: Berechnen einer neuen Variablen (*PERSONEN4A.SAV*)

Abbildung 4.35: Datendatei mit neu berechneter Variablen (*PERSONEN4A.SAV*)

Abbildung 4.36: Haushaltsbezogene Daten vor Gewichtung (in *AGGR.SAV*)

Abbildung 4.37: Dialogfeld zur Gewichtung

Abbildung 4.38: Fehlender Gewichtungswert (*AGGR_A.SAV*)

Abbildung 4.39: SPSS-Hinweis bei fehlenden Werten der Gewichtung...

Abbildung 4.40: Dialogbox für DATEN|FÄLLE AUSWÄHLEN (*AGGR.SAV* )

Abbildung 4.41: Aktivierung der Fallauswahl in der Dialogbox DATEN|FÄLLE AUSWÄHL...

Abbildung 4.42: Angabe der Bedingung bei DATEN|FÄLLE AUSWÄHLEN (*AGGR.S*...

Abbildung 4.43: Datensatz mit Filtervariable *filter_$* und markierten nicht a...

Abbildung 4.44: Neuer Datensatz nur mit den ausgewählten Fällen (*AGGR_B.SAV*?) (1...

Abbildung 4.45: Neuer Datensatz nur mit den ausgewählten Fällen (*AGGR_B.SAV*) (2)

Abbildung 4.46: Löschen der nicht ausgewählten Fälle (*AGGR.SAV*)

Abbildung 4.47: Beispiels-Transformation von Euro- in DM-Werte in einem Beispiel...

Abbildung 4.48: Variablen-Transformation für einen gefilterten Datensatz (*AGGR.S*...

Abbildung 4.49: Variante »Zufallsstichprobe« bei DATEN|FÄLLE AUSWÄHLEN

Abbildung 4.50: Ergebnis einer Zufallsauswahl bei DATEN|FÄLLE AUSWÄHLEN...

Abbildung 4.51: Variante »Fallbereich« bei DATEN|FÄLLE AUSWÄHLEN

Abbildung 4.52: Ergebnis einer »Fallbereichsauswahl« bei DATEN|FÄLLE AUSWÄHLEN...

Abbildung 4.53: Doppelter Fall in SPSS-Datendatei *PERSONEN4.SAV*

Abbildung 4.54: Dialogfeld zur Ermittlung doppelter Fälle

Abbildung 4.55: SPSS-Ergebnis der Ermittlung doppelter Fälle (*PERSONEN4.SAV*)

Abbildung 4.56: Löschen eines doppelten Falls (*PERSONEN4.SAV*)

## Kapitel 5

Abbildung 5.1: Der SPSS-Syntaxeditor

Abbildung 5.2: Beispieldatendatei *CORRELATIONS.SAV* für eine eingeschränkte Korrel...

Abbildung 5.3: Eine eingeschränkte (syntaxgesteuerte) Korrelationsmatrix als Outp...

Abbildung 5.4: Eine umfangreiche (menügesteuerte) Korrelationsmatrix als Output (

Abbildung 5.5: Syntax in einem Ausgabefenster (bezüglich *PERSONEN4A.SAV*)

Abbildung 5.6: Kopieren von Syntax in einem Ausgabefenster (bezüglich *PERSONEN4A.*...

Abbildung 5.7: Option EINFÜGEN in einer SPSS-Dialogbox (*PERSONEN4A.SAV*)

Abbildung 5.8: Im Syntaxeditor automatisch eingefügter Befehlstext (*PERSONEN4A.S*...

Abbildung 5.9: Syntaxfehler im Syntaxeditor (*PERSONEN4A.SAV*)

Abbildung 5.10: OPTIONEN-Einstellungen für den Syntaxeditor

Abbildung 5.11: Syntaxfehler-Hinweis im Ausgabefenster (*PERSONEN4A.SAV*)

Abbildung 5.12: Datendateibezeichnung bei mehreren geöffneten Datendateien

Abbildung 5.13: Auswahl einer Datendatei (*CORRELATIONS.SAV*) im Syntaxeditor bei m...

Abbildung 5.14: Output eines spezifischen SPSS-Befehls (für *CORRELATIONS.SAV*)

Abbildung 5.15: Leerzeile innerhalb eines SPSS-Kommandos

Abbildung 5.16: Unvollständiger Output bei einer Leerzeile innerhalb eines SPSS-...

Abbildung 5.17: Beendigung eines Kommentars mit einem Punkt

Abbildung 5.18: Beendigung eines Kommentars mit einer Leerzeile

Abbildung 5.19: Mehrzeiliger Kommentar

Abbildung 5.20: Leerzeile innerhalb eines mehrzeiligen Kommentars

Abbildung 5.21: Output bei einer Leerzeile innerhalb eines mehrzeiligen Kommentar...

Abbildung 5.22: Punkt als Satzzeichen innerhalb eines Kommentars

Abbildung 5.23: Problemloser Output bei einem Punkt als Satzzeichen innerhalb ei...

Abbildung 5.24: Der SPSS-Syntaxeditor als Zugangsweg

Abbildung 5.25: Syntaxhilfe über das »Fragezeichen-Symbol« im Syntaxeditor

Abbildung 5.26: Online-Syntaxhilfe zu einem bestimmten Befehl (hier bezüglich GET...

Abbildung 5.27: Öffnen eines bereits existierenden SPSS-Syntaxfiles

Abbildung 5.28: Speicherort für die SPSS-Journaldatei STATISTICS.JNL

Abbildung 5.29: Datenansicht nach Datentransformationen mit EXECUTE-Befehl unter...

Abbildung 5.30: Datenansicht nach Datentransformationen mit EXECUTE-Befehl unter...

Abbildung 5.31: Neue Datendatei nach Anwendung des GET-FILE-Befehls zusammen mit...

Abbildung 5.32: Neue Datendatei nach Anwendung des GET-FILE-Befehls zusammen mit...

Abbildung 5.33: Datendatei *DATEN_ALLE_INVERTIERT.SAV* mit den Variablen in den Zei...

Abbildung 5.34: Invertierte Datendatei *DATEN_ALLE_INVERTIERT.SAV* – FLIP-Befehl (...

Abbildung 5.35: Invertierte Datendatei *DATEN_ALLE_INVERTIERT2.SAV* – FLIP-Befehl ...

Abbildung 5.36: Nach Fällen gleiche, aber nach Variablen unterschiedliche zusamm...

Abbildung 5.37: Nach Variablen gleiche, aber nach Fällen unterschiedliche zusamm...

Abbildung 5.38: Verschiedene Datentransformationen mit Bezugnahme auf *PERSONEN4A*...

Abbildung 5.39: Zuweisung von Variablen- und Werteetiketten – VARIABLE LABELS, V...

Abbildung 5.40: Häufigkeitstabelle bei Auswahl einzelner Fälle in *PERSONEN5.SAV* ...

Abbildung 5.41: Auswahl von Fällen auf Basis von *PERSONEN5.SAV* – SELECT IF ohne ...

Abbildung 5.42: Filtern eines Datensatzes (*PERSONEN5.SAV*) – FILTER BY

Abbildung 5.43: Sortierung von Daten mithilfe von SORT CASES (*PERSONEN5.SAV*)

Abbildung 5.44: Häufigkeitstabellen für Männer und Frauen getrennt (auf Basis vo...

Abbildung 5.45: Haushaltsbezogene Beitragssummen nach der Aggregation von der Pe...

Abbildung 5.46: Gesamtbeitragssumme nach der Aggregation von der Personenebene (*A*...

Abbildung 5.47: Zur Anwendung des COUNT-Befehls (*EINKOMMEN_VERMOEGEN.SAV*)

Abbildung 5.48: Auflisten ausgewählter Variablen und Fälle – LIST (*PERSONEN4A.SAV*

Abbildung 5.49: Zufallsauswahl mit dem SAMPLE-Befehl (*PERSONEN4A.SAV*)

Abbildung 5.50: Variablen- und Dateninformationen – SYSFILE INFO (*PERSONEN4A.SAV*

Abbildung 5.51: Zu lange Titelzeile – TITLE (*PERSONEN4A.SAV*)

Abbildung 5.52: Auf zwei Zeilen verteilter Titel – TITLE (*PERSONEN4A.SAV*)

Abbildung 5.53: Statistiken im Fall mit Gewichtung – WEIGHT BY (*AGGR.SAV*)

Abbildung 5.54: Statistiken im Fall ohne Gewichtung – WEIGHT OFF (*AGGR.SAV*)

Abbildung 5.55: Ersetzen der Verzeichnisstruktur des Autors durch Ihre eigene in...

Abbildung 5.56: Beispiel-Rohdatendatei *PERS1.SAV* (vier Variablen, Fälle 1–10)

Abbildung 5.57: Beispiel-Rohdatendatei *PERS2_FALSCH.SAV* (alle elf Variablen, Fäl...

Abbildung 5.58: Beispiel-Rohdatendatei *PERS3.SAV* (alle Variablen, Fälle 11–20)

Abbildung 5.59: Haushaltsbezogene Datendatei *HH_ALLE2.SAV* nach Durchführen aller ...

Abbildung 5.60: Personenbezogene FREQUENCIES-Ergebnisse in Form von Häufigkeitst...

Abbildung 5.61: Personenbezogene FREQUENCIES-Ergebnisse in Form statistischer Ke...

Abbildung 5.62: Mit SELECT IF selektierte »einkommensschwache Haushalte« mit sta...

Abbildung 5.63: Mit SPLIT BY aufgeteilte Datei nach Altersgruppen, das heißt: na...

Abbildung 5.64: Syntax-Beispiel für AUSFÜHREN|ALLE (in Bezug auf *AGGR.SAV*)

Abbildung 5.65: Ausführlicher Output im Syntax-Beispiel bei AUSFÜHREN|ALLE ...

Abbildung 5.66: Syntax-Beispiel für AUSFÜHREN|AUSWAHL (in Bezug auf *AGGR.SAV*

Abbildung 5.67: Reduzierter Output im Syntax-Beispiel bei AUSFÜHREN|AUSWAHL

Abbildung 5.68: Öffnen eines Ausgabefensters nach Öffnen einer Datendatei (*123.SA*...

Abbildung 5.69: Viewer-Einstellungen

Abbildung 5.70: Ausschalten des Syntax-Protokolls

Abbildung 5.71: SPSS-Output nach Ausschalten des Syntax-Protokolls (in Bezug auf...

Abbildung 5.72: Der Navigationsbereich des Viewers (in Bezug auf *AGGR.SAV*)

Abbildung 5.73: Rubrik »Titel« im Navigationsbereich des Viewers (in Bezug auf *A*...

Abbildung 5.74: Rubrik »Hinweise« im Navigationsbereich des Viewers (in Bezug au...

Abbildung 5.75: Eigentliche Auswertungsrubrik »Statistiken« im Navigationsbereic...

Abbildung 5.76: Optionen des Ausschneidens, Verschiebens oder Kopierens im Navig...

Abbildung 5.77: Warnung im Viewer-Inhaltsbereich (in Bezug auf *PERSONEN4A.SAV*)

Abbildung 5.78: Fehlermeldung im Viewer-Inhaltsbereich (in Bezug auf *CORRELATION*...

Abbildung 5.79: Pivottabelle im Viewer-Inhaltsbereich (in Bezug ...

Abbildung 5.80: Doppelklick zum Verändern einer Pivottabelle im Viewer-Inhaltsbe...

Abbildung 5.81: Veränderung einer Pivottabelle im Viewer-Inhaltsbereich (hier in...

Abbildung 5.82: Erweiterte Verarbeitung-smöglichkeiten einer Pivottabelle im View...

Abbildung 5.83: Diagramm im Viewer-Inhaltsbereich (hier in Bezug auf *CORRELATIONS*...

Abbildung 5.84: Öffnen des Diagrammeditors aus dem Viewer-Inhaltsbereich heraus ...

Abbildung 5.85: Verändern eines Diagramms im Diagrammeditor

Abbildung 5.86: Abspeichern von SPSS-Output in einer EXCEL-Datei (hier in Bezug ...

Abbildung 5.87: SPSS-Output in einer EXCEL-Datei (hier in Bezug auf *CORRELATIONS*...

Abbildung 5.88: Abspeichern von SPSS-Output in einer WORD-Datei (in Bezug auf *CO*...

Abbildung 5.89: SPSS-Output in einer WORD-Datei (hier in Bezug auf *CORRELATIONS*....

## Kapitel 6

Abbildung 6.1: Fenster für HÄUFIGKEITEN (hier für *KORBTREFFER.SAV*)

Abbildung 6.2: HÄUFIGKEITEN-Output für »Basketball-Beispiel« (*KORBTREFFER.SAV*)

Abbildung 6.3: Datensatz *KORBTREFFER_F.SAV* mit zwei fehlenden Werten

Abbildung 6.4: Häufigkeitstabelle für eine Variable mit (zwei) fehlenden Werten ...

Abbildung 6.5: Häufigkeitstabelle für sehr viele (1.000) Merkmalsausprägungen ei...

Abbildung 6.6: Fehleranalyse anhand einer Häufigkeitstabelle (*KORBTREFFER_F2.SAV*

Abbildung 6.7: Suche nach einer bestimmten Variablen in einem Datensatz (*KORBTREF*...

Abbildung 6.8: Suche nach einem bestimmten Wert einer bestimmten Variablen in ei...

Abbildung 6.9: »Ausreißer-Datensatz« *AUSREISSER.SAV*

Abbildung 6.10: Häufigkeitsauszählung in SPSS mit Ausreißerwert »100« (*AUSREISSE*...

Abbildung 6.11: Starten einer Ausreißer-Analyse mithilfe von EXPLORATIVE DATENAN...

Abbildung 6.12: Ergebnis einer tabellarischen Ausreißeranalyse mithilfe von EXPL...

Abbildung 6.13: Ergebnis einer grafischen Ausreißer-Analyse mithilfe von EXPLORA...

Abbildung 6.14: DIAGRAMME bei HÄUFIGKEITEN (in Bezug auf *KORBTREFFER.SAV?*...

Abbildung 6.15: Balkendiagramm als Output (*KORBTREFFER.SAV*)

Abbildung 6.16: Kreisdiagramm als Output (*KORBTREFFER.SAV*)

Abbildung 6.17: Histogramm als Output (*KORBTREFFER.SAV* )

Abbildung 6.18: Normalverteilungskurve anzeigen lassen (hier in Bezug auf *KORBTR*...

Abbildung 6.19: Histogramm und Normalverteilungskurve als Output (*KORBTREFFER.SA*...

## Kapitel 7

Abbildung 7.1: Statistische Funktion MEAN (für *KORBTREFFER4.SAV* )

Abbildung 7.2: Anwendung der MEAN-Funktion auf die Variable *korbtreffer2014_2016*

Abbildung 7.3: Neue Variable *korbtreffer2014_2016* auf Basis der MEAN-Funktion (*KO*...

Abbildung 7.4: Anwendung der SUM-Funktion auf die Variable *korbtreffer2014_2016* ...

Abbildung 7.5: Neue Variable *korbtreffer2014_2016* für die SUM-Funktion (*KORBTREFF*...

Abbildung 7.6: HÄUFIGKEITEN|STATISTIKEN für das »Basketball-Beispiel« (*KORBT*...

Abbildung 7.7: Unterdrückung der Häufigkeitstabelle bei HÄUFIGKEITEN|STATISTIKEN...

Abbildung 7.8: Statistische Kennziffern (FREQUENCIES) für das »Basketball-Beispi...

Abbildung 7.9: BOOTSTRAPPING DURCHFÜHREN bei HÄUFIGKEITEN (*KORBTREFFER*...

Abbildung 7.10: Bootstrapping-Output bei HÄUFIGKEITEN (*KORB*...

Abbildung 7.11: Menü-Maske für DESKRIPTIVE STATISTIK (in Bezug auf *KORBTREFFER.S*...

Abbildung 7.12: OPTIONEN bei DESKRIPTIVE STATISTIK (hier in Bezug auf *KORBTREFFE*...

Abbildung 7.13: Output bei DESKRIPTIVE STATISTIK (*KORBTREFFER.SAV*)

Abbildung 7.14: BOOTSTRAP bei DESKRIPTIVE STATISTIKEN (in Bezug auf *KORBTREFFER.S*...

Abbildung 7.15: BOOTSTRAP-Output bei DESKRIPTIVE STATISTIKEN (*KORBTREFFER.SAV*)

Abbildung 7.16: Start-Dialog bei EXPLORATIVE DATENANALYSE (*KORBTREFFER.SAV?*...

Abbildung 7.17: STATISTIKEN bei EXPLORATIVE DATENANALYSE (hier in Bezug auf *KORB*...

Abbildung 7.18: Statistik-Output bei EXPLORATIVE DATENANALYSE (*KORBTREFFER.SAV*)

Abbildung 7.19: BOOTSTRAP-Output bei EXPLORATIVE DATENANALYSE (*KORBTREFFER.SAV*?)...

Abbildung 7.20: BOOTSTRAP-Output bei EXPLORATIVE DATENANALYSE (*KORBTREFFER.SAV*?)...

Abbildung 7.21: Um die Variable *geschlecht* erweiterte Korbtreffer-Datendatei *KOR*...

Abbildung 7.22: MEANS-Berechnungen für *korbtreffer* in Abhängigkeit von *geschlecht*

Abbildung 7.23: MEANS-Statistiken für *korbtreffer* in Abhängigkeit von *geschlecht* ...

Abbildung 7.24: MEANS-Statistikoutput für *korbtreffer* in Abhängigkeit von *geschl*...

Abbildung 7.25: Die Datendatei *HARM_GEOM.SAV*

Abbildung 7.26: Geometrisches Mittel für *wachstumsrate* (MEANS) (*HA*...

Abbildung 7.27: Harmonisches Mittel für *km_pro_h* (MEANS) (*HARM_GEOM.SAV*)

Abbildung 7.28: BOOTSTRAP-Output bei MITTELWERTE VERGLEICHEN (*KORBTREFFER5.SAV*)

Abbildung 7.29: Alle Fälle berücksichtigen bei SUMMARIZE (hier in Bezug auf *KORB*...

Abbildung 7.30: SUMMARIZE-Statistikoutput für *korbtreffer* in Abhängigkeit von *ge*...

## Kapitel 8

Abbildung 8.1: Erweiterte »Basketball-Beispiel«-Datendatei (*KORBTREFFER2.SAV?*...

Abbildung 8.2: Kreuztabellierungsanweisung für *korbtreffer* und *alter_klasse* (*KOR*...

Abbildung 8.3: Auswahl der beobachteten Fallzahlen in einer Kreuztabelle (*KORBTR*...

Abbildung 8.4: Kreuztabellenoutput für die beobachteten Fallzahlen (*KORBTREFFER3*...

Abbildung 8.5: Auswahl der beobachteten und der erwarteten Fallzahlen in einer K...

Abbildung 8.6: Kreuztabellenoutput für die beobachteten und die erwarteten Fallz...

Abbildung 8.7: Auswahl der zeilenweise berechneten relativen Häufigkeiten in ein...

Abbildung 8.8: Kreuztabellenoutput für die zeilenweise berechneten relativen Häu...

Abbildung 8.9: Auswahl der spaltenweise berechneten relativen Häufigkeiten in ei...

Abbildung 8.10: Kreuztabellenoutput für die spaltenweise berechneten relativen H...

Abbildung 8.11: Auswahl der auf die Gesamtfallzahl bezogenen relativen Häufigkei...

Abbildung 8.12: Kreuztabellenoutput für auf die Gesamtfallzahl bezogene relative...

Abbildung 8.13: Auswahl von absoluten und relativen Häufigkeiten in einer Kreuzt...

Abbildung 8.14: Kreuztabellenoutput für absolute und relative Häufigkeiten (*KORB*...

Abbildung 8.15: Datendatei *GESCHL_FARBE2.SAV* mit den Variablen *geschlecht*, *augen*...

Abbildung 8.16: Anweisung für eine erweiterte Kreuztabelle: erste Unterteilung d...

Abbildung 8.17: Anweisung für eine erweiterte Kreuztabelle: zweite Unterteilung ...

Abbildung 8.18: Output für eine erweiterte Kreuztabelle mit zwei Unterteilungen ...

Abbildung 8.19: Datendatei *GESCHL_FARBE.SAV* mit den Variablen *geschlecht* und *aug*...

Abbildung 8.20: Kreuztabellen-Output mit den beobachteten und den erwarteten Fal...

Abbildung 8.21: Auswahl verschiedener Assoziationsmaße (*GESCHL_FARBE.SAV*)

Abbildung 8.22: SPSS-Output für verschiedene Assoziationsmaße (*GESCHL_FARBE.SAV*?...

Abbildung 8.23: Bootstrapping für Assoziationsmaße (hier in Bezug auf *GESCHL_FAR*...

Abbildung 8.24: Bootstrapping-Output für verschiedene Assoziationsmaße (*GESCHL_F*...

Abbildung 8.25: Voraussetzung von mindestens Ordinalskalenniveau bei Korrelation...

Abbildung 8.26: Datendatei *SPEERWERFEN.SAV*

Abbildung 8.27: Aufruf des Spearman'schen Rangkorrelationskoeffizienten in SPSS ...

Abbildung 8.28: Wert des Rangkorrelationskoeffizienten – Beispiel 1 (*SPEERWERFEN*...

Abbildung 8.29: Datendatei *SPEERWERFEN2.SAV* (mit Mehrfachbesetzungen)

Abbildung 8.30: Wert des Rangkorrelationskoeffizienten – Beispiel 2 (*SPEERWERFEN*...

Abbildung 8.31: Aufruf von Kendalls Tau-Koeffizienten b in SPSS (*SPEERWERFEN.SAV*

Abbildung 8.32: Wert des Tau-b-Koeffizienten – Beispiel 1 (*SPEERWERFEN.SAV*)

Abbildung 8.33: Wert des Tau-b-Koeffizienten – Beispiel 2 (*SPEERWERFEN2.SAV*)

Abbildung 8.34: Aufruf verschiedener (weiterer) Rangkorrelationskoeffizienten in...

Abbildung 8.35: Werte verschiedener Rangkorrelationskoeffizienten (und des Brava...

Abbildung 8.36: Eine perfekt-positive lineare Korrelation zwischen zwei metrisch...

Abbildung 8.37: Keine lineare, aber eine nichtlineare Korrelation zwischen zwei ...

Abbildung 8.38: Aufruf des Bravais/Pearson-Korrelationskoeffizienten über das SP...

Abbildung 8.39: Ergebnis des Bravais/Pearson-Korrelationskoeffizienten nach Aufr...

Abbildung 8.40: Aufruf (auch) des Bravais/Pearson-Korrelationskoeffizienten über...

Abbildung 8.41: Ergebnis (auch) des Bravais/Pearson-Korrelationskoeffizienten na...

Abbildung 8.42: Zum Problem der verdeckten Korrelation (Quelle: Faik, Jürgen (20...

Abbildung 8.43: Bootstrapping bei nominalskalierten Variablen für verschiedene A...

Abbildung 8.44: Bootstrapping bei nominalskalierten Variablen für verschiedene A...

Abbildung 8.45: Bootstrapping-Output bei nominalskalierten Variablen für verschi...

Abbildung 8.46: Bootstrapping bei mindestens ordinalskalierten Variablen für ver...

Abbildung 8.47: Bootstrapping-Output bei mindestens ordinalskalierten Variablen ...

Abbildung 8.48: Bootstrapping bei mindestens ordinalskalierten Variablen für ver...

Abbildung 8.49: Bootstrapping-Output bei mindestens ordinalskalierten Variablen ...

## Kapitel 9

Abbildung 9.1: Lineare Regressionsfunktion und Punktewolke (Quelle: Faik, Jürgen...

Abbildung 9.2: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LIN*...

Abbildung 9.3: Lineare Einfachregression zwischen *ausgaben* und *einkommen* in der ...

Abbildung 9.4: Ergebnis der linearen Einfachregression zwischen *ausgaben* und *ein*...

Abbildung 9.5: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LIN*...

Abbildung 9.6: Ergebnis der linearen Einfachregression zwischen *ausgaben* und *ein*...

Abbildung 9.7: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LIN*...

Abbildung 9.8: Ergebnis der linearen Einfachregression zwischen *ausgaben* und *ein*...

Abbildung 9.9: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LIN*...

Abbildung 9.10: Ergebnis der linearen Einfachregression zwischen *ausgaben* und *ei*...

Abbildung 9.11: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LI*...

Abbildung 9.12: Ergebnis der linearen Einfachregression zwischen *ausgaben* und *ei*...

Abbildung 9.13: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LI*...

Abbildung 9.14: Vollständiges Ergebnis der linearen Einfachregression zwischen *a*...

Abbildung 9.15: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LI*...

Abbildung 9.16: Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben* ...

Abbildung 9.17: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* s...

Abbildung 9.18: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LI*...

Abbildung 9.19: Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben_*...

Abbildung 9.20: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben_l*...

Abbildung 9.21: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LI*...

Abbildung 9.22: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben_l*...

Abbildung 9.23: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LI*...

Abbildung 9.24: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben_l*...

Abbildung 9.25: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LI*...

Abbildung 9.26: Anfordern verschiedener linearer Regressionsbeziehungen zwischen...

Abbildung 9.27: Ergebnisse verschiedener linearer Regressionsbeziehungen zwische...

Abbildung 9.28: Ergebnisse verschiedener linearer Regressionsbeziehungen zwische...

Abbildung 9.29: Scatterplot zwischen *ausgaben* und *alter* in der Datendatei *LINEAR*...

Abbildung 9.30: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* u...

Abbildung 9.31: Ausgaben-Durchschnittswerte für zwei Altersgruppen auf Basis von...

Abbildung 9.32: Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben* ...

Abbildung 9.33: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* s...

Abbildung 9.34: Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben* ...

Abbildung 9.35: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* s...

Abbildung 9.36: Scatterplot zwischen *ausgaben* und *jahr* in der Datendatei *LINEARE*...

Abbildung 9.37: Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben* ...

Abbildung 9.38: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* u...

Abbildung 9.39: Anfordern verschiedener Trendfunktionen für *ausgaben* in *LINEARE_*...

Abbildung 9.40: Ergebnis der Schätzung verschiedener Trendfunktionen für *ausgabe*...

Abbildung 9.41: Ergebnis der Schätzung verschiedener Trendfunktionen für *ausgabe*...

Abbildung 9.42: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* u...

Abbildung 9.43: Anfordern einer linearen Regressionsbeziehung zwischen *ausg_diff*

Abbildung 9.44: Ergebnis einer linearen Regressionsbeziehung zwischen *ausg_diff* ...

Abbildung 9.45: Anfordern von Konfidenzintervallen für die geschätzten Regressio...

Abbildung 9.46: Konfidenzintervalle für die geschätzten Regressionsparameter als...

Abbildung 9.47: BOOTSTRAP-Anfordern von Konfidenzintervallen für die geschätzten...

Abbildung 9.48: BOOTSTRAP-Konfidenzintervalle für die geschätzten Regressionspar...

Abbildung 9.49: Anfordern der einfachen Residuen bei einer linearen Regression z...

Abbildung 9.50: Neue Variable *res_1* für die einfachen Residuen bei einer lineare...

Abbildung 9.51: Streudiagramm für die einfachen Residuen in Abhängigkeit von *ein*...

Abbildung 9.52: Standardisierte vorhergesagte Werte der abhängigen Variablen bei...

Abbildung 9.53: Streudiagramm für die standardisierten Residuen in Abhängigkeit ...

Abbildung 9.54: Um zwei logarithmierte Variablen erweiterter Datensatz *LINEARERE*...

Abbildung 9.55: Streudiagramm für die standardisierten Residuen in Abhängigkeit ...

Abbildung 9.56: Aufruf einer WLS-Schätzung für den Zusammenhang zwischen *ausgabe*...

Abbildung 9.57: Ergebnis einer WLS-Schätzung für den Zusammenhang zwischen *ausga*...

Abbildung 9.58: Anfordern der Prüfung auf Normalverteilung in Bezug auf die Resi...

Abbildung 9.59: Histogramm und Normalverteilung(-skurve) in Bezug auf die Residu...

Abbildung 9.60: Bivariate Korrelationen zwischen *ausgaben* einerseits sowie *einko*...

Abbildung 9.61: Methoden zur Berechnung einer linearen Mehrfachregression am Bei...

Abbildung 9.62: Ergebnis einer Regression zwischen *ausgaben* sowie *einkommen*, *hhg*...

Abbildung 9.63: Ergebnis einer Regression zwischen *ausgaben* sowie *einkommen*, *hhg*...

Abbildung 9.64: Ergebnis einer Regression zwischen *ausgaben* sowie *einkommen*, *hhg*...

Abbildung 9.65: Ergebnis einer Regression zwischen *ausgaben* sowie *einkommen*, *hhg*...

Abbildung 9.66: Zur Multikollinearitätsproblematik (Quelle: Faik, Jürgen (2015):...

Abbildung 9.67: Anfordern einer Kollinearitätsdiagnose für eine Regression zwisc...

Abbildung 9.68: Ergebnis einer Kollinearitätsdiagnose für eine Regression zwisch...

Abbildung 9.69: Ergebnis einer Kollinearitätsdiagnose für eine Regression zwisch...

Abbildung 9.70: Ergebnis einer Kollinearitätsdiagnose für eine Regression zwisch...

## Kapitel 10

Abbildung 10.1: Um die Variable *arm* erweiterter neuer Datensatz *REGRESSION_LOGIT*...

Abbildung 10.2: Anfordern einer Logit-Regressionsschätzung zwischen *arm* sowie *al*...

Abbildung 10.3: Ergebnis einer Logit-Regressionsschätzung zwischen *arm* und *alter*

Abbildung 10.4: Um die Variable *bereich* erweiterter neuer Datensatz *REGRESSION_O*...

Abbildung 10.5: Anfordern einer ordinalen Regressionsschätzung zwischen *bereich* ...

Abbildung 10.6: Ergebnis einer ordinalen Regressionsschätzung zwischen *bereich* u...

Abbildung 10.7: Anfordern einer multinomialen Regressionsschätzung zwischen *regi*...

Abbildung 10.8: Ergebnis einer multinomialen Regressionsschätzung zwischen *berei*...

## Kapitel 11

Abbildung 11.1: Datensatz DISKRIMINANZ1.SAV

Abbildung 11.2: Auswahl der Gruppierungsvariablen *pisa* mit Werten 1 und 2 (DISKR...

Abbildung 11.3: Auswahl der unabhängigen Diskriminanzvariablen: *klassengroesse* u...

Abbildung 11.4: Auswahl von Diskriminanz-Statistiken (hier in Bezug auf DISKRIMI...

Abbildung 11.5: Auswahl der (gleichen) Gruppengröße im Rahmen der Diskriminanzan...

Abbildung 11.6: Auswahl zu speichernder Variablen im Rahmen der Diskriminanzanal...

Abbildung 11.7: Gespeicherte Variablen im Datensatz DISKRIMINANZ1.SAV

Abbildung 11.8: Grundlegende Beurteilungskennziffern für die mit DISKRIMINANZ1.S...

Abbildung 11.9: Kernergebnisse für die mit DISKRIMINANZ1.SAV durchgeführte Diskr...

Abbildung 11.10: Datendatei CLUSTER1.SAV

Abbildung 11.11: Variablenauswahl bei der zweistufigen Clusteranalyse (CLUSTER1....

Abbildung 11.12: Outputsetzungen bei der zweistufigen Clusteranalyse (CLUSTER1.S...

Abbildung 11.13: Um die Variable *tsc_982* ergänzte Datendatei CLUSTER1.SAV bei de...

Abbildung 11.14: Output bei der zweistufigen Clusteranalyse (CLUSTER1.SAV)

Abbildung 11.15: Variablenauswahl mit der Vorgabe von zwei Clustern bei der K-Mi...

Abbildung 11.16: Abzuspeichernde Variablen bei der K-Mittelwerte-Clusteranalyse ...

Abbildung 11.17: Auswahl statistischer Kennziffern bei der K-Mittelwerte-Cluster...

Abbildung 11.18: Um die Variablen *qcl_1* und *qcl_2* ergänzte Datendatei CLUSTER1.S...

Abbildung 11.19: Output bei der K-Mittelwerte-Clusteranalyse (CLUSTER1.SAV)

Abbildung 11.20: Variablenauswahl bei der hierarchischen Clusteranalyse (CLUSTER...

Abbildung 11.21: Auswahl eines Dendrogramms bei der hierarchischen Clusteranalys...

Abbildung 11.22: Output der hierarchischen Clusteranalyse (CLUSTER1.SAV) (1)

Abbildung 11.23: Output der hierarchischen Clusteranalyse (CLUSTER1.SAV): Dendro...

Abbildung 11.24: Datendatei FAKTOREN1.SAV

Abbildung 11.25: Variablenauswahl bei der Faktorenanalyse (FAKTOREN1.SAV)

Abbildung 11.26: Statistiken bei der Faktorenanalyse (FAKTOREN1.SAV)

Abbildung 11.27: Wahl der konkreten Faktorenanalyse und »Faktoren-Entscheidungsr...

Abbildung 11.28: Auswahl eines Rotationsverfahrens (VARIMAX) bei der Faktorenana...

Abbildung 11.29: Faktorwerte für die einzelnen Fälle bei der Faktorenanalyse (FA...

Abbildung 11.30: Faktorenanalyse: Optionen (hier in Bezug auf FAKTOREN1.SAV)

Abbildung 11.31: Output der Faktorenanalyse (FAKTOREN1.SAV) (1)

Abbildung 11.32: Output der Faktorenanalyse (FAKTOREN1.SAV) (2)

Abbildung 11.33: Um zwei Variablen erweitere Datendatei FAKTOREN1.SAV im Rahmen ...

Abbildung 11.34: Variablenveränderungen in FAKTOREN1A.SAV

Abbildung 11.35: Anfordern und Erstellen einer Tabelle in SPSS (FAKTOREN1A.SAV)

Abbildung 11.36: Fallbezogene arithmetische Faktoren-Mittelwerte als Output eine...

Abbildung 11.37: Anforderung eines Streudiagramms zwischen den beiden Faktoren e...

Abbildung 11.38: Unbearbeitetes Streudiagramm zwischen den beiden Faktoren (FAKT...

Abbildung 11.39: Bearbeitetes Streudiagramm zwischen den beiden Faktoren: Linien...

Abbildung 11.40: Bearbeitetes Streudiagramm zwischen den beiden Faktoren: Datenb...

Abbildung 11.41: Beschriftetes und in den Nullpunkt verschobenes Streudiagramm z...

## Kapitel 12

Abbildung 12.1: Additive Zeitreihenverknüpfung: Konstante Saisonfigur (Quelle: F...

Abbildung 12.2: Multiplikative Zeitreihenverknüpfung: Variable Saisonfigur (Quel...

Abbildung 12.3: Datendateien *ZEITREIHE_ADDITIV1.SAV* (links) und *ZEITREIHE_MULTIP*...

Abbildung 12.4: Festlegung der Berechnung von gleitenden 4er-Durchschnitten für

Abbildung 12.5: Berechnung von gleitenden 4er-Durchschnitten für *variable_y* in d...

Abbildung 12.6: Glättung der Zeitreihe von *variable_y* in *ZEITREIHE_ADDITIV1.SAV* ...

Abbildung 12.7: Definition der relevanten Datumskomponente für die Datendatei *ZE*...

Abbildung 12.8: Von SPSS in die Datendatei *ZEITREIHE_ADDITIV1.SAV* eingefügte Dat...

Abbildung 12.9: Anfordern einer additiven Zeitreihen-Komponentenzerlegung (*ZEITR*...

Abbildung 12.10: Um weitere vier Variablen erweiterte Datendatei *ZEITREIHE_ADDIT*...

Abbildung 12.11: Saisonziffern bei konstanter Saisonfigur in Bezug auf *variable_*...

Abbildung 12.12: Anfordern einer multiplikativen Zeitreihen-Komponentenzerlegung...

Abbildung 12.13: Um weitere vier Variablen erweiterte Datendatei *ZEITREIHE_MULTI*...

Abbildung 12.14: Saisonziffern bei variabler Saisonfigur in Bezug auf *variable_y*

Abbildung 12.15: Datendatei *ÜBERLEBENSFUNKTION3.SAV*

Abbildung 12.16: Aufruf einer Sterbetafelanalyse (*ÜBERLEBENSFUNKTION3.SAV*)

Abbildung 12.17: Ergebnis einer Sterbetafelanalyse (*ÜBERLEBENSFUNKTION3.SAV*

Abbildung 12.18: Aufruf der Kaplan/Meier-Methode (*ÜBERLEBENSFUNKTION3.SAV*?)...

Abbildung 12.19: Aufruf der Kaplan/Meier-Methode (*ÜBERLEBENSFUNKTION3.SAV*?)...

Abbildung 12.20: Ergebnisse einer Kaplan/Meier-Analyse (*ÜBERLEBENSFUNKTION3.SAV*...

Abbildung 12.21: Ergebnisse einer Kaplan/Meier-Analyse (*ÜBERLEBENSFUNKTION3.SAV*...

Abbildung 12.22: Aufruf einer Cox-Regression (*ÜBERLEBENSFUNKTION3.SAV*) (1)

Abbildung 12.23: Aufruf einer Cox-Regression (*ÜBERLEBENSFUNKTION3.SAV*) (2)

Abbildung 12.24: Ergebnisse einer Cox-Regression (*ÜBERLEBENSFUNKTION3.SAV*)

## Kapitel 13

Abbildung 13.1: Beispiele für einen zweiseitigen Hypothesentest (oben) sowie für...

Abbildung 13.2: Datendatei *BAUTEILE.SAV*

Abbildung 13.3: Anfordern eines t-Mittelwerttests für die Variable *größe*...

Abbildung 13.4: Ergebnis eines t-Mittelwerttests für *größe* (*BAUTEILE.S*...

Abbildung 13.5: Datendatei *T-UNABHÄNGIG1.SAV*

Abbildung 13.6: Aufruf des t-Tests bei unabhängigen Stichproben mit *ergebnis* als...

Abbildung 13.7: Ergebnis des t-Tests bei unabhängigen Stichproben mit *ergebnis* a...

Abbildung 13.8: Welch-Test bei unabhängigen Stichproben mit *ergebnis* als Testvar...

Abbildung 13.9: Datendatei *MWU-UNABHÄNGIG1.SAV*

Abbildung 13.10: Eingabe der Testvariablen (*leistungsklasse*) und der Gruppierung...

Abbildung 13.11: Ergebnis des Mann/Whitney-U-Tests mit der Testvariablen *leistun*...

Abbildung 13.12: Datendatei *T-ABHÄNGIG1.SAV*

Abbildung 13.13: Aufruf des t-Tests bei abhängigen Stichproben mit *ergebnis2016* ...

Abbildung 13.14: Ergebnis des t-Tests bei abhängigen Stichproben mit *ergebnis201*...

Abbildung 13.15: Datendatei *WILCOXON-ABHÄNGIG1.SAV*

Abbildung 13.16: Aufruf des Wilcoxon-Tests bei abhängigen Stichproben mit *ergebn*...

Abbildung 13.17: Ergebnis des Wilcoxon-Tests bei abhängigen Stichproben mit *erge*...

Abbildung 13.18: Ergebnis des Vorzeichentests bei abhängigen Stichproben mit *erg*...

Abbildung 13.19: Datendatei *ANOVA_OHNE1.SAV*

Abbildung 13.20: Aufruf der einfaktoriellen ANOVA für *ergebnis* mit *qualifikation*

Abbildung 13.21: Zusätzlicher Aufruf der BONFERRONI-Korrektur bei der einfaktori...

Abbildung 13.22: Ergebnis der einfaktoriellen ANOVA (inklusive Bonferroni-Korrek...

Abbildung 13.23: Datendatei *KW-OHNE1.SAV*

Abbildung 13.24: Aufruf des Kruskal/Wallis-Tests bei unabhängigen Stichproben (h...

Abbildung 13.25: Aufruf des Kruskal/Wallis-Tests bei unabhängigen Stichproben mi...

Abbildung 13.26: Ergebnis des Kruskal/Wallis-Tests bei unabhängigen Stichproben ...

Abbildung 13.27: Datendatei *ANOVA_MIT1.SAV*

Abbildung 13.28: Aufruf von mehreren t-Tests bei Messwiederholung und drei Grupp...

Abbildung 13.29: Ergebnis von drei t-Tests bei Messwiederholung und drei Gruppen...

Abbildung 13.30: Datendatei *FRIEDMAN-MIT1.SAV*

Abbildung 13.31: Aufruf des Friedman-Tests mit den Testvariablen *ergebnis2015*, *e*...

Abbildung 13.32: Ergebnis des Friedman-Tests mit den Testvariablen *ergebnis2015*,...

## Kapitel 14

Abbildung 14.1: Datendatei *GLEICHVERTEILUNG.SAV*

Abbildung 14.2: Aufruf des Chi-Quadrat-Verteilungstests auf Gleichverteilung für...

Abbildung 14.3: Ergebnis des Chi-Quadrat-Verteilungstests auf Gleichverteilung f...

Abbildung 14.4: Datendatei *UNGLEICHVERTEILUNG.SAV*

Abbildung 14.5: Histogramm bezüglich der klassifizierten Variablen *einkommensmit*...

Abbildung 14.6: Ergebnis des Chi-Quadrat-Verteilungstests auf Gleichverteilung f...

Abbildung 14.7: Aufruf des Chi-Quadrat-Verteilungstests auf Normalverteilung für...

Abbildung 14.8: Ergebnis des Chi-Quadrat-Verteilungstests auf Normalverteilung f...

Abbildung 14.9: Ergebnis des Chi-Quadrat-Verteilungstests auf Normalverteilung f...

Abbildung 14.10: Aufruf des Kolmogorov/Smirnov-Verteilungstests auf Gleichvertei...

Abbildung 14.11: Ergebnis des Kolmogorov/Smirnov-Verteilungstests auf Gleichvert...

Abbildung 14.12: Aufruf des Kolmogorov/Smirnov-Verteilungstests auf Gleich- und ...

Abbildung 14.13: Ergebnis des Kolmogorov/Smirnov-Verteilungstests auf Gleich- un...

Abbildung 14.14: Aufruf des Kolmogorov/Smirnov- und des Shapiro/Wilk-Verteilungs...

Abbildung 14.15: Ergebnisauszug bezüglich des Kolmogorov/Smirnov- und des Shapir...

## Kapitel 15

Abbildung 15.1: Aufruf des Chi-Quadrat-Tests (*GESCHL_FARBE.SAV*)

Abbildung 15.2: Chi-Quadrat-Kontingenzkoeffiziententest für den Zusammenhang zwi...

Abbildung 15.3: Aufruf des Rangkorrelationskoeffizienten nach Spearman (*SPEERWER*...

Abbildung 15.4: Test für den Zusammenhang zwischen *weite2016* und *weite2017* beim ...

Abbildung 15.5: Aufruf des Korrelationskoeffizienten nach Bravais/Pearson (*SPEER*...

Abbildung 15.6: t-Test für den Zusammenhang zwischen *weite2016* und *weite2017* bei...

## Kapitel 16

Abbildung 16.1: Download-Möglichkeiten für die R-Essentials aus dem Internet (`ht`...

Abbildung 16.2: ZIP-Datei für R-Essentials (SPSS 28, auch für SPSS 29 passend)

Abbildung 16.3: Desktop-Button für R 4.2.1

Abbildung 16.4: RStudio-Maske (hier für R 4.2.1)

Abbildung 16.5: In SPSS installierte (R- beziehungsweise) Python-Anwendungen

Abbildung 16.6: Suche nach in SPSS installierten R-Anwendungen

Abbildung 16.7: Im SPSS-Erweiterungshub (unter UNTERSUCHEN) gefundene R-Anwendun...

Abbildung 16.8: Voraussetzungen für die Funktion RREGRIDGE (Beispiel)

Abbildung 16.9: Abruf der R-Erweiterung RREGRIDGE (Beispiel)

Abbildung 16.10: Erfolgreiche Installation der R-Erweiterung RREGRIDGE (Beispiel...

Abbildung 16.11: Aufruf von RREGRIDGE (Beispiel) in SPSS

Abbildung 16.12: Installation eines R-Pakets (hier von MASS) innerhalb von SPSS

Abbildung 16.13: Koeffizientenwerte für den Regressionszusammenhang zwischen *aus*...

Abbildung 16.14: Koeffizientenwerte für den Regressionszusammenhang zwischen *aus*...

## Kapitel 17

Abbildung 17.1: Ausgangsdatensatz *ANGST1.SAV* vor der Umskalierung von bestehende...

Abbildung 17.2: Recodierter Datensatz (Umskalierung in denselben Variablen) (*ANG*...

Abbildung 17.3: Recodierter Datensatz (in neue Variablen) (*ANGST1.SAV*)

Abbildung 17.4: Klassenbildung in DO-REPEAT-Schleife (*DO-REPEAT2.SAV*)

Abbildung 17.5: Bildung von neuen Variablen mittels LOOP-Sequenz (*ZUFRIEDENHEIT1*...

Abbildung 17.6: Datensatz mit vier Variablen vor LOOP-Prozedur (*LOOP_AUSGANG4_1*....

Abbildung 17.7: Datensatz mit zwei Variablen nach LOOP-Prozedur (*LOOP_ZIEL1_4.SA*...

Abbildung 17.8: Datensatz nach DO-IF-Schleife (*DO-REPEAT3.SAV*)

Abbildung 17.9: Ausgangsdatensatz *REGION.SAV* mit einer Variablen vor einer DO-IF...

Abbildung 17.10: Datensatz *REGION.SAV* mit drei Variablen nach einer DO-IF-Schlei...

Abbildung 17.11: Datendatei *ZUFRIEDEN.SAV* mit fehlenden Angaben

Abbildung 17.12: Datendatei *ZUFRIEDEN.SAV* mit fehlerhafter Zählvariablen *zufried*...

Abbildung 17.13: Datendatei *ZUFRIEDEN.SAV* mit korrekter Zählvariablen *zufrieden_*...

## Kapitel 18

Abbildung 18.1: Installation des Makros PROCESS in SPSS

Abbildung 18.2: In SPSS eingebettetes Makro PROCESS

Abbildung 18.3: Datendatei *ZUFRIEDENHEIT1_TO_5.SAV* mit fünf kategorialen Variabl...

Abbildung 18.4: Datendatei mit fünf kategorialen Variablen und aus ihnen abgelei...

Abbildung 18.5: Datensatz *DO-REPEAT3.SAV* für Prozess-Schleifen-Makro

Abbildung 18.6: Output nach Prozess-Schleifen-Makro (*DO-REPEAT3.SAV*)

Abbildung 18.7: Erstellen eines Makrosymbols (1)

Abbildung 18.8: Erstellen eines Makrosymbols (2)

Abbildung 18.9: Erstellen eines Makrosymbols (3)

Abbildung 18.10: Erstellen eines Makrosymbols (4)

## Kapitel 19

Abbildung 19.1: Datendateien *VERMOEGEN_A.SAV*, *VERMOEGEN_B.SAV*, *VERMOEGEN_C.SAV* u...

Abbildung 19.2: Fiktive Armutsquoten für die Länder A, B, C und D (*VERMOEGEN_A.S*...

Abbildung 19.3: (Personenbezogener) Ungleichheitsoutput (Gesamtvermögen) für die...

Abbildung 19.4: (Personen- und haushaltsbezogener) Ungleichheitsoutput (Gesamt-,...

Abbildung 19.5: (Personenbezogener) Ungleichheitsoutput (Gesamtvermögen) für dre...

# Einführung

Durch die Entwicklung umfassender statistischer Berichtssysteme im vergangenen Jahrhundert – zum Beispiel in der Volkswirtschaftslehre: Entwicklung der Volkswirtschaftlichen Gesamtrechnungen – erlebte das Arbeiten mit (Massen-)Daten, das heißt die empirische Forschung, einen ungeheuren Auftrieb. Ermöglicht wurde dies durch gewaltige Sprünge der Hardware (gerade in Bezug auf Prozessorleistung und Speicherkapazitäten). In diesem Zusammenhang sind auch die statistischen Softwarehilfestellungen ausgebaut worden.

Eines der bekanntesten Statistikprogramme ist SPSS®. Die Abkürzung SPSS kommt von »Statistical Package for the Social Sciences« (zeitweise stand SPSS im Übrigen auch mal für »Superior Performing Software System«). Zwischen 2009 und 2010 hieß SPSS kurzzeitig PAW = Predictive Analysis SoftWare.

Das inzwischen von IBM® angebotene SPSS wurde von seinen Funktionen her ständig erweitert beziehungsweise verbessert. Dies gilt nicht zuletzt für die Benutzerfreundlichkeit des Programms. Ich habe anfangs der 1990er-Jahre die empirischen (massenstatistischen) Auswertungen im Rahmen meiner Doktorarbeit noch mit einer frühen SPSS-Version auf einem Großrechner – im »Vor-Windows-Zeitalter« – bewerkstelligt (SPSS$^{X}$ 2.0). Hierbei musste ich höllisch auf Programmierfallen aufpassen. Heutzutage brauchen Sie zur Nutzung von SPSS im Grunde genommen keine (größeren) Programmierkenntnisse mehr, um mittels der SPSS-Bedienoberfläche in einer Geschwindigkeit, die noch vor 30 Jahren undenkbar schien, komplizierte statistische Berechnungen durchzuführen.

Dieser Gewinn an Benutzerfreundlichkeit ist aber zugleich eine Gefahr, verleitet dies doch mitunter zum »Quick-and-dirty-Arbeiten«, soll heißen: zu Auswertungen gewissermaßen »auf Knopfdruck«, ohne dass die Auswertungsergebnisse inhaltlich immer besonders sinnvoll wären. Mit meinem Buch möchte ich helfen, solche »Nonsens-Auswertungen« zu vermeiden. Außerdem möchte ich Hilfestellungen anbieten, wie die

mithilfe von SPSS gewonnenen Ergebnisse inhaltlich-sinnvoll interpretiert werden können. Hierbei habe ich mich bemüht, die Funktionalität von SPSS so umfassend wie möglich darzustellen, damit Sie aus diesem Buch den größtmöglichen Nutzen ziehen.

Das Buch bezog sich in seiner ersten Auflage auf die SPSS-Version 24.0 und in dieser aktualisierten, zweiten Auflage auf die SPSS-Version 29.0 mit möglichst vielen Programmpaketen. Trotz des Umfangs dieses Buchs musste eine Auswahl aus den vielfältigen SPSS-Programmelementen getroffen werden. Das hat dazu geführt, dass Erweiterungen in Form bestimmter SPSS-Zusatzmodule – wie zum Beispiel AMOS (= Analysis of Moment Structures) zur Schätzung von Strukturgleichungsmodellen – im Folgenden (eher) nicht behandelt werden.

Ich würde mich sehr freuen, wenn ich bei Ihnen durch das Buch beziehungsweise insbesondere durch diese Neuauflage des Buchs zumindest ein bisschen Begeisterung auch für solche statistischen Zusammenhänge erwecken würde, die Ihnen am Anfang vielleicht ziemlich kompliziert und eher langweilig erschienen sind.

# *Über dieses Buch*

Sie leben in einer Zeit der Informationsflut: Tagtäglich werden Sie mit vielen Informationen konfrontiert, das heißt mit Daten aus Politik, Wirtschaft, Kultur und so weiter, aber auch aus dem eigenen Alltagserleben. Um in diesem »Wirrwarr« den Überblick zu behalten, gilt es, diese Informationen zu sortieren, das heißt zu strukturieren, um daraus wesentliche Schlussfolgerungen ziehen und um diese Informationen interpretieren sowie für Ihre Zwecke sinnvoll nutzen zu können. Bei all diesen Dingen ist die Statistik ein unentbehrliches Mittel zum Zweck. Aufgrund der Masse an verfügbaren Daten brauchen Sie aus meiner Sicht leistungsfähige Statistiksoftware. Eine solche ist unzweifelhaft das Statistikprogramm SPSS, um das es in diesem Buch geht. Dabei knüpft dieses Buch an das erfolgreiche, schon bestehende Wiley-Buch »SPSS 26 für Dummies« von Felix Brosius an, indem es über die dort vorrangig präsentierten SPSS-»Basics« hinaus gerade auch

für spezielle Statistikfragestellungen SPSS-Lösungswege anbietet. Dies betrifft in erster Linie die umfangreiche Behandlung der statistischen Verfahren Regressionsanalyse und Testtheorie, aber auch das Thema der Automatisierung von SPSS-Befehlsfolgen. Das lesenswerte Buch von Felix Brosius wird allerdings nicht unbedingt benötigt, um dieses Buch hier zu verstehen.

Dem Vorgenannten Folge leistend, werde ich im Buch Folgendes mit Ihnen durchgehen:

- ✔ Es werden zunächst Grundlagen der statistischen Datenverarbeitung vermittelt, natürlich mit speziellem Blick auf SPSS.
- ✔ Danach geht es um das Beschreiben und die Darstellung statistischer Sachverhalte im Rahmen der sogenannten deskriptiven (= beschreibenden) Statistik.
- ✔ Demgegenüber beschäftigt sich die sogenannte induktive (= schließende) Statistik mit dem Schließen aus Stichprobenbefunden auf die jeweilige statistische Gesamtheit, wobei wahrscheinlichkeitstheoretische Grundzusammenhänge bestehen.
- ✔ Um Ihnen den breiten Anwendungsbereich von SPSS beispielhaft zu verdeutlichen und Ihnen das Erstellen eigener SPSS-Programme zu erleichtern, biete ich Ihnen außerdem auch noch – gewissermaßen als Muster – einige spezielle, selbst geschriebene SPSS-Programme für bestimmte Zusammenhänge (zum Beispiel für die Messung sozialer Ungleichheit) an. Dabei wird ausführlich auf die Automatisierung von SPSS-Befehlsfolgen – vor allem in Form sogenannter Makros – eingegangen.

Sie müssen im Übrigen die Kapitel in dem Buch nicht in einer bestimmten Reihenfolge lesen. Wenn Sie an einem Sachverhalt besonders interessiert sind, können Sie an der entsprechenden Textstelle unmittelbar nachschlagen. In diesem Sinne ist das Buch auch eher als Nachschlagewerk und weniger als Lehrbuch konzipiert.

# Törichte Annahmen über meine Leser

Dieses Buch bietet sich vor allem für Studierende, wissenschaftliche Mitarbeiter und dergleichen an Hochschulen (Universitäten, Fachhochschulen, Berufsakademien), aber auch in (Politik-)Beratungs- und Forschungsinstitutionen an. Darüber hinaus ist es aber durchaus auch für Praktiker geeignet, die mit nicht immer ganz einfachen empirischen Fragestellungen konfrontiert sind. Hierbei denke ich an Praktiker aus Industrie, Handel, Verwaltungen und Behörden, aus dem Gesundheitswesen oder aus dem Medienbereich. Auch diese Personenkreise, so ist meine sicherlich nicht unbegründete Hoffnung, sollen durch dieses Buch mit den für ihre Tätigkeiten wichtigen statistischen Zusatzkenntnissen ausgestattet werden.

Sie, liebe Leserinnen und Leser, müssen dabei gar nicht so viel Vorwissen mitbringen. Es ist im Grunde genommen lediglich ein gewisses statistisches Grundverständnis vonnöten. Zur Wiederholung derartiger Grundlagen sei auf mein Statistikbuch »Statistik für Wirtschafts- und Sozialwissenschaftler für Dummies«, Weinheim 2015: Wiley, verwiesen. Allerdings bin ich überzeugt, dass Sie auch ohne nennenswerte Statistikkenntnisse nachfolgend gut zurechtkommen werden, weil ich mich bemüht habe, an den einzelnen Textstellen kurz zum Verständnis der entsprechenden Zusammenhänge die jeweils notwendigen Statistikausführungen zu machen.

# Wichtige Konventionen in diesem Buch

Fachbegriffe werden im Text **fett** hervorgehoben. Zum Teil werden etwas weniger bedeutsame Begriffe beziehungsweise Formulierungen auch *kursiv* hervorgehoben. Die entsprechenden Begriffe finden sich im Übrigen auch im Sachwortverzeichnis, wodurch ein schnelles Auffinden zentraler Begriffe im Text ermöglicht wird.

SPSS-Menüelemente werden in Kapitälchen (das heißt: in kleinen Großbuchstaben) dargestellt, zum Beispiel: DATEI|SPEICHERN. Das heißt: Die übergeordnete Menüebene findet sich vor dem senkrechten Strich, die nachgeordnete Menüebene danach. Ebenfalls großgeschrieben sind im Text SPSS-Syntaxbefehle, zum Beispiel: COMPUTE. Demgegenüber werden Variablenbezeichnungen (zum Beispiel: *variable1*) kleingeschrieben und zudem kursiv dargestellt. Ebenfalls in kursiver Schrift dargestellt, aber großgeschrieben sind im Folgenden Verzeichnisangaben (zum Beispiel: *C:/SPSS-BUCH/DATEIEN*) und Dateiangaben (zum Beispiel: *123.SAV*).

# Wie dieses Buch aufgebaut ist

Das vorliegende Buch setzt sich aus sechs Teilen zusammen. Zunächst einmal werden Sie in Teil I mit eher allgemeinen Hilfestellungen der statistischen Datenverarbeitung versorgt, ehe Sie in Teil II die Anwendungsmächtigkeit von SPSS in Bezug auf deskriptive Statistikfragestellungen kennenlernen. Ergänzend kommen dann in Teil III statistische Aspekte zur Sprache, die sowohl der deskriptiven als auch der induktiven Statistikebene zugeordnet werden können (bei Bezugnahme auf die Grundgesamtheit: deskriptive Statistik, bei Stichproben: induktive Statistik). »Reine« Themen der induktiven Statistik werden in Bezug auf SPSS anschließend in Teil IV angesprochen. In Teil V werden unter anderem einige spezielle, selbst erstellte SPSS-Programme diskutiert. Zuletzt folgt, wie bei jedem Dummies-Buch, der Top-Ten-Teil. Letzterer bietet Ihnen Kurzübersichten über SPSS in Form von To-do-Listen, Tipps und so weiter.

## Teil I: Grundlagen der statistischen Datenverarbeitung und von SPSS

Hier geht es nach einigen Vorbemerkungen zur Bedeutung von Daten und ihrer Skalierung um die grundlegende Benutzerführung/-steuerung in SPSS. Das heißt: Es werden das Einlesen und die Verarbeitung von

Daten mit SPSS angesprochen. Dabei spielen verschiedene Datei- und Variablenoperationen eine Rolle.

## *Teil II: Deskriptive SPSS-Statistiken*

Gegenstand dieses Teils sind beschreibende (deskriptive) SPSS-Statistikoperationen. Hierbei wird auf eindimensionale Häufigkeitsverteilungen inklusive ihrer grafischen Darstellungsmöglichkeiten eingegangen. Es werden wichtige statistische Kennziffern wie Mittelwerte oder Streuungsmaße behandelt.

## *Teil III: Deskriptive und induktive SPSS-Statistiken gemischt*

Teil III beschäftigt sich mit statistischen Zusammenhängen zwischen zwei beziehungsweise noch mehr Variablen. Dies geschieht unter Bezugnahme auf die Korrelations- und die Regressionsmethodik. Betrachtungen unter anderem zur Cluster- und zur Zeitreihenanalyse schließen Teil III ab.

## *Teil IV: Induktive SPSS-Statistiken*

In Teil IV geht es im Wesentlichen um das statistische Testen. Dabei werden sowohl Tests auf (Mittelwert-)Unterschiede zwischen Variablen als auch Tests auf bestimmte Verteilungsformen der interessierenden Variablen präsentiert.

## *Teil V: SPSS-Syntaxprogrammen für Fortgeschrittene*

Zunächst werden in Teil V Ausführungen zu der Einbindung von Elementen aus den Programmen R und Python in SPSS gemacht. Danach wird auf die Automatisierung von SPSS-Syntaxbefehlen – vor allem in Form von Makros – Bezug genommen. Nach weiteren allgemeinen (zum Teil wiederholenden) Anmerkungen zur Struktur von SPSS-Syntaxprogrammen wird beispielhaft auf SPSS-Syntaxprogramme bezüglich der Armuts-/Reichtumsmessung beziehungsweise allgemein der Ungleichheitsmessung eingegangen.

## *Teil VI: Top-Ten-Teil*

Hinzu kommt noch der Top-Ten-Teil, in dem die Highlights von SPSS gelistet angegeben sind. Hierbei wird auf die zehn wichtigsten SPSS-Befehle Bezug genommen, und es werden die zehn wichtigsten SPSS-Tipps gegeben.

## *Symbole in diesem Buch*

In diesem Buch wird eine Reihe von Symbolen genutzt. Diese Symbole stelle ich Ihnen im Folgenden kurz vor.

Statistik verleitet leicht zu Fehlinterpretationen. Sie kennen sicherlich die Frage: »Wie lügt man am besten mit Statistik?«, oder Sie kennen den Spruch: »Traue keiner Statistik, die du nicht selbst gefälscht hast.« Beim »Vorsicht-Icon« geht es entsprechend vor allem darum, mit den dort von mir gegebenen Informationen zu vermeiden, dass Sie in entsprechende Statistikfallen tappen.

Manchmal ist das »Vorsicht-Icon« noch zu vorsichtig angelegt. Dies gilt etwa dann, wenn Sie durch Unachtsamkeit Gefahr laufen, Ihre kompletten Daten zu vernichten. In solchen »Katastrophen«-Fällen ist eine besonders nachhaltige Warnung angebracht.

Um Sie nicht nur zur Vorsicht zu mahnen (mittels der beiden eben genannten Symbole), sondern um Ihnen auch Tipps zu geben, wie Sie bestimmte Fallstricke mit SPSS umgehen beziehungsweise wie Sie möglichst effizient mit SPSS arbeiten, werden Ihnen durch das »Tipp-Icon« hoffentlich hilfreiche Anmerkungen seitens meiner Wenigkeit gegeben.

Viele Dinge lernen sich leichter, wenn sie mit Beispielen unterlegt werden. Dies veranschaulicht an den entsprechenden Stellen die dortigen Zusammenhänge. Hierzu dienen im Buch viele mehr oder

weniger kleine SPSS-Beispieldatensätze. Sie sind im Internet unter http://www.wiley-vch.de/ISBN978-3527-72068-2 abrufbar. Damit das »Beispiel-Symbol« angesichts der vielen Beispiele im Buch nicht inflationär eingesetzt wird, finden Sie es nur an besonders prägnanten Stellen, wo komplexe Zusammenhänge unbedingt einer zusätzlichen beispielhaften, eher speziellen Erläuterung bedürfen.

Unter diesem Symbol werden wichtige statistische Sachverhalte beziehungsweise Kennziffern wie zum Beispiel Mittelwerte erläutert, damit die zugehörigen SPSS-Operationen besser nachvollzogen werden können. Wem das aber zu viel »theoretischer Ballast« ist, der kann über diese theoretischen Hintergrundinformationen hinweglesen, ohne dass seine »Ausbildung« zum SPSS-Experten dadurch grundsätzlich infrage gestellt wäre.

# *Wie geht es weiter im Buch?*

Es folgt Teil I. In diesem Teil werden die Grundlagen der statistischen Datenverarbeitung mit SPSS vermittelt. Zunächst wird dort der Aufbau von SPSS-Datendateien beschrieben, ehe das Zusammenlegen und die Bearbeitung von Datendateien skizziert werden. In diesem Zusammenhang wird auch auf spezielle SPSS-Syntaxbefehle eingegangen.

# Teil I

# Grundlagen der statistischen Datenverarbeitung und von SPSS

# IN DIESEM TEIL …

- ✔ Vorbemerkungen zu SPSS
- ✔ Daten als Basis der SPSS-Analyse
- ✔ Datendateien zusammenbringen
- ✔ Datendateien bearbeiten
- ✔ Zur SPSS-Syntax

# Kapitel 1

# Vorbemerkungen zu SPSS

## IN DIESEM KAPITEL

Bedeutung von Statistik

Bedeutung von Statistikprogrammen

Relevanz von SPSS

Start von SPSS

Ziel des ersten Kapitels ist es, die Bedeutung von Statistik beziehungsweise von Statistikprogrammen zu veranschaulichen. Hierbei wird naheliegenderweise insbesondere auf das Statistikprogramm SPSS eingegangen.

## *Bedeutung von Statistik*

Sie kennen bestimmt den mehr oder weniger originellen Spruch: »Wie lügt man am besten mit **Statistik**?« Da Sie zu diesem Buch gegriffen haben, gehe ich jetzt einfach mal davon aus, dass Sie sich trotzdem für Statistik interessieren.

Vielleicht führen Sie täglich eine Statistik über Ihr Körpergewicht oder Ihre Blutzuckerwerte oder Sie schreiben Ihre Siege beziehungsweise Niederlagen beim Schach regelmäßig auf. Oder: Sie führen Buch über das Wachstum einer Pflanze und so weiter und so fort. Vielleicht müssen Sie sich ja aber auch von Berufs wegen mit Statistiken beschäftigen, etwa weil Sie in der Personalabteilung Ihrer Firma die Gehaltsstatistik führen müssen.

All dies sind statistische Anwendungsgebiete. Im Grunde genommen können Sie alle Bereiche des Lebens statistisch erfassen. Durch die

beachtlichen technologischen Sprünge der vergangenen 50 Jahre in der **Informationsverarbeitung** stehen Ihnen hierzu mittlerweile maschinelle Hilfen (in Form leistungsfähiger Rechner) ebenso zur Verfügung wie vielfältige **Statistiksoftware**. Dadurch werden Ihre Aktivitäten als Statistiker erheblich erleichtert.

# *Bedeutung von Statistikprogrammen*

Braucht man dazu aber wirklich ein mehr oder weniger teures Statistikprogramm? Kann man das »bisschen Statistik« nicht im Kopf oder schlimmstenfalls mit einem (programmierbaren) Taschenrechner erledigen?

»Ja«, könnte man antworten, wenn man es nur mit wenigen Zahlen zu tun hat und auch nur »einfache« Statistiken benötigt. Die Lehrerin, die für ihre 20 Schüler eine Notenübersicht über die jüngste Mathearbeit und, sozusagen als Höchstes der Gefühle, einen Mittelwert aus den einzelnen Noten berechnen möchte, braucht gewiss nicht mehr als einen Taschenrechner. Vielleicht – wenn sie im Kopfrechnen gut ist, was bei einer Mathe-Lehrerin ja durchaus vorkommen soll – benötigt sie noch nicht einmal einen solchen.

# *Tabellenkalkulationsprogramme als Alternative*

Stellen Sie sich aber einmal vor, Sie hätten 2.000 benotete Mathearbeiten vorliegen und müssten diese statistisch verarbeiten (also etwa einen Mittelwert berechnen), dann ist die Situation schon etwas anders. Selbst mit einem Taschenrechner wäre hier die Erstellung des Notenspiegels, eventuell zuzüglich der Ermittlung eines Mittelwerts, deutlich aufwendiger. Liegen die betreffenden Daten elektronisch vor, bräuchte man im Grunde genommen aber immer noch kein Statistikprogramm wie SPSS, da man in diesem Fall zum Beispiel mit einem **Tabellenkalkulationsprogramm** wie EXCEL® wohl bis hierher auch zurechtkäme.

Gleichwohl wird bei (sehr) großen Datenmengen die Nutzung eines Tabellenkalkulationsprogramms recht schnell unübersichtlich. Hinzu kommt, dass weiterführende Statistikanwendungen wie zum Beispiel bestimmte (auf spezielle Fragestellungen zugeschnittene) statistische Testverfahren nicht beziehungsweise nur teilweise und abgespeckt in EXCEL zur Verfügung stehen.

Wollen oder können Sie – nicht jeder ist schließlich ein Experte in Sachen **Informationstechnologie** – die entsprechenden Statistikprozeduren nicht selbst in einer Ihnen geläufigen Programmiersprache programmieren, bietet sich in diesen Fällen der Rückgriff auf (fertig programmierte, mehr oder weniger benutzerfreundliche) Statistiksoftware an.

Das Vorstehende heißt nicht, dass Statistiksoftware bei allen statistischen Fragestellungen einem Tabellenkalkulationsprogramm oder Ähnlichem überlegen ist. Statistiksoftware ist nämlich auf die Verarbeitung empirischer Datensätze spezialisiert.

Für eher modelltheoretisch interessante Fragestellungen im Sinne von Gedankenexperimenten – etwa aus dem Gebiet der Wahrscheinlichkeitsrechnung beziehungsweise genauer: der Kombinatorik – reicht der Bezug auf Programme wie EXCEL aus beziehungsweise ist aus meiner Sicht sogar vorteilhaft. Ein Beispiel ist die Entnahme einer bestimmten Anzahl von Kugeln unterschiedlicher Farbe aus einem Behältnis (»Urne«) ohne Zurücklegen der gezogenen Kugeln (sogenanntes »Urnenmodell ohne Zurücklegen« = »Hypergeometrische Verteilung«). Hier kann man die betreffenden Wahrscheinlichkeiten gewissermaßen »auf Knopfdruck« in EXCEL berechnen. Will man etwa die Wahrscheinlichkeit für sechs Richtige im Lotto »6 aus 49« ermitteln, muss man in EXCEL einfach eingeben: HYPGEOM.VERT(6;6;6;49;0). Daraufhin erhält man das für Lottospieler ernüchternde Ergebnis: 7,15112E-08 (also: 0,0000000715112 oder in Prozent: 0,00000715112 Prozent).

# *Verschiedene Statistikprogramme*

An dieser Stelle haben Sie indes immer noch die »Qual der Wahl«, aus der Vielzahl an statistischer »Spezialsoftware« ein geeignetes Statistikprogramm auszuwählen. In den vergangenen Jahrzehnten sind nämlich zur Umsetzung der vorstehend genannten Formen der Datenauswertung viele leistungskräftige **Statistikprogramme** entwickelt worden.

Ohne Anspruch auf Vollständigkeit können etwa folgende Statistikprogramme genannt werden:

- ✔ STATA®,
- ✔ R®,
- ✔ SAS®,
- ✔ EVIEWS®,
- ✔ LIMDEP®,
- ✔ RATS® und
- ✔ eben SPSS®.

## *STATA*

**STATA** zum Beispiel weist einen großen Funktionsumfang auf. Allerdings wird dem Programm eine gewisse Trägheit in Bezug auf die Berücksichtigung neuer Methoden ebenso nachgesagt wie eine vergleichsweise schwierige Vereinbarkeit mit anderer Statistiksoftware.

## *R*

**R** ist eine sogenannte Open-Source-Statistiksoftware; das heißt: R steht kostenlos zur Verfügung. Diese Software erfordert relativ gute Methodenkenntnisse und eine längere Einarbeitungszeit. Durch die zur Verfügung gestellte Entwicklungsumgebung **RStudio** ist das Arbeiten mit R allerdings deutlich benutzerfreundlicher ausgestaltet worden. R besticht durch die große Anzahl an Statistikpaketen, die vergleichsweise

schnell und aktuell der Nutzergemeinde zur Verfügung gestellt werden. Zum Teil allerdings ist dies mit dem Nachteil verbunden, dass nicht alle Pakete ganz ausgereift sind. Vorteilhaft an R ist, dass es sehr gut in andere Statistikprogramme integrierbar ist. Beispielsweise gibt es auch eine Schnittstelle zu SPSS. Auf diese Art und Weise kann SPSS um R-Prozeduren ergänzt werden. Dies gilt beispielsweise im Zusammenhang mit der sogenannten Ereignisanalyse, die nicht direkt in SPSS zu finden ist. Auf die Verknüpfung von SPSS mit R wird in Teil V noch näher eingegangen.

### *SAS*

**SAS** findet vor allem in der Biometrie, der klinischen Forschung, aber auch im Bankensektor Anwendung und ist für die Verarbeitung (sehr) großer Datensätze sehr gut geeignet. Darüber hinaus enthält SAS eine Vielzahl an (statistischen) Modellen, ist allerdings in der Anschaffung nicht ganz billig.

### *EVIEWS, RATS, LIMDEP*

EVIEWS, RATS und LIMDEP schließlich sind ökonometrische Spezialsoftware. Hierbei haben **EVIEWS** und **RATS** ihren Fokus auf Zeitreihendaten beziehungsweise -modellen. **LIMDEP** fokussiert auf logistische Regressionen für diskrete abhängige Variablen, wenngleich auch hier für Zeitreihenanalysen (nicht zuletzt für Panelanalysen) wirkmächtige Modelle zur Verfügung stehen.

## *Zur Relevanz von SPSS*

Alle vorgenannten Programme enthalten also die gängigen deskriptiven und induktiven Statistikprozeduren. In diesem Buch geht es – wie Sie sicherlich schon vermutet haben – im Besonderen um das Programm **SPSS**.

Was macht nun ausgerechnet SPSS im Vergleich zu anderer Statistiksoftware so besonders? Der Preis kann es eigentlich nicht sein, denn im Vergleich zum kostenlosen R und etwa gegenüber STATA ist die SPSS-Lizenz in der Regel nach einer gewissen Zeit (etwa jedes Jahr)

für einiges Geld zu verlängern. Daher verwundert es nicht, dass SPSS in besonderem Maße an (öffentlichen) Hochschulen verbreitet ist.

In letzter Zeit ist mitunter von einer statistischen Alternativsoftware zu SPSS die Rede. Die Software **PSPP** wird ausdrücklich als Kontrapunkt zu SPSS gesehen. Dies wird schon allein daran deutlich, dass die drei S und das eine P von SPSS bei PSPP genau vertauscht worden sind. Die Idee von PSPP ist der (kosten-)freie Ersatz von SPSS. Allerdings stehen die entsprechenden statistischen Funktionen in PSPP bislang nur sehr begrenzt zur Verfügung, soll heißen, dass sich PSPP sozusagen immer noch in der Entwicklungsphase befindet.

Warum aber sollten Sie dennoch ausgerechnet auf SPSS zurückgreifen?

Da ist zum einen die große Verbreitung von SPSS in den Sozialwissenschaften und der Psychologie, (zumindest früher) zum Teil verbunden mit der Vorgabe, wissenschaftliche Empiriearbeiten in wissenschaftlichen Journals auf SPSS zu basieren. Außerdem ist SPSS – etwa im Gegensatz zu R (trotz der Erweiterung von R um RStudio) – in (sehr) leicht zu erlernender Weise mittels Benutzermenü zu bedienen (wobei die einzelnen Prozeduren derselben Logik folgen), und es werden bei SPSS eher geringe Methodenkenntnisse vorausgesetzt. Eigentlich ist es in SPSS nur wichtig, dass man weiß, was man berechnen möchte und wie die Ergebnisse zu interpretieren sind. Den »Rest« macht sozusagen SPSS als ein einfach zu bedienendes statistisches Hilfsmittel von alleine.

In der Leichtigkeit der Bedienung von SPSS liegt allerdings eine nicht zu unterschätzende Gefahr: Statistisch eher unbedarfte Nutzer von SPSS könnten dazu verleitet werden, per Knopfdruck statistische Analysen »gewissermaßen ohne Sinn und Verstand« durchzuführen. Beispiele hierfür sind etwa **Nonsenskorrelationen** – also unsinnige statistische Zusammenhänge – wie etwa zwischen der Anzahl an Betonmischmaschinen und der Anzahl der Urlaube von Haushalten in einem bestimmten Land und in einem

bestimmten Jahr oder aber auch die Anwendung nicht passender statistischer Tests auf bestimmte Sachverhalte (zum Beispiel eines einfachen Mittelwerttests, wenn man eigentlich an der Streuung von Variablen interessiert ist).

# *Start von SPSS*

Damit Sie aber überhaupt mit SPSS arbeiten können, müssen Sie es trivialerweise erst einmal mit der betreffenden **Anwendungsdatei** *SETUP.EXE* auf Ihrem Rechner installieren, wobei Sie sich natürlich auf die Erfordernisse Ihres **Betriebssystems** (Windows®, Mac OS® oder Linux®) beziehen müssen. Wenn Sie eine Lizenzversion erworben haben, müssen Sie im Zuge des **Installationsprozesses** unter anderem die betreffende Lizenznummer angeben. Sie erhalten nach erfolgreicher Lizenzierung dann eine Mitteilung, wie sie in Abbildung 1.1 (dort für die SPSS-Version 24, die in der ersten Auflage dieses Buchs verwendet wurde) enthalten ist. (Ich darf Ihnen im Übrigen versichern, dass Abbildung 1.1 die erste und einzige Abbildung in diesem neu aufgelegten Buch ist, die auf SPSS 24 Bezug nimmt. Alle anderen SPSS-bezogenen Abbildungen in diesem Buch beziehen sich auf die für die Neuauflage verwendete SPSS-Version 29.)

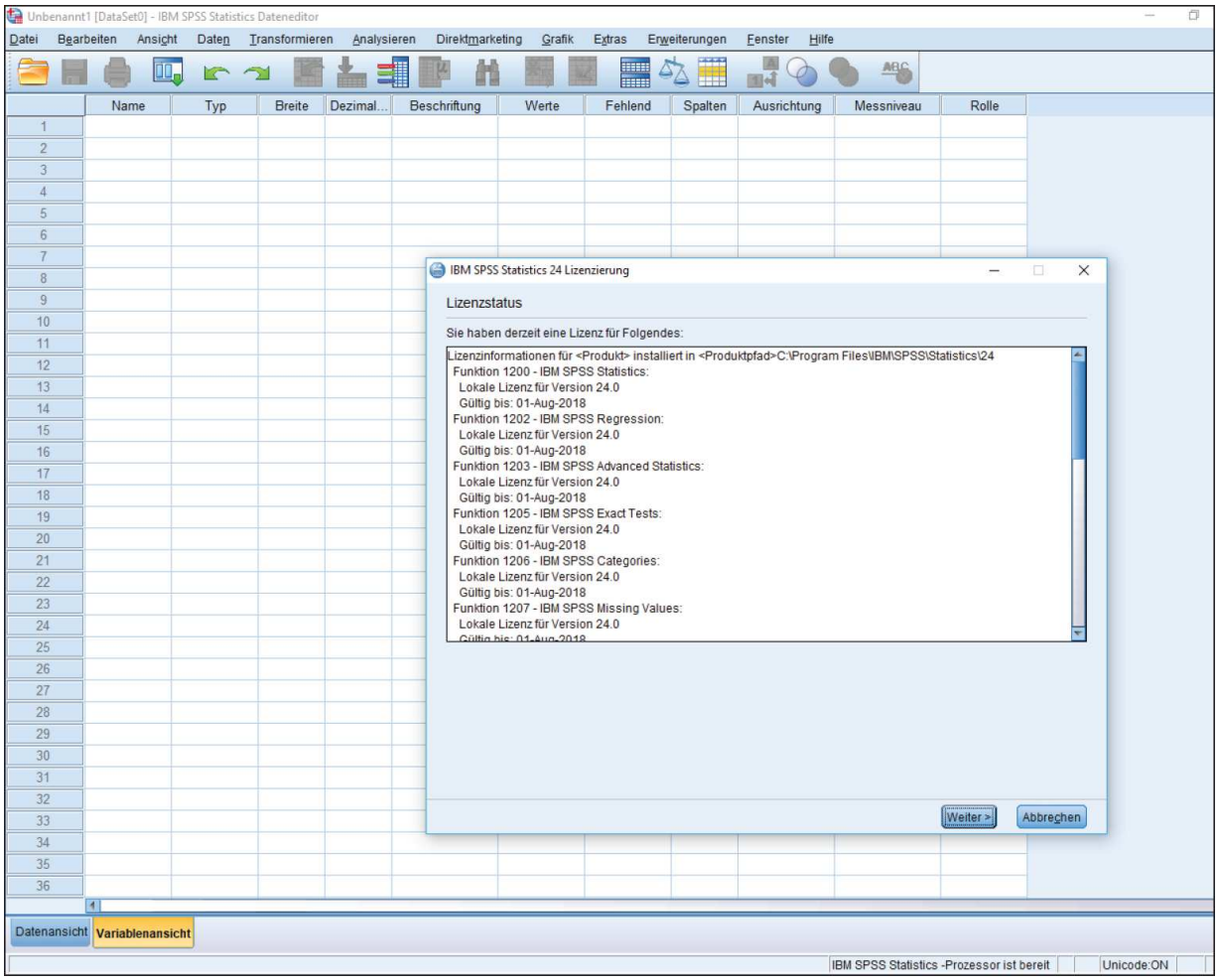



**Abbildung 1.1:** Mitteilung nach erfolgreicher Lizenzierung (Auszug; hier bezogen auf SPSS 24)

# *Der Eingangsbildschirm*

Wenn Sie nun SPSS auf Ihrem Rechner installiert haben, möchten Sie sicherlich damit arbeiten. Dazu starten Sie das SPSS-Icon, das Sie zum Beispiel auf Ihrem Desktop platziert haben, und es erscheint der **Eingangsbildschirm**. Es wird Ihnen, wie aus Abbildung 1.2 zu entnehmen ist, ein sogenanntes **Dialogfeld** (alternative Bezeichnung: Dialogbox) mit zuletzt bearbeiteten Dateien angeboten. (Wenn Sie SPSS 29 das erste Mal öffnen, werden hier natürlich keine Dateien angezeigt. Ich hatte – wie Sie aus Abbildung 1.2 unschwer erkennen – offenkundig schon ein wenig mit SPSS 29 gearbeitet …) Dieses spezielle Dialogfeld können Sie allerdings durch Anklicken des Felds DIESES

DIALOGFELD NICHT MEHR ANZEIGEN in der unteren linken Ecke des Dialogfelds für künftige SPSS-Anwendungen ausblenden.

Das Fenster, das sich in Abbildung 1.2 geöffnet hat, ist also ein sogenanntes Dialogfeld. Dialogfelder sind typisch für die Kommunikation mit SPSS. Dort können Sie unter anderem Einstellungsveränderungen vornehmen oder zum Beispiel statistische Operationen durchführen, worauf in den Teilen II und III des Buchs noch ausgiebig eingegangen wird.

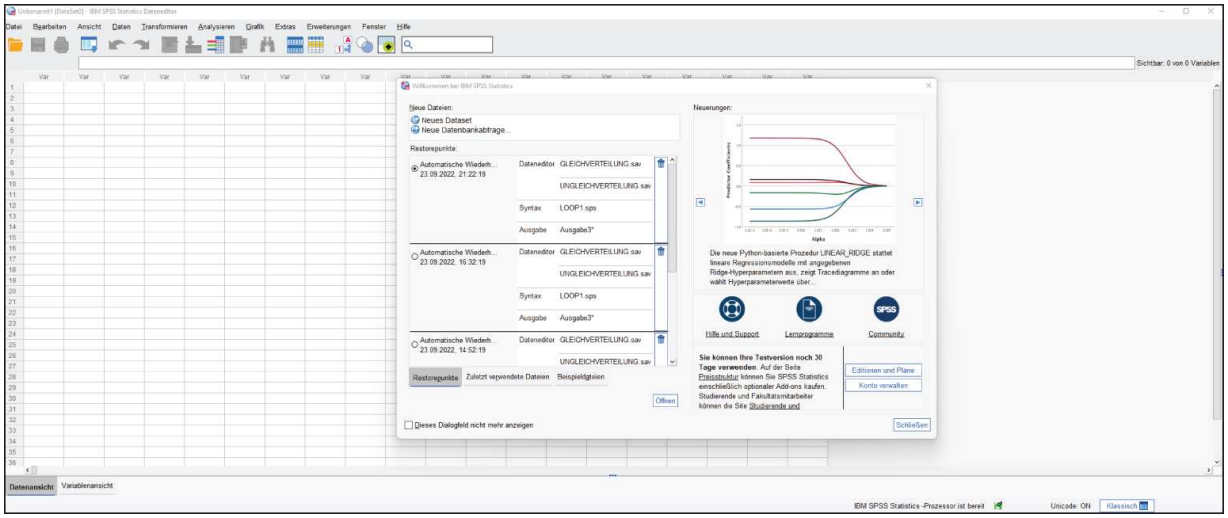



**Abbildung 1.2:** Der Start-Bildschirm von SPSS 29

Unten im Bildschirm befindet sich die **Statusleiste**. In Abbildung 1.2 ist zu erkennen, dass zwischen der Daten- und der Variablenansicht gewechselt werden kann. In unserem Fall ist DATENANSICHT voreingestellt (und der entsprechende Button entsprechend angeklickt). Des Weiteren wird in der Statusleiste in Abbildung 1.2 angegeben, dass der Statistikprozessor bereit ist, um prinzipiell irgendwelche statistische Auswertungen zum Laufen zu bringen. Außerdem wird ganz rechts in der Statuszeile die Sprach- beziehungsweise Zeichensatzeinstellung angezeigt, und zwar ist erkennbar UNICODE hier aktiviert. Damit ist sichergestellt, dass ohne Probleme länderübergreifend Zeichensätze in SPSS korrekt erkannt und verarbeitet werden können (zum Beispiel die Zeichen »ä«, »ö«, »ü«).

# *Die SPSS-Datenansicht*

Sie befinden sich jetzt automatisch in der SPSS-**Datenansicht**. In unserem Fall werden keine Daten angezeigt, was natürlich nicht verwundert, da Sie ja bislang auch noch keine Daten eingegeben beziehungsweise eingelesen haben. Alternativ können Sie im Start-Bildschirm unten links in der Statuszeile von der Daten- in die **Variablenansicht** wechseln, wo Sie für bestehende Variablen deren Namen oder deren Formatierung ändern können. Doch dazu in Kapitel 4 mehr!

In der oberen **Menüleiste** finden Sie, wie aus Abbildung 1.3 hervorgeht, das gesamte Spektrum an SPSS-Funktionen, gegliedert nach den Reitern:

- ✔ **DATEI**: Hier geht es nicht zuletzt um das Einlesen und das Abspeichern von Daten – dazu gleich mehr!
- ✔ **BEARBEITEN**: Hiermit können Sie zum Beispiel Fälle oder Variablen einfügen. Auch finden Sie hier die Dialogbox OPTIONEN, mit der Sie verschiedene Einstellungen in SPSS vornehmen (verändern) können.

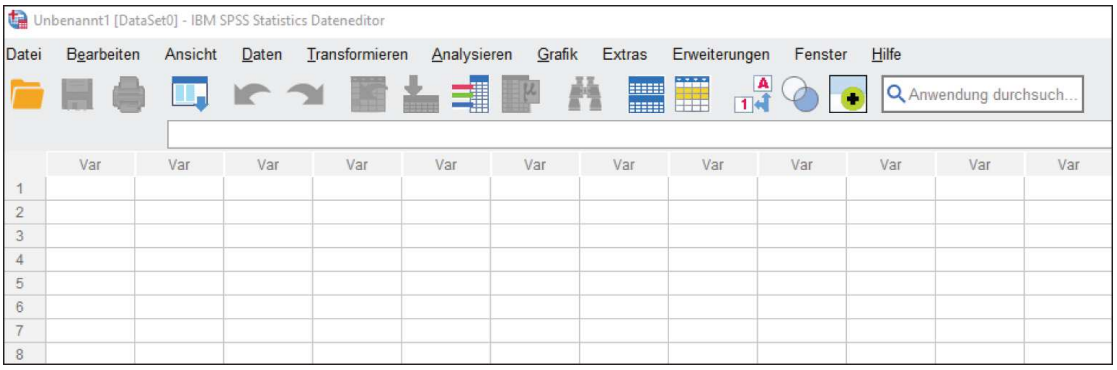



**Abbildung 1.3:** Das SPSS-Menü

- ✔ **ANSICHT**: Dieser Bereich dient in erster Linie dazu, dass Sie die Bildschirmansicht nach eigenem Gutdünken einstellen können, zum Beispiel in Bezug auf die dargestellte Schriftart. Da es sich hierbei

eher um »Luxusprobleme« handelt, werde ich auf diesen Bereich im Folgenden nicht näher eingehen.

- **DATEN**: Dies ist vor allem deshalb interessant, weil Sie hier bestimmte Fälle auswählen können. Zum Beispiel könnten Sie in einer Datei, die Daten für Frauen und Männer umfasst, die Daten für die Gruppe der Männer ausblenden und nur für die Frauen statistische Auswertungen durchführen. Eine weitere wichtige Unterrubrik in DATEN ist das Gewichten Ihrer Daten. Das ist etwa dann bedeutsam, wenn Sie für die deutsche Wohnbevölkerung aus Kostengründen eine Stichprobe gezogen haben, Ihre Ergebnisse aber bevölkerungsrepräsentativ auf die Gesamtbevölkerung Deutschlands hochrechnen möchten. Hinzu kommen in dieser Rubrik vor allem noch das Sortieren und das Aggregieren von Daten.
- **TRANSFORMIEREN**: Hier geht es um die Veränderung (das ist mit dem Fremdwort »Transformieren « gemeint) bestehender Variablen, indem Sie diese entweder in einer neuen Variablen umdefinieren oder die Werte der betreffenden Variablen verändern.
- **ANALYSIEREN**:Dies ist das Kernstück von SPSS. Hier finden Sie die in SPSS nutzbaren statistischen Methoden. Letzteres umfasst einfache deskriptive Statistiken wie Mittelwert- oder Streuungsberechnungen, aber auch komplexe statistische Testverfahren.
- **GRAFIK**: Wie der Name verrät, können Sie in dieser Rubrik Diagramme erstellen. Damit sind zum Beispiel Häufigkeitsdarstellungen gemeint.
- **EXTRAS**: Hier lassen sich Makros für Variablen definieren, Tabellen zensieren (das heißt: kürzen) und ähnliche »Scherze« durchführen. Auf Makros werde ich in Teil V zwar noch kommen, dann aber über die SPSS-Syntax und nicht über den hier erwähnten Menüpunkt. Der Grund dafür ist, dass das Vorgehen über die SPSS-Syntaxsprache an dieser Stelle flexibler ist.
- **ERWEITERUNGEN**:Um diese Rubrik nutzen zu können, müssen Sie Ihr SPSS-Grundmodul um zusätzliche SPSS-Module (wie das oben in der Einleitung bereits erwähnte AMOS = Analysis of

Moment Structures) erweitern, die Sie aber extra installieren müssen. Auch können Sie hier zum Beispiel Ihr SPSS-Programmpaket um Programmelemente des Statistikprogramms R erweitern. (Dazu in Teil V mehr!)

- **FENSTER**: Im Rahmen dieses Reiters erhalten Sie zum einen eine Übersicht über die von Ihnen geöffneten Fenster, und zum anderen können Sie Ihren Bildschirm teilen beziehungsweise geöffnete Fenster minimieren. SPSS-Fenster sind dabei grundsätzlich gekennzeichnet durch eine Kopfzone mit Menüzeile und Symbolleisten und eine Fußzone mit der Statuszeile, die Informationen über Programmzustände enthält.
- **HILFE**: Hier werden Ihnen Hilfestellungen zu bestimmten Stichwörtern ebenso angeboten wie der Zugang zu SPSS-Internetforen und Ähnlichem.

Standardmäßig wird bei SPSS als weitere Hilfestellung ein PDF-**Benutzerhand-buch** mitgeliefert, und zwar in der Datei *SYNTAXREFERENCE.PDF*. Auf immerhin 2.419 Seiten werden Ihnen im Rahmen der SPSS-Version 29 alle entsprechenden SPSS-Befehle dargelegt. Wenn Sie also mal unsicher sind, was Sie in SPSS tun können beziehungsweise was Sie in SPSS besser lassen sollten, empfehle ich Ihnen das Durchblättern dieses Handbuchs.

Bitte beachten Sie, dass bei leerem Datenfenster die meisten der oben genannten Menüoptionen nicht zur Verfügung stehen. Zum Beispiel liegt es ja geradezu auf der Hand, dass Sie unter ANALYSIEREN bei leerem Datenfenster keine statistischen Auswertungen vornehmen können. Wo nichts ist, kann man auch nichts auswerten!

# *Was Sie so alles in SPSS öffnen können*

Grundsätzlich können Sie in SPSS Daten

- ✔ beschreiben,
- ✔ verändern,
- ✔ auswählen und
- ✔ gewichten.

Hierzu stehen Ihnen verschiedene Dateimöglichkeiten zur Verfügung. Klicken Sie bitte einmal auf DATEI, und wenn Sie anschließend noch einmal klicken, und zwar auf ÖFFNEN, wird Ihre Aufmerksamkeit unmittelbar auf die entsprechenden »Zutaten« von SPSS gelenkt (siehe Abbildung 1.4):

- ✔ Daten
- ✔ Syntax
- ✔ Ausgabe

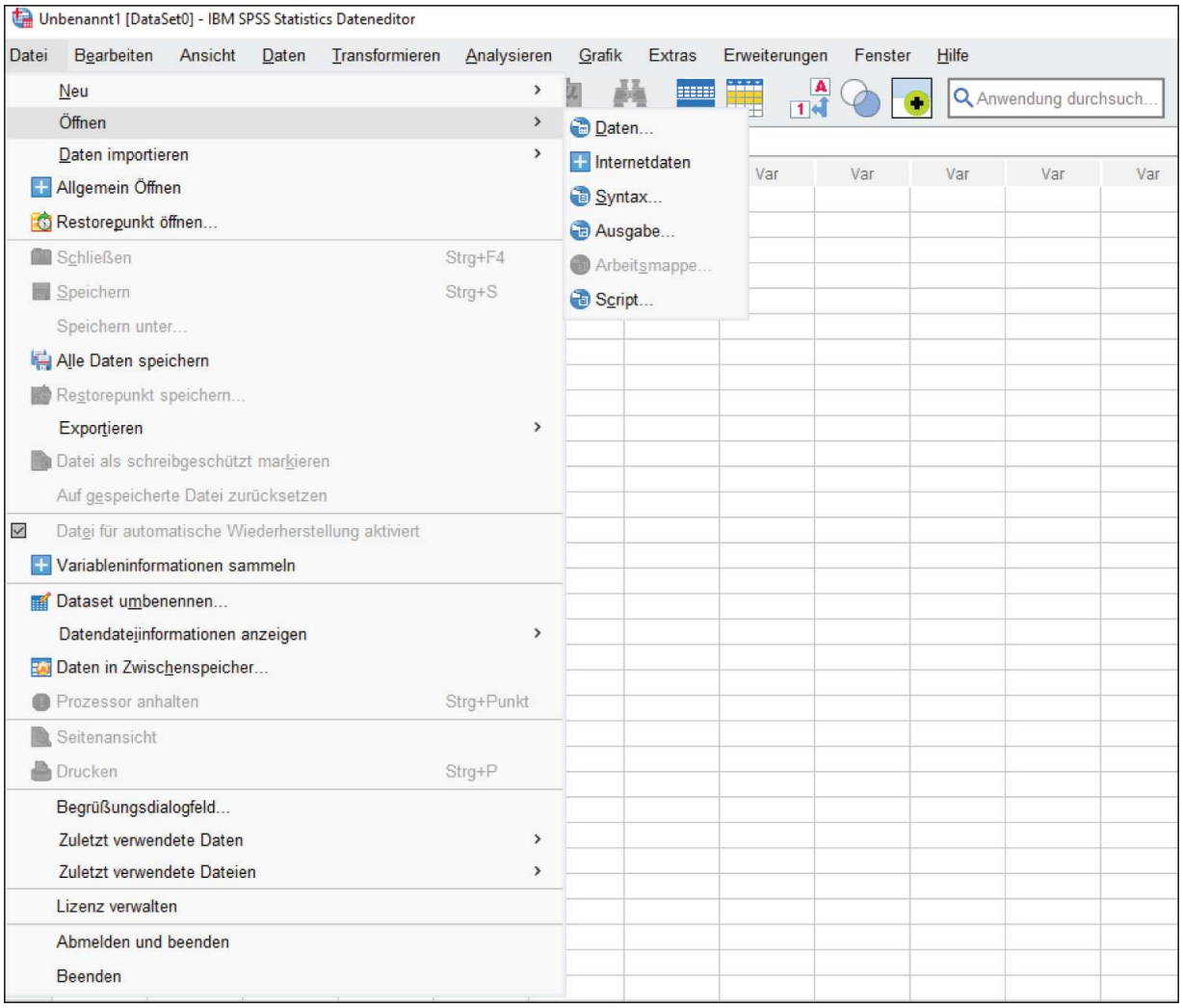



**Abbildung 1.4:** Datei in SPSS öffnen

Das heißt: Sie können ein File (= eine Datei) mit Daten (**Datenfile**) hochladen, ihre Berechnungen mittels Benutzermenü oder alternativ mit einem eigenen Programm in Form eines **Syntaxfiles** durchführen und schließlich das Berechnete in einem **Outputfile** dokumentieren. Hierbei sind Syntaxfiles immer dann vorteilhaft, wenn Sie längere Befehlsoperationen in SPSS abarbeiten beziehungsweise nach einiger Zeit diese Befehlsoperationen bequem und schnell wiederholen möchten. Auf die einzelnen File-Arten wird in den folgenden Kapiteln von Teil I noch ausführlich eingegangen.

Sie können mehrere Daten-, Syntax- und Outputfiles gleichzeitig in SPSS öffnen, sollten es aber damit nicht übertreiben, da Sie sonst leicht den Überblick verlieren. Sie sollten stets im Auge haben, welches das jeweils aktive Fenster ist, das Sie gerade für Ihre Arbeit benötigen.

Die **Extensionen**, das heißt die Dateikennzeichnungen/-ergänzungen der einzelnen Files sind aktuell

- ✔ beim Datenfile (ab SPSS-Version 7.5): *.SAV*
- ✔ beim Syntaxfile: *.SPS*
- ✔ beim Outputfile: *.SPV*

Wenn Sie also irgendwo eine Datei mit der Kennzeichnung *.SAV* sehen – zum Beispiel *123.SAV* –, handelt es sich um eine SPSS-Datei, die Daten in der für SPSS typischen Form (als sogenannte Rechteckmatrix; dazu vor allem in Kapitel 2 mehr!) enthält. Bei einer Datei mit der Kennzeichnung *.SPS* liegt eine SPSS-Syntaxdatei vor, das heißt eine Datei, in der verschiedene SPSS-Befehle stehen, die nacheinander abgearbeitet werden. Ihre Ergebnisse schließlich befinden sich in (Output-)Dateien mit der Kennzeichnung *.SPV* am Ende.

Ab SPSS-Version 16 sind frühere SPSS-Outputfiles in den neueren SPSS-Versionen nicht mehr unmittelbar in SPSS einlesbar, das heißt letztlich nicht mehr (unkompliziert) nutzbar: Die entsprechende Extension wurde seinerzeit von *.SPO* auf *.SPV* verändert. Nur mittels der Installation einer zusätzlichen Applikation – der sogenannten »**Legacy-Viewer**«**-Applikation** – ist an dieser Stelle eine Abwärtskompatibilität von SPSS in neueren Versionen als Version 16 gegenüber eben dieser Version 16 herstellbar. Auf der glücklichen Seite des statistischen Lebens ist derjenige, der früher die SPSS-Outputfiles, was möglich ist, als EXCEL-Datei (oder als WORD-Datei) abgespeichert hatte und sich

solcherart seine alten Ergebnisse in EXCEL anschauen (beziehungsweise eventuell weiterverarbeiten) kann.

Unter BEARBEITEN|OPTIONEN können Sie verschiedene SPSS-Anzeigen und dergleichen gewissermaßen nach eigenem Gutdünken an Ihre Vorstellungen anpassen. Wie aus Abbildung 1.5 ersichtlich, können Sie unter dem Menüpunkt BEARBEITEN|OPTIONEN in 12 Reitern Einstellungen verändern.

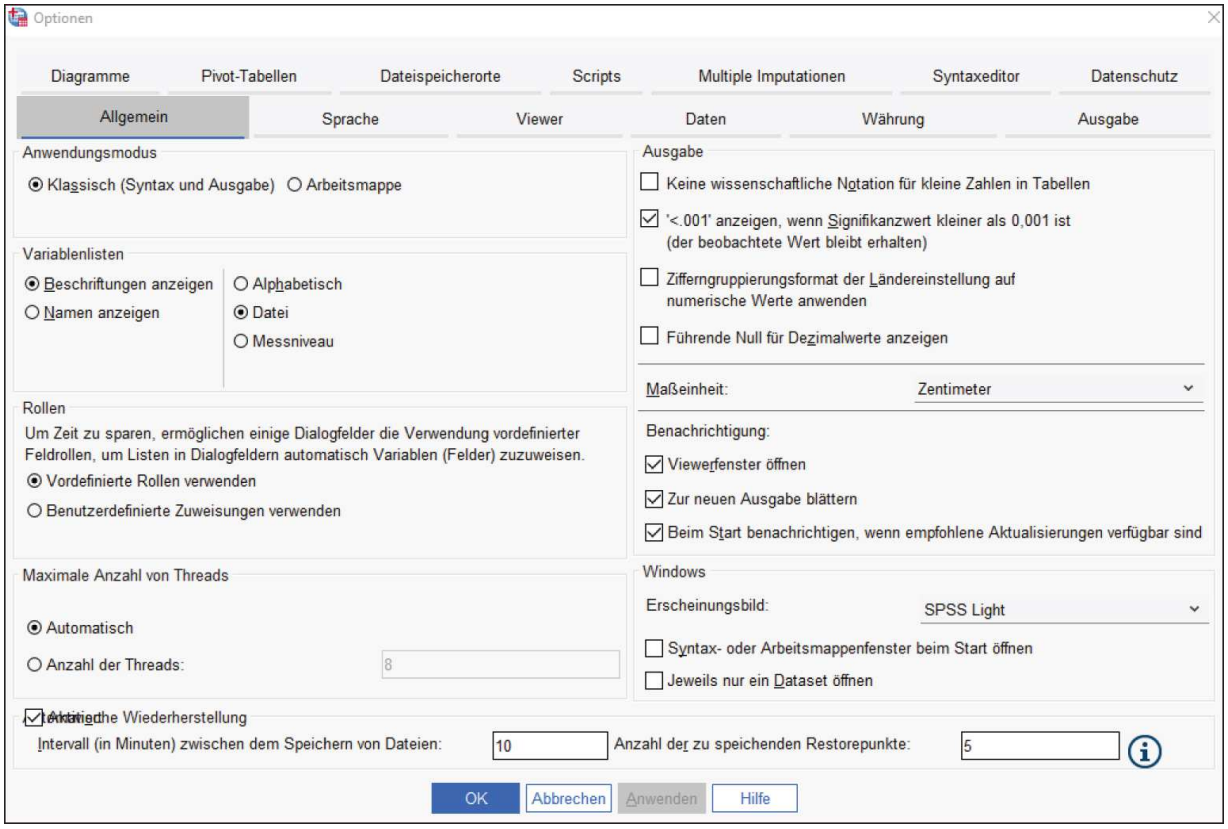



**Abbildung 1.5:** OPTIONEN bei SPSS

Wenn Sie genug von SPSS haben (oder wenn Sie, was positiver klingt, mit Ihren Berechnungen in SPSS fertig sind), können Sie eine SPSS-Sitzung für eine Datendatei einfach durch DATEI|BEENDEN schließen. Zuvor sollten Sie alles Erhaltenswerte (wie zum Beispiel Ihre veränderten Daten oder Ihr verändertes Syntaxfile oder Ihren Output) speichern, weil: Weg ist weg!

# Kapitel 2

# Daten als Mütter statistischer Analysen

## IN DIESEM KAPITEL

Bedeutung von Daten und Informationen

Regeln im Umgang mit Daten in SPSS

Einlesen von Daten in SPSS

Im Folgenden werden Daten als Grundlagen statistischer Analysen erörtert. Hierbei steht im Vordergrund, wie Daten in SPSS behandelt beziehungsweise überhaupt in SPSS eingelesen werden.

## *Bedeutung von Daten und Informationen*

Am Anfang stehen bei der statistischen Analyse natürlich die **Daten**. Ohne Daten sind Sie völlig blank, ohne diese gibt's einfach nichts zum Auswerten!

## *Vorgehen statistischer Untersuchungen*

Die Daten werden für **statistische Einheiten** erfasst. Als statistische Einheiten bezeichnet man die Untersuchungseinheiten. Das können Personen, Haushalte oder Organisationen, aber auch zum Beispiel Städte oder Länder sein. In den Naturwissenschaften zum Beispiel kommen

auch nicht-menschliche Untersuchungseinheiten infrage – zum Beispiel, wenn Sie Ihre Messergebnisse zum freien Fall von Steinen, die Sie an verschiedenen Tagen von einem Hochhaus geworfen haben (aber bitte in einer menschenleeren Gegend!), statistisch auswerten möchten.

Statistische Untersuchungen mit den für die Untersuchungseinheiten bedeutsamen Daten bestehen in der Regel aus drei Hauptaktivitäten:

- ✔ der Datenerhebung,
- ✔ der Datenverarbeitung und
- ✔ der Dateninterpretation.

Ganz grundsätzlich geht es dabei um die Erhebung und Verarbeitung von **Informationen**. Unter Informationen kann die Ordnung von Daten im Sinne rohen Faktenmaterials verstanden werden. In diesem Sinne sind Informationen auf bestimmte Sachverhalte gerichtetes, das heißt zweckbezogenes Wissen.

Von der **Datenerhebung** sehen Sie bitte an dieser Stelle zunächst einmal ebenso wie von der **Dateninterpretation** ab und beziehen sich bitte auf die **Datenverarbeitung**, naheliegenderweise speziell auf die in SPSS. Das bedeutet, dass zu einem bestimmten Sachverhalt Daten erhoben worden sind, die es nun in SPSS zu verarbeiten gilt. Beispielsweise könnten Sie in Ihrem Freundeskreis die jeweiligen Einkommen abgefragt haben, was möglicherweise nicht so gut in Ihrem Freundeskreis angekommen ist und Sie um ein paar Freunde ärmer gemacht hat; dafür sind Sie aber immerhin um ein paar Daten reicher geworden. Diese Daten müssen nun aber mit SPSS erst einmal »Bekanntschaft machen«. Das heißt: Die Daten müssen in SPSS zunächst einmal **eingelesen** werden.

# *Regeln im Umgang mit Daten in SPSS*

SPSS braucht hierzu ganz besondere Daten, und zwar solche, die in einer rechteckigen Datentabelle angeordnet sind. Die Beobachtungsfälle (**Merkmalsträger**) bilden die Zeilen und die Variablen (**Merkmale**) die Spalten einer solchen Datentabelle. In den Zellen derartiger Datentabellen stehen die **Merkmalsausprägungen** pro Merkmalsträger im Sinne der Schnittmenge aus Zeilen- und Spalteninformationen.

In Abbildung 2.1 sehen Sie, dass dort zwei Variablen *var00001* und *var00002* in den ersten beiden Spalten zu finden sind (siehe die Datendatei *123.SAV*, die Sie sich hochladen können). Des Weiteren können Sie erkennen, dass für drei Fälle 1 bis 3 in den ersten drei Zeilen über diese beiden Variablen Informationen vorliegen. Beispielsweise steht die Ausprägung der ersten Variablen *var00001* für den ersten Fall in der ersten Spalte und der ersten Zeile; es ist der Wert »1.00«. Für den dritten Fall ergibt sich klar ersichtlich – um ein weiteres Beispiel zu geben – als Ausprägung der zweiten Variablen *var00002* (zufälligerweise) ebenfalls der Wert »1.00« in der zweiten Spalte und der dritten Zeile.

Falls Sie sich im Übrigen wundern, dass in Abbildung 2.1 die Variablennamen in Großbuchstaben abgefasst sind, im Text hingegen in Kleinbuchstaben, das liegt daran, dass SPSS nicht case-sensitiv ist. Sie müssen also bei der Vergabe von Variablennamen in SPSS nicht auf die Groß- und Kleinschreibung achten – im Gegensatz etwa zum case-sensitiven Statistikprogramm STATA. Dies macht das Leben, wie ich finde, an dieser Stelle einfacher.

| | VAR00001 | VAR00002 | Var |
|---|---|---|---|
| 1 | 1,00 | 10,00 | |
| 2 | 2,00 | 5,00 | |
| 3 | 3,00 | 1,00 | |
| 4 | | | |

**Abbildung 2.1:** Beispiel für eine rechteckige Datentabelle (*123.SAV*)

# *Datenskalierungen*

Die Merkmalsausprägungen können recht verschiedenartig sein. Man spricht von unterschiedlichen **Datenskalierungen:**

- **Nominale Skalierung:** Hier stellen die Ausprägungen der Merkmale gewissermaßen nur Namensetikette dar, und sie sind völlig gleichwertig zueinander, können also nicht – ohne zusätzliche Werturteile – untereinander in eine Rangordnung gebracht werden. Hierunter fallen Stringvariablen (»Buchstaben-« beziehungsweise »Zeichenvariablen «) wie die Namen der Untersuchungseinheiten (»Müller«, »Meyer«, »Schulze« und so weiter), aber auch in Zahlen angegebene Variablen ohne natürliche Rangordnung wie zum Beispiel das hier (der Einfachheit halber) dichotom gegliederte Geschlecht mit den Untergliederungen 1 (für weiblich) und 2 (für männlich). Bei einer Nominalskalierung dürfen Sie streng genommen noch nicht einmal die Grundrechenarten nutzen und auch keine Größer-(gleich-) beziehungsweise Kleiner-(gleich-)Vergleiche anstellen. Sie können lediglich Vergleiche dahin gehend durchführen, ob zwei Ausprägungen gleich oder ungleich sind. Außerdem ist das Auszählen des Vorkommens der Ausprägungen einer Variablen – deren sogenannte *Häufigkeiten* – möglich.
- **Ordinale Skalierung:** Bei diesem Skalierungstyp können Sie eine Rangordnung zwischen den einzelnen Merkmalsausprägungen aufstellen; daher heißt dieser Skalierungstyp mitunter auch Rangskalierung. Was Sie hier aber nicht können, ist die sinnvolle Bestimmung der Abstände zwischen den Merkmalsausprägungen. Ein Beispiel an dieser Stelle sind Schulnoten von 1 bis 6. Hier interessiert nur, dass 1 die beste und 6 die schlechteste Note ist; die Abstände zwischen den Merkmalsausprägungen interessieren hingegen nicht beziehungsweise sind ohne Aussagekraft. Wenn Sie einfach jede mögliche Notenstufe mit 2 multiplizieren würden, würden sich ja die Abstände zwischen den Notenstufen vergrößern; dies wäre aber ohne inhaltliche Aussagekraft. Wichtig wäre auch dann allein, dass 2 die beste und 12 die schlechteste Note wäre. Sie

dürfen demnach bei Ordinalskalierung, streng genommen, zwar noch nicht einmal die Grundrechenarten verwenden, aber immerhin sind Größer-(gleich-) beziehungsweise Kleiner-(gleich-)Vergleiche möglich. Auch bei Ordinalskalierung können Sie Häufigkeiten für die Merkmalsausprägungen ermitteln.

- **Metrische Skalierung:** Die »Königsskalierung« ist die metrische Skalierung, da Sie hier alle möglichen mathematischen Aktionen auf die betreffenden Merkmalsausprägungen anwenden können. Man unterscheidet *diskrete metrische Merkmale* (ohne Nachkommastellen; zum Beispiel Fallzahlen von Personengruppen – also beispielsweise: 50 Personen) von *stetigen metrischen Merkmalen* (mit Nachkommastellen; zum Beispiel das Einkommen centgenau oder die Körpergröße – also beispielsweise: 1.983,42 Euro beziehungsweise 1,92 Meter). Datumsvariablen (»22.12.1977«, »23.02.1988« und so weiter) können ebenso wie Zeitvariablen (»14:18 Uhr«, »20:15 Uhr« und so weiter) insofern als metrische Merkmale interpretiert werden, als Sie die Abstände bei ihnen in Tagen, Minuten oder Sekunden angeben können. Die Bildung von Häufigkeiten für die einzelnen Merkmalsausprägungen bei metrischer Skalierung ist nur dann sinnvoll, wenn nicht allzu viele Merkmalsausprägungen vorliegen. Andernfalls empfiehlt es sich, die einzelnen Merkmalsausprägungen in Klassen zusammenzubringen.

Die nachfolgende Übersicht zeigt die verschiedenen Datentypen, wie eben diskutiert. Zusätzlich zu den vorstehenden Erörterungen sind in Abbildung 2.2 die metrischen Daten noch weiter differenziert, und zwar in intervallskalierte und kardinalskalierte Daten. Intervallskalierte metrische Daten haben keinen natürlichen Nullpunkt, sodass bei ihnen Relationen aus den einzelnen Ausprägungen im Grunde genommen ohne Aussagekraft sind. Dies gilt zum Beispiel für Temperaturangaben in Celsius oder in Fahrenheit. Demgegenüber können bei kardinalskalierten metrischen Daten entsprechende Verhältnisaussagen sinnvoll getroffen werden: So kann in der Einheit Kelvin die Aussage

getroffen werden, dass 40 Grad Kelvin viermal so heiß wie 10 Grad Kelvin sind, da sich die Temperaturangaben bei Kelvin auf den absoluten Nullpunkt der Temperatur beziehen. Auch bei Gewichts- oder Körpergrößenvergleichen sind entsprechende Aussagen sinnvoll möglich, weil jeweils ein natürlicher Nullpunkt von 0 Kilogramm beziehungsweise von 0 Zentimetern vorliegt (im jeweiligen »Null-Fall« ist eine Person halt noch nicht geboren). Eine Person mit einem Körpergewicht von 120 Kilogramm ist doppelt so schwer wie eine Person mit einem Körpergewicht von 60 Kilogramm, oder ein 200 Zentimeter großer Mensch ist doppelt so groß wie eine Person mit einer Körpergröße von 100 Zentimetern.

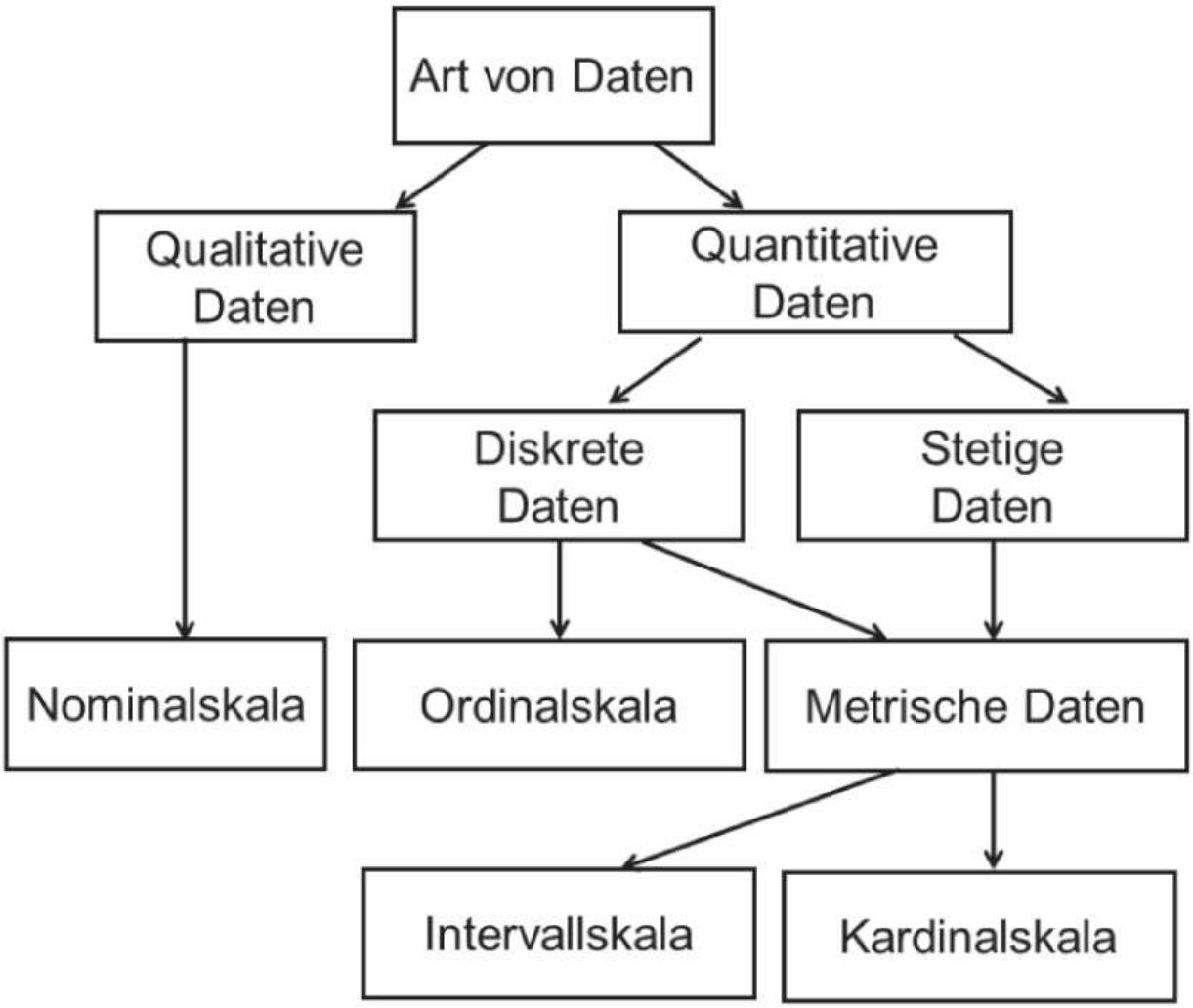


**Abbildung 2.2:** Verschiedene Datentypen (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley-VCH, S. 33)

Je nach **Skalierungsniveau** ergeben sich unterschiedliche statistische Auswertungsmöglichkeiten. Die folgende Tabelle stellt dies für einige ausgewählte statistische Kennziffern dar. Zugleich ist in der Tabelle angegeben, in welchem Kapitel des Buchs auf die jeweilige Kennziffer Bezug genommen wird.

Im Einzelnen bedeuten die in Tabelle 2.1 angegebenen Kennziffern Folgendes:

- ✔ *Absolute* beziehungsweise *relative Häufigkeit*: Anzahl des Vorkommens einer Merkmalsausprägung entweder in Fällen (absolute Häufigkeit) oder in Prozent (relative Häufigkeit).

| ***Kennziffer*** | ***Nominalskala*** | ***Ordinalskala*** | ***Metrische Skala*** |
|---|---|---|---|
| *Kapitel 6:* | | | |
| Absolute beziehungsweise relative Häufigkeit | Ja | Ja | Ja |
| *Kapitel 7:* | | | |
| Modus | Ja | Ja | Ja |
| Median | Nein | Ja | Ja |
| Arithmetisches Mittel | Nein | Nein | Ja |
| Geometrisches Mittel | Nein | Nein | Ja |
| Harmonisches Mittel | Nein | Nein | Ja |
| Spannweite | Nein | Nein | Ja |
| Varianz, Standardabweichung | Nein | Nein | Ja |
| Schiefe, Wölbung | Nein | Nein | Ja |
| *Kapitel 8:* | | | |
| Assoziationsmaße | Ja | Ja | Ja |

| *Kennziffer* | *Nominalskala* | *Ordinalskala* | *Metrische Skala* |
|---|---|---|---|
| Spearman'scher Rangkorrelationskoeffizient | Nein | Ja | Ja |
| Bravais/Pearson-Korrelationskoeffizient | Nein | Nein | Ja |

**Tabelle 2.1:** Unterschiedliche statistische Auswertungsmöglichkeiten bei unterschiedlichem Skalenniveau (Beispiele)

- *Modus*:Dies ist die am häufigsten vorkommende Merkmalsausprägung.
- *Median*:Hiermit ist die genau in der Mitte der aufsteigend geordneten Liste der Merkmalsausprägungen liegende Merkmalsausprägung gemeint (zum Beispiel: Bei fünf Ausprägungen 1, 2, 4, 7 und 9 ist der Wert 4 der in der Mitte liegende Wert, also der Median).
- *Arithmetischer Mittelwert*:Er wird auch als Durchschnitt bezeichnet. In diesem Sinne spiegelt er die Merkmalssumme dividiert durch die Fallzahl wider (zum Beispiel: Für die fünf Ausprägungen 1, 2, 4, 7 und 9 beträgt das arithmetische Mittel $\frac{1+2+4+7+9}{5} = 4{,}6$).
- *Geometrischer Mittelwert*: Vor allem für Veränderungs- beziehungsweise Wachstumsraten als Merkmalsausprägungen ist der arithmetische Mittelwert nicht geeignet. An dieser Stelle kommt der geometrische Mittelwert zum Einsatz, und zwar als N-te Wurzel aus dem Produkt der Veränderungsfaktoren (dabei hier N: Anzahl der Veränderungen insgesamt; zum Beispiel: Für die drei Veränderungsraten von 5 Prozent, 4 Prozent und 2 Prozent beträgt das geometrische Mittel für die Veränderungsfaktoren $\sqrt[3]{1{,}05 \cdot 1{,}04 \cdot 1{,}02} = 1{,}0366$, und damit ist die mittlere Veränderungsrate 3,66 Prozent).
- *Harmonischer Mittelwert*:Dieser Mittelwert ist für Verhältnisse aus zwei Größen definiert, wobei – im Unterschied zum

arithmetischen Mittel – die Zählergröße einen festen Wert hat und die Nennergröße variabel ist; das harmonische Mittel ist daher der Kehrwert des arithmetischen Mittelwerts der Kehrwerte der Merkmalsausprägungen (zum Beispiel: Berechnung der Fahrzeit bei gegebener Wegstrecke, also etwa 2 Stunden Fahrzeit für die ersten 100 Kilometer und dann 3 Stunden Fahrzeit für die weiteren 100 Kilometer Wegstrecke; das harmonische Mittel beträgt in diesem Fall:
$\frac{100\ km + 100\ km}{2\ h + 3\ h} = 40\ \frac{km}{h}$).

- ✔ *Spannweite*: Sie ist als Differenz aus der höchsten und der niedrigsten Merkmalsausprägung definiert.
- ✔ *Varianz, Standardabweichung*: Varianz und Standardabweichung geben Auskunft über die Streuung der Merkmalsausprägungen. Die Varianz ist dabei die durchschnittliche quadratische Abweichung der jeweiligen Merkmalsausprägungen vom arithmetischen Mittelwert. Die Standardabweichung ist die Quadratwurzel aus der Varianz.
- ✔ *Schiefe, Wölbung*: Anhand der Schiefe können Sie erkennen, ob eine Verteilung der Merkmalsausprägungen symmetrisch oder asymmetrisch ist. Die Wölbung gibt gegenüber einer Referenzverteilung an (üblicherweise wird hier die »Glockenkurve« der Normalverteilung unterstellt), ob die Verteilung der Merkmalsausprägungen von ihrer Form her gewölbter (»ausgebauchter«) ist oder nicht.
- ✔ *Assoziationsmaße*: Für zwei Merkmale X und Y wird untersucht, ob »irgendwie « ein statistischer Zusammenhang besteht.
- ✔ *Spearman'scher Rangkorrelationskoeffizient*:Hier wird für zwei (mindestens ordinalskalierte) Merkmale X und Y analysiert, ob ein positiver (= gleichgerichteter) oder ein negativer (= entgegengerichteter) statistischer Zusammenhang zwischen ihnen vorliegt. Dabei interessieren nicht die Abstände

zwischen den einzelnen Merkmalsausprägungen, sondern nur deren jeweilige Rangordnung (im Sinne von »höher« beziehungsweise »niedriger«).

- ✔ *Bravais/Pearson-Korrelationskoeffizient*: Für zwei metrisch skalierte Merkmale X und Y wird überprüft, ob ein positiver (= gleichgerichteter) oder ein negativer (= entgegengerichteter) statistischer Zusammenhang zwischen ihnen vorliegt. Dabei gehen auch die Abstände zwischen den einzelnen Merkmalsausprägungen in die Berechnung ein.

# *SPSS-Variablentypen*

Wie auch anhand der Ausführungen im vorstehenden Abschnitt deutlich geworden ist, können Daten prinzipiell in ganz verschiedenen Formen auftreten.

Merkmale werden insbesondere dann als **Variablen** bezeichnet, wenn sie durch numerische Ausdrücke charakterisiert sind. In diesem Sinne würde etwa die ordinale Variable *schulnote* aus den numerischen Ausdrücken »1« bis »6« bestehen, aber auch die nominale Variable *geschlecht* mit zum Beispiel »1« für »männlich« und »2« für »weiblich«. Variablen, die direkt beobachtet werden können (Beispiele: Augen- oder Haarfarbe einer Person), heißen **manifeste Variablen**, während Variablen (wie etwa eine solche für die individuelle Lebenszufriedenheit), die nicht direkt beobachtbar sind, als **latente Variablen** bezeichnet werden. Im letztgenannten Fall behilft man sich damit, dass man eine Variable im Sinne eines Indikators bildet – zum Beispiel bei der Lebenszufriedenheit eine Variable mit Werten von »0« (»völlig unzufrieden«) bis »10« (»völlig zufrieden«). Befragte Personen müssen dann jeweils einen Wert auf dieser Werteskala angeben.

Die verbreitetsten SPSS-Variablentypen sind:

- ✔ *Numerische Variablen* (zum Beispiel: Einkommen, Alter, Gewicht, Körpergröße, Haushaltsgröße)

- *Stringvariablen*, die alternativ auch als alphanumerische oder Zeichenkettenvariablen bezeichnet werden (zum Beispiel: Namen, Geschlecht, Haarfarbe, und zwar auch dann, wenn den Ausprägungen Zahlenwerte zugewiesen werden wie zum Beispiel »1« für »männlich« und »2« für »weiblich« beim Geschlecht)
- *Datums-/Zeitvariablen* (zum Beispiel: Geburtsdatum, Renteneintrittsdatum)

In Abgrenzung zu den anderen Variablenarten werden Stringvariablen vielfach auch als **qualitative Daten** und die anderen Variablenarten (Datums-/Zeitvariablen, numerische Variablen) als **quantitative Daten** bezeichnet.

Mit diesen Variablentypen sind unterschiedliche Möglichkeiten der statistischen Bearbeitung verbunden. Zum Beispiel können Sie mit einer Stringvariablen, die im Grunde genommen nichts anderes ist als ein Wort, nicht addieren, subtrahieren, multiplizieren und dividieren. Sie können hier also noch nicht einmal die Grundrechenarten anwenden.

Spitzfindige Leser werden an dieser Stelle den berechtigten Einwand anbringen, dass sie an eine Stringvariable ja Zeichen anhängen beziehungsweise von ihr Zeichen wegnehmen können. Erweitern Sie zum Beispiel das Wort »Ei« zu »Eier«, kann dies durchaus in einem weiteren Sinne als Addition interpretiert werden. Umgekehrt könnten Sie die Reduktion von »Eier« zu »Ei« als Subtraktion interpretieren. Sie werden mir aber doch recht geben, dass diese Erweiterung beziehungsweise Verminderung um Buchstaben etwas anderes ist als die herkömmliche Addition beziehungsweise Subtraktion.

Von den genannten Variablentypen (also: String-, numerische und Datums-/Zeitvariablen) kommen in praktischen Anwendungen numerische und Zeichenkettenvariablen deutlich am häufigsten vor. Hierbei haben Sie bei Variablen mit Nominalskalenniveau die Wahl zwischen numerischer und Zeichenketten-Codierung, zum Beispiel beim Geschlecht »1« für Männer und »2« für Frauen beziehungsweise

»männlich« für Männer und »weiblich« für Frauen. In der Arbeit mit SPSS empfiehlt sich auch bei nominalskalierten Variablen die numerische Codierung, zumal zum Beispiel in der Regressionsanalyse nominalskalierte Variablen numerisch codiert werden müssen (dort dann als sogenannte »0/1-Dummies« mit den Ausprägungen »0« und »1«).

# *Zum Nutzen quantitativer Ausdrücke in SPSS*

Grundsätzlich funktioniert SPSS – wie andere Statistikprogramme auch – bei numerischen Variablenausdrücken besonders gut. Dies verwundert eigentlich auch nicht, weil es für numerische beziehungsweise für metrisch skalierte Variablen die weitreichendsten statistischen Anwendungsmöglichkeiten gibt. Gleichwohl gibt es aber natürlich auch für Stringvariablen beziehungsweise für nominal- oder ordinalskalierte Merkmale einige SPSS-Verarbeitungsmöglichkeiten, was auch aus Tabelle 2.1 hervorgeht.

Angenommen, Sie möchten einen Notendurchschnitt ausrechnen (was allerdings, ganz streng genommen, so nicht zulässig ist, da ja eine Variable *note* keine metrische, sondern »lediglich« eine ordinale Variable ist; siehe hierzu die entsprechenden Angaben in Tabelle 2.1). Nehmen Sie des Weiteren an, die Noten lägen lediglich in Wörtern vor: »sehr gut«, »gut«, »befriedigend«, »ausreichend«, »mangelhaft« und »ungenügend«.

Wie sollte man aus diesen Buchstabenfolgen (= Wörtern) einen Notendurchschnitt bilden?

Es leuchtet ein, dass zu diesem Zweck numerische Variablenausprägungen vonnöten sind. Durch Überführung der »Buchstaben-Variablen« in die numerischen Ausdrücke 1, 2, … , 6 können Sie also Ihren Durchschnitt ausrechnen. Kommt zum Beispiel die Note 1 zweimal, die Note 2 fünfmal, die Note 3

zehnmal, die Note 4 fünfmal, die Note 5 dreimal und die Note 6 fünfmal vor, lässt sich daraus als Durchschnittswert berechnen:

$$Notendurchschnitt = \frac{1 \cdot 2 + 2 \cdot 5 + 3 \cdot 10 + 4 \cdot 5 + 5 \cdot 3 + 6 \cdot 5}{30} \approx 3{,}57.$$

Der Notendurchschnitt beträgt in diesem Beispiel also 3,57 (was nicht so toll ist, aber es ist ja nur ein Beispiel …).

In SPSS lässt es sich folglich besonders gut mit **quantitativen Daten** im Sinne von Zahlen arbeiten. Zwar können Sie auch für **qualitative Variablen** wie »Nachname« zum Beispiel Häufigkeiten ermitteln (im Sinne von: wie oft die jeweiligen Variablenausprägungen im Datensatz vorkommen) oder auch Mittelwerte wie den häufigsten Wert (= Modus) bestimmen, aber zum Beispiel einen Durchschnittswert können Sie wie vielfältige andere statistische Indikatoren (Streuungsmaße und dergleichen) für qualitative Informationen nicht berechnen. Das liegt schlicht und ergreifend daran, dass Sie – wie oben bereits skizziert – qualitative Informationen, die sozusagen in »Buchstaben-/Zeichenform« vorliegen, in einem strengen mathematischen Sinne nicht miteinander addieren, nicht voneinander subtrahieren beziehungsweise auch nicht miteinander multiplizieren oder durcheinander dividieren, also für diese noch nicht einmal die Grundrechenarten nutzen können. Ganz zu schweigen davon, dass Sie für solche Daten nicht sinnvoll einen Logarithmus, eine Potenz oder einen Wurzelausdruck beziehungsweise Tangens, Sinus, Cosinus und so weiter berechnen können! Einfach ausgedrückt: Sie können »Müller«, »Meyer« und »Schulze« als Ausprägungen einer nominalen Namensvariablen nicht addieren.

## *Umformung in Zahlenwerte*

Durch die Umformung (= **Transformation**) in Zahlenwerte können Sie aber sogar für nominalskalierte Merkmale mindestens einen arithmetischen Mittelwert berechnen, auch wenn das eigentlich für eine solche Skalierung, streng genommen, nicht okay ist. **Codieren** Sie beispielsweise die Ausprägungen »männlich« und »weiblich« der Variablen *geschlecht* mit »0« und »1« und haben in Ihrem Datensatz 5 Frauen und 95 Männer, wirft Ihnen SPSS als arithmetischen Mittelwert

0,95 aus. Anhand dieses Werts können Sie immerhin sofort erkennen, dass die Anzahl der Männer die der Frauen deutlich übersteigt und dass der Männeranteil 95 Prozent beträgt. Das ist doch auch schon mal was! (Allerdings handelt es sich hier, streng genommen, nicht um einen Durchschnittswert, sondern um eine relative Häufigkeit.)

Außerdem vereinfacht die Transformation in Zahlenwerte die Analyse von nominalen Variablen mit mehr als zwei Ausprägungen auch dahin gehend, dass Auswahlentscheidungen erleichtert werden. Nehmen Sie als Beispiel die Variable *Augenfarbe* mit den Ausprägungen »grün«, »blau«, »grau« und »braun«, und nehmen Sie zudem an, dass Sie sich nur für die Ausprägungen »grün« und »blau« interessieren, dann müssten Sie ohne Transformation jede einzelne von Ihnen ausgewählte Ausprägung angeben. Transformieren Sie indes »grün« = 1, »blau« = 2, »grau« = 3 und »braun« = 4, dann müssen Sie für die betreffende Auswahl in SPSS lediglich als Variablenwert »kleiner gleich 2« angeben. Das ist doch praktisch – oder? (Wie das Ganze genau geht, wird später in den Kapiteln 4 und 5 noch näher erläutert. Für Ungeduldige: Hier kommt unter anderem der RECODE-Befehl für die Variablentransformation zum Einsatz.)

Es kommt noch mindestens ein weiterer Nutzen einer Transformation in Zahlenwerte hinzu, und zwar in Form sogenannter **Dummy-Variablen** im Rahmen einer Regressionsanalyse (mehr dazu in den Kapiteln 9 und 10!). Hier können Sie in eine eigentlich metrisch skalierte statistische Beziehung zwischen der zu erklärenden Variablen und der oder den erklärenden Variablen auch ein oder mehrere nominalskalierte Merkmale (beziehungsweise auch ein oder mehrere ordinalskalierte Merkmale) einbeziehen. Zum Beispiel würden Sie – wie oben bereits angedeutet – das erklärende Merkmal Geschlecht mit den Ausprägungen 0 = weiblich und 1 = männlich (Sie können natürlich 0 und 1 auch genau umgekehrt definieren) berücksichtigen. Bezüglich der Augenfarbe mit den Ausprägungen »grün«, »blau« »grau« und »braun« könnten Sie vier Dummy-Variablen kreieren: *gruen* mit 1 = grün und 0 = nicht-grün, *blau* mit 1 = blau und 0 = nicht-blau, *grau* mit 1 = grau und 0 = nicht-grau sowie *braun* mit 1 = braun und 0 = nichtbraun. (Wie Sie noch in Kapitel 9 sehen werden, brauchen Sie von den eben genannten vier Dummies

bezüglich der Augenfarbe in Ihrer Regressionsanalyse nur drei zu berücksichtigen.)

# *Zum Problem fehlender Werte*

Bei der Variablencodierung kann – als Folge von Antwortverweigerungen bei Befragungen oder als Folge von Eingabefehlern – das Problem der **fehlenden** beziehungsweise ungültigen Werte auftreten. In derartigen Fällen sollten Sie in Ihrem Datensatz Werte vergeben, die (weit) außerhalb des zulässigen Wertebereichs einer Variablen liegen. Der Grund dafür ist, dass Sie mit den fehlenden Werten üblicherweise nicht rechnen möchten und daher nicht in die Verlegenheit kommen sollten, sie in die Berechnungen versehentlich einzubeziehen. Durch Zuweisung eines bestimmten, »exotischen« numerischen Ausdrucks für die fehlenden/ungültigen Werte können diese entweder automatisch oder durch das Filtern in SPSS von den Berechnungen ausgeschlossen werden.

Weist eine Variable zum Beispiel drei zulässige Ausprägungen »1«, »2« und »3« auf, wäre etwa die Ausprägung »9« (oder die »-9«) für eine fehlende (beziehungsweise ungültige) Variablenausprägung sinnvoll. Wäre hingegen die »9« selbst eine zulässige Ausprägung – wie beispielsweise bei der Angabe des Alters in Lebensjahren (von Kindern) –, wäre die »9« keine sinnvolle Wahl zur Kennzeichnung einer fehlenden (beziehungsweise ungültigen) Variablenausprägung. In diesem Fall könnte man zum Beispiel die »-9« zur Kennzeichnung einer fehlenden (beziehungsweise ungültigen) Variablenausprägung verwenden.

Die vorgenannten Fälle kennzeichnen **benutzerdefinierte fehlende Werte,** wenn Sie also selbst zum Beispiel die »9« zur Kennzeichnung eines fehlenden Werts vergeben haben. Demgegenüber handelt es sich um **systemfehlende Werte**, wenn diese von SPSS automatisch festgelegt werden, und zwar mit »SYSMIS« (als Abkürzung für SYStem-MISsing) beziehungsweise mit einem Punkt in Ihrer Datentabelle.

In folgenden Fällen setzt SPSS eine Variablenausprägung auf »SYSMIS«:

- wenn sich bei einer numerischen Variablen unzulässige Zeichen als Variablenausprägungen finden (zum Beispiel ein Buchstabe),
- wenn bei einer numerischen Variablen kein Eintrag vorliegt (wenn also zum Beispiel eine Frage von einem Interviewten nicht beantwortet wurde),
- wenn bei der Veränderung (= Transformation) von Variablen unzulässige mathematische Operationen durchgeführt werden: zum Beispiel eine Division durch eine Variablenausprägung in Höhe von null oder die Logarithmierung einer nicht-positiven Zahl. Zum Beispiel: Die dritte von Ihnen befragte Person hat ein Vermögen von 0 (*vermoegen* = 0) angegeben, ist also vermögenslos. Wenn Sie nun eine Einkommens/Vermögens-Relation bilden wollen im Sinne von *relation* = *einkommen*/*vermoegen*, würden Sie bei der dritten Person durch null dividieren, was mathematisch nicht zulässig ist, und daher würde *relation* für diesen dritten Fall auf SYSMIS gesetzt. Gleiches würde passieren, wenn Sie den (natürlichen) Logarithmus von *vermoegen* für diesen dritten Fall bilden möchten: *vermoegen_ln* = ln(*vermoegen*). Auch hier würde beim dritten Fall die Ausprägung jetzt für *vermoegen_ln* auf SYSMIS gesetzt werden. Im Datensatz erhielten Sie sowohl bei *relation* als auch bei *vermoegen_ ln* einen Punkt als Kennzeichnung eines fehlenden Werts für die dritte Person.

Das vorstehende Problem ist in Abbildung 2.3 visualisiert. Die der Abbildung zugrunde liegenden Daten finden sich in *EINKOMMEN_VERMOEGEN.SAV*.

Konkret können Sie die SYSMIS-Werte im SPSS-Dateneditor also einfach daran erkennen, dass in der betreffenden Zelle der Datenmatrix nur ein Punkt eingetragen ist.

| fall | einkommen | vermoegen | relation | vermoegen_ln |
|---|---|---|---|---|
| 1 | 1000,00 | 2000,00 | ,50 | 7,60 |
| 2 | 500,00 | 1000,00 | ,50 | 6,91 |
| 3 | 1000,00 | . | . | . |
| | | | | |

**Abbildung 2.3:** Zum Problem fehlender Werte bei mathematischen Variablenoperationen (*EINKOMMEN_VERMOEGEN.SAV*)

# *Einlesen von (massenstatistischen) Daten in SPSS*

Mittels Telefon-, Onlinebefragung und dergleichen **erhobene Daten** können in SPSS in unterschiedlicher Weise erfasst (= »eingelesen«) werden:

- ✔ manuell direkt über den SPSS-Dateneditor
- ✔ Erfassung über andere Programme wie EXCEL
- ✔ Erfassung mit einem spezifischen Datenerfassungsprogramm
- ✔ Erfassung über eine Erfassungsmaske in einem spezifischen Datenbankprogramm

Im **Dateneditor** (das heißt in der Datenansicht) werden Ihre Daten dabei zunächst *temporär* gespeichert. Verändern Sie Ihre Daten durch verschiedene Operationen, werden Sie beim Beenden Ihrer SPSS-Sitzung gefragt, ob Sie Ihre Daten *dauerhaft* sichern möchten.

Wichtig für SPSS ist in jedem Fall, dass es **Rechteckdatensätze** voraussetzt. Das ist in diesem Kapitel ja schon einmal erwähnt worden und meint einfach, dass in den Spalten einer SPSS-Datendatei die Merkmale und in den Zeilen die Fälle angegeben sind und in den Zellen die einzelnen Merkmalsausprägungen (pro Fall und pro Variable) stehen.

# *Manuelle Dateneingabe*

Eine Möglichkeit, Ihre Daten in SPSS einzulesen, ist die **manuelle Dateneingabe**. Sie können dabei im Dateneditor die einzelnen Daten entweder fall- oder variablenweise über Ihre PC-Tastatur und mithilfe der Cursor-Navigation eingeben. Konkret können Sie mithilfe von Rechtspfeil- oder Tabulator-Taste den eingetippten Wert bestätigen und eine Spalte nach rechts zur nächsten Variablen wechseln. Mithilfe der ↵-Taste bewegen Sie sich eine Zeile nach unten, das heißt zum nächsten Fall. In diesem Zusammenhang geht aus Abbildung 2.4 die Eingabe der drei Werte 1,00, 2,00 und 3,00 hervor, wobei in dieser Abbildung die Eingabe des drittgenannten Werts noch nicht ganz abgeschlossen ist. (Gegebenenfalls verlangt SPSS von Ihnen gemäß Ihren SPSS-Einstellungen im Übrigen an dieser Stelle bei den Werteingaben einen Punkt als Zeichen für die Dezimalstelle – gemäß der US-amerikanischen Notationseinstellung.)

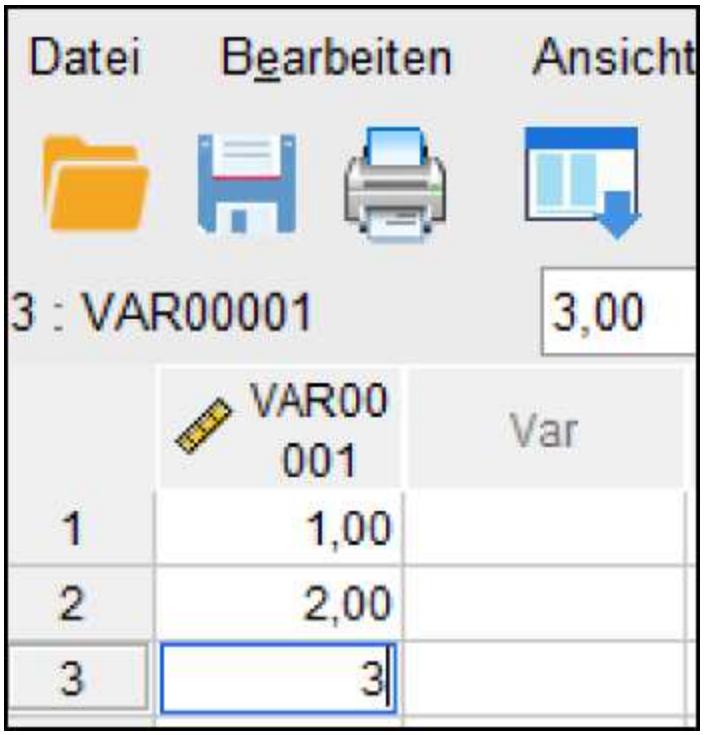



**Abbildung 2.4:** Manuelle Werteeingabe im Dateneditor

Vor der eben genannten Dateneingabe lagen in der Datenansicht ja noch keine Daten vor; hatten Sie aber Variablennamen (beziehungsweise eventuell auch noch Variablenbezeichnungen) schon (im Variableneditor, das heißt in der **Variablenansicht**) angegeben, finden Sie in den Zellen der ersten Zeile ausschließlich jeweils einen Punkt vor. Diese Punkte stehen für systemfehlende Werte, für die sogenannten SYSMIS-Werte. Sie sind sozusagen ein Anzeichen dafür, dass SPSS für diese Variablen auf eine Werteeingabe wartet.

Damit Sie sich Ihre Arbeit erleichtern beziehungsweise übersichtlicher gestalten, sollten Sie bereits an dieser Stelle unten in der Statusleiste vom Daten- in den Variableneditor wechseln und dort zumindest die Variablen mit den für Sie relevanten Namen bezeichnen. In Abbildung 2.5 ist die von SPSS so im Daten-editor genannte Variable *var00001* in *fall* im Variableneditor umbenannt worden.

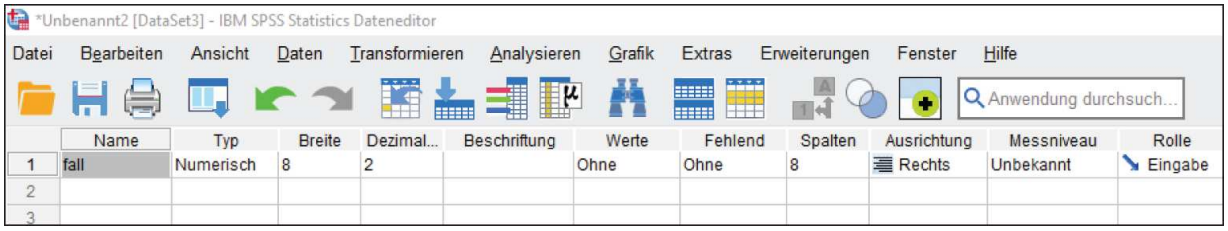



**Abbildung 2.5:** Variablenumbenennung im Variableneditor

Haben Sie sich bei der Werteeingabe vertippt, können Sie das relativ leicht korrigieren, indem Sie die betreffende Zelle markieren und anschließend einfach den neuen Wert eintippen; dadurch überschreiben Sie den alten, fehlerhaften Wert. Wenn Sie einen Wert nur teilweise verändern möchten, klicken Sie doppelt auf die betreffende, fehlerhafte Zelle.

Möchten Sie einen neuen **Fall** innerhalb eines Datensatzes **einfügen**, bewegen Sie den Cursor auf die SPSS-Zeilennummer des Falls, vor dem dieser neue Fall eingefügt werden soll, und setzen einen rechten

Mausklick. Aus dem Kontextmenü wählen Sie anschließend FÄLLE EINFÜGEN. In Abbildung 2.6 soll vor Fall Nummer 3 ein neuer Fall eingefügt werden. Wie Sie sehen, öffnet sich zunächst das genannte Kontextmenü. (Sie hätten sich die Option FÄLLE EINFÜGEN im Übrigen auch ausschließlich über das Menü mittels BEARBEITEN| FÄLLE EINFÜGEN erschließen können.)

**Abbildung 2.6:** Kontextmenü für FÄLLE EINFÜGEN

Wie erwartet, erhält dieser neue Fall für alle Variablen (in unserem Beispiel gibt es ja nur eine Variable *fall*) zunächst einmal jeweils den SYSMIS-Wert in Form eines Punkts (siehe hierzu Abbildung 2.7).

Es kann aber auch die umgekehrte Situation relevant sein; das heißt, dass Sie einen **Fall löschen** möchten. Hierzu markieren Sie (mit einem rechten Mausklick) die SPSS-Zeilennummer des zu löschenden Falls und wählen einfach aus dem Kontextmenü LÖSCHEN aus. Gemäß Abbildung 2.8 soll der eingefügte Fall in der dritten Zeile wieder gelöscht werden. Entsprechend markieren Sie zunächst diese dritte Zeile, klicken mit der rechten Maustaste auf diese Zeile und wählen anschließend im Kontextmenü LÖSCHEN.

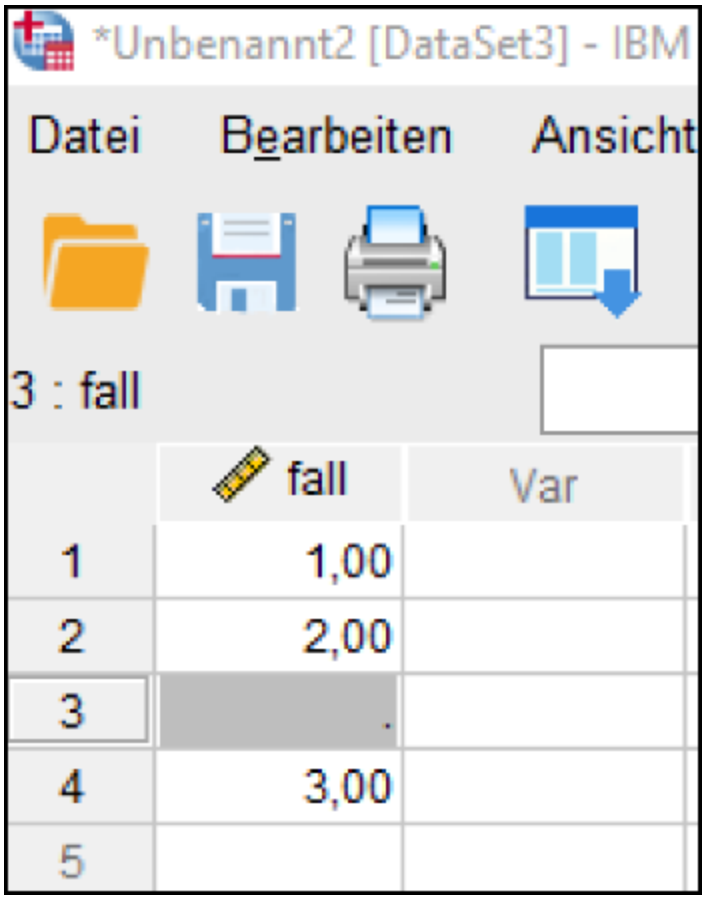



**Abbildung 2.7:** Eingefügter Fall

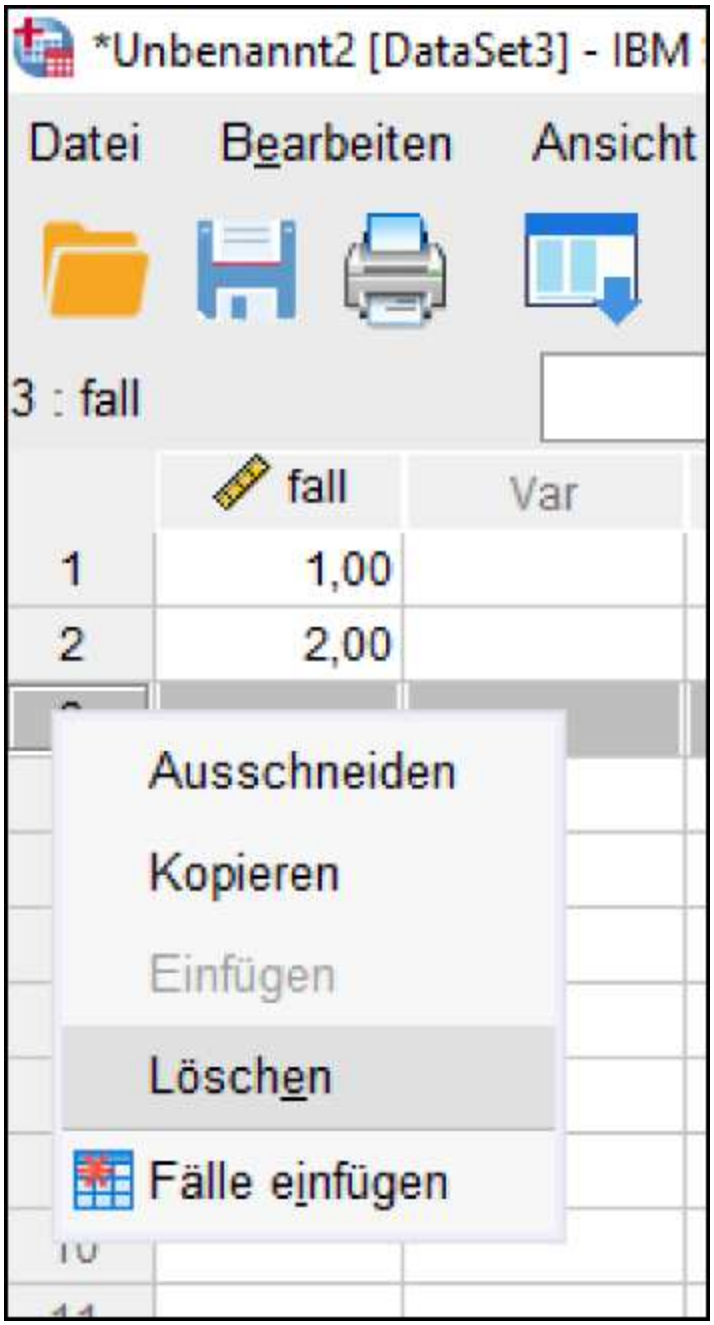



**Abbildung 2.8:** Kontextmenü für das Löschen von Fällen

Klicken Sie nunmehr im Kontextmenü auf LÖSCHEN, erhalten Sie als Ergebnis, dass Fall Nummer 3 wieder weg ist (siehe Abbildung 2.9).

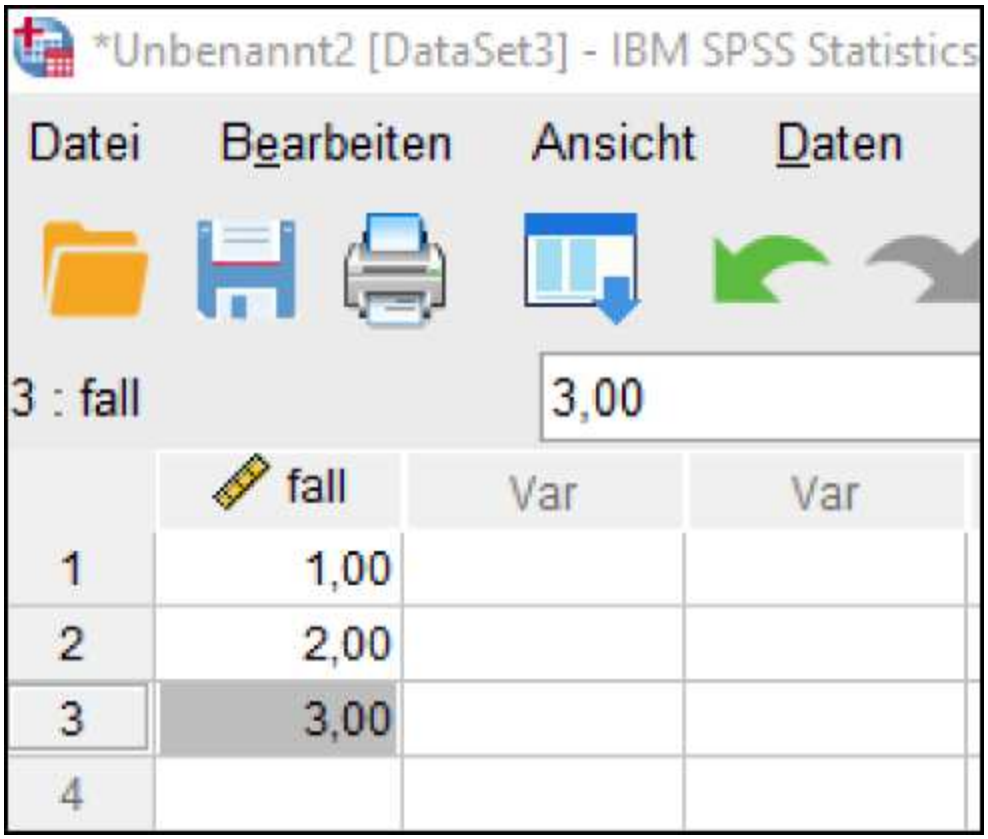



**Abbildung 2.9:** Datendatei nach Löschen eines Falls

Haben Sie bei den vorstehenden Vorgängen (aber auch bei allen anderen SPSS-Vorgängen) etwas falsch gemacht, können Sie mittels BEARBEITEN|RÜCKGÄNGIG (oder einfach über die Ihnen aus Windows Office sicherlich bekannte Tastenkombination Strg+Z) den letzten Vorgang (= die letzte Änderung) wieder rückgängig machen.

Wenn Sie mit der Dateneingabe fertig sind, sollten Sie Ihre Datendatei über das Menü DATEI|SPEICHERN UNTER abspeichern, das heißt für spätere Bearbeitungen sichern. Am besten wählen Sie – wie in Abbildung 2.10 mit *FALL.SAV* – eine SPSS-Datendatei mit der Dateikennzeichnung *SAV* aus.

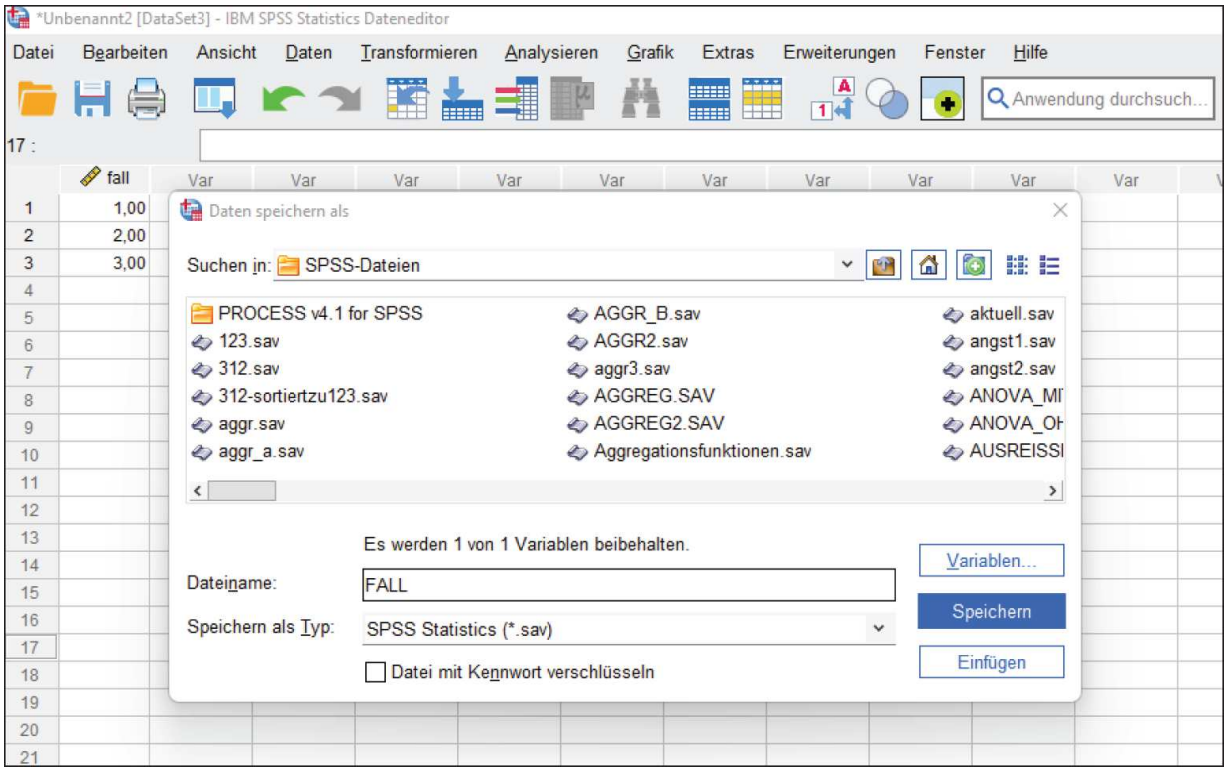



**Abbildung 2.10:** Abspeichern einer manuell erstellten Datendatei als SAV-File (*FALL.SAV*)

# *Vorhandene Datendateien*

Einfacher ist das Eingeben beziehungsweise Einlesen von Daten naheliegenderweise dann, wenn die Daten bereits in »irgendeiner« Datei vorhanden sind. Sehr einfach ist das Ganze, wenn die Daten bereits in einer *SAV*-Datei vorliegen, denn in diesem Fall haben Sie es ja bereits mit dem von SPSS gewünschten Dateiformat zu tun. Sie können dann über das Menü DATEI|ÖFFNEN|DATEN aus Ihrem Fundus bestehender *SAV*-Files das gewünschte auswählen. In diesem Zusammenhang können Sie eine existierende *SAV*-Datendatei, die zuletzt genutzt wurde, auch über DATEI|ZULETZT VERWENDETE DATEN öffnen. (Ähnliches gilt übrigens auch für zuletzt genutzte Syntax- und Outputfiles; auf diese Dateitypen komme ich aber später in Kapitel 5 noch ausführlicher zurück.)

Ihre bereits vorhandenen Daten können aber auch als *TXT-, XLS(X)-, CSV*-Dateien, aber auch als STATA- oder SAS-Files – um nur die

wichtigsten Formate zu nennen – eingelesen werden. Abbildung 2.11 gibt einen Überblick über die einlesbaren Datenfile-Formate.

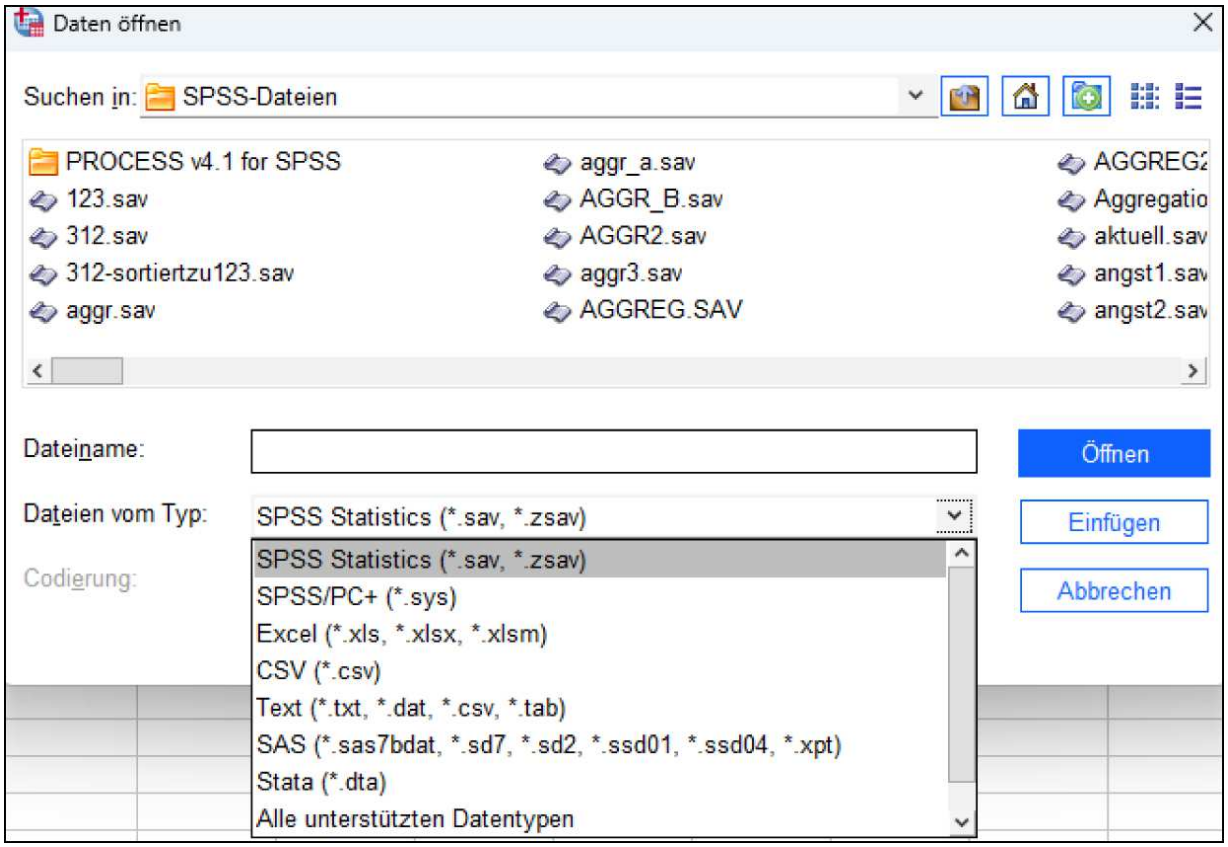



**Abbildung 2.11:** In SPSS einlesbare Datenfile-Formate

# *Einlesen von TXT-Files*

Lassen Sie uns einen Blick auf das Einlesen von Daten werfen, die als **Textfile** vorliegen.

## *Trennzeichen in TXT-Files*

Wenn die einzelnen Merkmale pro Fall durch **Trennzeichen** (Blank – also: Leerzeichen –, Semikolon, Komma, Tabulator) voneinander getrennt sind, erkennt SPSS beim Einlesen automatisch die betreffenden Variablen. In den beiden folgenden Beispielen (Abbildung 2.12 und Abbildung 2.13) ist in einem Texteditor zum einen eine Trennung durch Semikolon und zum anderen eine Tabulatortrennung bei den Daten aus einer Textdatei (man spricht auch von **ASCII-Daten**) vorgenommen

worden. »ASCII« steht dabei für »American Standard Code for Information Interchange« beziehungsweise auf Deutsch: »Amerikanischer Standard-Code für den Informationsaustausch« und ist die in diesem Zusammenhang bedeutendste Zeichencodierung.

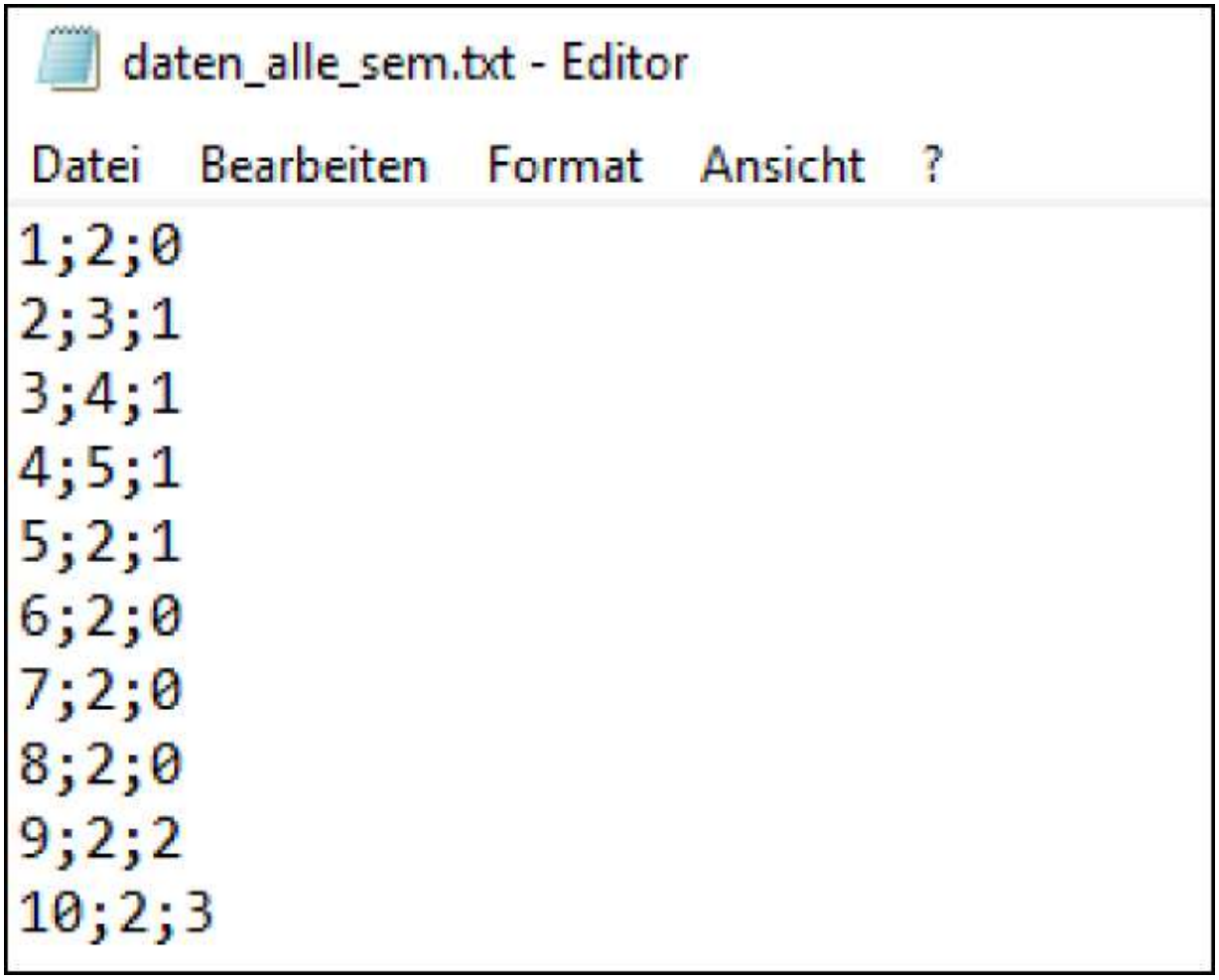



**Abbildung 2.12:** Semikolon-Trennung bei ASCII-Daten (DATEN_ALLE_SEM.TXT)

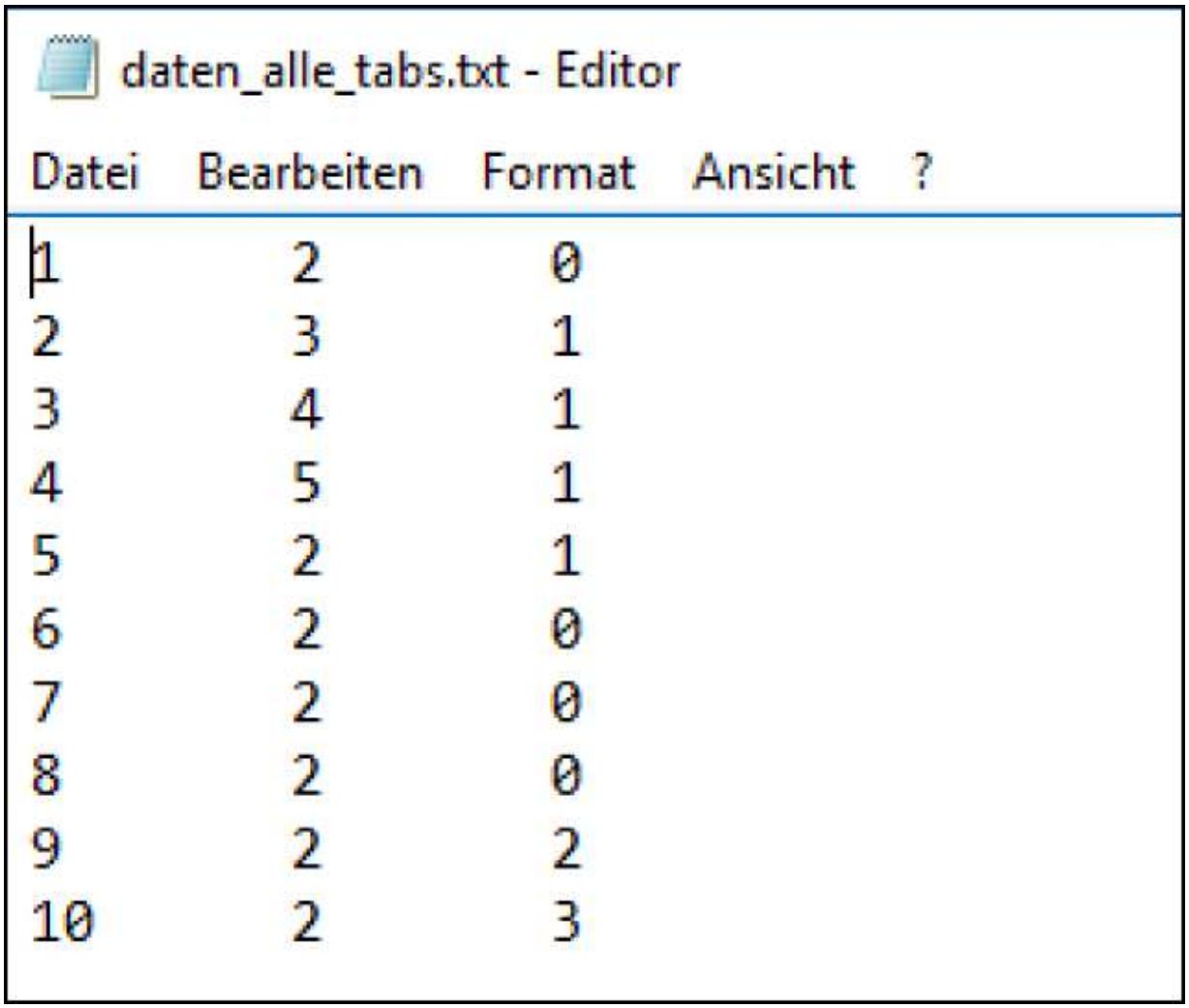



**Abbildung 2.13:** Tabulator-Trennung bei ASCII-Daten (DATEN_ALLE_TABS.TXT)

Die ASCII-Daten können nun in SPSS menügesteuert entweder über DATEI|ÖFFNEN (wie bald illustriert) oder über DATEI|DATEN IMPORTIEREN… eingelesen werden. Das Prozedere wird am Beispiel der mithilfe von Tabulatoren getrennten ASCII-Daten geschildert.

Zunächst öffnen Sie also die gewünschte ASCII-Datei, die ich – zugegebenermaßen wenig originell – als *DATEN_ALLE_TABS.TXT* bezeichnet habe (Abbildung 2.14).

Anschließend werden Sie menügesteuert mithilfe eines sogenannten **Textassistenten** in sechs Schritten mit entsprechenden Abfragen letztlich hin zum endgültigen Einlesen der ASCII-Daten in SPSS

geführt. Dies geschieht in einer Form, in der die Daten dann in SPSS auswertbar sind.

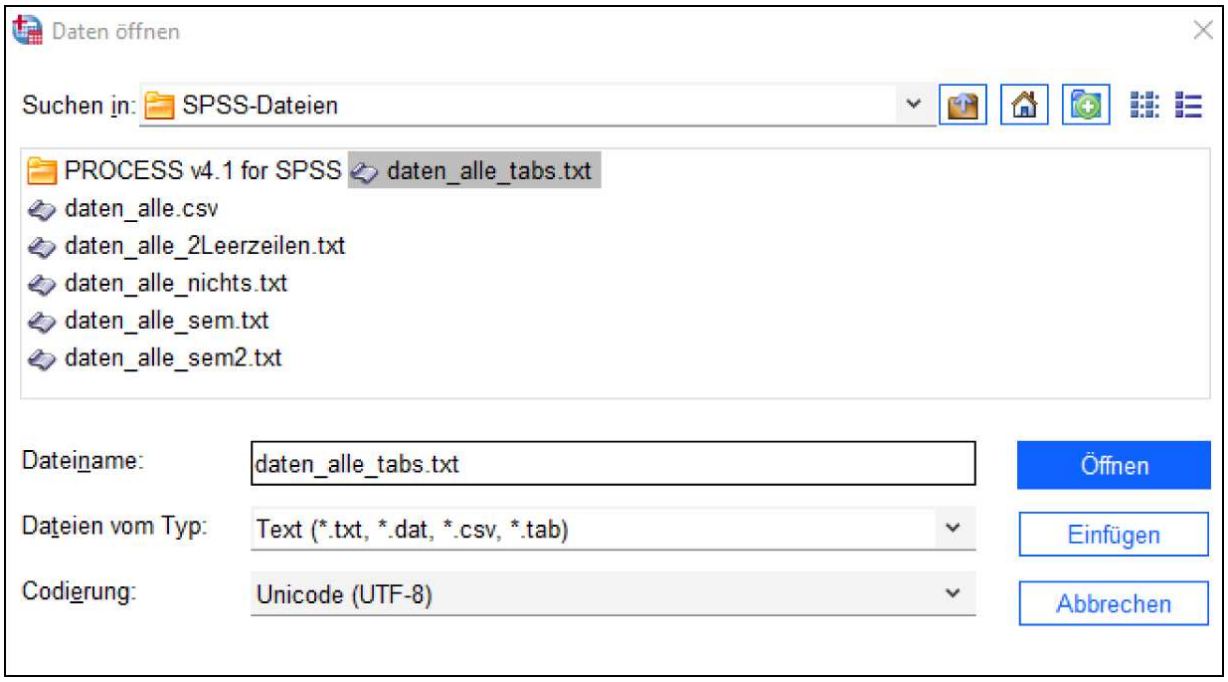



**Abbildung 2.14:** Öffnen einer ASCII-Datei (mit Trennzeichen) in SPSS (DATEN_ALLE_TABS.TXT)

Im ersten Schritt (siehe Abbildung 2.15) werden Sie gefragt, ob Ihre Textdatei ein vordefiniertes **Format** aufweist oder nicht. Das heißt: Haben Sie Nachkommastellen für Ihre Variablen angegeben oder nicht, haben Sie bei Ihren Variablen zwischen numerischen und Stringvariablen unterschieden und so weiter?

In unserem Fall liegt kein vordefiniertes Format vor; deshalb wird hier die Voreinstellung NEIN beibehalten. Außerdem zeigt der erste Bildschirm beim Assistenten für den Textimport schon einmal in einer Vorschau die Daten so an, wie sie anhand der Tabulator-Trennung erkannt werden. Das sieht doch schon einmal ganz gut aus!

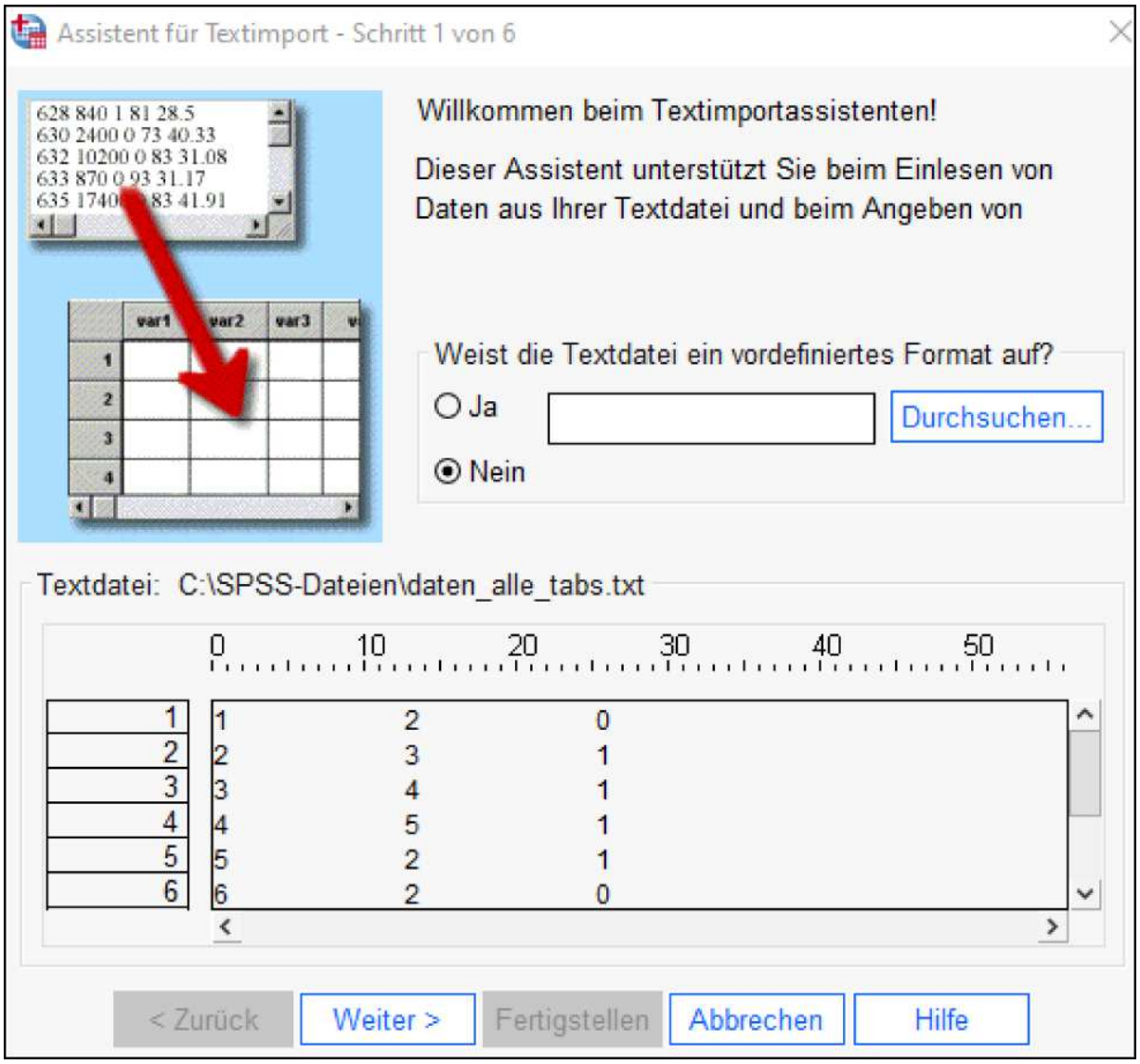



**Abbildung 2.15:** Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 1

Im zweiten Schritt des Einleseprozesses (siehe Abbildung 2.16) wird gefragt, ob die Variablen durch ein Trennzeichen getrennt sind, was ja hier der Fall ist, oder ob sie eine feste Breite aufweisen. Was bei fester Breite zu beachten ist, werde ich gleich etwas weiter unten in diesem Kapitel im Unterabschnitt »Keine Trennzeichen in ASCII-Files« darstellen.

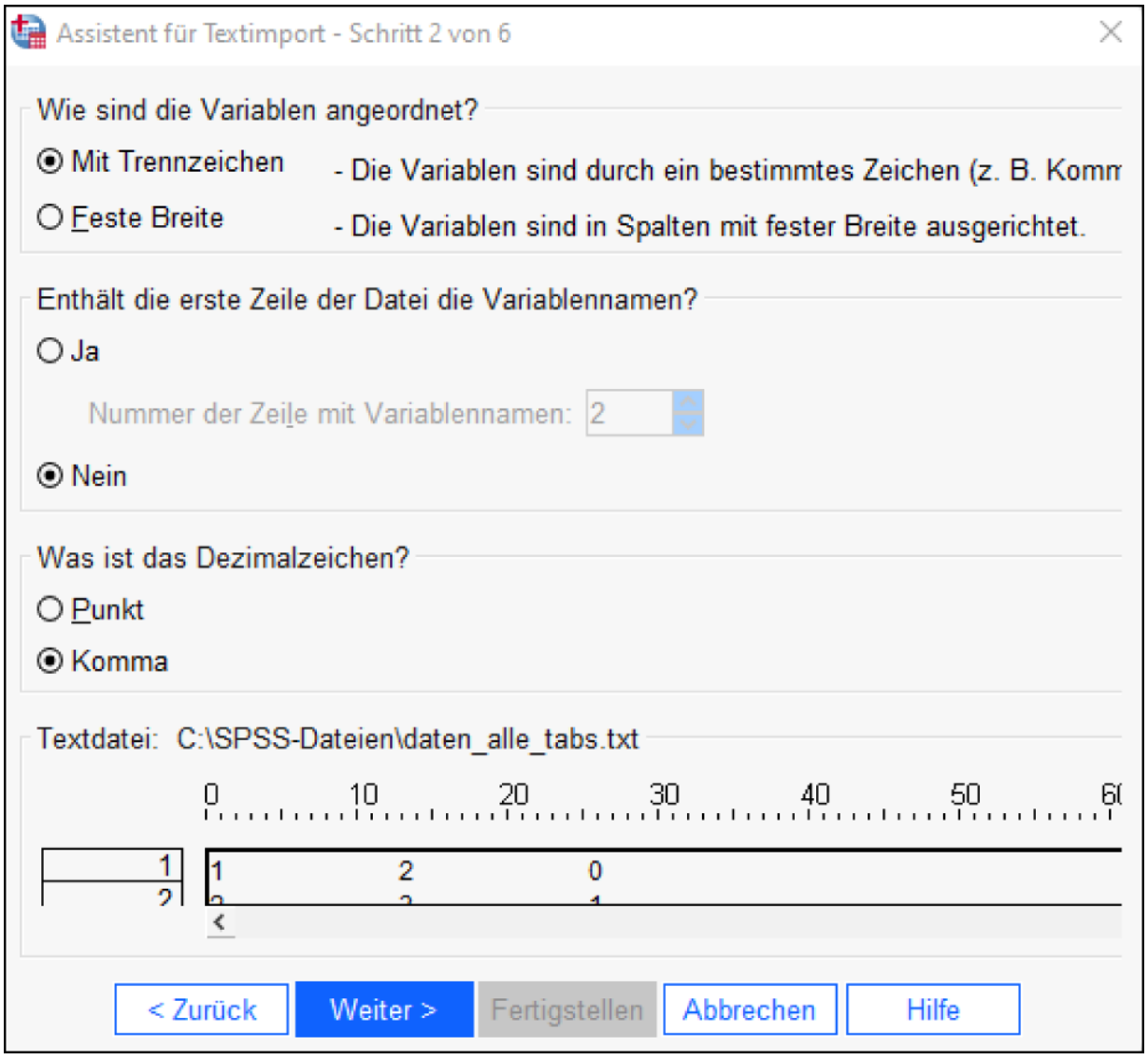



**Abbildung 2.16:** Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 2

Außerdem wird in Schritt 2 gefragt, ob die erste Zeile der Datei die Variablennamen enthält, was im hier diskutierten Beispiel nicht der Fall ist, sodass Sie bitte auch an dieser Stelle die Voreinstellung NEIN übernehmen. Hätten Sie hingegen in der ASCII-Datei in der ersten Zeile Variablennamen hingeschrieben – in unserem Fall zum Beispiel für die drei Variablen *var1*, *var2* und *var3* –, hätten Sie an dieser Stelle natürlich JA anklicken müssen.

Schließlich werden Sie im zweiten Schritt auch noch gefragt, ob Sie bei der Angabe von Dezimalstellen in SPSS einen Punkt oder ein Komma für Ihre ASCII-Daten verwenden. In unserem Fall sind ja in der ASCII-

Datei keinerlei Kommata verwendet worden, sodass hier die freie Auswahl besteht und die Entscheidung zugunsten des Kommas als Dezimalzeichen (und dem Punkt als Tausendertrennzeichen) ausgefallen ist. Alternativ (sogenannte »amerikanische Notation«) können Sie auch den Punkt als Dezimalzeichen (und das Komma dann als Tausendertrennzeichen) verwenden.

Beim dritten Schritt (siehe Abbildung 2.17) werden Sie zunächst gefragt, in welcher Zeile sich der erste Fall befindet. Da Sie ja in der ersten Zeile in der ASCII-Datei keine Variablenbezeichnungen stehen haben und somit der erste Fall in der ersten Zeile steht, lassen Sie als Antwort die »1« stehen.

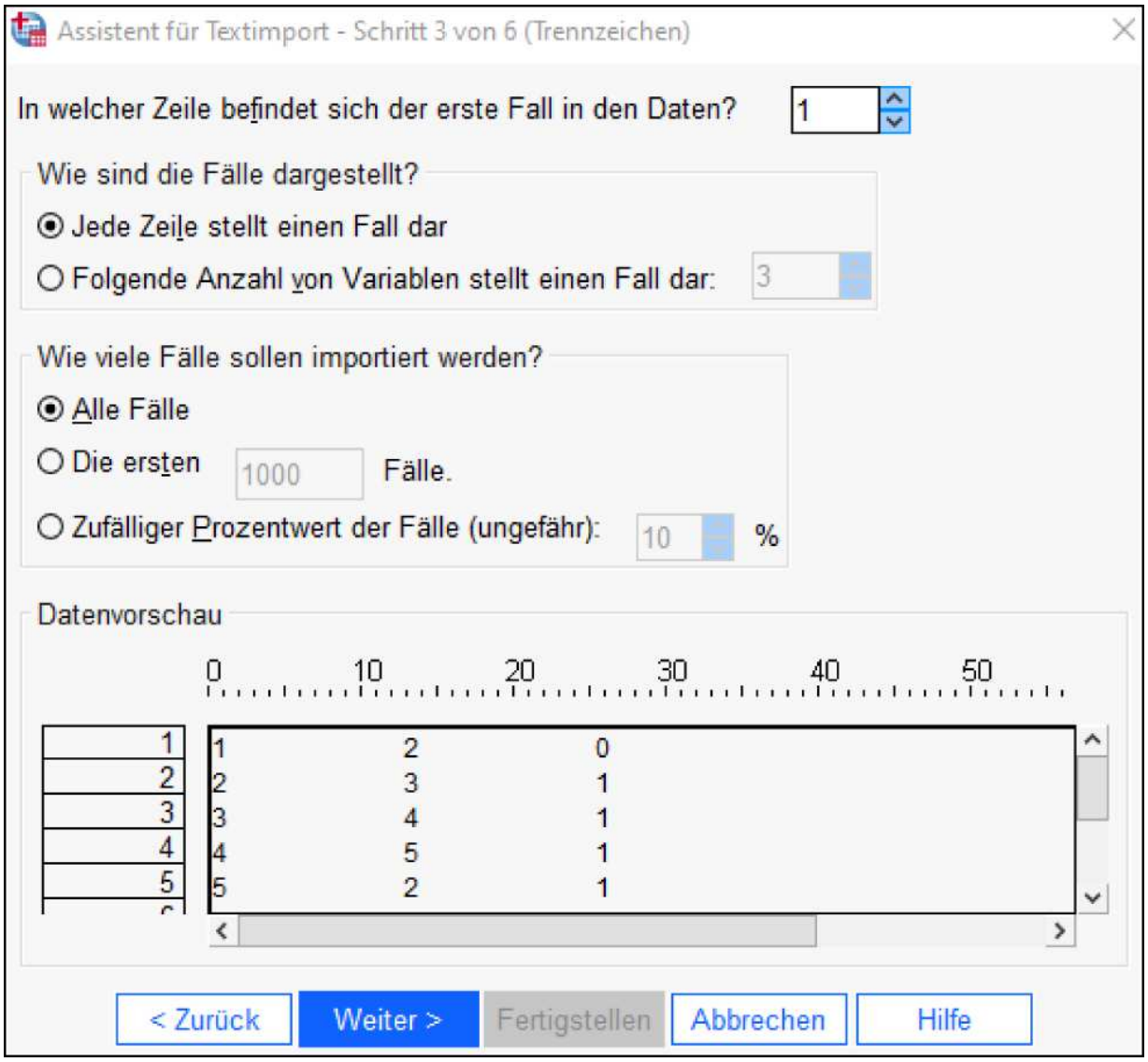


**Abbildung 2.17:** Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 3

Des Weiteren antworten Sie bitte auf die Frage WIE SIND DIE FÄLLE DARGESTELLT? mit: JEDE ZEILE STELLT EINEN FALL DAR. Sie hätten aber an dieser Stelle auch die andere Option ankreuzen können, dass drei Variablen einen Fall darstellen, denn unser ASCII-Datensatz enthält bekanntlich drei Variablen pro Fall. Dies hätte hier dasselbe Ergebnis erbracht.

Im dritten Schritt wird Ihnen eine weitere Frage gestellt, nämlich wie viele Fälle SPSS importieren soll. An dieser Stelle können Sie sich entscheiden, zwischen a) allen Fällen (was in unserem Fall angeklickt worden ist), b) einer bestimmten Anzahl von Fällen mit Beginn beim ersten Fall und c) einer X-Prozent-Stichprobe aus Ihren Fällen.

Spaßeshalber führen Sie bitte mit dem obigen Beispieldatensatz *DATEN_ALLE_TABS.TXT* beim Einlesen eine **Stichprobenziehung** durch, indem Sie beim dritten Schritt eine 20-Prozent-Stichprobe aus Ihren zehn Fällen ziehen (siehe Abbildung 2.18).

Tatsächlich wird wegen dieser zufälligen Stichprobenziehung nur noch ein Teil aller Fälle eingelesen. Allerdings sind nicht – wie bei einer 20-Prozent-Stichprobe zu erwarten war – zwei Fälle, sondern vier Fälle von SPSS ausgewählt worden (siehe Abbildung 2.19). Die Stichprobenziehung von SPSS ist hier nämlich nur ungefährer Natur, wie bereits in Abbildung 2.18 durch den Klammerzusatz UNGEFÄHR angezeigt, und erst bei (sehr) großen Stichproben spiegelt der Auswahlsatz auch mit relativ großer Präzision die gewünschte Anzahl der ausgewählten Fälle wider.

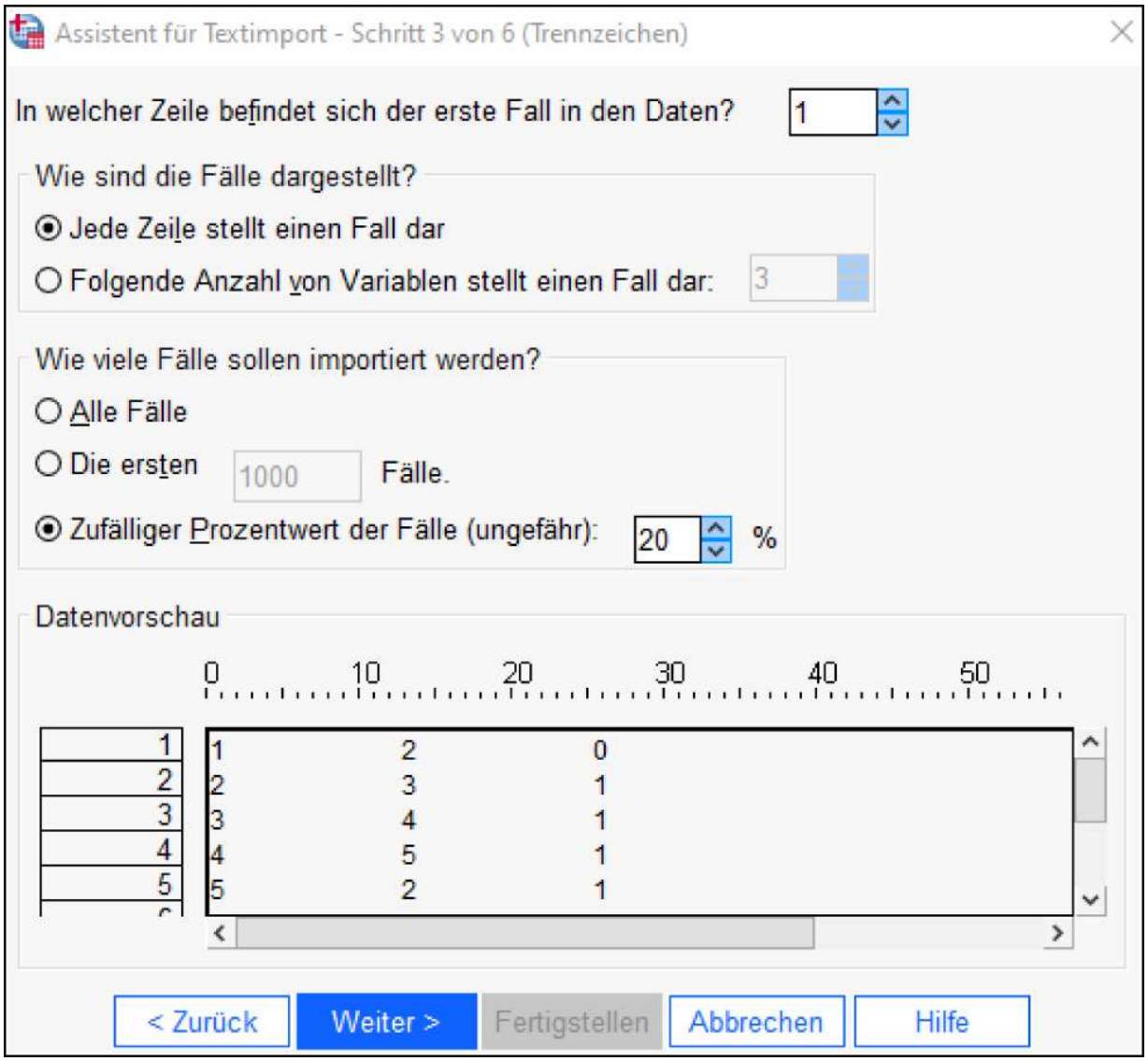



**Abbildung 2.18:** Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 3 mit Stichprobenziehung

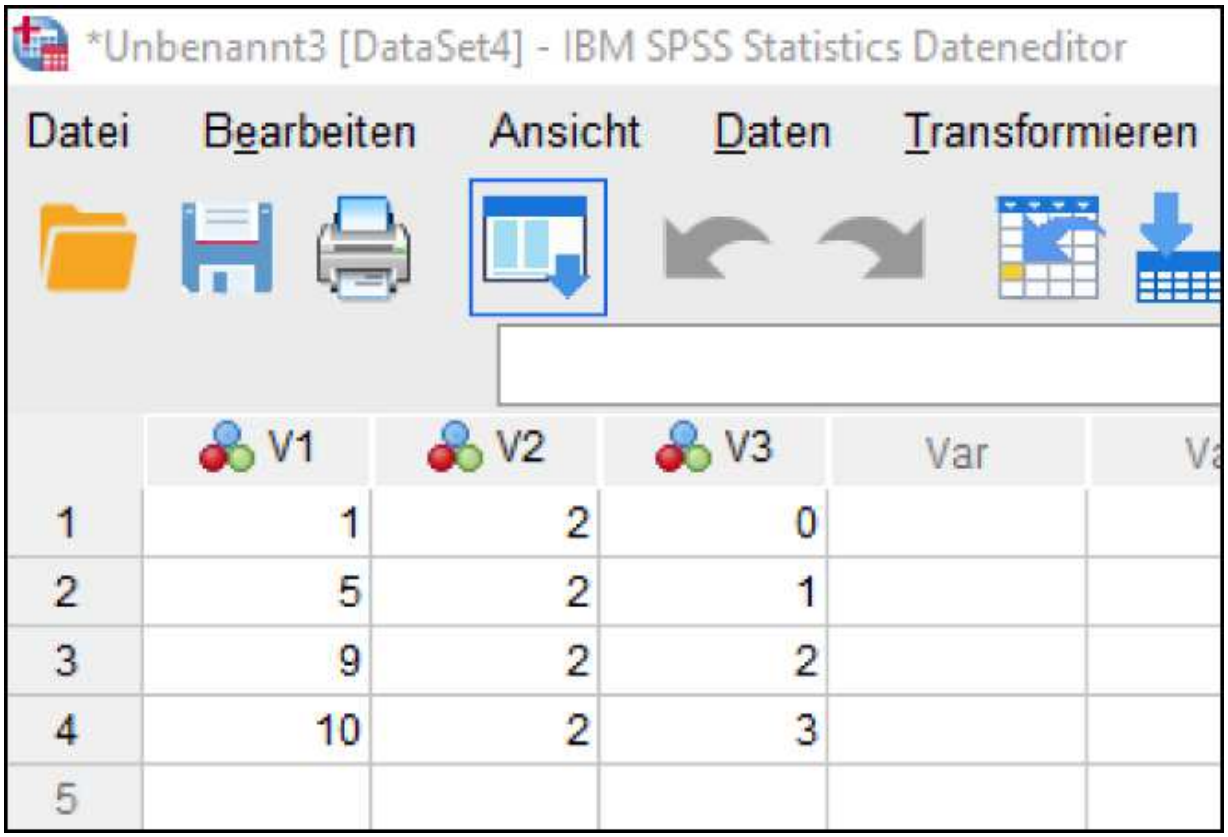



**Abbildung 2.19:** Eingelesene Stichprobe (1)

Wiederholen Sie den Einlesevorgang der zehn Fälle als sogenannte 20-Prozent-Stichprobe, wird eine neue **Zufallsstichprobe** gezogen, und Sie erhalten zum Beispiel die in Abbildung 2.20 angegebene Datentabelle, die sich von den Fällen her von der in Abbildung 2.19 dargestellten Datentabelle unterscheidet. Dies macht den Zufallscharakter der Stichprobenauswahl deutlich.

In der Praxis kann es sinnvoll sein, bei sehr großen Datensätzen nur eine bestimmte Anzahl von Fällen in die Analyse einzubeziehen – allein schon deshalb, um die Rechnerlaufzeiten in einem vertretbaren Rahmen zu halten. Ich hatte in einem früheren Job mal die Aufgabe, circa 20 Millionen Fälle mit jeweils rund 100 Variablen auszuwerten. Dies geschah letztlich zwar auch in SPSS, aber mittels einer Ein-Prozent-Stichprobe, die relativ leicht zu bearbeiten war. Sie sehen: Manchmal ist eine Stichprobenziehung sinnvoll.

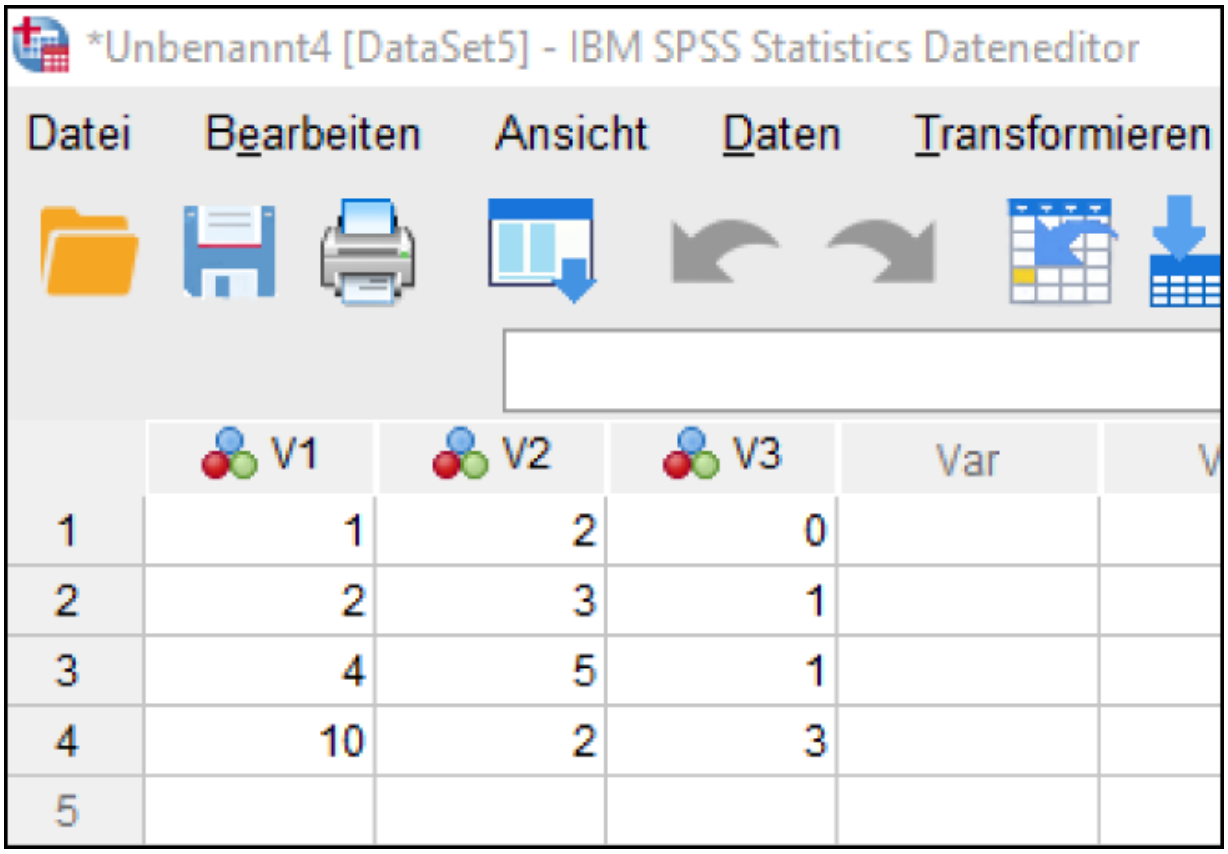



**Abbildung 2.20:** Eingelesene Stichprobe (2)

Der vierte Schritt (siehe Abbildung 2.21) ist vergleichsweise unspektakulär. Hier wird das Trennzeichen zwischen den einzelnen Variablen angegeben (hier via Häkchen bei TABULATOR, wegen des zugrunde liegenden ASCII-Files mit Tabulatortrennung). Überdies könnten Sie bei einer Stringvariablen Leerzeichen am Anfang und/oder am Ende aus dem Datensatz entfernen, was aber im Beispiel hier keine Rolle spielt, da keine der Variablen eine Stringvariable ist.

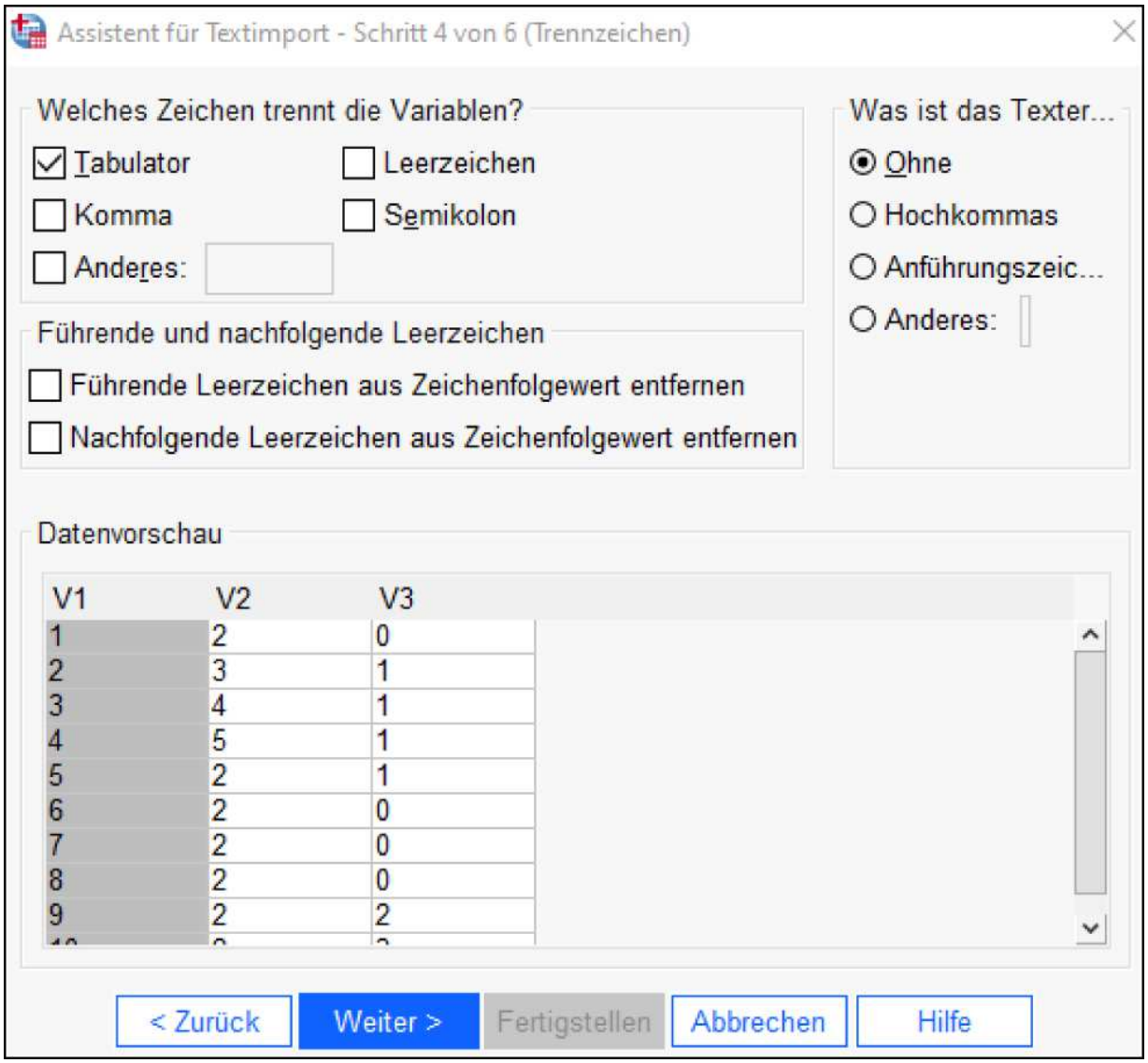



**Abbildung 2.21:** Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 4

Schritt 5 (siehe Abbildung 2.22) ist wie schon der vierte Schritt wenig interessant. Hier können Sie für die in der Datenvorschau ausgewählten Variablen das Datenformat (und den Variablennamen) festlegen. Dies geht aber in der **Variablenansicht** viel einfacher, worauf ich noch ausführlich in Kapitel 4 kommen werde.

Der Einlesevorgang endet schließlich mit dem sechsten Schritt (siehe Abbildung 2.23). Hier müssen Sie nur noch ganz unten auf FERTIGSTELLEN klicken, und fertig ist die Laube!

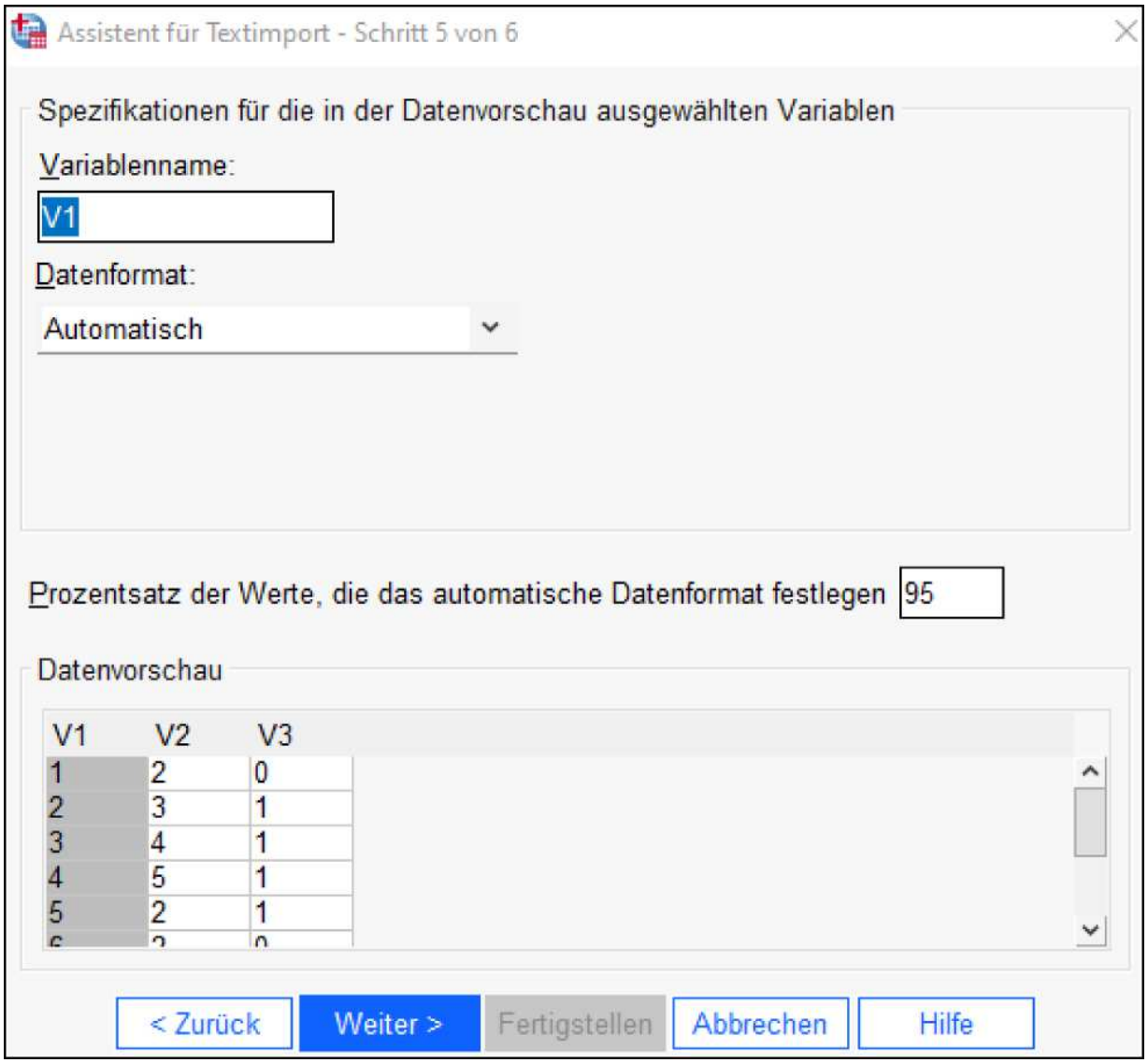



**Abbildung 2.22:** Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 5

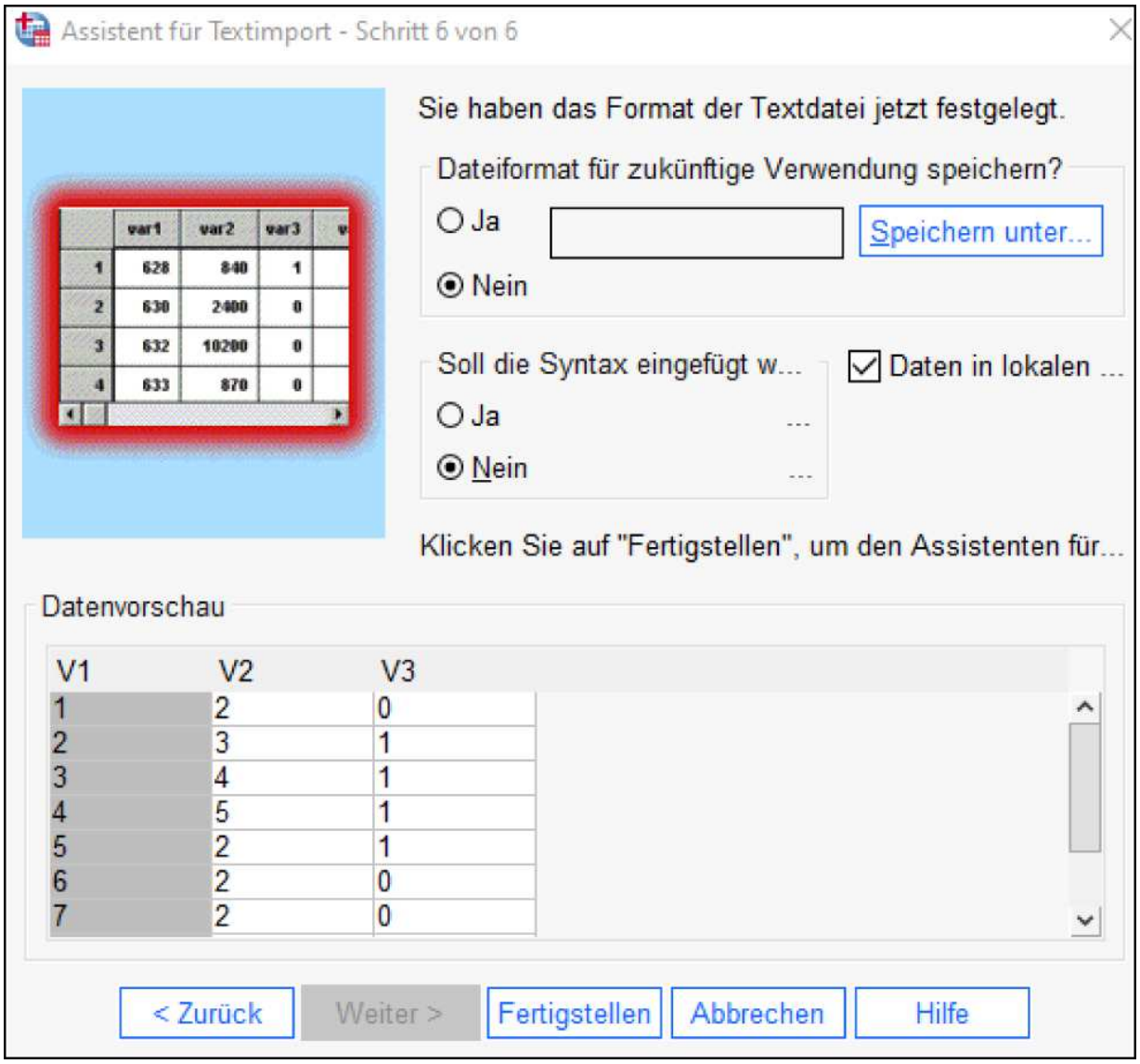
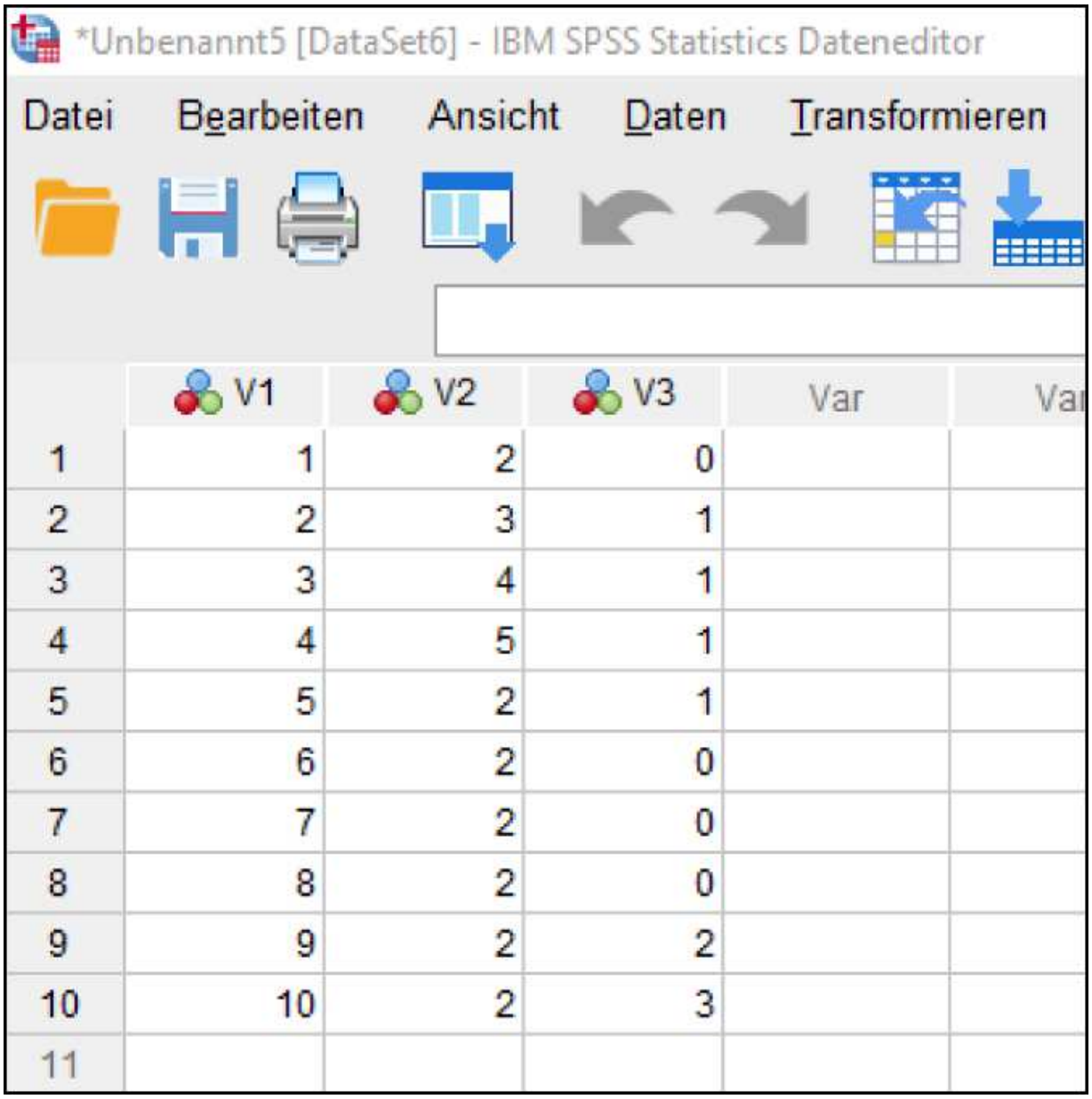



**Abbildung 2.23:** Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 6

Jetzt haben Sie also Ihre Daten in SPSS implementiert, und in SPSS werden im **Dateneditor** Ihre Daten in Form der drei Variablen und der zehn Fälle angezeigt, wobei SPSS die Variablen als *v1*, *v2* und *v3* bezeichnet hat (siehe Abbildung 2.24). Im Dateneditor-Fenster befindet sich also die Datenmatrix.



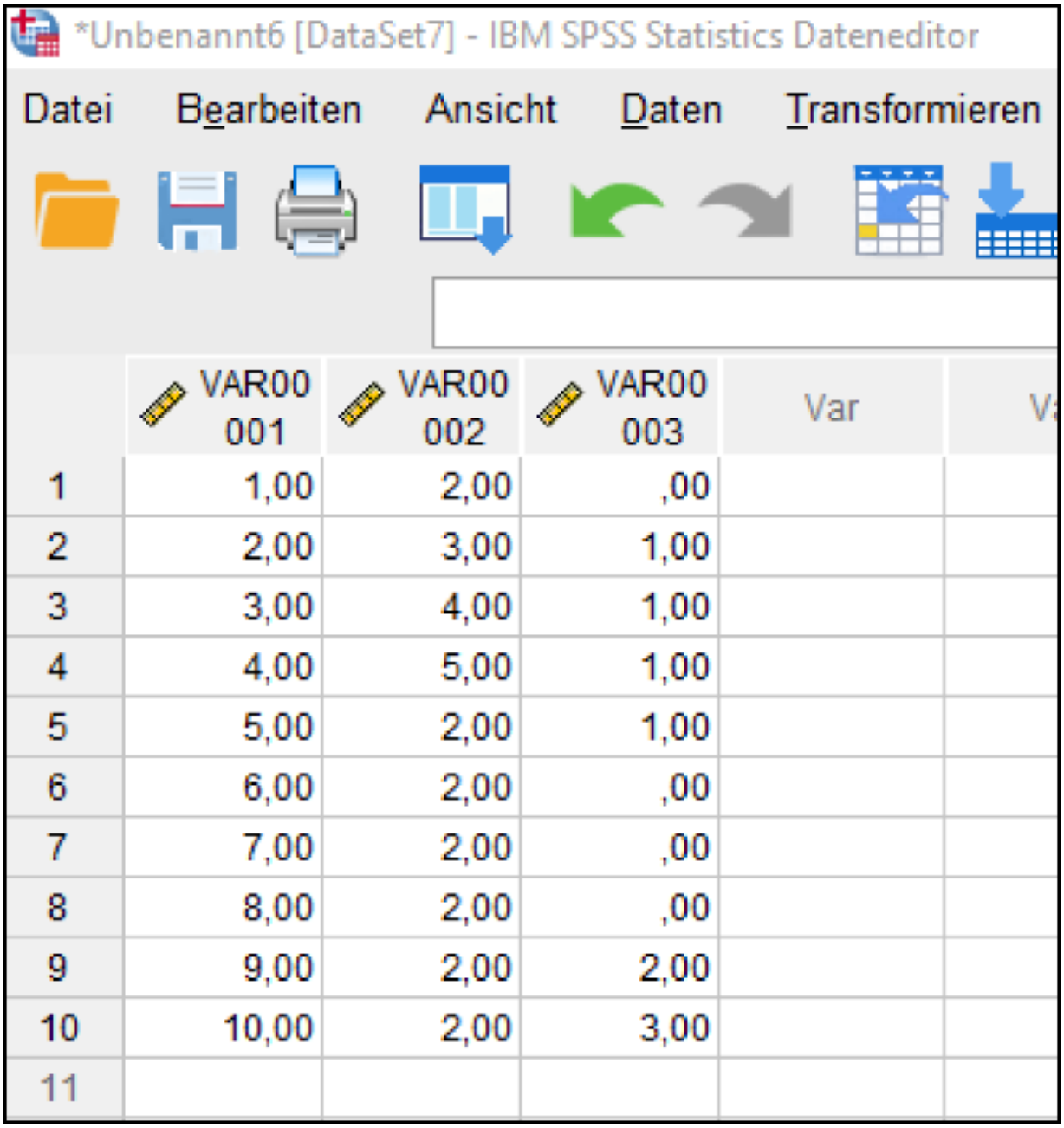

| | V1 | V2 | V3 | Var | Var |
|---|---|---|---|---|---|
| 1 | 1 | 2 | 0 | | |
| 2 | 2 | 3 | 1 | | |
| 3 | 3 | 4 | 1 | | |
| 4 | 4 | 5 | 1 | | |
| 5 | 5 | 2 | 1 | | |
| 6 | 6 | 2 | 0 | | |
| 7 | 7 | 2 | 0 | | |
| 8 | 8 | 2 | 0 | | |
| 9 | 9 | 2 | 2 | | |
| 10 | 10 | 2 | 3 | | |
| 11 | | | | | |

**Abbildung 2.24:** In SPSS mittels des Textassistenten eingelesene ASCII-Daten mit Trennzeichen

Damit haben Sie es fast geschafft – im Grunde genommen ist nur eine Sache noch zu beachten: Damit die eingelesenen Daten Ihnen erhalten bleiben, speichern Sie sie als **SPSS-Datenfile** mit der Extension – also mit der Dateikennzeichnung – *SAV* ab (zum Beispiel als *MUSTERDATEI.SAV)*, und zwar über das Menü DATEI|SPEICHERN (UNTER).

Im Fall der Variablen-Trennung durch Tabulator in einem ASCII-File können Sie die betreffenden ASCII-Daten auch viel einfacher mittels Copy-and-Paste nach SPSS »bringen«. Dazu kopieren Sie die Daten im ASCII-File einfach, markieren im SPSS-Dateneditor die Zelle mit den Koordinaten »Erste Zeile/erste Spalte« (also die Zelle ganz oben links) und fügen die kopierten Daten ein. Dies bietet sich insbesondere bei wenigen Variablen und wenigen Fällen an.

Das Ergebnis (siehe Abbildung 2.25) sieht im Grunde genommen genauso aus wie in Abbildung 2.24 – mit dem einzigen, klitzekleinen und unerheblichen Unterschied, dass die Variablen leicht anders bezeichnet worden sind durch SPSS (*var00001*, *var00002* und *var00003* anstelle von *v1*, *v2* und *v3*).

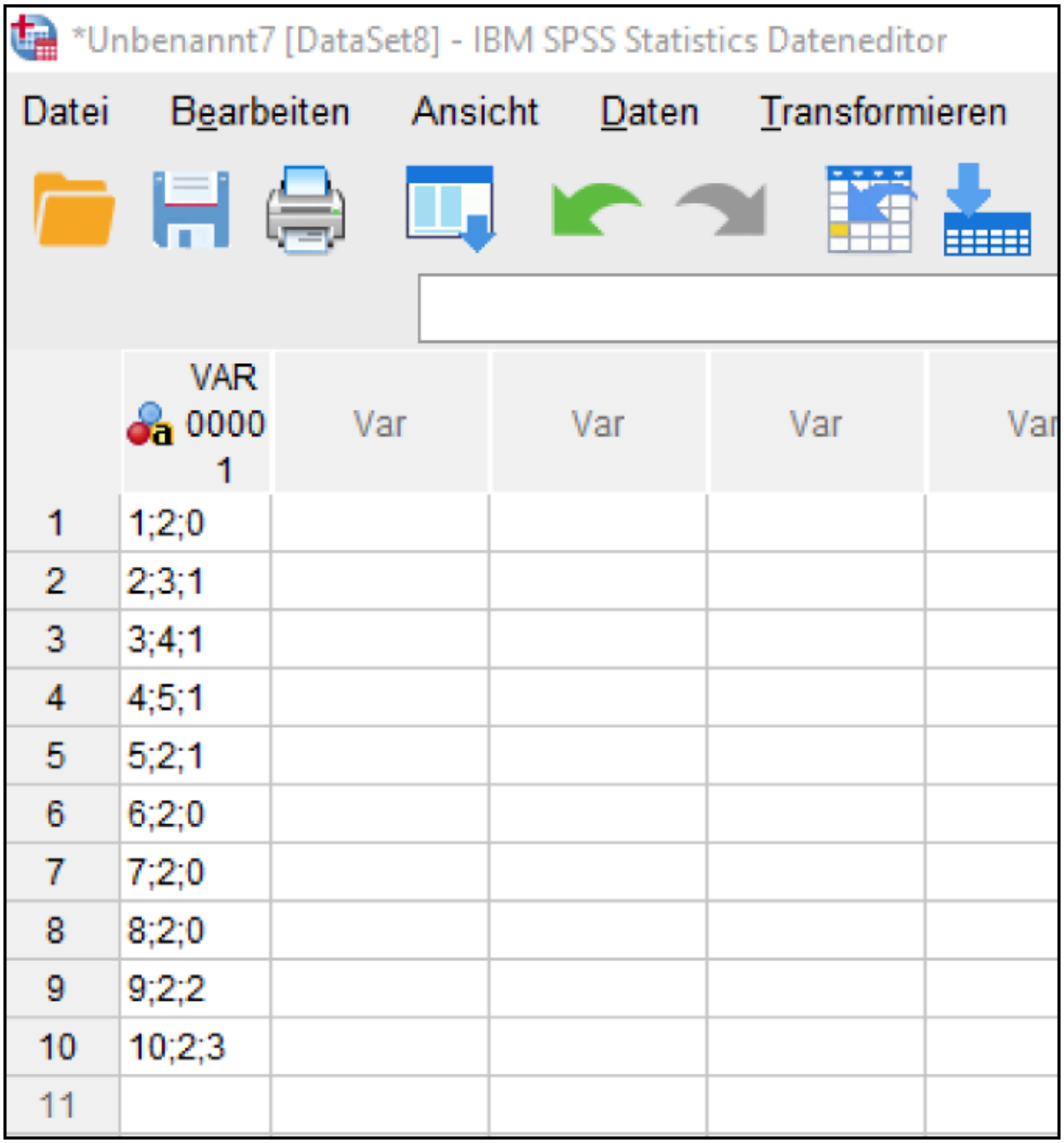



| | VAR00001 | VAR00002 | VAR00003 | Var | |
|---|---|---|---|---|---|
| 1 | 1,00 | 2,00 | ,00 | | |
| 2 | 2,00 | 3,00 | 1,00 | | |
| 3 | 3,00 | 4,00 | 1,00 | | |
| 4 | 4,00 | 5,00 | 1,00 | | |
| 5 | 5,00 | 2,00 | 1,00 | | |
| 6 | 6,00 | 2,00 | ,00 | | |
| 7 | 7,00 | 2,00 | ,00 | | |
| 8 | 8,00 | 2,00 | ,00 | | |
| 9 | 9,00 | 2,00 | 2,00 | | |
| 10 | 10,00 | 2,00 | 3,00 | | |
| 11 | | | | | |

**Abbildung 2.25:** In SPSS mittels Copy-and-Paste eingelesene ASCII-Daten (Tabulator-Trennung)

Allerdings geht dieses Copy-and-Paste in anderen Fällen nicht so gut von der Hand, sodass an dieser Stelle eine massive Warnung angebracht ist.

Wenn Sie das Vorstehende (= Copy-and-Paste) mit einem ASCII-Datenfile versuchen, in dem die Variablen durch Semikolon voneinander getrennt sind (also zum Beispiel mit *DATEN_ALLE_SEM.TXT*), werden Sie Schiffbruch erleiden. Dann wird Ihr Ergebnis so unbefriedigend wie in Abbildung 2.26 aussehen. Das heißt: In diesem Fall werden alle Daten pro Fall in eine Variable geschrieben. Daher sollten Sie in diesem Fall unbedingt den Textassistenten wie oben nutzen. Oder weniger vornehm ausgedrückt: Lassen Sie bitte in diesem Fall die Finger von Copy-and-Paste!




| | VAR 00001 | Var | Var | Var | Var |
|---|---|---|---|---|---|
| 1 | 1;2;0 | | | | |
| 2 | 2;3;1 | | | | |
| 3 | 3;4;1 | | | | |
| 4 | 4;5;1 | | | | |
| 5 | 5;2;1 | | | | |
| 6 | 6;2;0 | | | | |
| 7 | 7;2;0 | | | | |
| 8 | 8;2;0 | | | | |
| 9 | 9;2;2 | | | | |
| 10 | 10;2;3 | | | | |
| 11 | | | | | |

**Abbildung 2.26:** In SPSS mittels Copy-and-Paste eingelesene ASCII-Daten (Semikolon-Trennung)

## *Keine Trennzeichen in ASCII-Files*

Etwas schwieriger wird es mit dem Einlesen von ASCII-Daten auch dann, wenn Sie im Ursprungs-ASCII-Datenfile kein Trennzeichen haben, sondern nur eine feste Variablenlänge vorliegt. Zwischen den Variablen steht im ASCII-Datenfile also nichts! In unserem

Beispieldatensatz weist die erste Variable zwei Stellen auf (mit den Zahlen von 1 bis 10, wobei 1 bis 9 jeweils eine führende Null zugewiesen wurde), und die zweite sowie die dritte Variable haben jeweils eine Stelle (siehe Abbildung 2.27). Pro Fall sind also im Datensatz in diesem Beispiel vier Stellen vergeben worden. Man spricht daher von einer **festen Satzlänge**.

Wenn Sie nun versuchen, diesen Datensatz *DATEN_ALLE_NICHTS.TXT* in SPSS einzulesen, kommt leider gleich am Anfang des Textassistenten die **Warnung**, dass Sie das Einlesen anderweitig betreiben müssen (siehe Abbildung 2.28).

Das Problem, warum die ASCII-Daten mit fester Breite nicht eingelesen werden, besteht laut der vorstehend genannten SPSS-Warnung darin, dass die Daten in **Unicode** codiert sind. Das Ziel einer Codierung via Unicode liegt grundsätzlich darin, dass Text in verschiedenen Sprachen über Ländereinstellungen hinweg problemlos genutzt werden kann. Unicode kann von SPSS (ab Version 16) verlustfrei bearbeitet werden. Länderspezifische Codierungen, die sich zum Beispiel in SAS- oder Textdateien finden, werden beim Import in SPSS in das Format UTF-8 umgewandelt beziehungsweise beim Export von SPSS in andere Dateiformate wieder in den ursprünglichen Code zurückverwandelt (wobei UTF-8 für »Universal-Character-Set Transformation Format 8 Bit« steht). (In SPSS-Versionen vor SPSS 16 verwenden Sie im Übrigen für zu öffnende Dateien den sogenannten Codepage-Modus.)

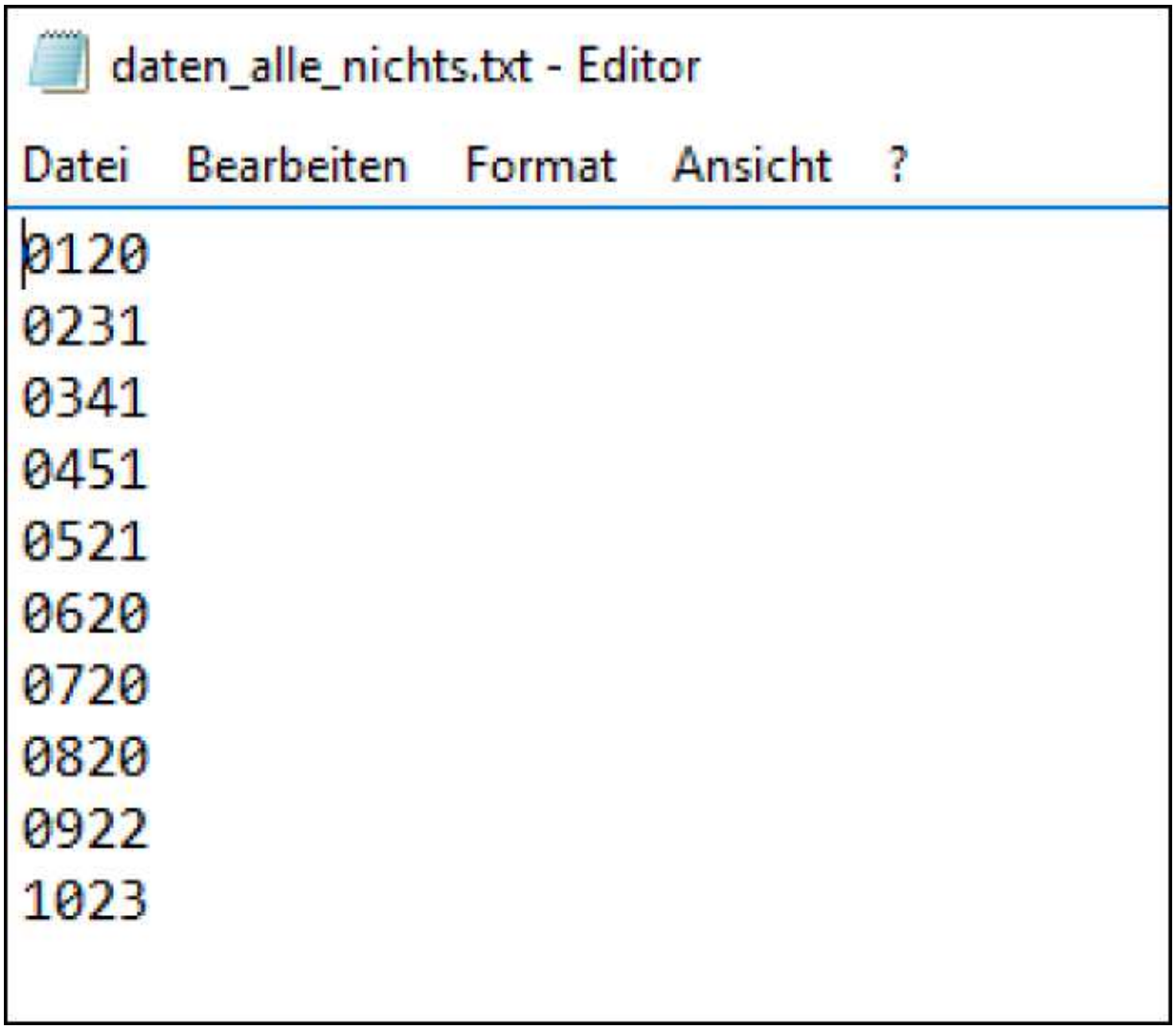



**Abbildung 2.27:** ASCII-Daten mit fester Satzlänge (DATEN_ALLE_NICHTS.TXT)

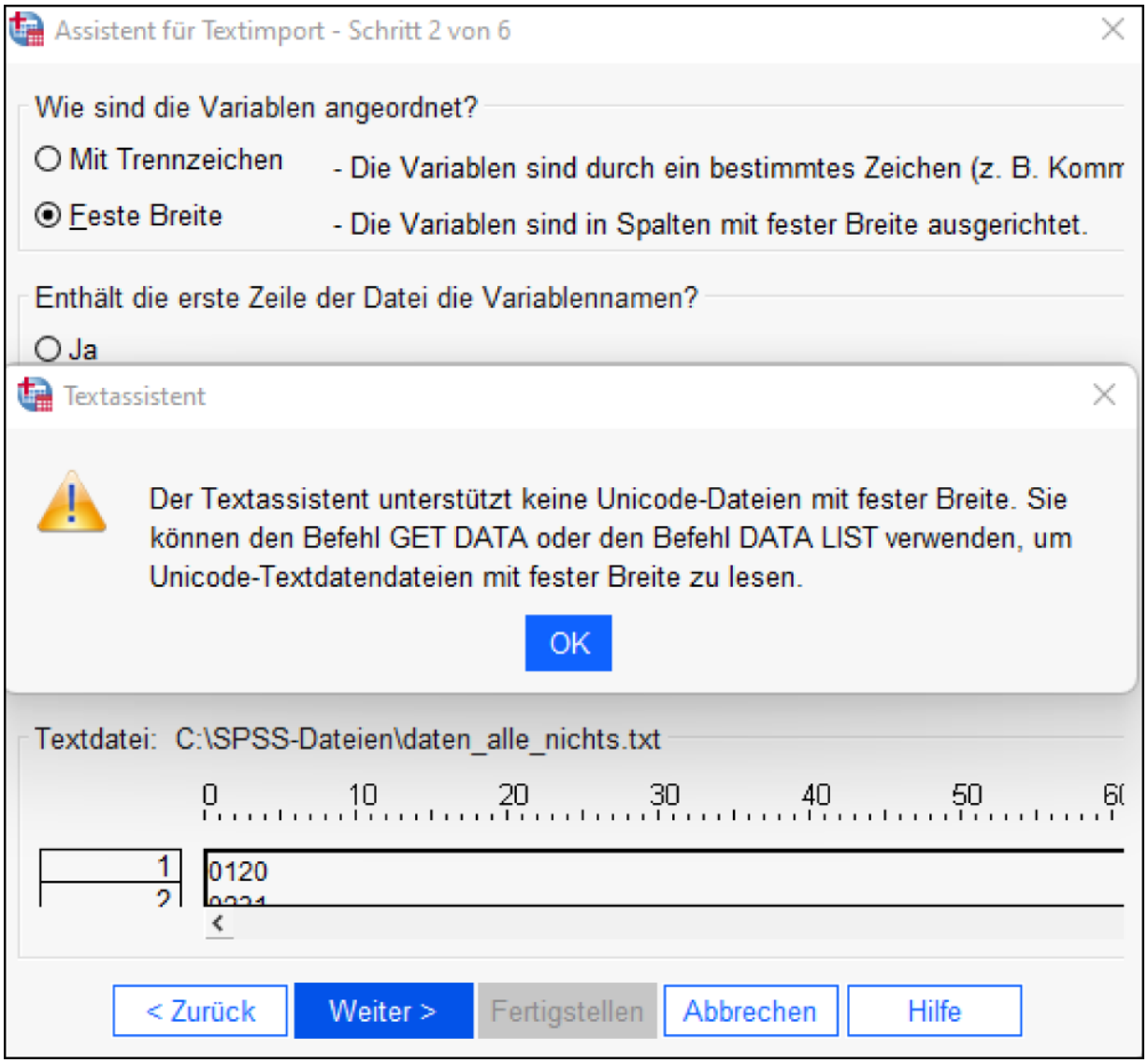



**Abbildung 2.28:** SPSS-Warnung beim Versuch des Einlesens von ASCII-Daten mit fester Satzlänge (Unicode-Format) mittels des Textassistenten

In unserem Fall verhindert der Unicode aber leider offenkundig das Einlesen Ihrer Daten mittels Textassistenten. Um Ihre Daten dennoch leicht einlesen zu können, müssen Sie beim Öffnen der betreffenden ASCII-Datei bei der Codierung auf LOKALE CODIERUNG wechseln (siehe Abbildung 2.29).

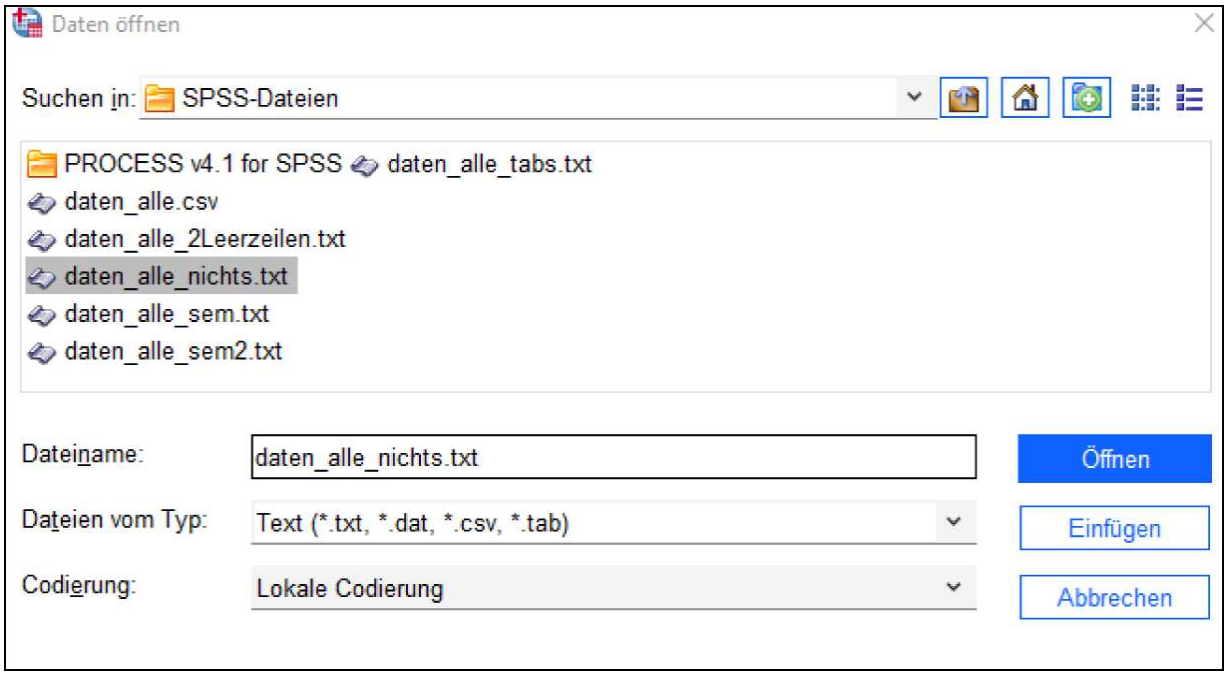



**Abbildung 2.29:** Lokale Codierung von ASCII-Daten

Wie Sie Abbildung 2.30 entnehmen können, werden bei Schritt 1 im Rahmen des Einleseprozesses die ASCII-Daten mit fester Satzlänge derart dargestellt, dass die Merkmalsausprägungen pro Fall wie im Rohdatenfile ohne jegliches Trennzeichen direkt hintereinander abgebildet sind (siehe Abbildung 2.30), was jetzt vielleicht nicht ganz unerwartet kommt.

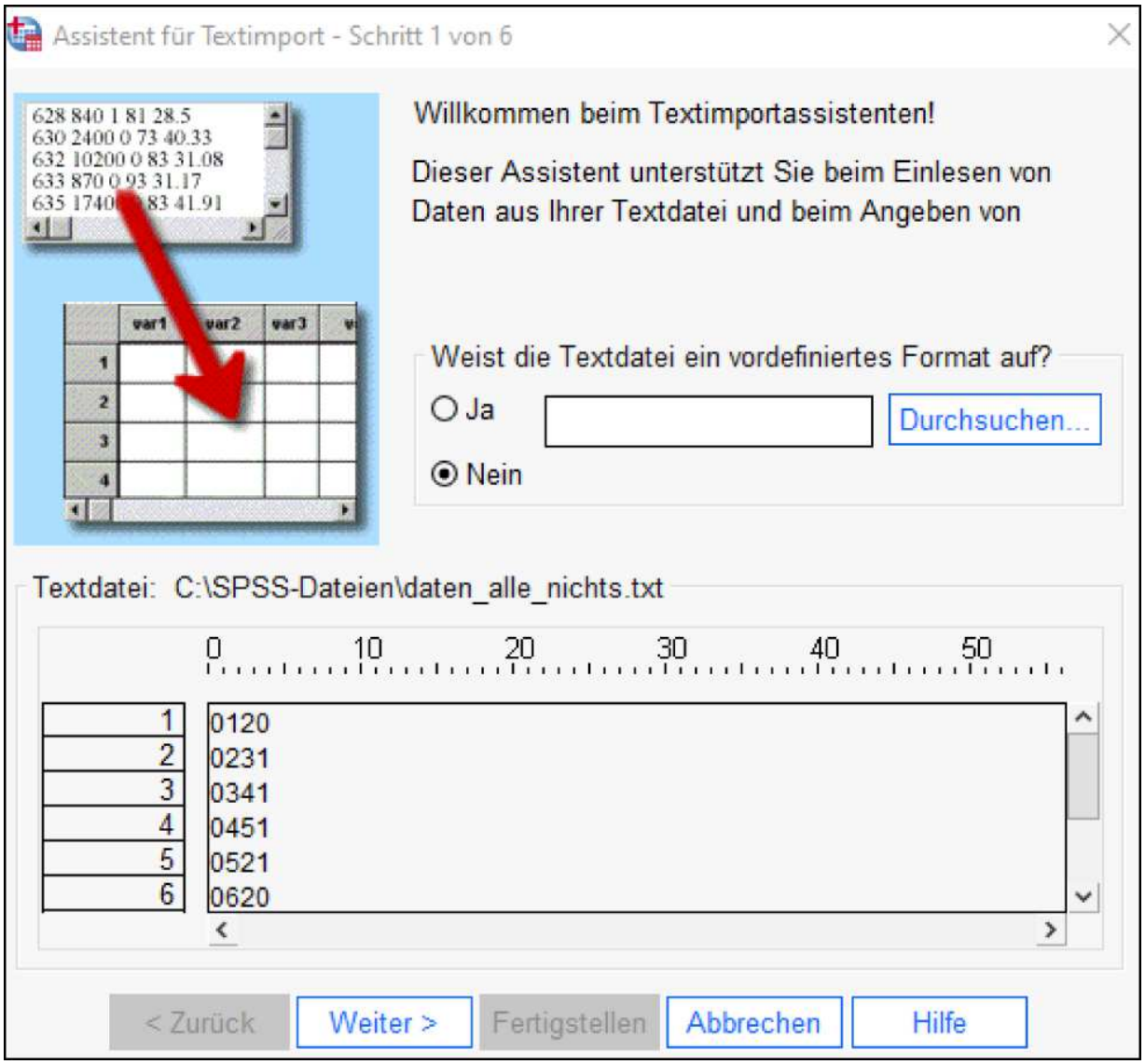



**Abbildung 2.30:** Textassistent zum Einlesen von ASCII-Daten ohne Trennzeichen, mit fester Satzlänge – Schritt 1

In Schritt 2 wird nunmehr naheliegenderweise der Modus FESTE BREITE eingestellt (siehe den obigen Teil von Abbildung 2.31).

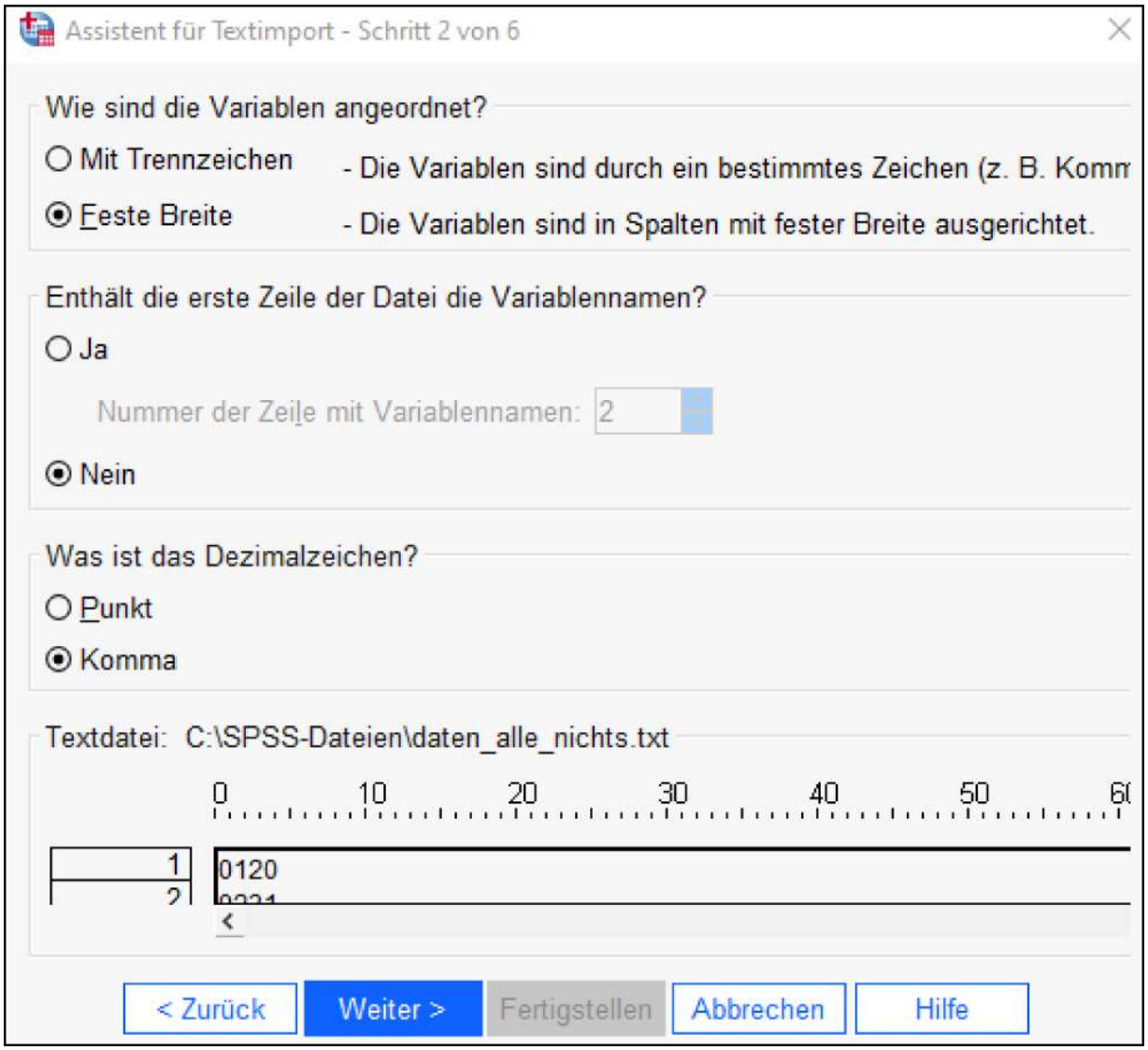



**Abbildung 2.31:** Textassistent zum Einlesen von ASCII-Daten ohne Trennzeichen, mit fester Satzlänge – Schritt 2

Schritt 3 sieht wie oben bei ASCII-Daten mit Trennzeichen aus (siehe hierzu eventuell nochmals Abbildung 2.17), sodass ich ihn hier nicht noch einmal behandele. Demgegenüber ist Schritt 4 bei ASCII-Daten ohne Trennzeichen, aber mit fester Satzlänge sehr wichtig. Hier müssen Sie nämlich manuell zwischen den einzelnen Variablen – jeweils zur Abgrenzung der Variablen voneinander – einen senkrechten Trennstrich ziehen (siehe Abbildung 2.32). Das heißt in unserem Datenbeispiel: Ziehen Sie den ersten senkrechten Trennstrich nach der zweiten Spalte und den zweiten nach der dritten Spalte, wobei Sie bitte jeweils auf TRENNUNG EINFÜGEN klicken; nach der vierten Spalte müssen Sie

übrigens keinen Trennstrich mehr ziehen, weil dort ja das Ende der Variablenliste erreicht ist.

Schritt 5 ist nicht sonderlich spannend: Hier habe ich lediglich die Standard-Variablen-bezeichnungen von *f1*, *f2* und *f3* in *v1*, *v2* und *v3* geändert (einfach nur deshalb, weil der erste Buchstabe von »Variable« nun mal das »V« und nicht das »F« ist … ;-)). Diese (pingelige) Veränderung meinerseits ist in Abbildung 2.33 angegeben.

Schritt 6 ist hier nicht dargestellt, weil er dem obigen sechsten Schritt bei ASCII-Daten mit Trennzeichen wie ein Ei dem anderen gleicht (siehe hierzu eventuell nochmals Abbildung 2.23). Wenn Sie diesen sechsten Schritt auch noch absolviert haben, erhalten Sie – zum Glück! – schließlich die gewohnte und vor allem erwartete Datendarstellung im SPSS-Dateneditor (siehe Abbildung 2.34).

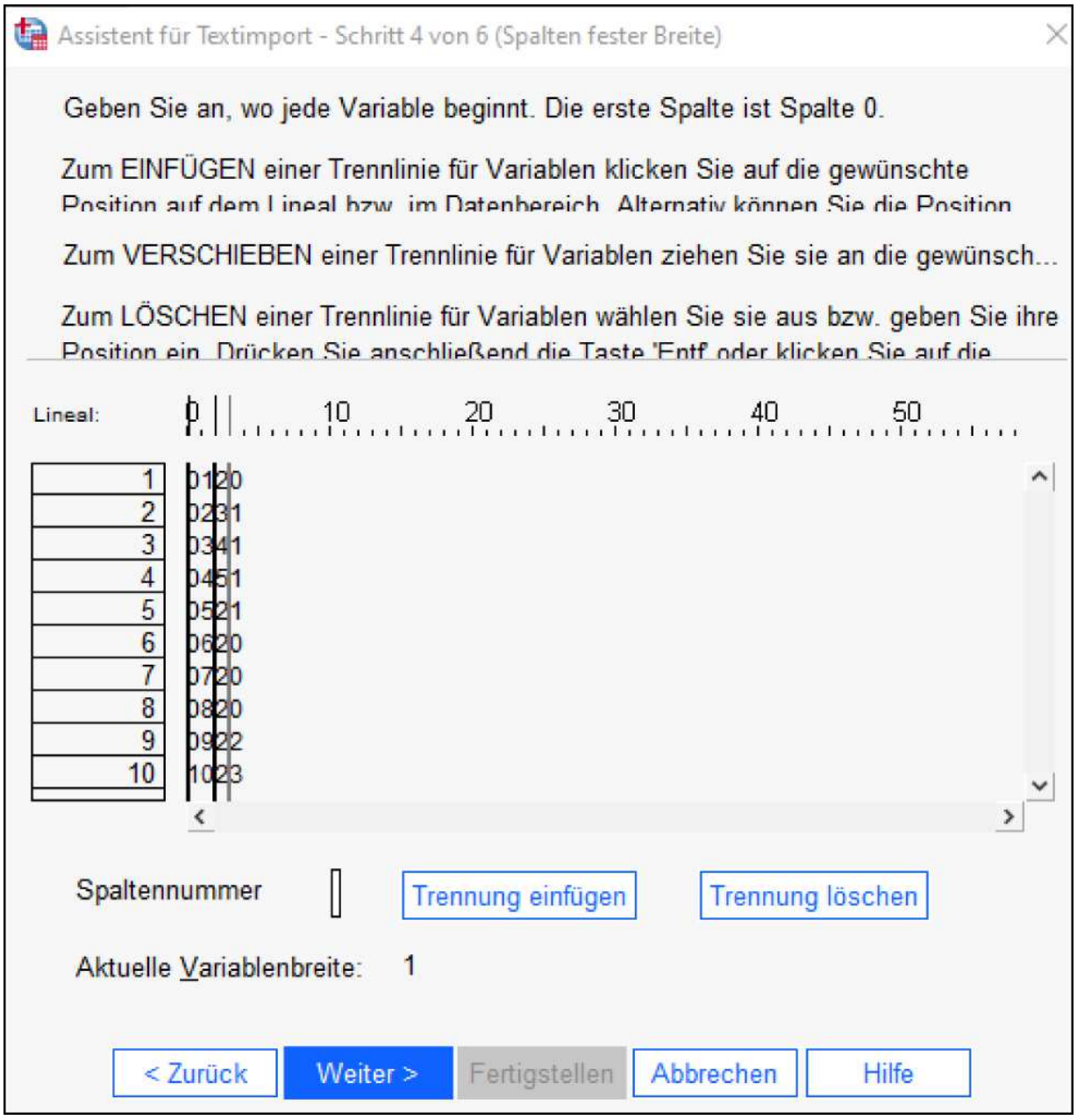



**Abbildung 2.32:** Textassistent zum Einlesen von ASCII-Daten ohne Trennzeichen, mit fester Satzlänge – Schritt 4

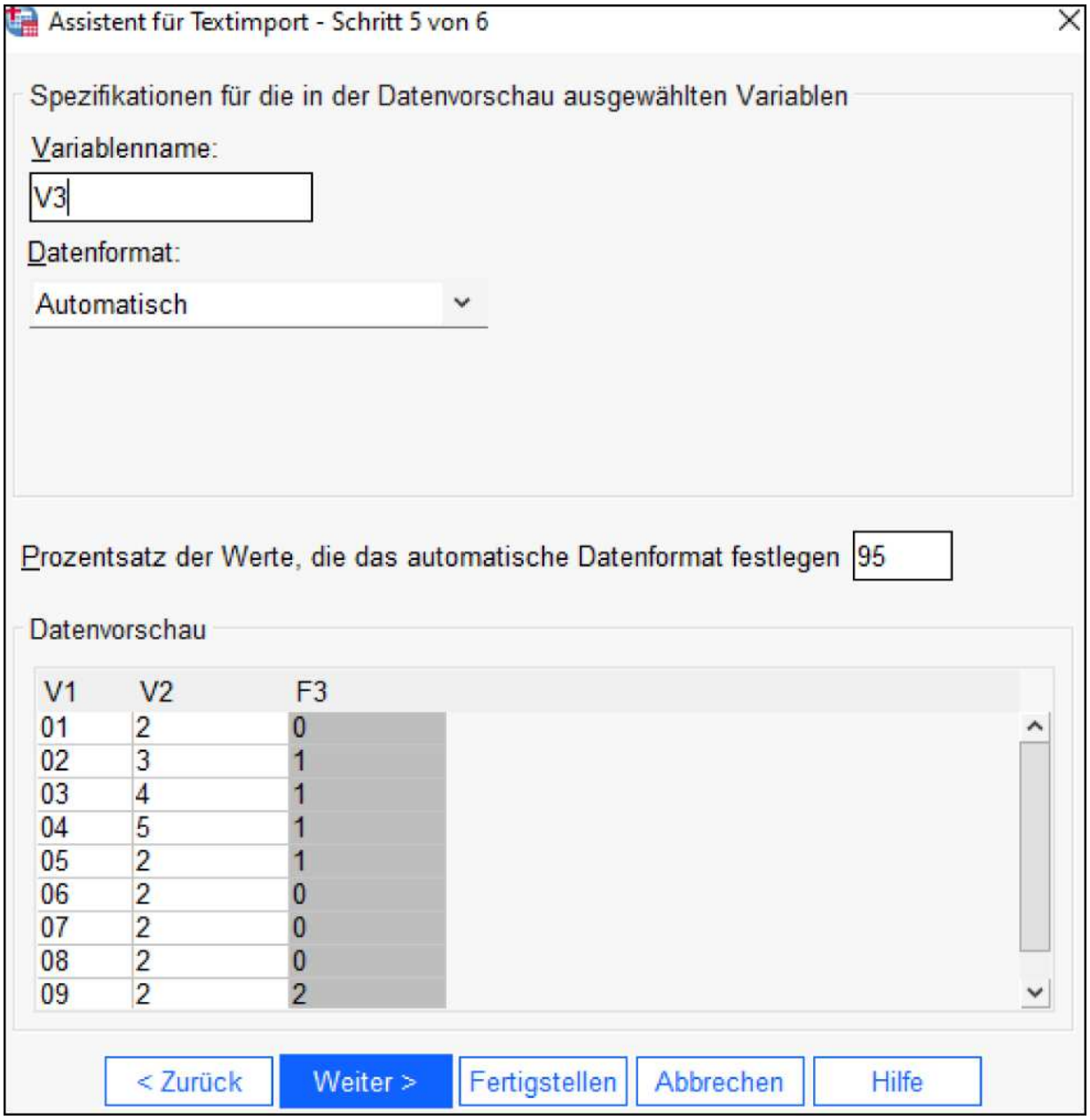



**Abbildung 2.33:** Textassistent zum Einlesen von ASCII-Daten ohne Trennzeichen, mit fester Satzlänge – Schritt 5

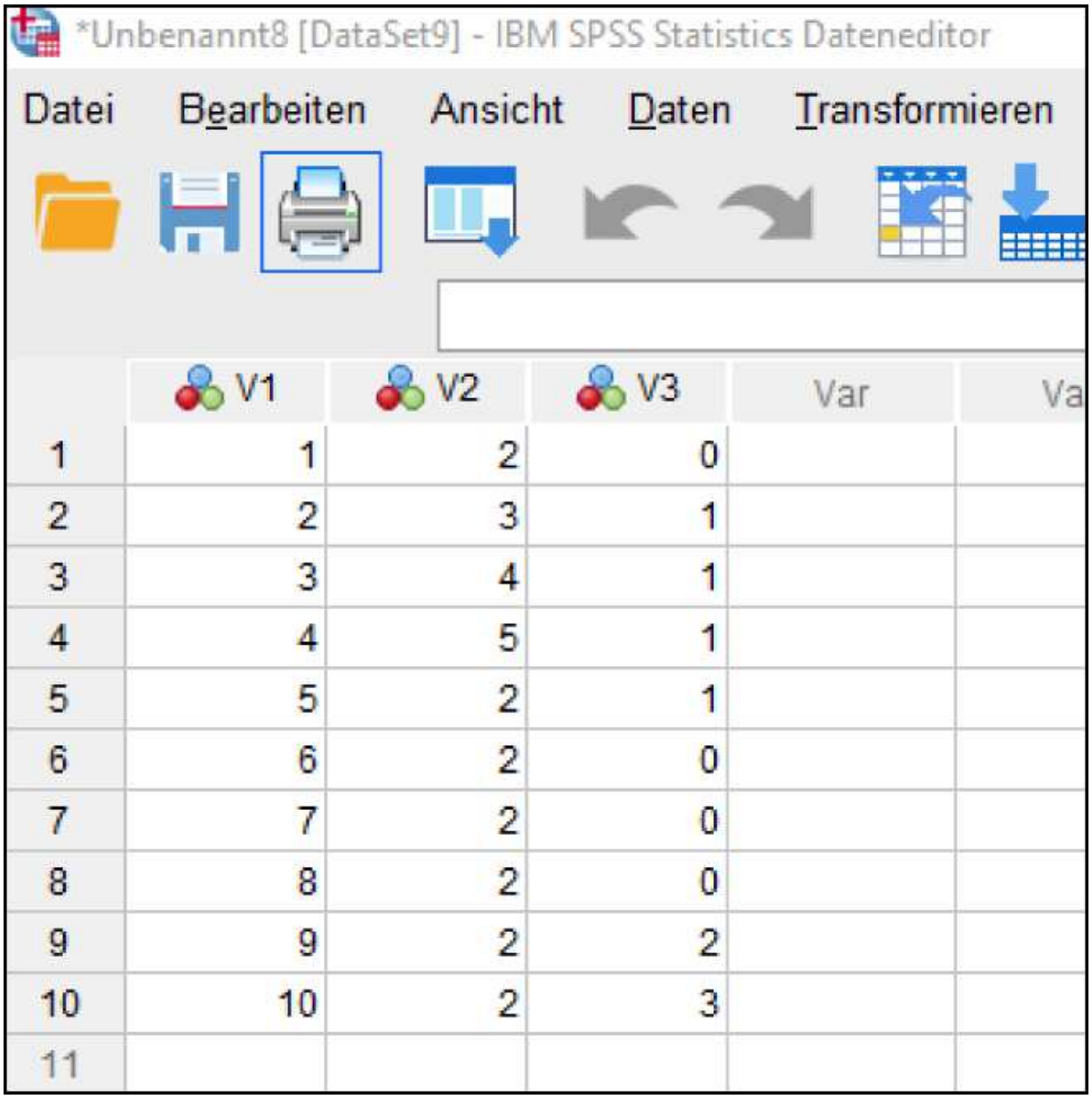


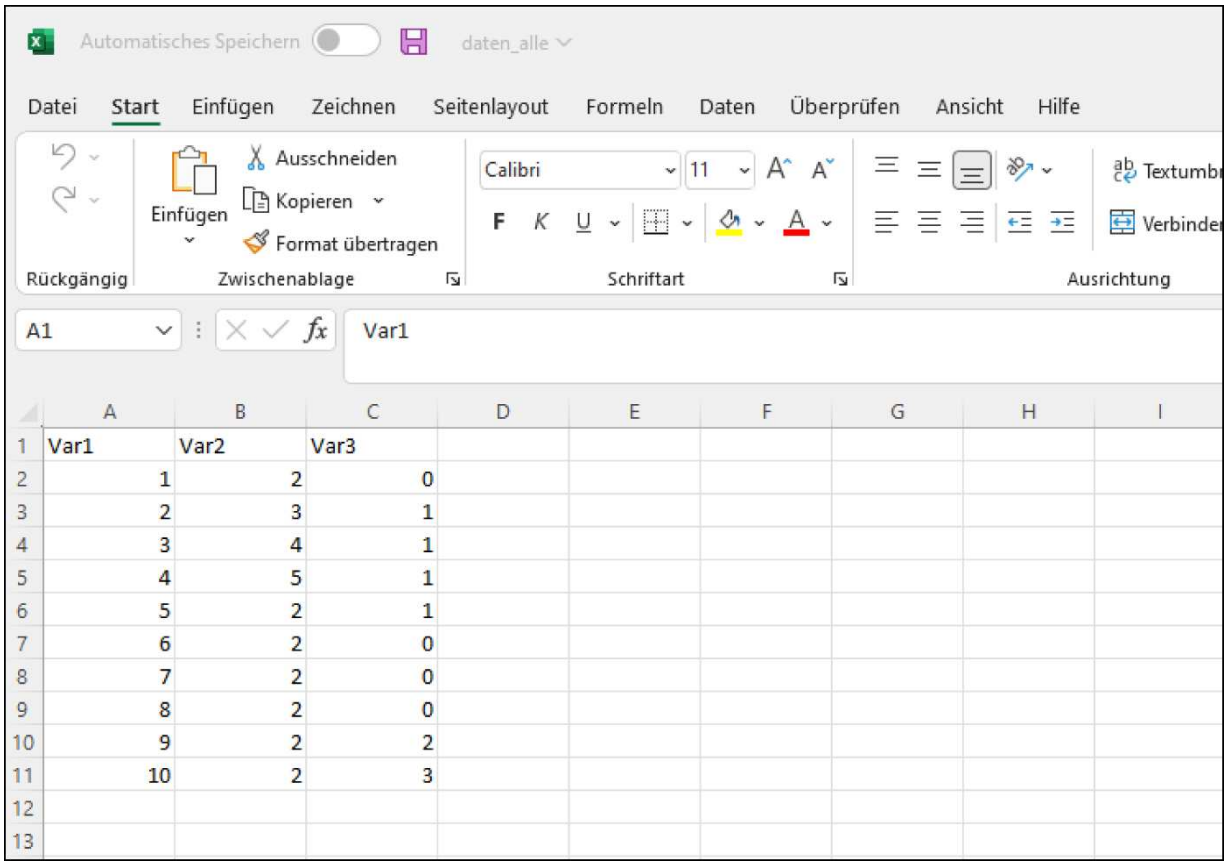

**Abbildung 2.34:** In SPSS mittels Textassistenten eingelesene ASCII-Daten ohne Trennzeichen und mit fester Satzlänge

# *Einlesen von EXCEL-, STATA- und SAS-Daten*

Sind Ihre Daten in **EXCEL** gespeichert, können Sie auch diese XLS(X)-Daten benutzergeführt nach SPSS »bringen«. Sie können aber auch hier – insbesondere bei eher kleinen Datendateien in Bezug auf Fall-

und/oder Variablenanzahl – via Copy-and-Paste die betreffenden Daten nach SPSS kopieren.

Im Folgenden sollen die in Abbildung 2.35 dargestellten EXCEL-Daten nach SPSS eingelesen werden. Diese EXCEL-Daten finden Sie in *DATEN_ALLE.XLSX*.




| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Var1 | Var2 | Var3 | | | | | | |
| 2 | 1 | 2 | 0 | | | | | | |
| 3 | 2 | 3 | 1 | | | | | | |
| 4 | 3 | 4 | 1 | | | | | | |
| 5 | 4 | 5 | 1 | | | | | | |
| 6 | 5 | 2 | 1 | | | | | | |
| 7 | 6 | 2 | 0 | | | | | | |
| 8 | 7 | 2 | 0 | | | | | | |
| 9 | 8 | 2 | 0 | | | | | | |
| 10 | 9 | 2 | 2 | | | | | | |
| 11 | 10 | 2 | 3 | | | | | | |
| 12 | | | | | | | | | |
| 13 | | | | | | | | | |

**Abbildung 2.35:** EXCEL-Beispieldaten (DATEN_ALLE.XLSX)

Gerade bei EXCEL kann es vorkommen, dass die EXCEL-Variablennamen Leerzeichen enthalten. Das ist allerdings insofern nicht weiter dramatisch, als in diesem Fall SPSS die Leerzeichen in Unterstriche umwandelt. Sie müssen sich nur dessen bewusst sein, wenn Sie die betreffenden Variablen in SPSS ansprechen möchten.

Der Import dieser Daten in SPSS wird durch den in Abbildung 2.36 dargestellten Bildschirm in die Wege geleitet. Diesen Bildschirm

erhalten Sie, wenn Sie zuvor im SPSS-Menü entweder DATEI|ÖFFNEN oder DATEI|DATEN IMPORTIEREN|EXCEL… angeklickt haben.

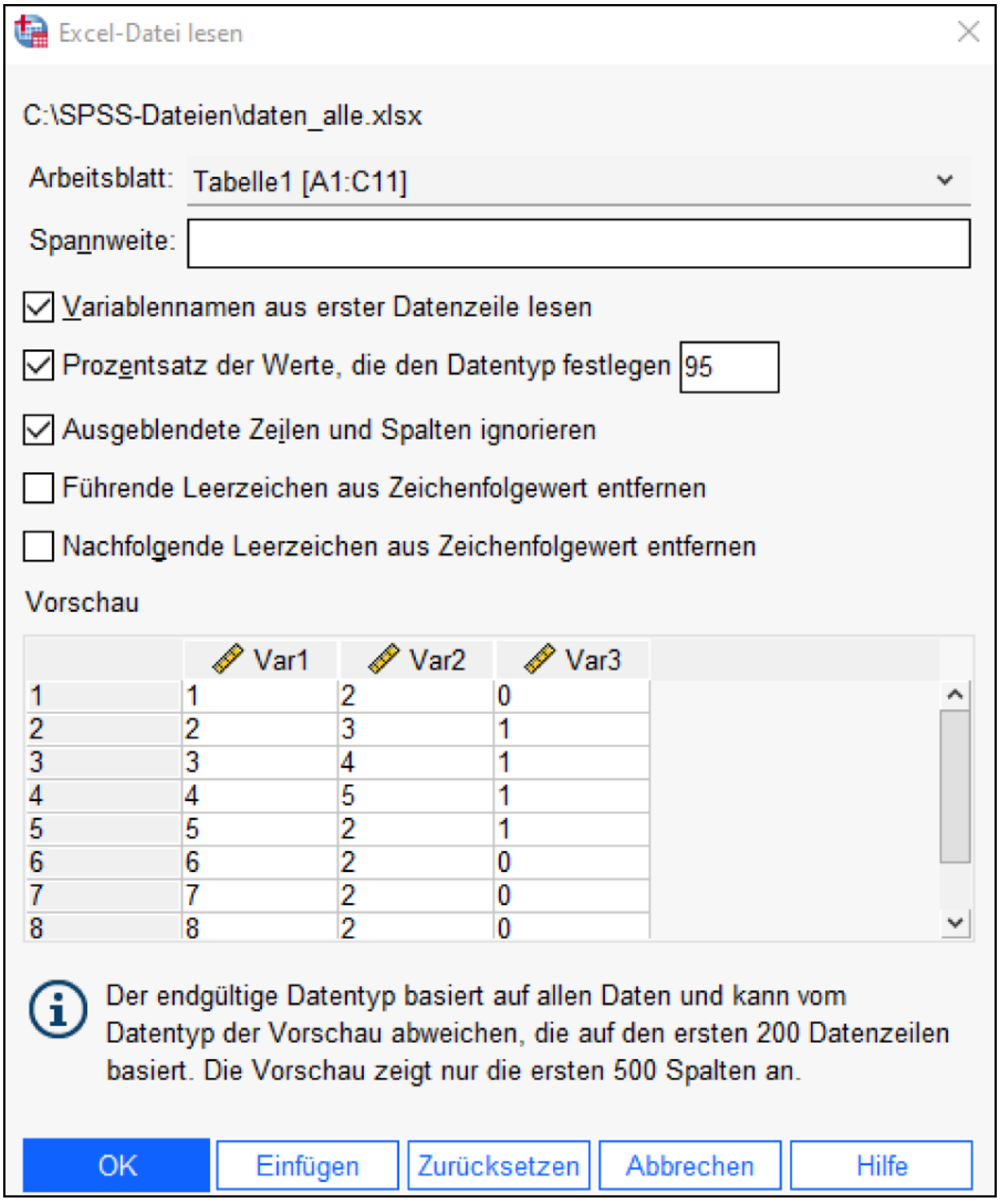
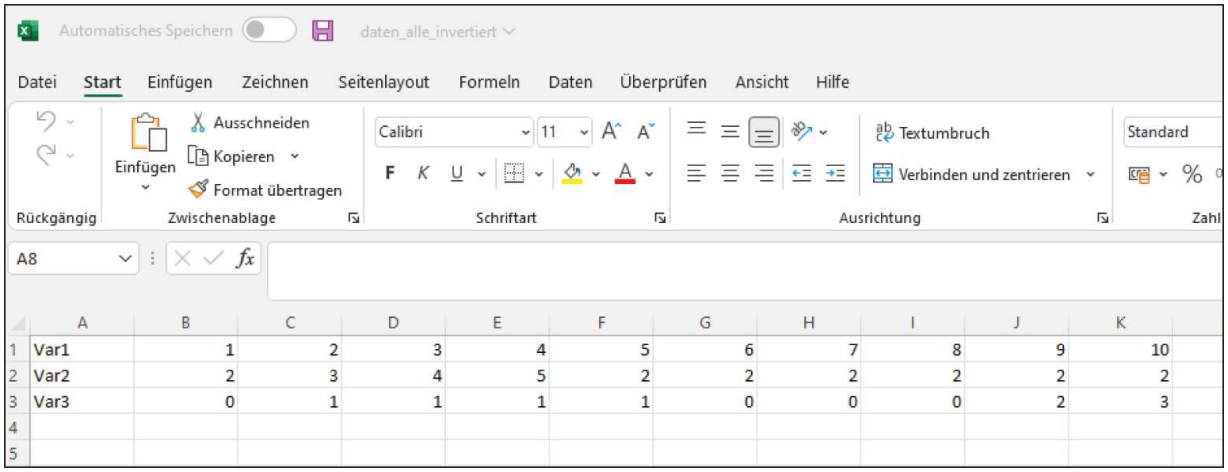



**Abbildung 2.36:** Bildschirm zum Einlesen von EXCEL-Daten in SPSS (DATEN_ALLE.XLSX)

Wollen Sie **CSV-Daten** in SPSS einlesen, geht dies grundsätzlich genauso vonstatten, wie oben in diesem Kapitel bei den ASCII-Daten beschrieben, wobei als Trennzeichen hier das Semikolon angenommen wird (ursprünglich bedeutete CSV: Comma-Seperated Values).

Schließlich können Sie auch **STATA-Datendateien** (Extension *.DTA*) oder zum Beispiel auch **SAS-Datendateien** (Extension unter anderem *.SAS7BDAT* ) direkt in SPSS einlesen. Dies bewerkstelligen Sie am besten über DATEI|DATEN IMPORTIEREN und dann über STATA beziehungsweise SAS.

Manchmal kommt es gerade bei EXCEL-Dateien vor, dass dort – wie in Abbildung 2.37 – die Fälle in den Spalten und die Variablen in den Zeilen stehen. Die zugrunde liegende EXCEL-Datei heißt *DATEN_ALLE_INVERTIERT.XLSX*.




| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Var1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 2 | Var2 | 2 | 3 | 4 | 5 | 2 | 2 | 2 | 2 | 2 | 2 |
| 3 | Var3 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 2 | 3 |
| 4 | | | | | | | | | | | |
| 5 | | | | | | | | | | | |

**Abbildung 2.37:** EXCEL-Datei mit Fällen in den Spalten und Variablen in den Zeilen (DATEN_ALLE_INVERTIERT.XLSX)

Dies mag ja SPSS eigentlich gar nicht: SPSS setzt standardmäßig voraus, dass die Fälle in den Zeilen und die Variablen in den Spalten stehen, und so kommt nach dem Import der in Abbildung 2.37 angegebenen EXCEL-Datei das SPSS-Bild der Abbildung 2.38 zustande.

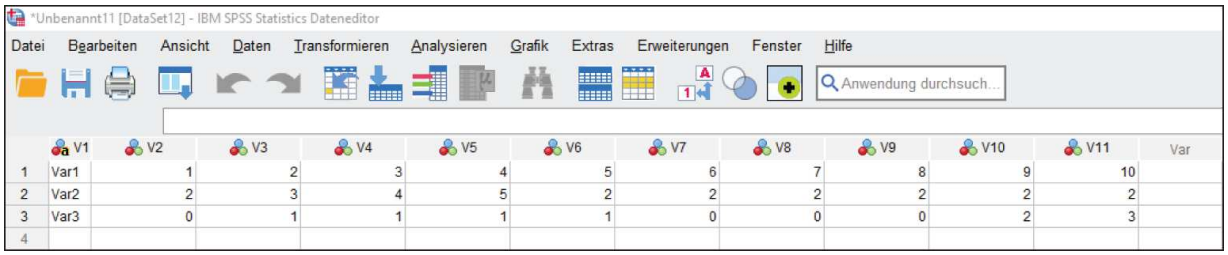



**Abbildung 2.38:** SPSS-Anzeige für EXCEL-Datei DATEN_ALLE_INVERTIERT.XLSX mit Fällen in den Spalten und Variablen in den Zeilen

Das ist keine sonderlich schöne Ansicht – Sie hätten sicherlich gerne genau die umgekehrte Ansicht; das heißt: Spalten und Zeilen müssen vertauscht werden. Zum Glück bietet SPSS hier eine Lösung an, nämlich das sogenannte **Transponieren**. Klicken Sie also bitte auf DATEN|TRANSPONIEREN. Anschließend müssen Sie zehn der vermeintlichen elf Variablen in die Rubrik VARIABLE(N) und darüber hinaus die eine vermeintliche Variable *v1*, in der die Namen der drei »richtigen« Variablen stehen, in die Rubrik NAMENSVARIABLE verschieben, wie in Abbildung 2.39 geschehen.

Wenn Sie anschließend auf OK klicken, erhalten Sie fast schon das erhoffte Bild mit unseren drei vertrauten Variablen und den zugehörigen Ausprägungen (siehe Abbildung 2.40).

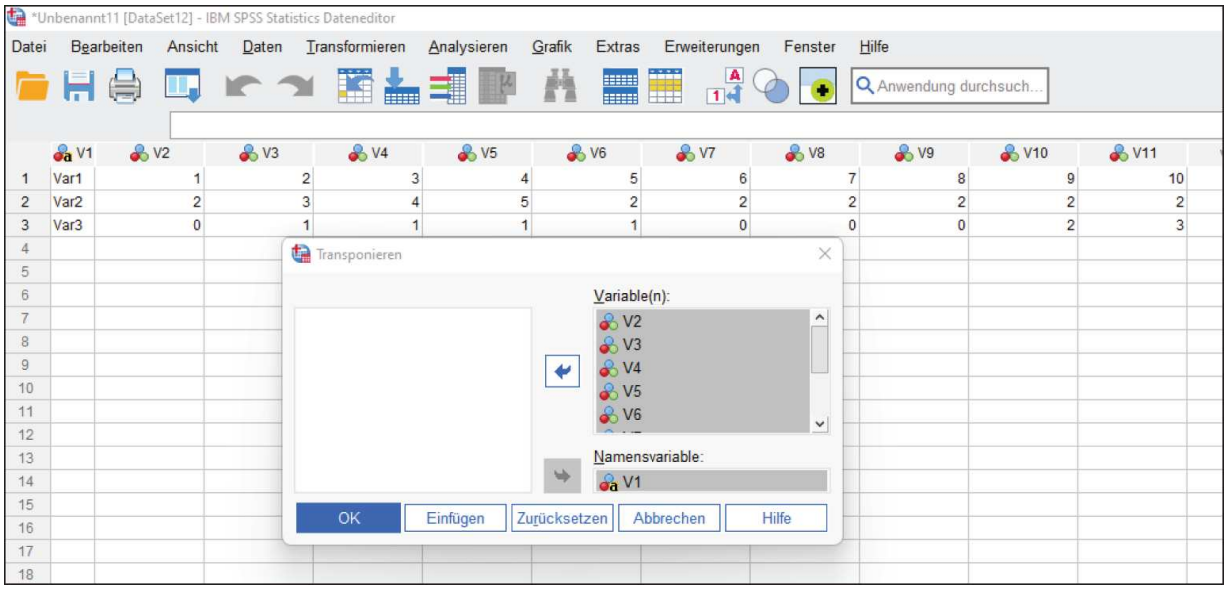
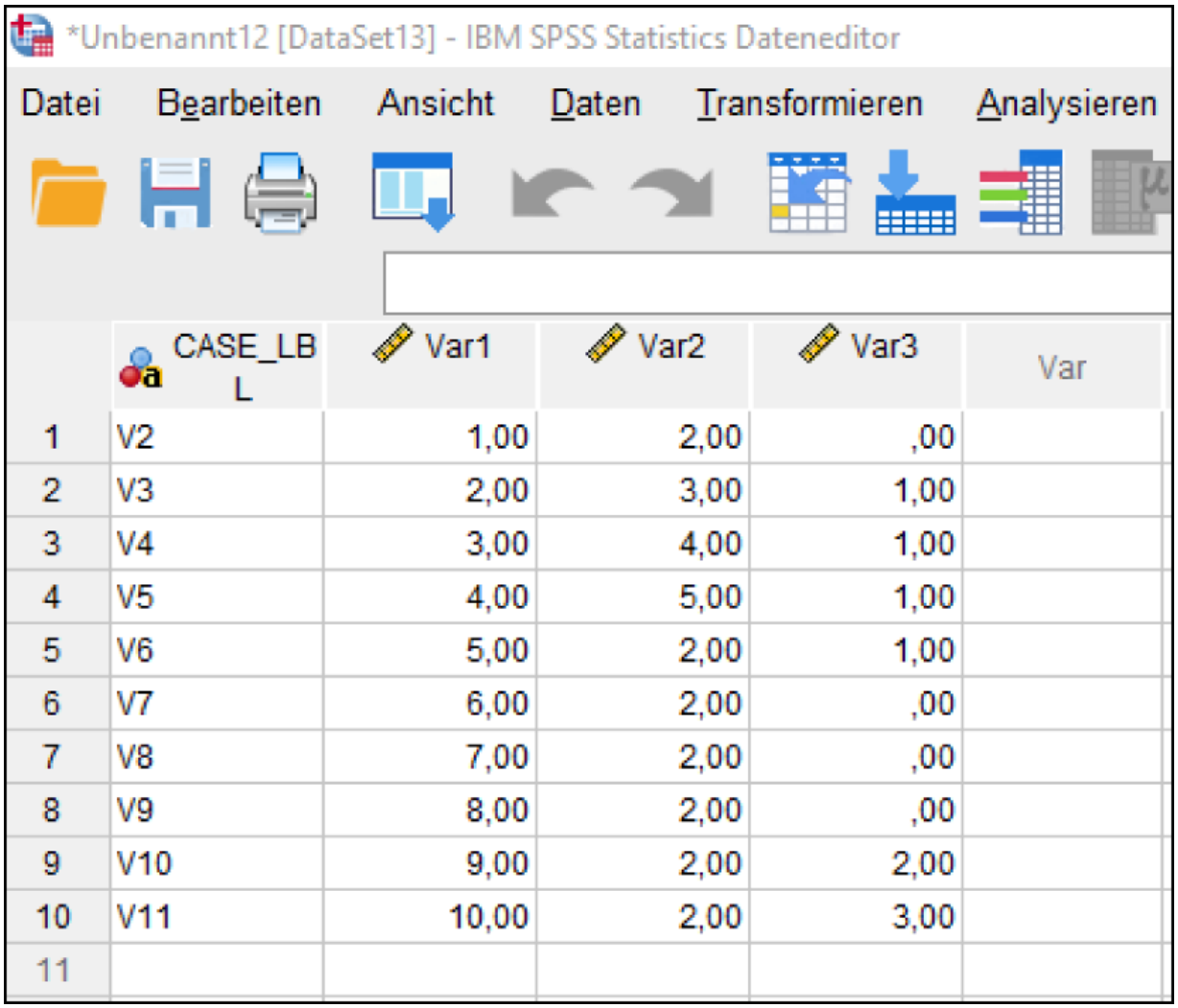



**Abbildung 2.39:** Zwischenschritt beim Transponieren von SPSS-Daten



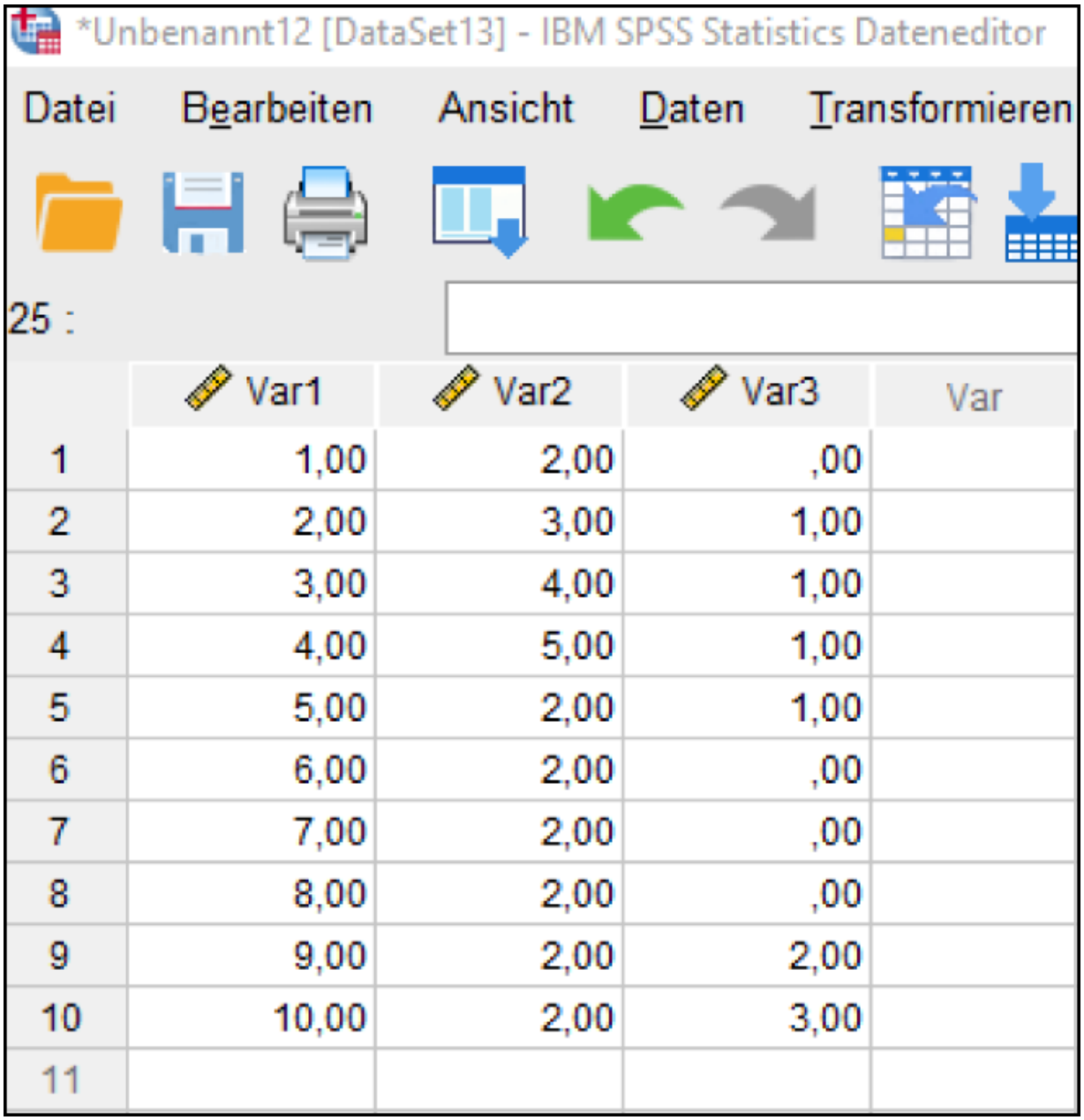

| | CASE_LBL | Var1 | Var2 | Var3 | Var |
|---|---|---|---|---|---|
| 1 | V2 | 1,00 | 2,00 | ,00 | |
| 2 | V3 | 2,00 | 3,00 | 1,00 | |
| 3 | V4 | 3,00 | 4,00 | 1,00 | |
| 4 | V5 | 4,00 | 5,00 | 1,00 | |
| 5 | V6 | 5,00 | 2,00 | 1,00 | |
| 6 | V7 | 6,00 | 2,00 | ,00 | |
| 7 | V8 | 7,00 | 2,00 | ,00 | |
| 8 | V9 | 8,00 | 2,00 | ,00 | |
| 9 | V10 | 9,00 | 2,00 | 2,00 | |
| 10 | V11 | 10,00 | 2,00 | 3,00 | |
| 11 | | | | | |

**Abbildung 2.40:** Transponierte SPSS-Daten

An der Darstellung von Abbildung 2.40 stört nur noch die erste Spalte, die mit *case_lbl* überschrieben ist und die vormaligen »falschen« Variablenbezeichnungen *v2* bis *v11* enthält, was Sie aber ja nicht sonderlich zu interessieren braucht. Was tun? Ganz einfach: Die genannte Spalte manuell löschen, und Sie erhalten das gewünschte Bild (siehe Abbildung 2.41) – geschafft!!!

Nachdem Sie die verschiedenen *TXT*-, *XLS(X)*-, *CSV*-, *DTA*- oder *SAS7BDAT*-Files erfolgreich eingelesen haben, vergessen Sie bitte nicht, diese als *SAV*-Datenfile abzuspeichern, damit Sie die betreffenden Daten für weitere Anwendungen nutzen können. Das

Abspeichern in SPSS geht bekanntlich ganz einfach über DATEI|SPEICHERN (UNTER).




| | Var1 | Var2 | Var3 | Var |
|---|---|---|---|---|
| 1 | 1,00 | 2,00 | ,00 | |
| 2 | 2,00 | 3,00 | 1,00 | |
| 3 | 3,00 | 4,00 | 1,00 | |
| 4 | 4,00 | 5,00 | 1,00 | |
| 5 | 5,00 | 2,00 | 1,00 | |
| 6 | 6,00 | 2,00 | ,00 | |
| 7 | 7,00 | 2,00 | ,00 | |
| 8 | 8,00 | 2,00 | ,00 | |
| 9 | 9,00 | 2,00 | 2,00 | |
| 10 | 10,00 | 2,00 | 3,00 | |
| 11 | | | | |

**Abbildung 2.41:** Transponierte und um die Vorspalte bereinigte SPSS-Daten

Den ganzen beschriebenen Aufwand (beim Einlesen von *TXT*-, *XLS(X)*-, *CSV*-, *DTA*- oder *SAS7BDAT*-Files) haben Sie im Übrigen selbstverständlich nicht, wenn Sie Ihre Daten direkt in SPSS eingeben (dafür haben Sie dann allerdings eine andere Form von Aufwand …)

beziehungsweise wenn Ihnen die auszuwertenden Daten unmittelbar als *SAV*-Datei zur Verfügung gestellt worden sind.

# Kapitel 3

# Hier wächst zusammen, was zusammengehört: Datendateien zusammenbringen

## IN DIESEM KAPITEL

Fälle zusammenbringen

Variablen zusammenbringen

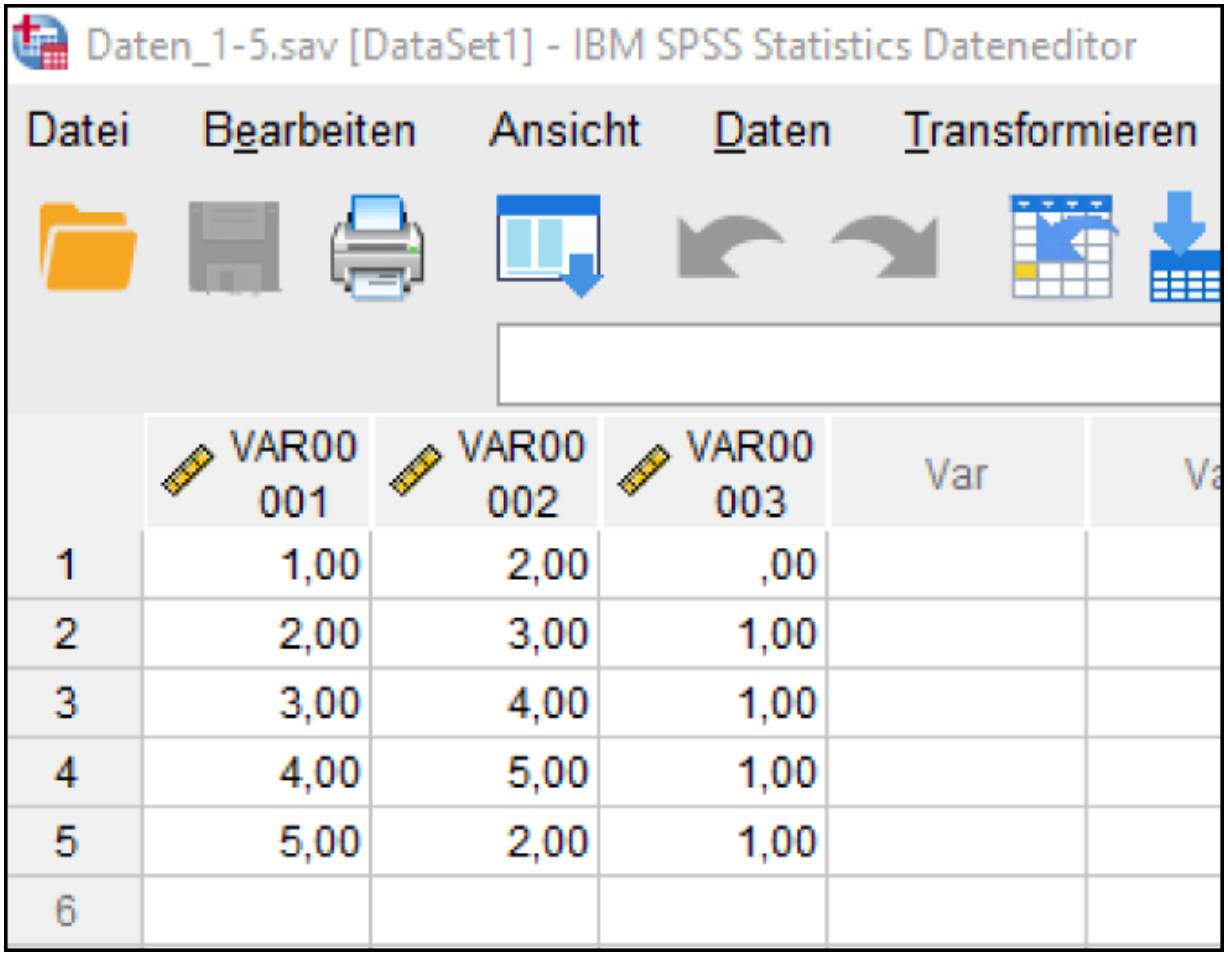
In diesem Kapitel geht es zum einen darum, mehrere Datendateien mit gleichen Variablen, aber unterschiedlichen Fällen fallweise zu verknüpfen. Zum anderen werden nachfolgend Datendateien mit gleicher Fall-, aber unterschiedlicher Variablenanzahl zusammengefügt.

## *Fälle zusammenbringen*

In der Praxis liegen nicht selten Datensätze mit den gleichen Variablen für unterschiedliche Fälle getrennt voneinander vor. Sie sind aber manchmal daran interessiert, diese unterschiedlichen Datensätze mit gleichen Variablen, aber unterschiedlichen **Fällen zusammenzubringen**. Stellen Sie sich zum Beispiel vor, Sie hätten Daten zur sozialen Lage und zur Gesundheit der Bevölkerungen in mehreren Ländern Europas erhoben – so wie dies zum Beispiel in der sozialwissenschaftlichen Studie SHARE (= **S**urvey of **H**ealth, Ageing and **RE**tirement) der Fall ist – und wollten nun die länderspezifischen Datensätze zu einem einzigen Datensatz zusammenfügen.

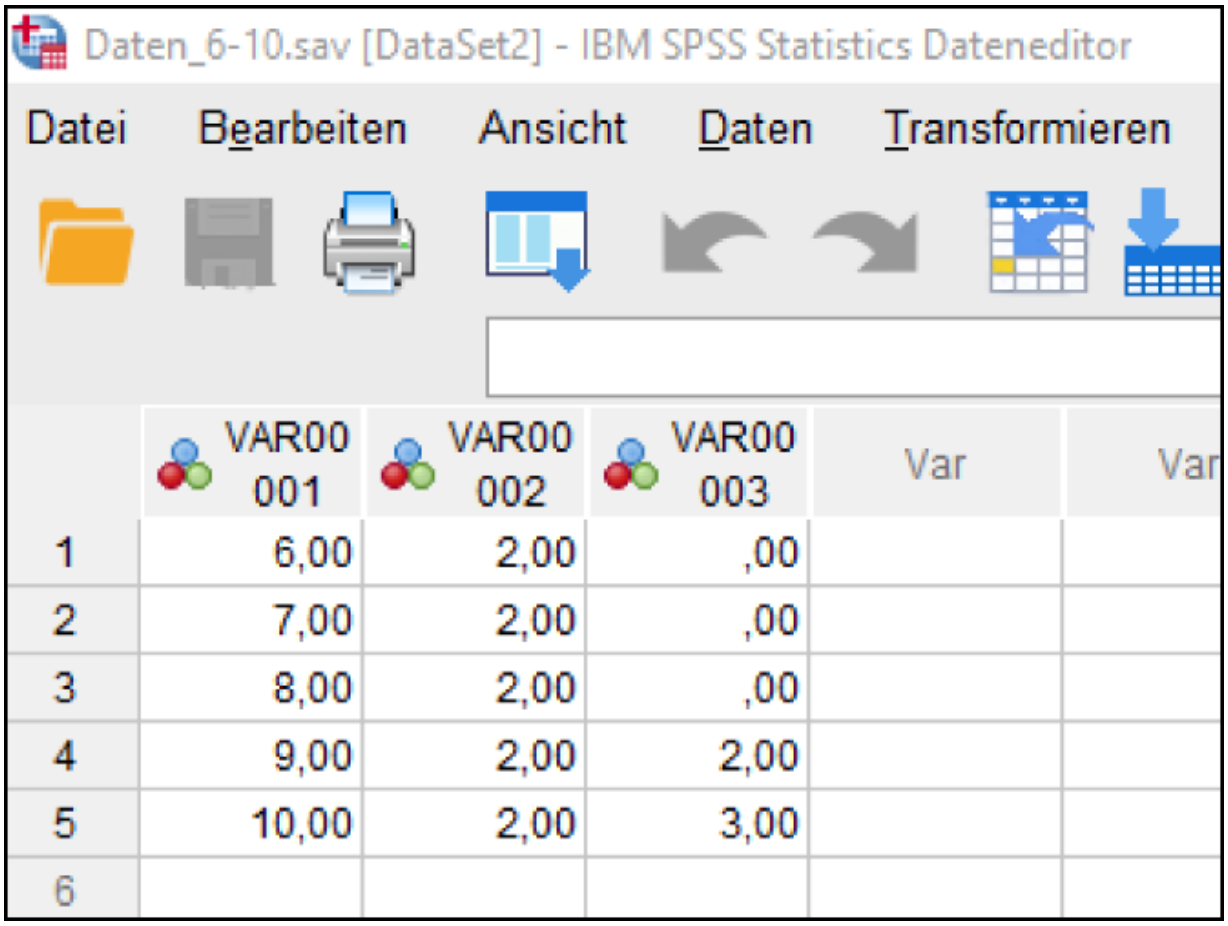
Dann stellt sich die Situation vom Grundsatz her wie in unserem folgenden Beispiel dar. Es existieren zwei Dateien mit den drei

annahmegemäß inhaltlich übereinstimmenden Variablen 1, 2 und 3. Die Fallnummer steht hierbei jeweils bei der ersten Variablen. Wie aus Abbildung 3.1 und Abbildung 3.2 hervorgeht, enthält die erste Datei die Fälle 1 bis 5 und die zweite die Fälle 6 bis 10.




| | VAR00001 | VAR00002 | VAR00003 | Var | |
|---|---|---|---|---|---|
| 1 | 1,00 | 2,00 | ,00 | | |
| 2 | 2,00 | 3,00 | 1,00 | | |
| 3 | 3,00 | 4,00 | 1,00 | | |
| 4 | 4,00 | 5,00 | 1,00 | | |
| 5 | 5,00 | 2,00 | 1,00 | | |
| 6 | | | | | |

**Abbildung 3.1:** SPSS-Datei mit den Fällen 1–5 (siehe *var00001*; *DATEN_1-5.SAV*)




| | VAR00001 | VAR00002 | VAR00003 | Var | Var |
|---|---|---|---|---|---|
| 1 | 6,00 | 2,00 | ,00 | | |
| 2 | 7,00 | 2,00 | ,00 | | |
| 3 | 8,00 | 2,00 | ,00 | | |
| 4 | 9,00 | 2,00 | 2,00 | | |
| 5 | 10,00 | 2,00 | 3,00 | | |
| 6 | | | | | |

**Abbildung 3.2:** SPSS-Datei mit den Fällen 6–10 (siehe *var00001*; *DATEN_6-10.SAV*)

Diese beiden Dateien sollen nunmehr zusammengefügt werden.

Wichtig ist natürlich beim vorstehenden Beispiel, dass die Variablen 1 bis 3 jeweils die gleichen Inhalte haben. Würde im Datensatz aus Abbildung 3.1 zum Beispiel *var00002* die Anzahl der gekauften Äpfel und im Datensatz aus Abbildung 3.2 hingegen die Anzahl der verkauften Birnen kennzeichnen, würde SPSS dennoch unter *var00002* die Fälle 1 bis 5 mit den Fällen 6 bis 10 zusammenbringen, obgleich hier also Äpfel mit Birnen verglichen beziehungsweise miteinander kombiniert würden.

Damit die Fälle in der zusammengefügten Datei am Ende in der Rangfolge 1–10 stehen, öffnen Sie bitte zunächst die Datei mit den

Fällen 1–5 (*DATEN_1-5.SAV*) und hängen an diese die Datei mit den Fällen 6–10 (*DATEN_6-10.SAV*).

In Abbildung 3.3 ist dargestellt, wie an die geöffnete Datei *DATEN_1-5.SAV* die andere Datei angehängt wird, und zwar über DATEN|DATEIEN ZUSAMMENFÜGEN|FÄLLE HINZUFÜGEN.

**Abbildung 3.3:** Fälle hinzufügen (1)

Es werden nunmehr alle drei Variablen in ein neues Datenset überführt (siehe Abbildung 3.4), und Sie erhalten am Ende die gewünschte zusammengefügte Datei (siehe Abbildung 3.5).

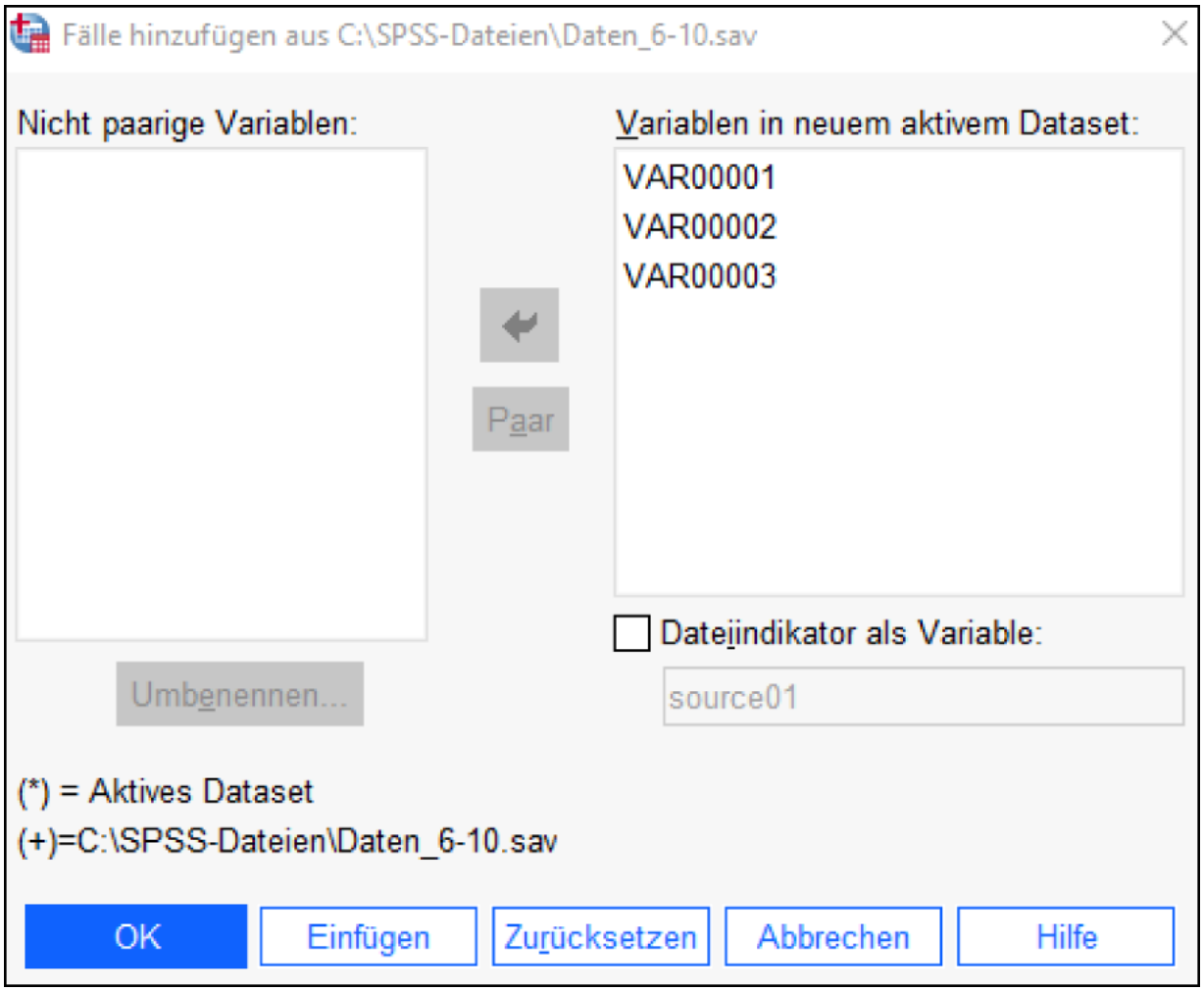
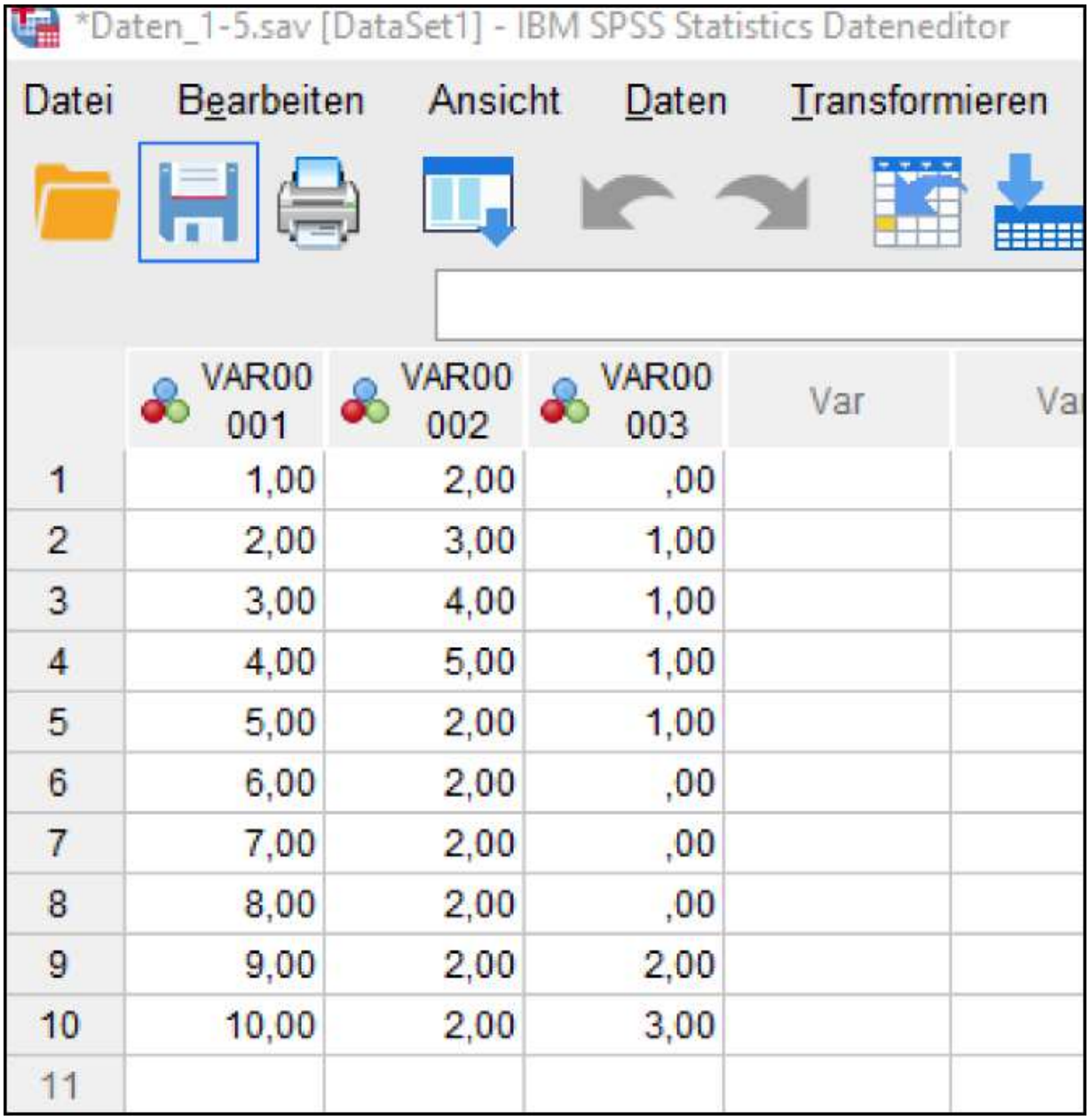



**Abbildung 3.4:** Fälle hinzufügen (2)



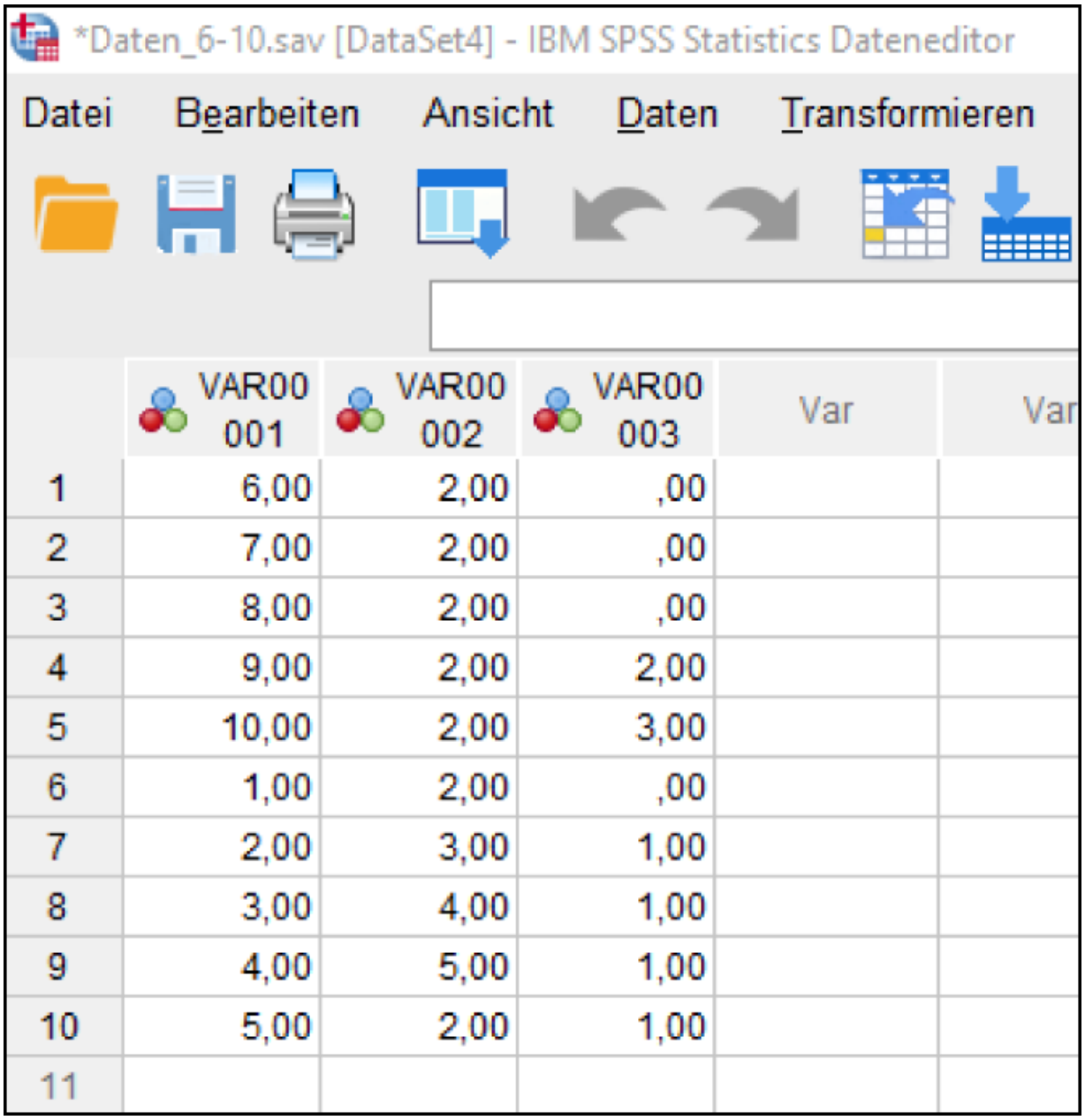

| | VAR00001 | VAR00002 | VAR00003 | Var | Va |
|---|---|---|---|---|---|
| 1 | 1,00 | 2,00 | ,00 | | |
| 2 | 2,00 | 3,00 | 1,00 | | |
| 3 | 3,00 | 4,00 | 1,00 | | |
| 4 | 4,00 | 5,00 | 1,00 | | |
| 5 | 5,00 | 2,00 | 1,00 | | |
| 6 | 6,00 | 2,00 | ,00 | | |
| 7 | 7,00 | 2,00 | ,00 | | |
| 8 | 8,00 | 2,00 | ,00 | | |
| 9 | 9,00 | 2,00 | 2,00 | | |
| 10 | 10,00 | 2,00 | 3,00 | | |
| 11 | | | | | |

**Abbildung 3.5:** Nach Fällen zusammengeführte Datei (Variante 1)

Diese sollten Sie – möglichst unter einem neuen Namen – anschließend als SPSS-Datendatei abspeichern (wieder ganz einfach über DATEI|SPEICHERN (UNTER)). Ein Grund hierfür ist, dass Sie ansonsten Ihre alte Datendatei vom Namen her überschreiben würden und dass solcherart der ursprüngliche Dateninhalt verloren ginge. Es kann nun aber sein, dass Sie später

einmal gerade diesen ursprünglichen Dateninhalt für bestimmte Auswertungszwecke benötigen.

Hätten Sie zuerst *DATEN_6-10.SAV* geöffnet und an diese Datei *DATEN_1-5.SAV* angehängt, hätte die fallbezogene Reihenfolge – anhand der ersten Variablen – 6–10 in den ersten fünf Zeilen und dann 1–5 in den folgenden fünf Zeilen gelautet (siehe Abbildung 3.6). Wenn Sie das stört, ist es aber überhaupt kein Problem, dies zu ändern. Sie müssten einfach über das Menü DATEN|FÄLLE SORTIEREN den Datensatz nach der Variablen *var00001* aufsteigend **sortieren** lassen, wie das in Abbildung 3.7 angedeutet ist, und unmittelbar danach erhalten Sie den Datensatz in der gewünschten Anordnung der Fälle.




| | VAR00001 | VAR00002 | VAR00003 | Var | Var |
|---|---|---|---|---|---|
| 1 | 6,00 | 2,00 | ,00 | | |
| 2 | 7,00 | 2,00 | ,00 | | |
| 3 | 8,00 | 2,00 | ,00 | | |
| 4 | 9,00 | 2,00 | 2,00 | | |
| 5 | 10,00 | 2,00 | 3,00 | | |
| 6 | 1,00 | 2,00 | ,00 | | |
| 7 | 2,00 | 3,00 | 1,00 | | |
| 8 | 3,00 | 4,00 | 1,00 | | |
| 9 | 4,00 | 5,00 | 1,00 | | |
| 10 | 5,00 | 2,00 | 1,00 | | |
| 11 | | | | | |

**Abbildung 3.6:** Nach Fällen zusammengeführte Datei (Variante 2)

Es könnte übrigens der Fall eintreten, dass in den beiden von den Fällen unterschiedlichen, aber von den Variablen her gleichen Dateien die Variablen unterschiedlich angeordnet gewesen wären. Wäre zum Beispiel in der einen Datei die Anordnung *var00001*, *var00002*, *var00003* und in der anderen Datei die Anordnung *var00003*, *var00002*, *var00001* gewesen, hätte die zusammengefügte Datei dennoch so

ausgesehen wie in Abbildung 3.5 (oder auch in Abbildung 3.6). Probieren Sie es doch einfach mal aus!

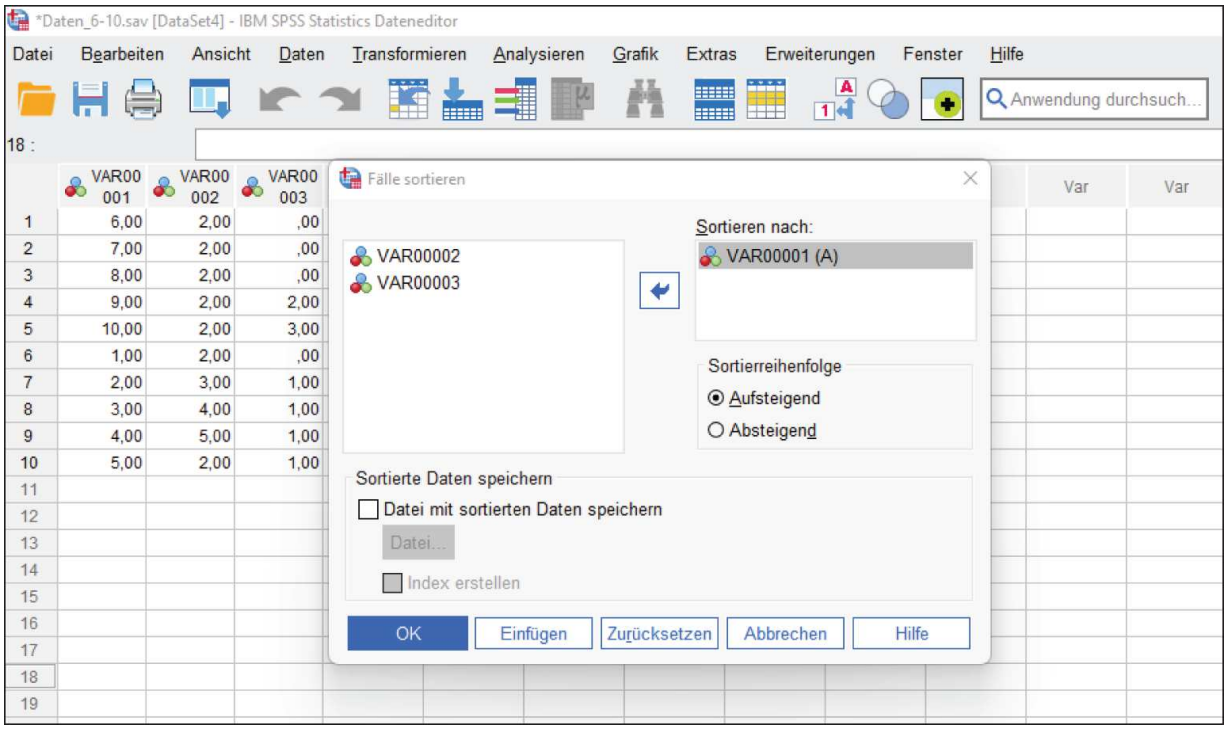


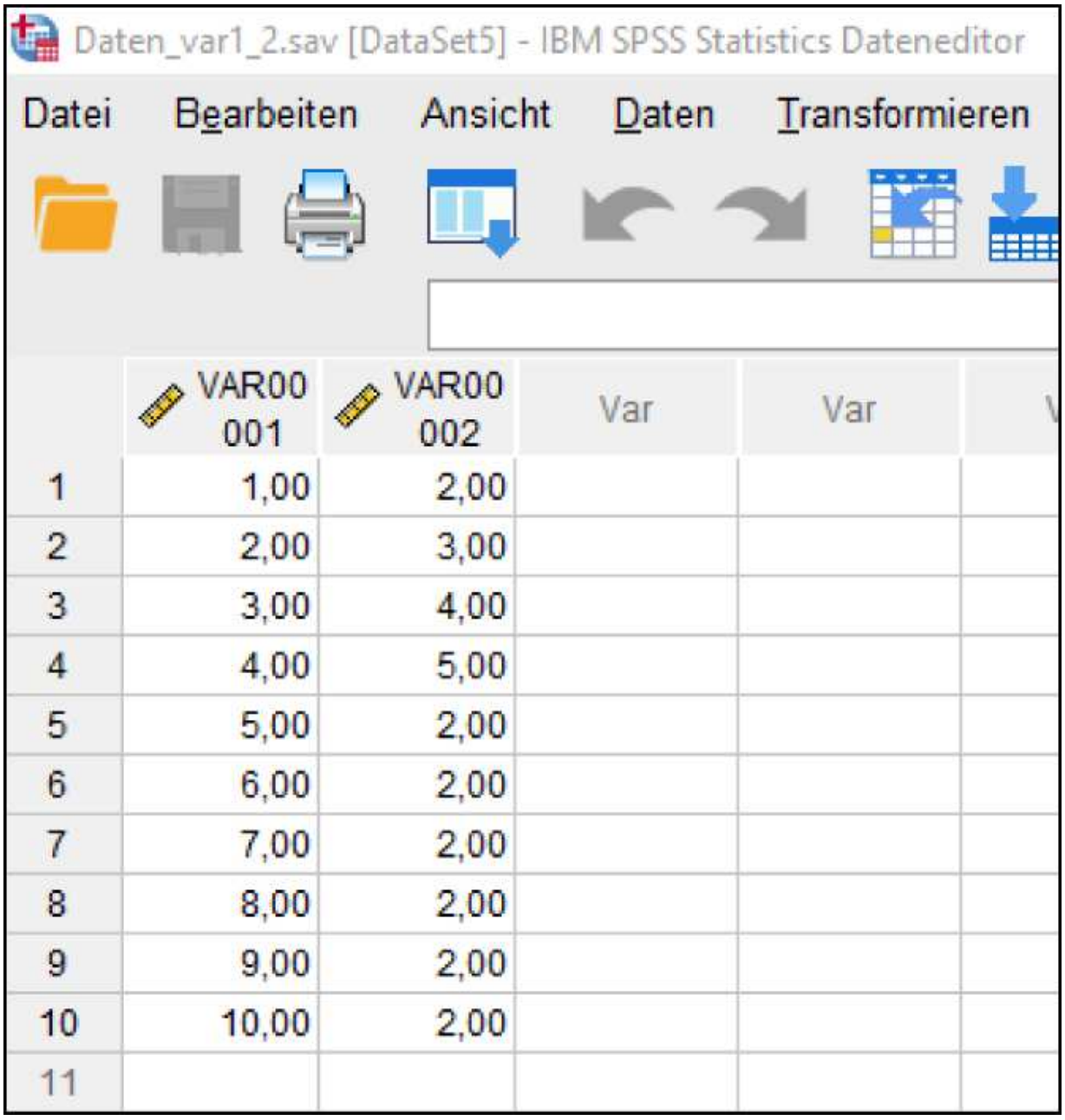

**Abbildung 3.7:** Sortieren von Fällen in einem Datenset

# *Variablen zusammenbringen*

Eventuell haben Sie wieder zwei unterschiedliche Datendateien vorliegen, die sich aber jetzt nicht nach den Fällen, sondern nach den Variablen unterscheiden. Ihre Aufgabe besteht nunmehr darin, die nach Fällen gleichen Dateien hinsichtlich der unterschiedlichen **Variablen zusammenzufügen**.

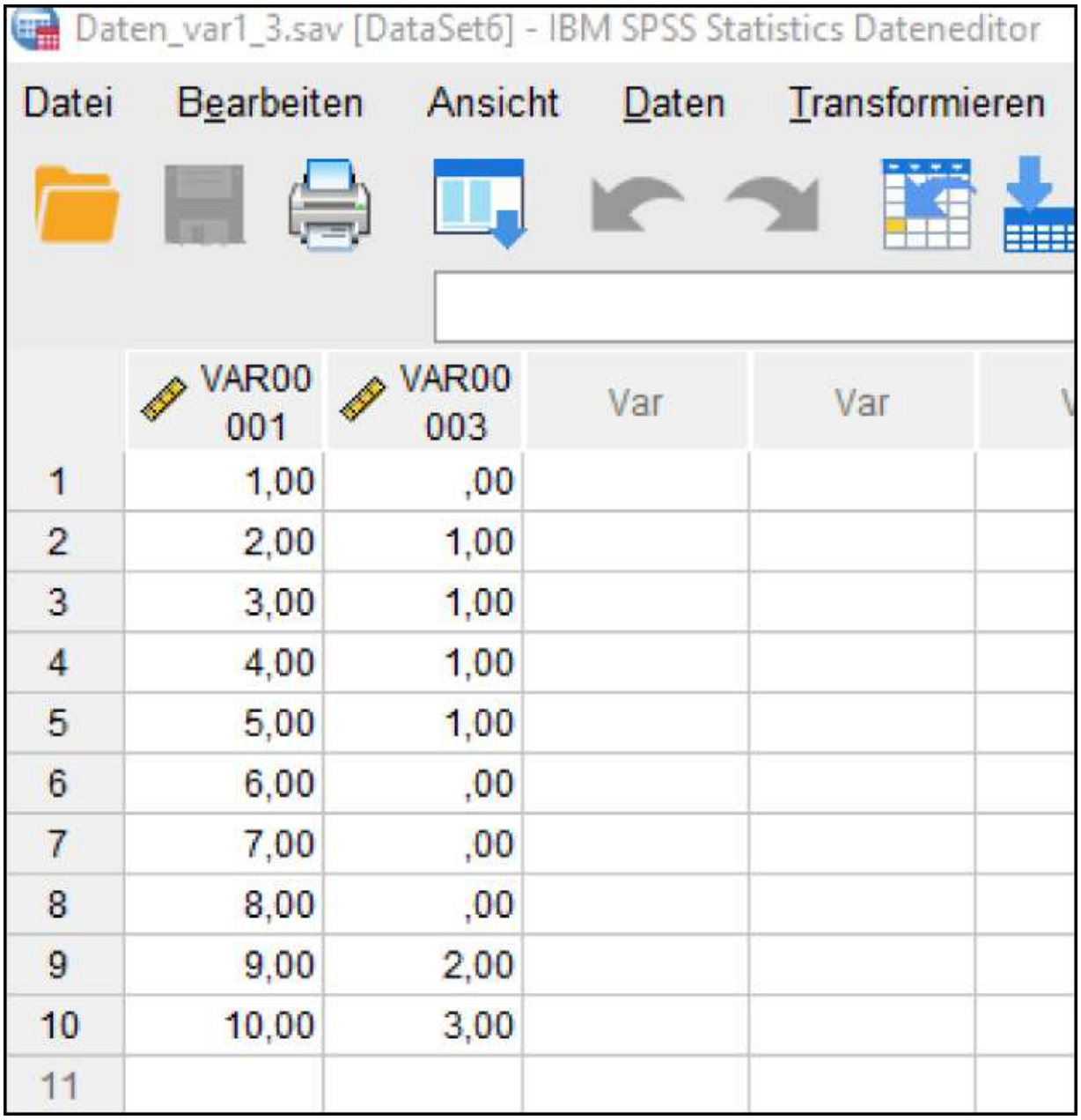
Um diesen Fall zu illustrieren, werden nachfolgend die Dateien *DATEN_VAR1_2.SAV* und *DATEN_VAR1_3.SAV* betrachtet. Während die erstgenannte Datei die Variablen 1 und 2 beinhaltet, finden sich in der zweitgenannten Datei die Variablen 1 und 3 (siehe Abbildung 3.8 und Abbildung 3.9). Als **Schlüsselvariable** soll *var00001* fungieren. Bei der Schlüsselvariablen handelt es sich um eine Variable, die in beiden Datensätzen mit denselben Ausprägungen vorkommt. Im vorliegenden Fall enthält *var00001* jeweils die Ausprägungen 1,00 bis 10,00. Man

könnte die Schlüsselvariable auch als »Verknüpfungsvariable« bezeichnen. Im nachfolgend behandelten Datensatz spiegelt die Schlüsselvariable *var00001* einfach die jeweilige Fallnummer wider.




| | VAR00001 | VAR00002 | Var | Var |
|---|---|---|---|---|
| 1 | 1,00 | 2,00 | | |
| 2 | 2,00 | 3,00 | | |
| 3 | 3,00 | 4,00 | | |
| 4 | 4,00 | 5,00 | | |
| 5 | 5,00 | 2,00 | | |
| 6 | 6,00 | 2,00 | | |
| 7 | 7,00 | 2,00 | | |
| 8 | 8,00 | 2,00 | | |
| 9 | 9,00 | 2,00 | | |
| 10 | 10,00 | 2,00 | | |
| 11 | | | | |

**Abbildung 3.8:** SPSS-Datei mit den Variablen 1 und 2 (*DATEN_VAR1_2.SAV*)




| | VAR00001 | VAR00003 | Var | Var | |
|---|---|---|---|---|---|
| 1 | 1,00 | ,00 | | | |
| 2 | 2,00 | 1,00 | | | |
| 3 | 3,00 | 1,00 | | | |
| 4 | 4,00 | 1,00 | | | |
| 5 | 5,00 | 1,00 | | | |
| 6 | 6,00 | ,00 | | | |
| 7 | 7,00 | ,00 | | | |
| 8 | 8,00 | ,00 | | | |
| 9 | 9,00 | 2,00 | | | |
| 10 | 10,00 | 3,00 | | | |
| 11 | | | | | |

**Abbildung 3.9:** SPSS-Datei mit den Variablen 1 und 3 (*DATEN_VAR1_3.SAV*)

Über DATEN|DATEIEN ZUSAMMENFÜGEN|VARIABLEN HINZUFÜGEN bringen Sie die beiden Dateien zusammen. In diesem Zusammenhang geht aus Abbildung 3.10 hervor, dass hier an die Datei *DATEN_VAR1_2.SAV* (mit den Variablen 1 und 2) die Datei *DATEN_VAR1_3.SAV* (mit den Variablen 1 und 3) angefügt wird.

**Abbildung 3.10:** Variablen hinzufügen (1)

Aus Abbildung 3.11 ist zu erkennen, dass *var00001* – wie erwähnt – als Schlüsselvariable fungiert. Die Zusammenführung der beiden Datendateien erfolgt ersichtlich über Anklicken der 1:1-ZUSAMMENFÜHRUNG AUF BASIS DER SCHLÜSSELWERTE und auch noch durch das Setzen eines Häkchens dahingehend, dass die Dateien vor dem Zusammenführen sortiert werden.

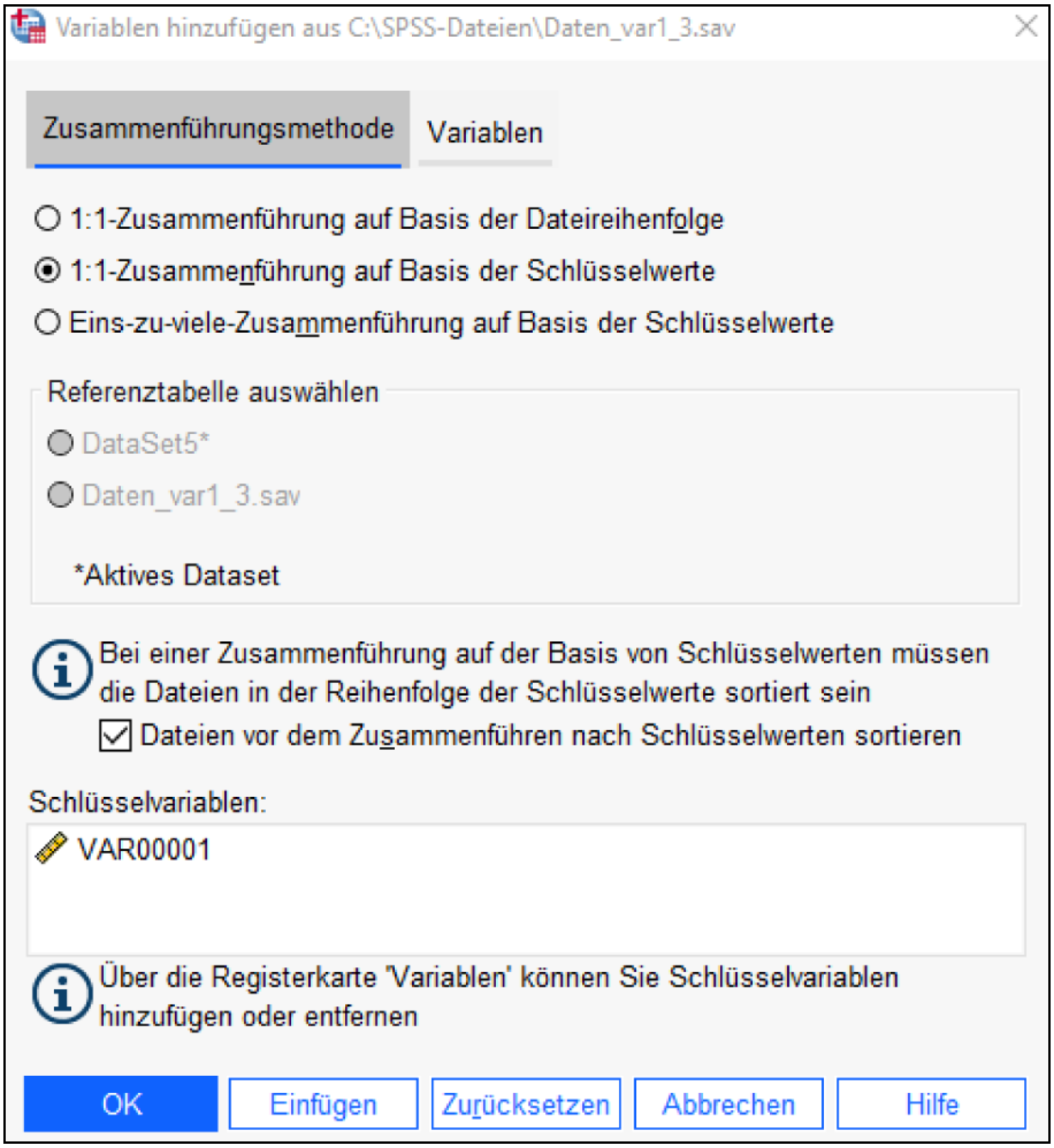
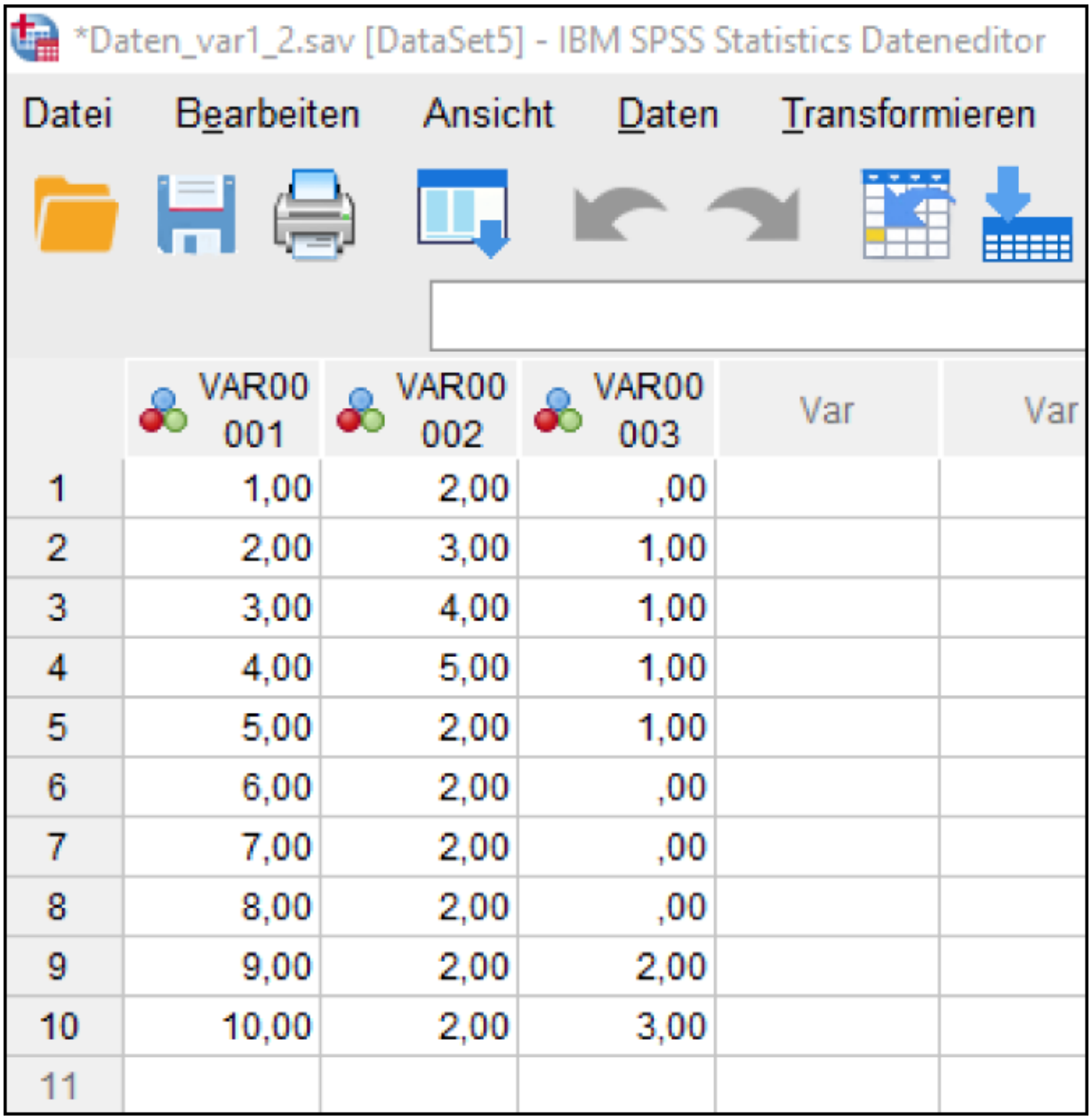



**Abbildung 3.11:** Variablen hinzufügen (2)

Als Gesamtergebnis erhalten Sie schließlich, wie gewünscht, für die zehn Fälle die aus drei Variablen bestehende neue Gesamtdatei (siehe Abbildung 3.12).

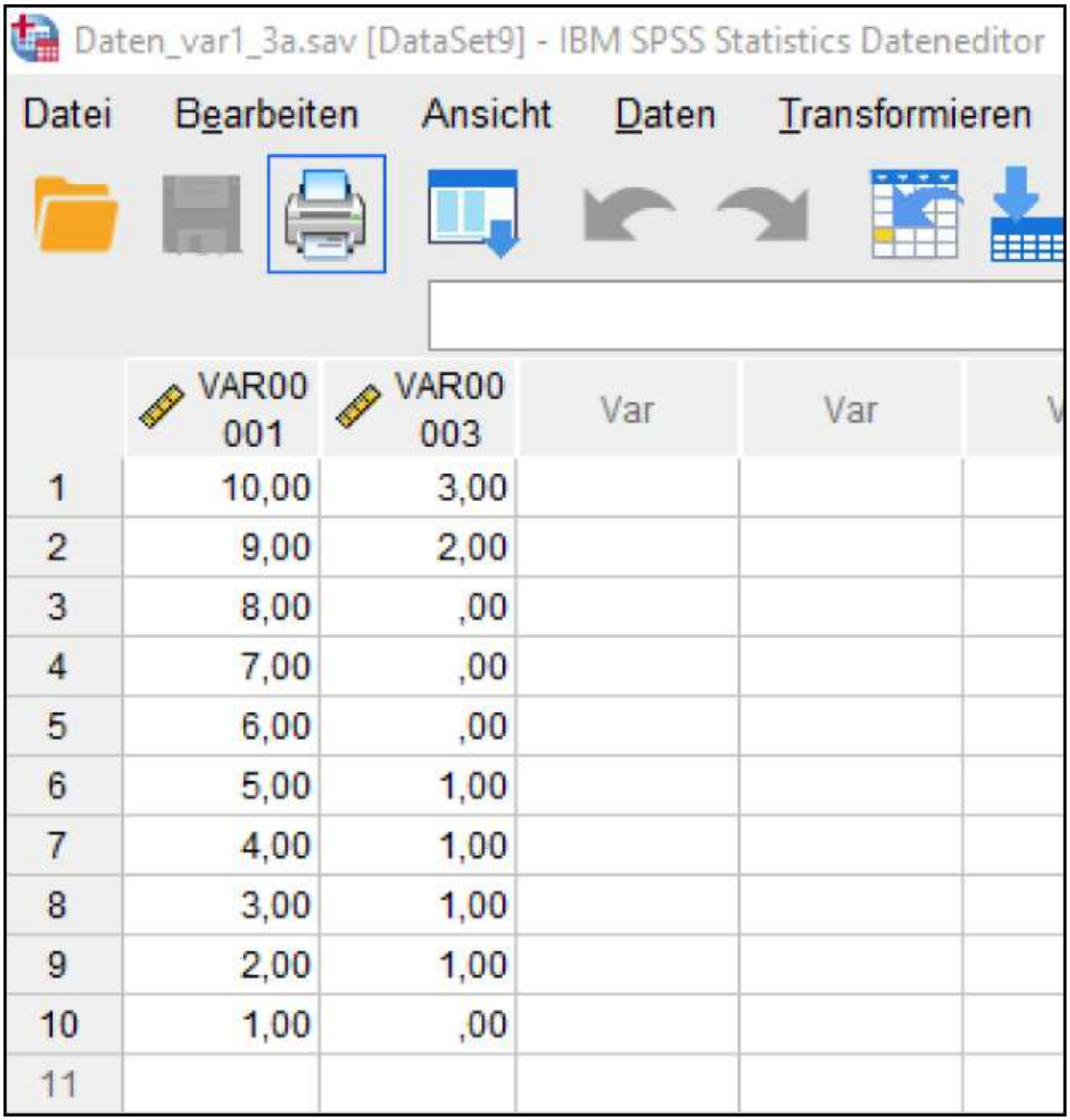



| | VAR00001 | VAR00002 | VAR00003 | Var | Var |
|---|---|---|---|---|---|
| 1 | 1,00 | 2,00 | ,00 | | |
| 2 | 2,00 | 3,00 | 1,00 | | |
| 3 | 3,00 | 4,00 | 1,00 | | |
| 4 | 4,00 | 5,00 | 1,00 | | |
| 5 | 5,00 | 2,00 | 1,00 | | |
| 6 | 6,00 | 2,00 | ,00 | | |
| 7 | 7,00 | 2,00 | ,00 | | |
| 8 | 8,00 | 2,00 | ,00 | | |
| 9 | 9,00 | 2,00 | 2,00 | | |
| 10 | 10,00 | 2,00 | 3,00 | | |
| 11 | | | | | |

**Abbildung 3.12:** Nach Variablen zusammengeführte Datei

Die vorstehende Zusammenführung nach Variablen über 1:1-ZUSAMMENFÜHRUNG AUF BASIS DER SCHLÜSSELWERTE funktioniert nur dann so problemlos wie eben angegeben, wenn die beiden Dateien in gleicher Weise sortiert sind: in unserem Fall bezüglich *var00001* jeweils aufsteigend von 1 bis 10.

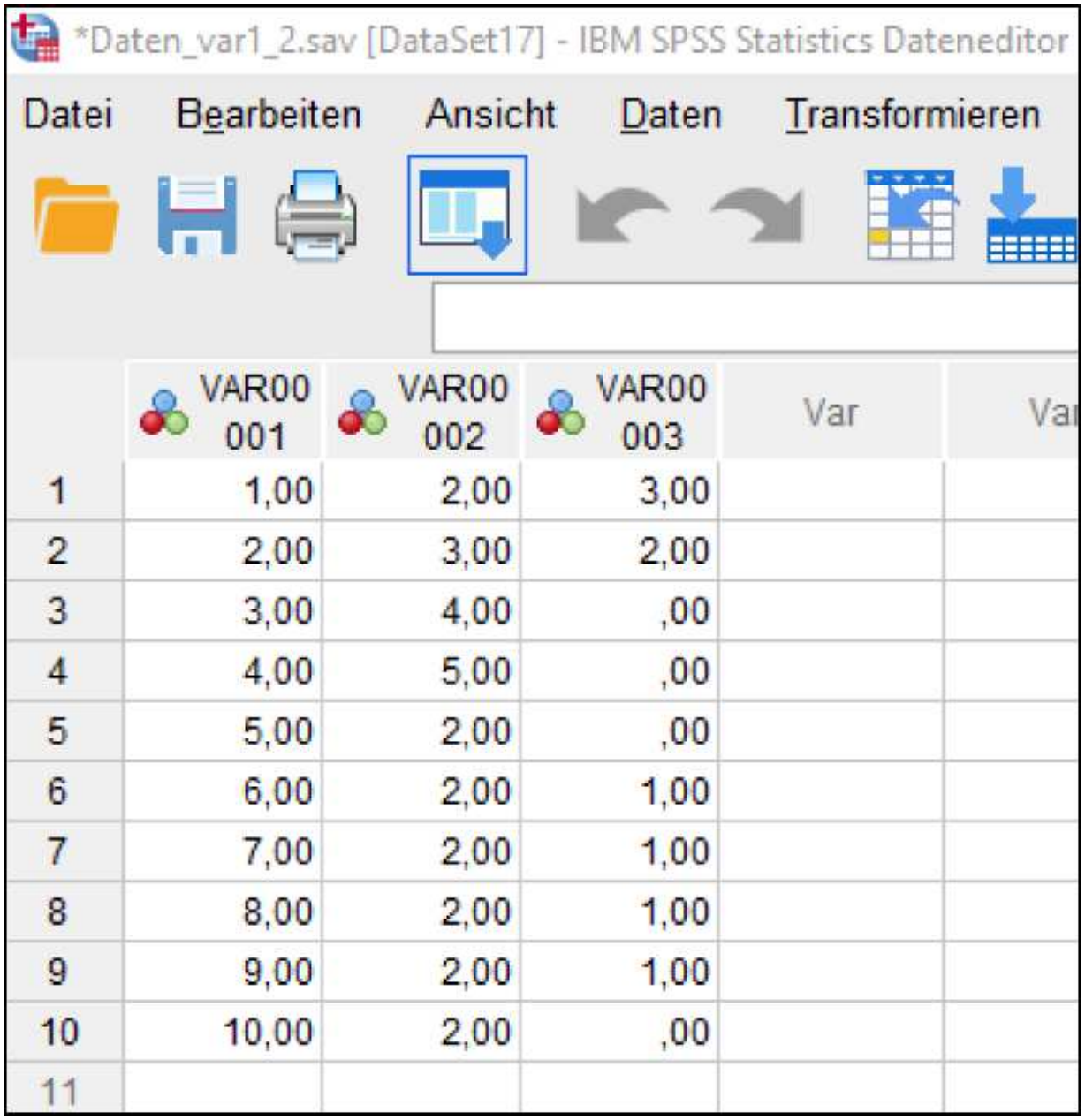
Wäre hingegen die Datei mit den Variablen 1 und 3 in absteigender Sortierung der Schlüsselvariablen gegeben (siehe Abbildung 3.13; also in Zeile 1 Ausprägung »10,00«, in Zeile 2 Ausprägung »9,00« und so weiter) und soll diese mit der aus Abbildung 3.8 bekannten, aufsteigend sortierten Datei *DATEN_VAR1_2.SAV* zusammengeführt werden (also mit Ausprägung »1,00« in Zeile 1, Ausprägung »2,00« in Zeile 2 und so weiter), würden Sie dann eine unliebsame Überraschung erleben, wenn Sie nicht auf 1:1-ZUSAMMENFÜHRUNG AUF BASIS DER SCHLÜSSELWERTE, sondern auf 1:1-ZUSAMMENFÜHRUNG AUF BASIS DER DATEIREIHENFOLGE klicken würden (siehe hierzu gegebenenfalls nochmals Abbildung 3.11). Die »falsch herum« angeordnete und in Abbildung 3.13 dargestellte Datendatei finden Sie übrigens unter *DATEN_VAR1_3A.SAV*.




| | VAR00001 | VAR00003 | Var | Var | V |
|---|---|---|---|---|---|
| 1 | 10,00 | 3,00 | | | |
| 2 | 9,00 | 2,00 | | | |
| 3 | 8,00 | ,00 | | | |
| 4 | 7,00 | ,00 | | | |
| 5 | 6,00 | ,00 | | | |
| 6 | 5,00 | 1,00 | | | |
| 7 | 4,00 | 1,00 | | | |
| 8 | 3,00 | 1,00 | | | |
| 9 | 2,00 | 1,00 | | | |
| 10 | 1,00 | ,00 | | | |
| 11 | | | | | |

**Abbildung 3.13:** Datei mit den Variablen 1 und 3, absteigend nach Variable 1 sortiert (*DATEN_VAR1_3A.SAV*)

Die genannte unliebsame Überraschung ist in Abbildung 3.14 dargestellt: Beispielsweise enthält Fall 1 (*var00001* = 1,00) die fehlerhafte *var00003*-Ausprägung 3,00 (richtig wäre: *var00003* = 0,00), weil dies in der Datei aus Abbildung 3.13 die *var00003*-Ausprägung in der ersten Zeile ist, die aber eigentlich zum zehnten Fall (*var00001* = 10,00) gehört!




| | VAR00001 | VAR00002 | VAR00003 | Var | Var |
|---|---|---|---|---|---|
| 1 | 1,00 | 2,00 | 3,00 | | |
| 2 | 2,00 | 3,00 | 2,00 | | |
| 3 | 3,00 | 4,00 | ,00 | | |
| 4 | 4,00 | 5,00 | ,00 | | |
| 5 | 5,00 | 2,00 | ,00 | | |
| 6 | 6,00 | 2,00 | 1,00 | | |
| 7 | 7,00 | 2,00 | 1,00 | | |
| 8 | 8,00 | 2,00 | 1,00 | | |
| 9 | 9,00 | 2,00 | 1,00 | | |
| 10 | 10,00 | 2,00 | ,00 | | |
| 11 | | | | | |

**Abbildung 3.14:** Fehlerhafte, nach den Variablen zusammengeführte Gesamtdatei

Im vorstehenden Fall ist also die nach *var00001* absteigend sortierte Datei vor der Zusammenführung mittels DATEN|FÄLLE SORTIEREN ebenso wie die andere Datei mit den Variablen 1 und 2 aufsteigend zu sortieren (siehe Abbildung 3.15; die neue Sortierung ist in *DATEN_VAR1_3A_AUFSTEIGEND.SAV* abgespeichert), und alles ist auch bei einer 1:1-ZUSAMMENFÜHRUNG AUF BASIS DER DATEIREIHENFOLGE in Butter (siehe Abbildung 3.16)!

Beispiele für das Zusammenführen von Variablen finden sich in der sozialwissenschaftlichen Praxis etwa beim Zusammenfügen deutscher Befragungsdaten aus der internationalen Erhebung SHARE (= **S**urvey of **H**ealth, **A**geing and **RE**tirement) einerseits und mit prozessproduzierten Daten der Deutschen Rentenversicherung, genauer ausgedrückt: aus der Rentenversicherungsstatistik, andererseits. Aus der Rentenversicherungsstatistik sind in diesem Zusammenhang unter anderem Informationen zu den Rentenanwartschaften und den soziodemografischen beziehungsweise -ökonomischen SHARE-Informationen zugespielt worden. Der neue Datensatz wird als SHARE-RV bezeichnet (nähere Informationen hierzu finden Sie im Internet beim Forschungsdatenzentrum der Rentenversicherung (FDZ-RV) unter http://fdz-rv.de; Zugriff am 20.09.2022).

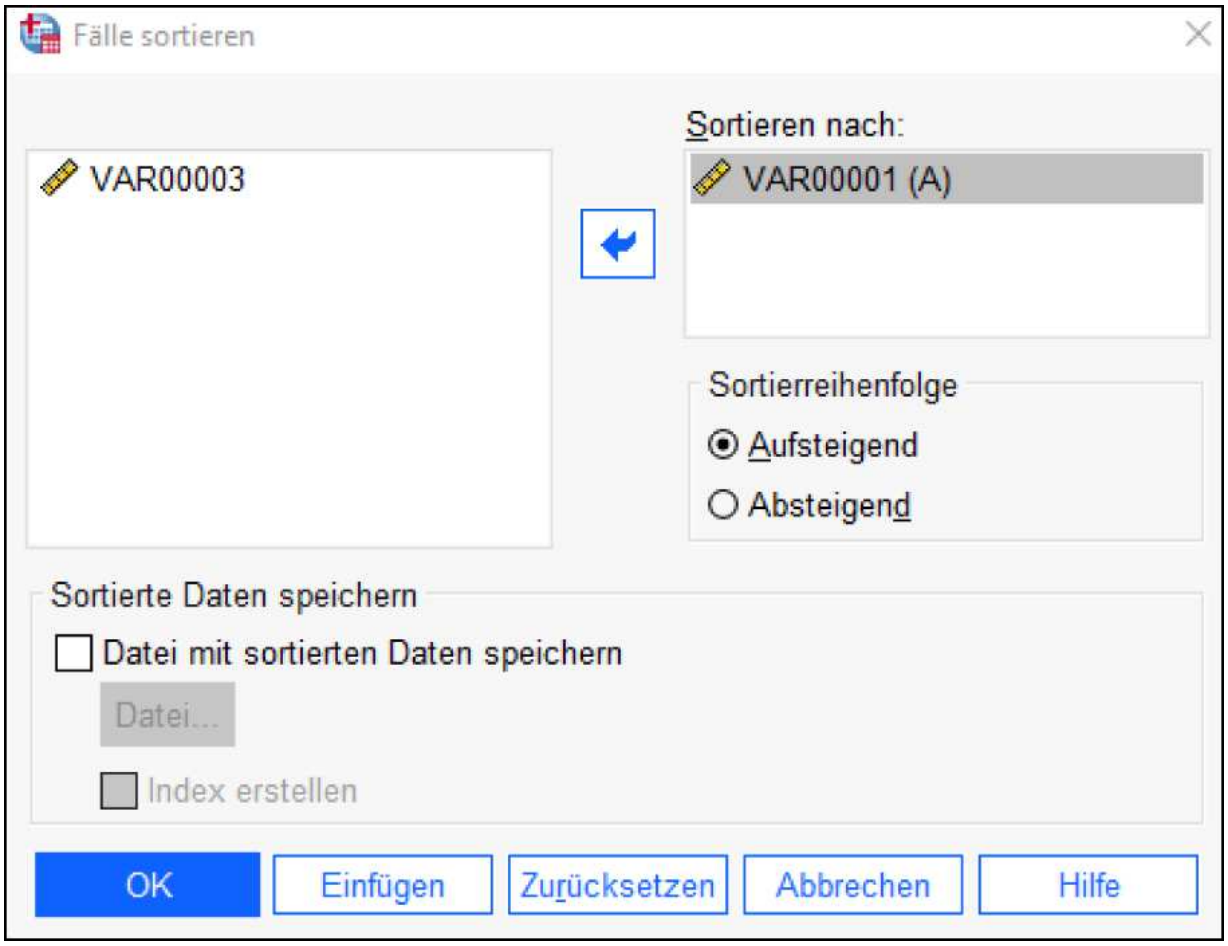
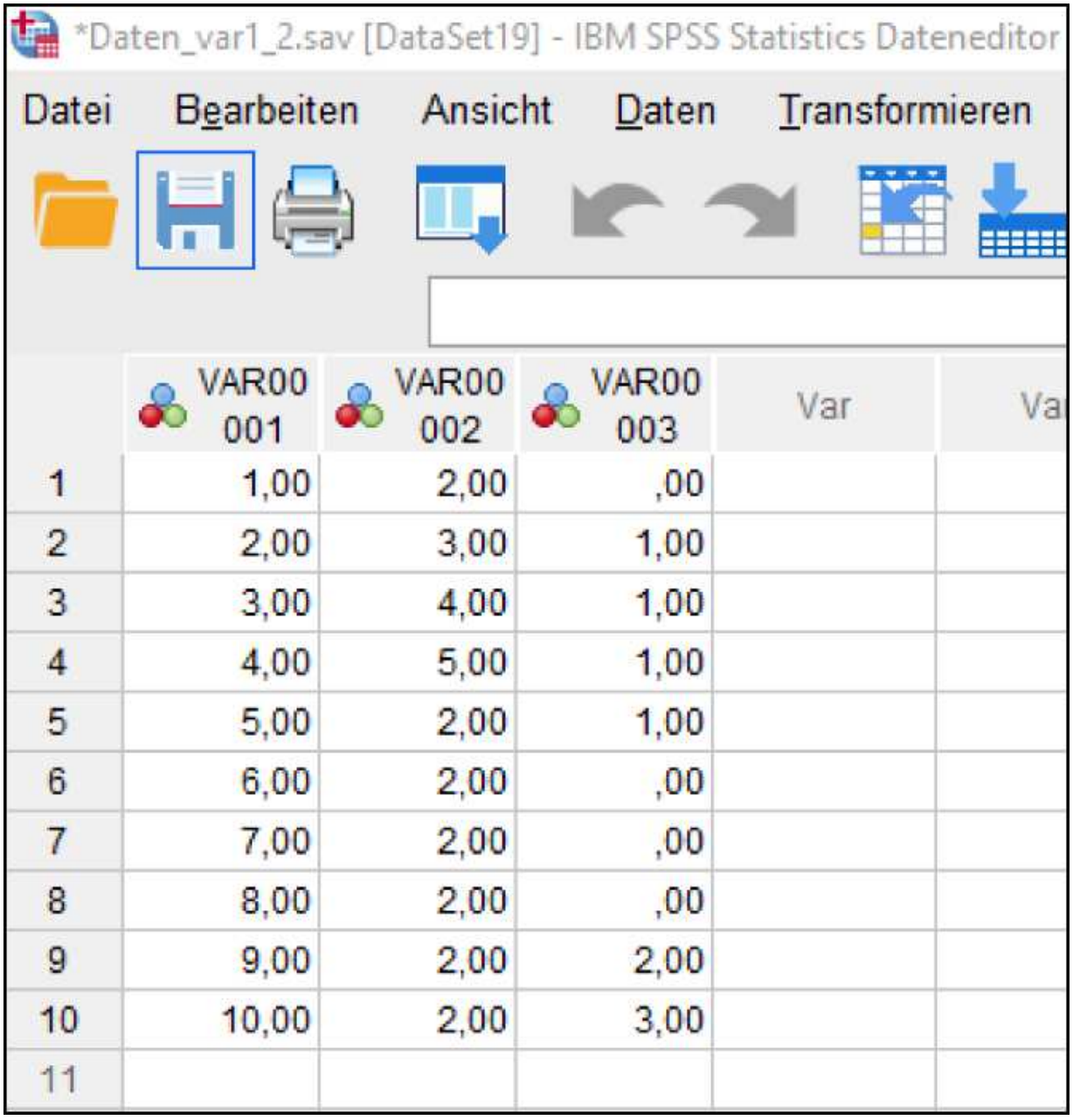


**Abbildung 3.15:** Umsortieren einer Datei




| | VAR00001 | VAR00002 | VAR00003 | Var | Va |
|---|---|---|---|---|---|
| 1 | 1,00 | 2,00 | ,00 | | |
| 2 | 2,00 | 3,00 | 1,00 | | |
| 3 | 3,00 | 4,00 | 1,00 | | |
| 4 | 4,00 | 5,00 | 1,00 | | |
| 5 | 5,00 | 2,00 | 1,00 | | |
| 6 | 6,00 | 2,00 | ,00 | | |
| 7 | 7,00 | 2,00 | ,00 | | |
| 8 | 8,00 | 2,00 | ,00 | | |
| 9 | 9,00 | 2,00 | 2,00 | | |
| 10 | 10,00 | 2,00 | 3,00 | | |
| 11 | | | | | |

**Abbildung 3.16:** Korrekte, nach den Variablen zusammengeführte Gesamtdatei – nach Umsortieren einer der beiden Dateien

Ändern Sie im Übrigen in einem Datenblatt einen Wert, erscheint oben in der obersten Menüleiste bei der Angabe des Dateinamens vor diesem ein Sternchen. Ändern Sie doch einfach mal in der zuletzt diskutierten Datendatei *DATEN_VAR1_3A_AUFSTEIGEND.SAV* in der achten Zeile bei *var00003* den Wert »0,00« in »1,00«, wie in Abbildung 3.17

dargestellt. Dann erscheint ganz oben links am Bildschirm vor dem Dateinamen das besagte Sternchen *. Das können Sie ebenfalls Abbildung 3.17 entnehmen.

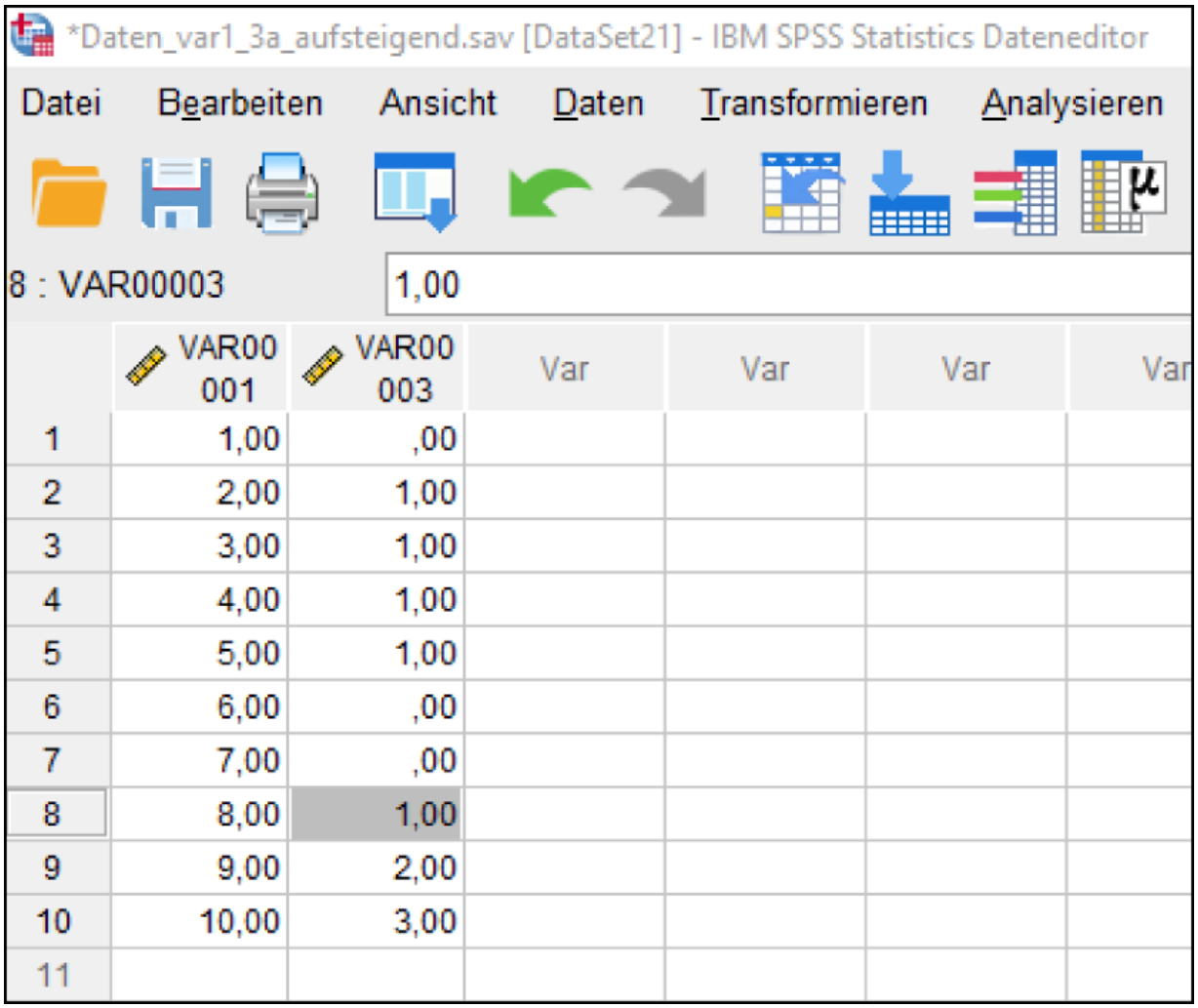



**Abbildung 3.17:** Änderung der in SPSS abgebildeten Dateikennzeichnung nach Änderung eines Datenfelds

Solche Sternchen vor dem Dateinamen sind im Übrigen in diesem Kapitel unter anderem auch schon in Abbildung 3.12, Abbildung 3.14 und Abbildung 3.16 zu erkennen gewesen. Dies liegt dort daran, dass die gezeigten Dateien jeweils das Ergebnis zusammengeführter Dateien sind, also auch *geänderte* Datensätze darstellen.

# Kapitel 4

# Datendateien in Form bringen

## IN DIESEM KAPITEL

Daten aggregieren

Daten (um-)formatieren

In Dateien Fälle bearbeiten (gewichten, filtern)

Kapitel 4 beschäftigt sich damit, Variablen von einer untergeordneten Datenebene aus auf einer übergeordneten Datenebene zu aggregieren. Zum Beispiel geht es darum, die persönlichen Einkommen einzelner Haushaltsmitglieder (untergeordnete Personenebene) per Addition zu einem Haushaltseinkommen (übergeordnete Haushaltsebene) zu aggregieren. Des Weiteren wird in Kapitel 4 gezeigt, wie Sie Variablen formatieren beziehungsweise wie Sie in Dateien Fälle bearbeiten, konkret: wie Sie sie gewichten und filtern können.

## *Daten aggregieren*

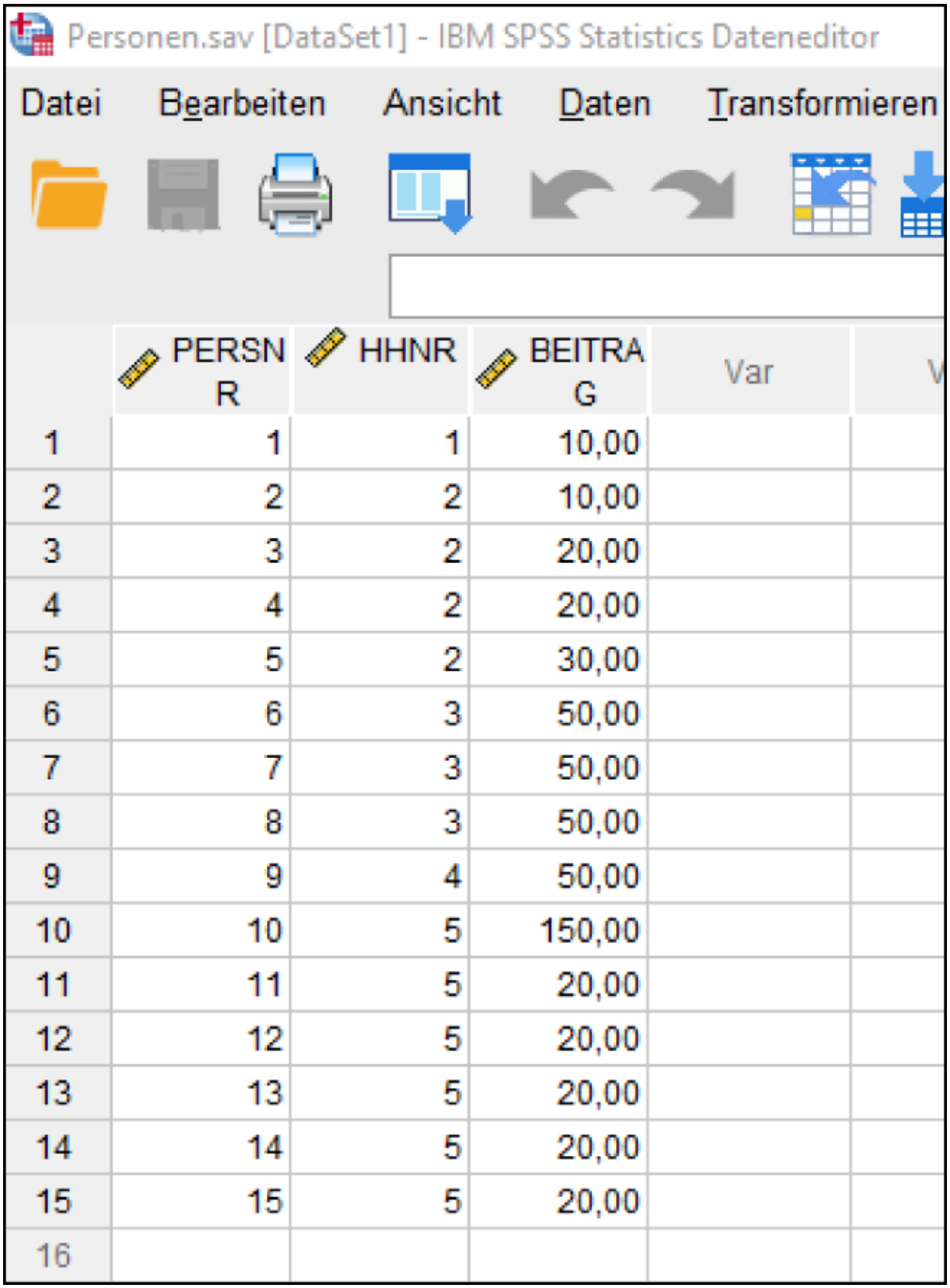
Manchmal liegen Daten vor, bei denen eine Untersuchungseinheit (zum Beispiel ein privater Haushalt) sozusagen hierarchisch in kleinere Untereinheiten (zum Beispiel in Haushaltsmitglieder) gegliedert ist. Die entsprechende Untersuchungseinheit repräsentiert die übergeordnete Datenebene, und die Untereinheiten stellen – Nomen est omen! – die untergeordnete Datenebene dar. Die Elemente einer untergeordneten Datenebene sind also die Elemente einer übergeordneten Datenebene. Es kann nun Ihr Ziel sein, die Daten der untergeordneten Datenebene zusammenzufassen, und zwar zu Einheiten der übergeordneten Datenebene, das heißt: zu **aggregieren**.

## *Aggregation mit Breakvariable*

Was hier so technisch klingt, ist eigentlich ganz einfach: Wenn Sie zum Beispiel an der Zusammenfassung von Informationen in einem Haushalt zusammenlebender Personen interessiert sind, können Sie den betreffenden Personendatensatz zu einem Haushaltsdatensatz aggregieren.

So könnten Sie daran interessiert sein, die persönlichen Einkommen der einzelnen Haushaltsmitglieder zu einem Haushaltseinkommen zusammenzufassen, das heißt: zu addieren. Vielleicht sind Sie aber auch daran interessiert, die jeweilige Haushaltsgröße mit den Personeninformationen zu ermitteln. Um all dies tun zu können, brauchen Sie eine Verknüpfungsvariable zwischen untergeordneter Ebene (Personendatensatz) und übergeordneter Ebene (Haushaltsdatensatz). Diese Variable wird in SPSS **Breakvariable** genannt und wäre in unserem Beispiel die Kennzeichnungsnummer der übergeordneten (Haushalts-)Ebene, im unten stehenden Beispielfall *hhnr*.

Das entsprechende Beispiel, das ich mit Ihnen durchgehen möchte, ist folgendes (Sie finden den Datensatz unter *PERSONEN.SAV* ): Stellen Sie sich vor, 15 Personen leisteten für eine kulturelle Veranstaltung jeweils einen freiwilligen finanziellen Obolus. Letzterer ist in der nachfolgend behandelten Personendatei unter *beitrag* verbucht. Mit *persnr* werden die 15 Personen durchnummeriert, und mit *hhnr* ist die Haushaltskennzeichnungsnummer gemeint. Beispielsweise gehören die Personen mit den Personennummern *persnr* = 2 bis *persnr* = 5 dem Haushalt 2, also jenem Haushalt mit *hhnr* = 2, an (siehe Abbildung 4.1).




| | PERSNR | HHNR | BEITRAG | Var | V |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 10,00 | | |
| 2 | 2 | 2 | 10,00 | | |
| 3 | 3 | 2 | 20,00 | | |
| 4 | 4 | 2 | 20,00 | | |
| 5 | 5 | 2 | 30,00 | | |
| 6 | 6 | 3 | 50,00 | | |
| 7 | 7 | 3 | 50,00 | | |
| 8 | 8 | 3 | 50,00 | | |
| 9 | 9 | 4 | 50,00 | | |
| 10 | 10 | 5 | 150,00 | | |
| 11 | 11 | 5 | 20,00 | | |
| 12 | 12 | 5 | 20,00 | | |
| 13 | 13 | 5 | 20,00 | | |
| 14 | 14 | 5 | 20,00 | | |
| 15 | 15 | 5 | 20,00 | | |
| 16 | | | | | |

**Abbildung 4.1:** Beispiels-SPSS-Personendatendatei (*PERSONEN.SAV*)

Wie also aus Abbildung 4.1 ersichtlich, leben die 15 Personen (mit den Personennummern *persnr* = 1 bis 15) in fünf Haushalten, da die Haushaltsnummerierung in Form der Variablen *hhnr* von 1 bis 5 läuft. Sie können die in der Variablen *beitrag* abgespeicherten finanziellen Beiträge der einzelnen Personen in einem Haushalt durch simple Addition ermitteln. Für Haushalt 1 (*hhnr* = 1) kommen Sie auf 10,00 Euro (nur eine Person), für Haushalt 2 (*hhnr* = 2) auf 80,00 Euro (=10,00 Euro + 2·20,00 Euro + 30,00 Euro) und so weiter.

In diesem Beispiel können Sie die Aggregation leicht manuell durchführen. Bei mehreren Tausend Datensätzen ist aber die Datenverarbeitung via SPSS deutlich angenehmer. Wie diese Aggregation prinzipiell in SPSS vonstattengeht, illustriere ich Ihnen anhand des Beispiels aus Abbildung 4.1.

Nachdem mit DATEN|AGGREGIEREN in SPSS der Aggregationsvorgang gestartet worden ist (siehe Abbildung 4.2), soll zum einen die Summe aus den individuellen Beiträgen pro Haushalt berechnet und zum anderen die jeweilige Haushaltsgröße ermittelt werden. Hierzu wählen Sie im in Abbildung 4.3 dargestellten Fenster aus der bei FUNKTIONEN angebotenen Vielfalt mathematischer Funktionen bei der zu bearbeitenden Variablen *beitrag* die Summenfunktion SUM und bei *persnr* die Funktion NU aus. Die letztgenannte Funktion NU steht dabei für die Anzahl der (ungewichteten) Fälle.

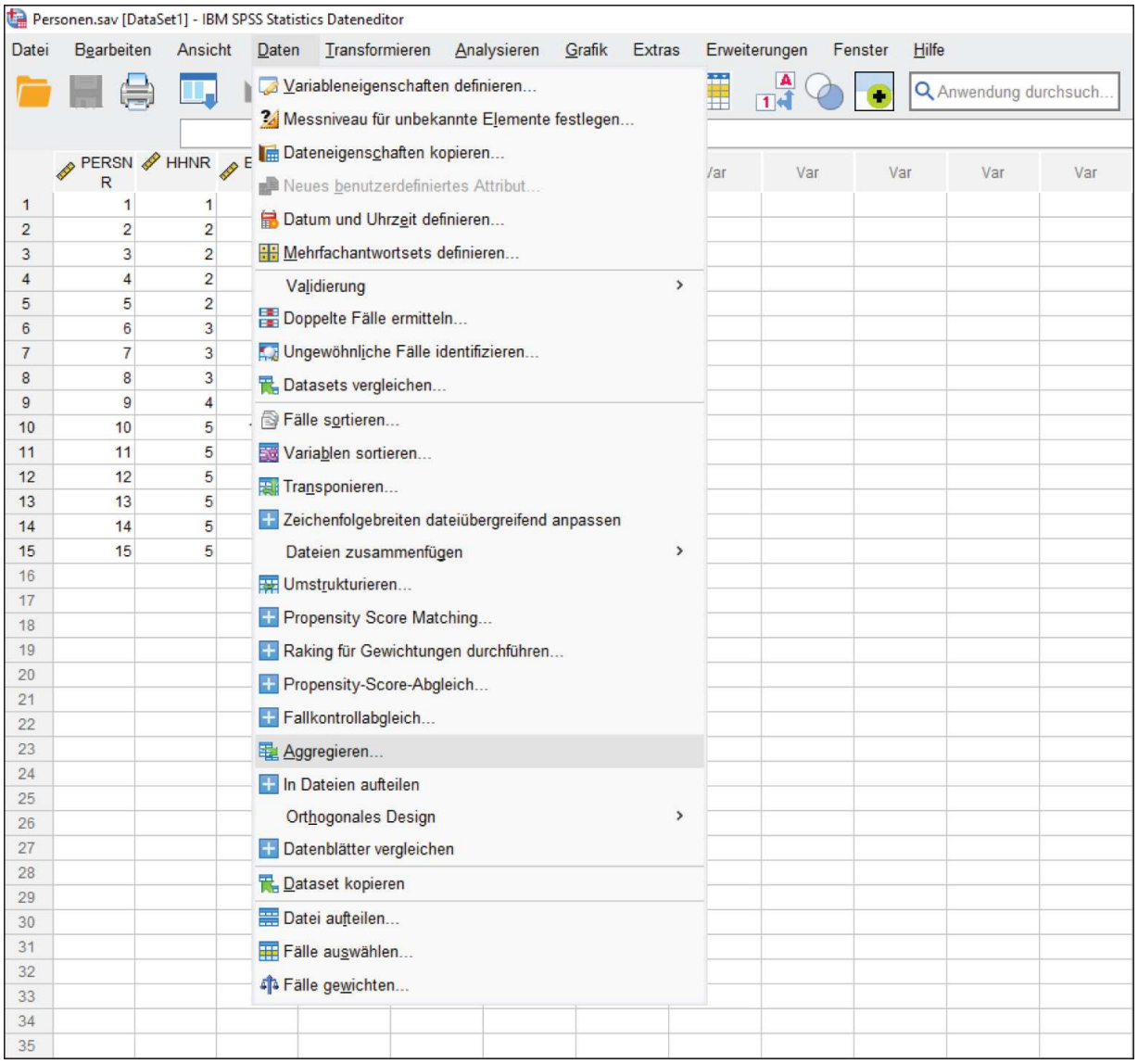



**Abbildung 4.2:** Daten-Aggregation in SPSS starten

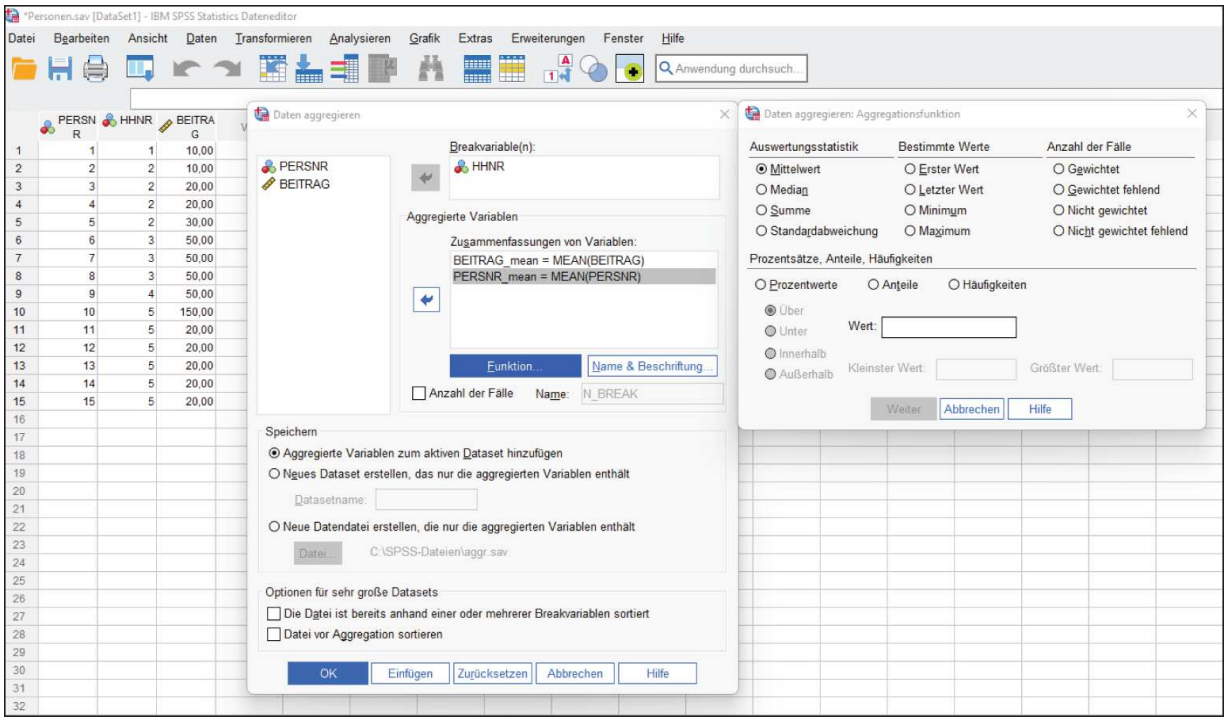



**Abbildung 4.3:** Mögliche Funktionen im Zusammenhang mit dem Aggregieren von Daten

Die Funktionen, die Ihnen an dieser Stelle über den Button FUNKTIONEN angeboten werden, sind:

- ***AUSWERTUNGSSTATISTIK:***

  ***MITTELWERT:*** Berechnung des arithmetischen Mittelwerts über die untergeordneten Einheiten hinweg (zum Beispiel für die vier Personen Nummer 2 bis 5 in Haushalt Nummer 2 mit Beiträgen in Höhe von 10, 20, 20 und 30 Euro:

$$\frac{10\,\text{Euro} + 20\,\text{Euro} + 20\,\text{Euro} + 30\,\text{Euro}}{4\,\text{Personen}} = 20\,\frac{\text{Euro}}{\text{Person}});$$

  ***MEDIAN:*** Berechnung des genau in der Mitte liegenden Werts über die untergeordneten Einheiten hinweg (zum Beispiel für die vier Personen Nummer 2 bis 5 in Haushalt Nummer 2 mit Beiträgen in Höhe von 10, 20, 20 und 30 Euro:

$$\frac{20\,\text{Euro} + 20\,\text{Euro}}{2} = 20\,\text{Euro}\,);$$

***SUMME:*** Berechnung der Wertesumme über die untergeordneten Einheiten hinweg (zum Beispiel für die vier Personen Nummer 2 bis 5 in Haushalt Nummer 2 mit Beiträgen in Höhe von 10, 20, 20 und 30 Euro: 10 Euro + 20 Euro + 20 Euro + 30 Euro = 80 Euro);

***STANDARDABWEICHUNG:*** Berechnung der Streuung um den arithmetischen Mittelwert über die untergeordneten Einheiten hinweg, mit Bezugnahme auf (N-1) Fälle der untergeordneten Einheiten (zum Beispiel für die vier Personen Nummer 2 bis 5 in Haushalt Nummer 2 mit Beiträgen in Höhe von 10, 20, 20 und 30 Euro:

$$\sqrt{\frac{1}{3}\cdot\left[(10-20)^2+(20-20)^2+(20-20)^2+(30-20)^2\right]}\approx 8{,}16\,(\text{Euro})$$

).

- ***BESTIMMTE WERTE:***

  ***ERSTE:*** Es wird nur der erste Wert der untergeordneten Einheiten berücksichtigt (zum Beispiel für die vier Personen Nummer 2 bis 5 in Haushalt Nummer 2 mit Beiträgen in Höhe von 10, 20, 20 und 30 Euro: 10 Euro);

  ***LETZTE:*** Es wird nur der letzte Wert der untergeordneten Einheiten berücksichtigt (zum Beispiel für die vier Personen Nummer 2 bis 5 in Haushalt Nummer 2 mit Beiträgen in Höhe von 10, 20, 20 und 30 Euro: 30 Euro);

  ***MINIMUM:*** Es wird nur der niedrigste Wert der untergeordneten Einheiten berücksichtigt (zum Beispiel für die vier Personen Nummer 2 bis 5 in Haushalt Nummer 2 mit Beiträgen in Höhe von 10, 20, 20 und 30 Euro: 10 Euro);

  ***MAXIMUM:*** Es wird nur der höchste Wert der untergeordneten Einheiten berücksichtigt (zum Beispiel für die vier Personen Nummer 2 bis 5 in Haushalt Nummer 2 mit Beiträgen in Höhe von 10, 20, 20 und 30 Euro: 30 Euro).

- ***ANZAHL DER FÄLLE:***

  ***GEWICHTET:*** Es wird die Anzahl der jeweils untergeordneten Einheiten bei Gewichtung des Datensatzes angegeben (zum Beispiel

für die vier Personen Nummer 2 bis 5 in Haushalt Nummer 2 mit Beiträgen in Höhe von 10, 20, 20 und 30 Euro und da keine Gewichtung hier vorliegend: 4 Personen);

***GEWICHTET FEHLEND:*** Es wird die Anzahl der jeweils untergeordneten Einheiten bei Gewichtung des Datensatzes angegeben, für die in unserem Beispiel keine Personennummer vorliegt (zum Beispiel für die vier Personen Nummer 2 bis 5 in Haushalt Nummer 2 mit Beiträgen in Höhe von 10, 20, 20 und 30 Euro und da keine Gewichtung, aber auch keine fehlende Personennummer hier vorliegend: 0 Personen);

***NICHT GEWICHTET:*** Es wird die Anzahl der jeweils untergeordneten Einheiten ohne Gewichtung des Datensatzes angegeben (zum Beispiel für die vier Personen Nummer 2 bis 5 in Haushalt Nummer 2 mit Beiträgen in Höhe von 10, 20, 20 und 30 Euro: 4 Personen);

***NICHT GEWICHTET FEHLEND:*** Es wird die Anzahl der jeweils untergeordneten Einheiten ohne Gewichtung des Datensatzes angegeben, für die in unserem Beispiel keine Personennummer vorliegt (zum Beispiel für die vier Personen Nummer 2 bis 5 in Haushalt Nummer 2 mit Beiträgen in Höhe von 10, 20, 20 und 30 Euro und da keine fehlende Personennummer hier vorliegend: 0 Personen).

Falls Sie die einzelnen Funktionen nachvollziehen möchten, können Sie sich das Ganze in der Datei *AGGREGATIONSFUNKTIONEN.SAV* anschauen. Dort habe ich nämlich alle obigen Funktionsvarianten für *beitrag* durchgespielt.

Lassen Sie mich zu unserer Ausgangsfragestellung der Summenbildung für *beitrag* und der ungewichteten Fallzahlen für *persnr* zurückkommen: Um sicherzustellen, dass die Beitragssumme ebenso im Haushaltszusammenhang wie die Fallzahl über alle Haushaltsmitglieder hinweg ermittelt wird, muss in unserem Beispiel auf Basis von *PERSONEN.SAV* die Haushaltskennzeichnungsnummer *hhnr* zur Breakvariablen deklariert werden (siehe Abbildung 4.4).

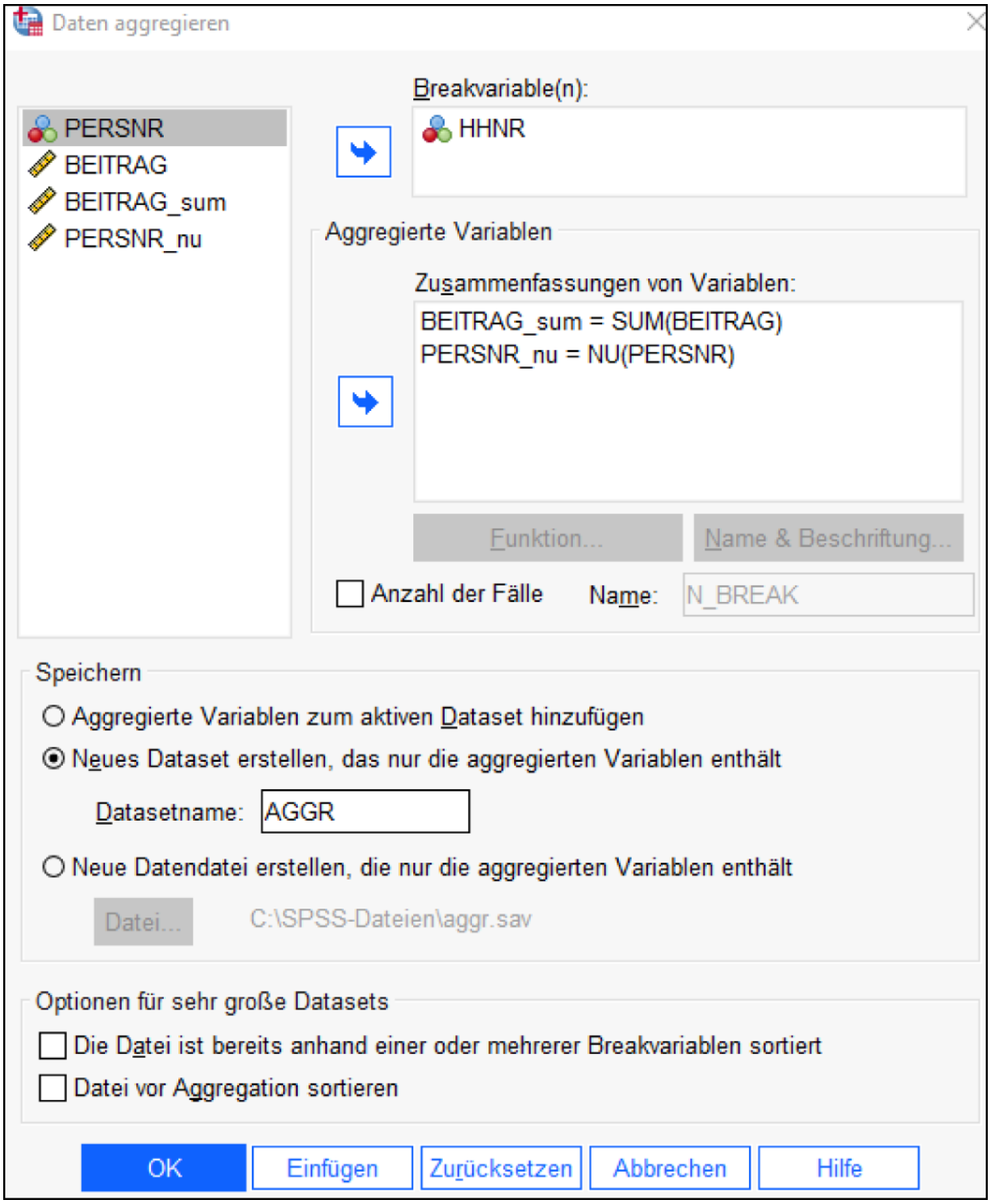


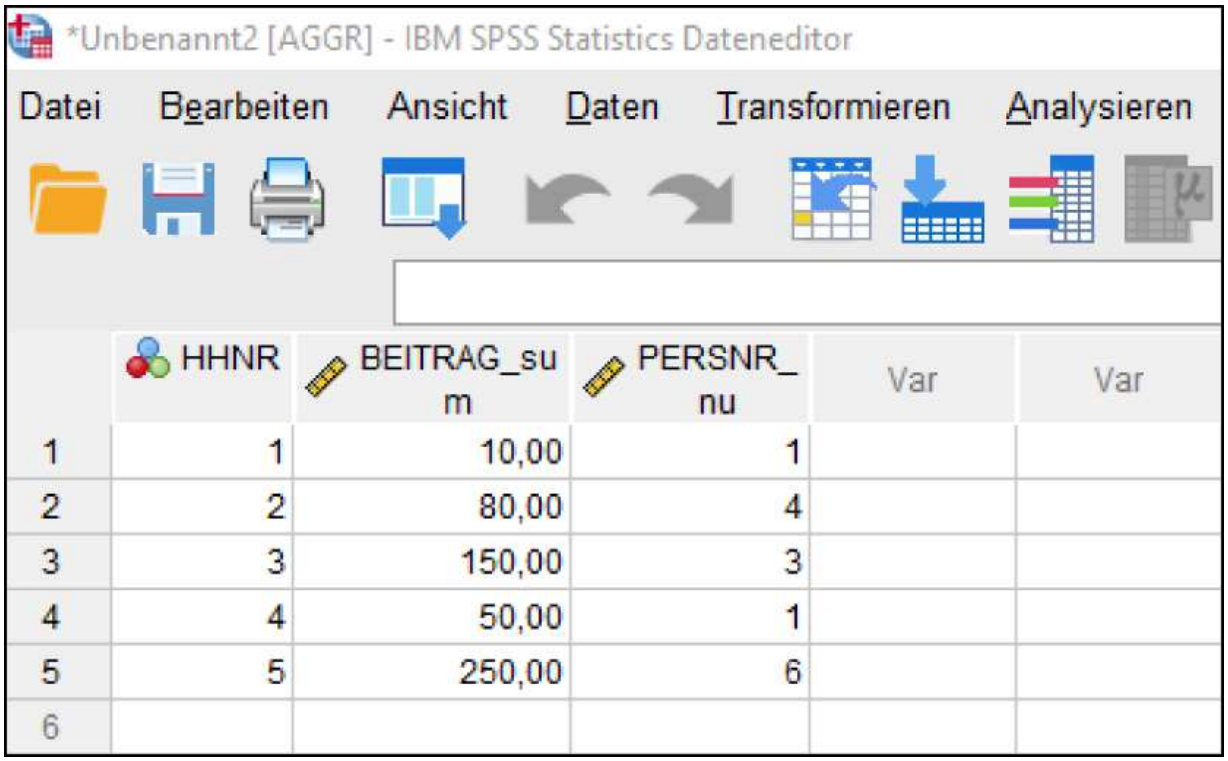

**Abbildung 4.4:** Durchführen des Aggregierens in SPSS

Die aus Abbildung 4.4 ersichtliche Einstellung, ein neues Datenset mit ausschließlich den aggregierten Variablen zu erstellen (unter dem Namen *AGGR*), liefert als neuen Datensatz die gewünschte aggregierte Datei,

die Sie jetzt unter einem eigenen Namen als *SAV*-Datei abspeichern sollten (zum Beispiel als *AGGR.SAV* ). Das Datenset *AGGR* beziehungsweise die Datendatei *AGGR.SAV* sind notwendig, damit Daten im Haushalts- und *nicht* im Personendatenset abgespeichert werden.

Für die Beitragssummen erhalten Sie in unserem Beispiel die erwarteten Beträge von 10,00 Euro (Haushalt 1), 80,00 Euro (Haushalt 2), 150,00 Euro (Haushalt 3), 50,00 Euro (Haushalt 4) und 250,00 Euro (Haushalt 5), und als Haushaltsgrößen resultieren 1 Person (Haushalt 1), 4 Personen (Haushalt 2), 3 Personen (Haushalt 3), 1 Person (Haushalt 4) und 6 Personen (Haushalt 5; siehe Abbildung 4.5).

In der Forschungspraxis stellt sich häufig die Aufgabe, Datensätze zu aggregieren. Beispielsweise liegen die Daten des Sozio-oekonomischen Panels einerseits auf der Personen-, andererseits aber auch auf der Haushaltsebene vor. Sie müssen also zumindest zum Teil bei haushaltsbezogenen Auswertungen Informationen, die individuell auf der Personenebene vorliegen, auf die Haushaltsebene durch Aggregation – in der Weise, wie ich das eben beschrieben habe – »bringen«.




| | HHNR | BEITRAG_sum | PERSNR_nu | Var | Var |
|---|---|---|---|---|---|
| 1 | 1 | 10,00 | 1 | | |
| 2 | 2 | 80,00 | 4 | | |
| 3 | 3 | 150,00 | 3 | | |
| 4 | 4 | 50,00 | 1 | | |
| 5 | 5 | 250,00 | 6 | | |
| 6 | | | | | |

**Abbildung 4.5:** Aggregierter Datensatz

## *Aggregation ohne Breakvariable*

Wählen Sie im Unterschied zu den vorstehenden Aggregationen mit Breakvariablen *keine* Breakvariable aus, aggregiert SPSS in seinen neueren Versionen (so auch in der hier genutzten Version 29) über *alle* Untersuchungseinheiten, das heißt auch, über alle untergeordneten Einheiten für den gesamten Datensatz hinweg und gibt solcherart Summenwerte und dergleichen an. Das ist doch sehr bequem – oder?

Öffnen Sie bitte zunächst wieder die Ausgangs-Datendatei *PERSONEN.SAV*. Im nächsten Schritt führen Sie im von Ihnen so benannten Datenset *AGGR2* die funktionalen Zuordnungen *beitrag_sum* = sum(*beitrag*) und *persnr_nu* = nu(*persnr*) aus, jetzt aber ohne Angabe einer Breakvariablen. Das heißt: Lassen Sie bitte das Feld BREAKVARIABLE(N) leer (siehe Abbildung 4.6).

Im Ergebnis erhalten Sie eine Datendatei mit einem Fall, der als aggregierte Variablenwerte für *beitrag_sum* 540,00 (Euro) und für *persnr_nu* 15 (Personen) ausweist (siehe Abbildung 4.7).

Während Sie also sogar über alle Fälle eines Datensatzes hinweg problemlos aggregieren dürfen, ist demgegenüber die Aggregation von SPSS-Zufallsstichproben nicht sinnvoll. Der Grund hierfür ist, dass eine SPSS-Zufallsstichprobe je nach konkreter (Zufalls-)Ziehung jeweils anders zusammengesetzt ist und daher der aggregierte Datensatz je nach Ziehung andere Elemente (in unserem Beispiel: Personen) enthält. Hierbei werden zum Beispiel bei der Zufallsziehung von Personen einige Mitglieder eines bestimmten Haushalts aus dem sich ergebenden Datensatz ausgeschlossen. Dann würde bei der Aggregation hin zur Haushaltsebene die Anzahl der Personen in diesem Haushalt nicht korrekt ausgewiesen, weil einige Haushaltsmitglieder bei der Aggregation nicht berücksichtigt werden können.

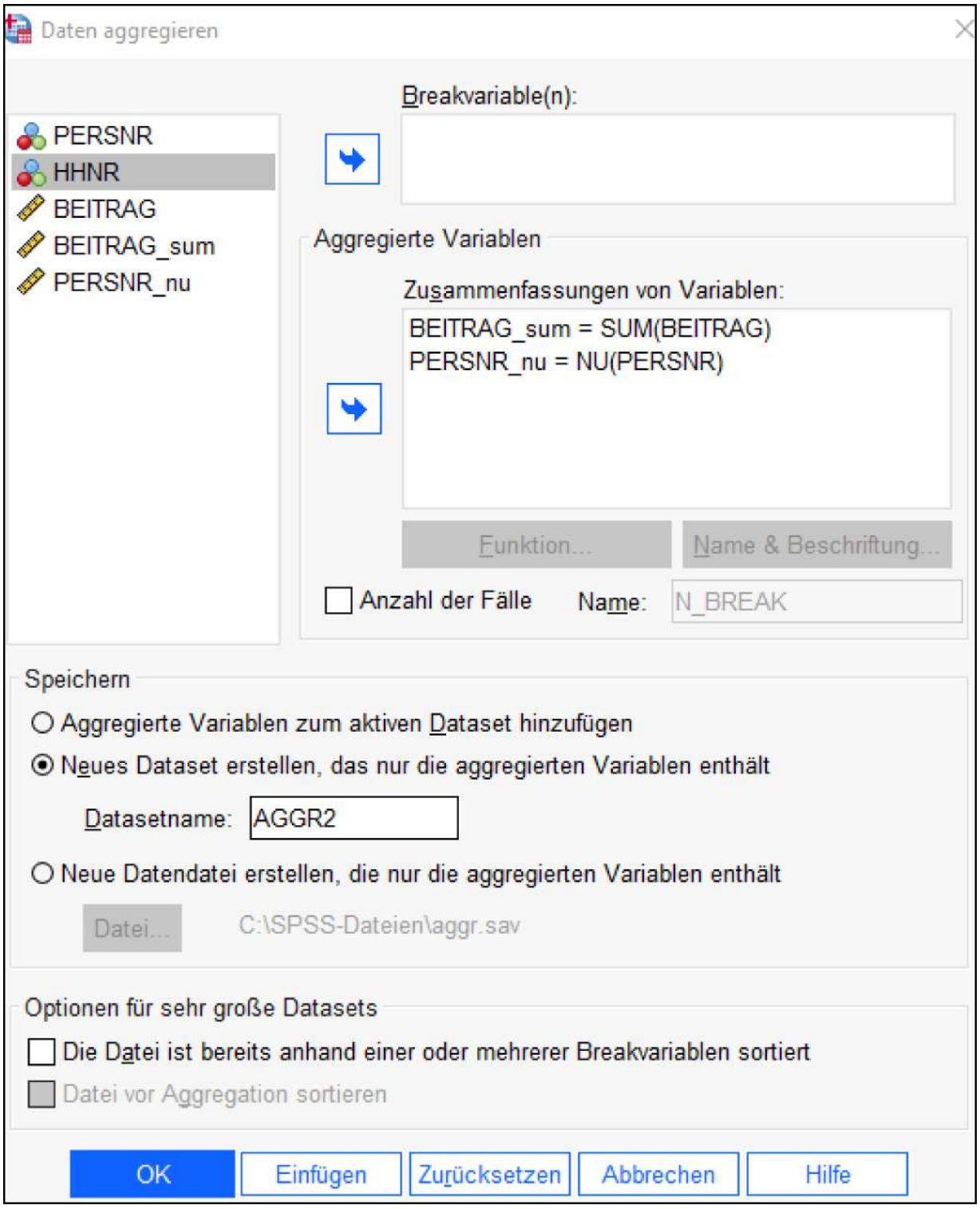



**Abbildung 4.6:** Daten-Aggregation ohne Breakvariable

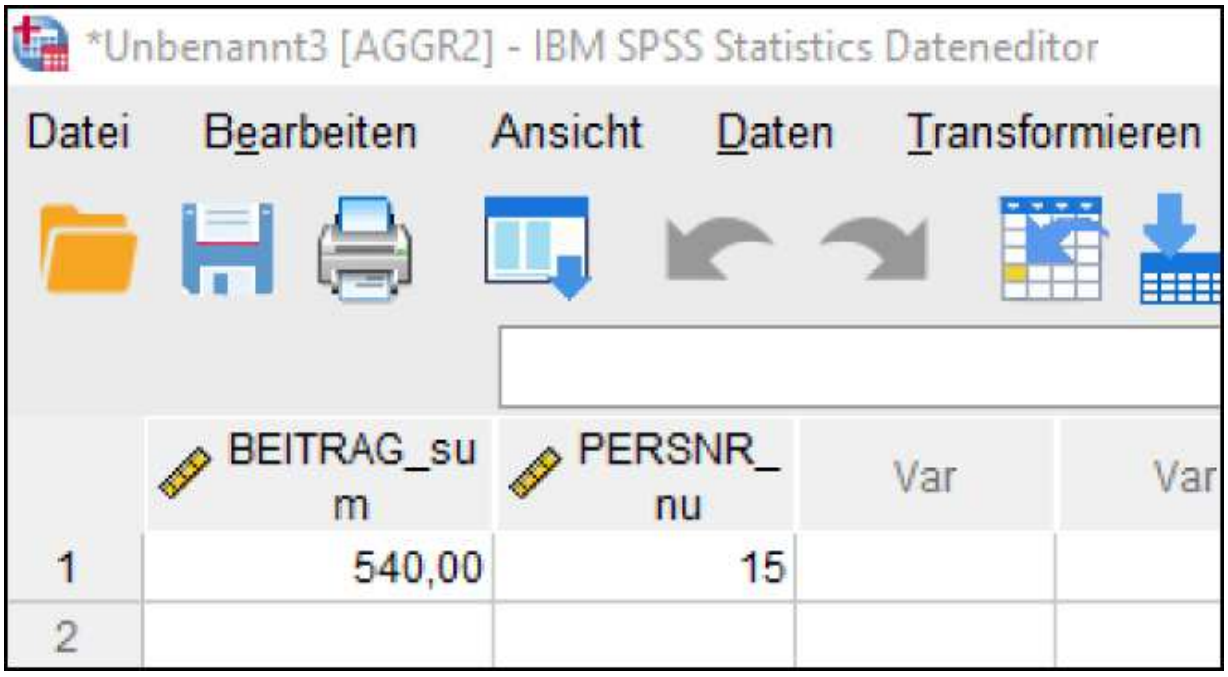
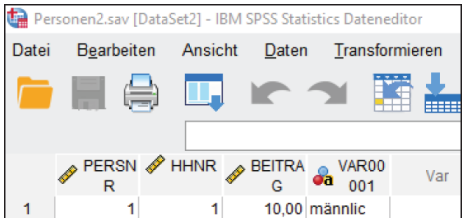



**Abbildung 4.7:** Ergebnis der Daten-Aggregation ohne Breakvariable

# *Daten (um-)formatieren*

## *Variableneditor*

Bestehende Variablen in einem Datensatz können Sie recht bequem in der Variablenansicht, also im **Variableneditor**, in Form bringen, soll heißen: formatieren. Hierzu müssen Sie nach dem Öffnen einer SPSS-Datendatei nur von der **Daten-** in die **Variablenansicht** wechseln. Angenommen, Sie haben die in Abbildung 4.8 dargestellte SPSS-Datendatei *PERSONEN2.SAV* geöffnet und möchten hierbei die Variable in der letzten Spalte – offensichtlich inhaltlich die Variable *geschlecht* – etwas anschaulicher bezeichnen als durch das schnöde vorgegebene *var00001*, dann sollten Sie unten in der linken Fensterecke auf VARIABLENANSICHT klicken, denn in Abbildung 4.8 gilt als Einstellung ja noch DATENANSICHT.



| | | | | | |
|---|---|---|---|---|---|
| 2 | 2 | 2 | 10,00 | weiblich | |
| 3 | 3 | 2 | 20,00 | männlic | |
| 4 | 4 | 2 | 20,00 | männlic | |
| 5 | 5 | 2 | 30,00 | weiblich | |
| 6 | 6 | 3 | 50,00 | männlic | |
| 7 | 7 | 3 | 50,00 | männlic | |
| 8 | 8 | 3 | 50,00 | weiblich | |
| 9 | 9 | 4 | 50,00 | männlic | |
| 10 | 10 | 5 | 150,00 | weiblich | |
| 11 | 11 | 5 | 20,00 | weiblich | |
| 12 | 12 | 5 | 20,00 | weiblich | |
| 13 | 13 | 5 | 20,00 | männlic | |
| 14 | 14 | 5 | 20,00 | weiblich | |
| 15 | 15 | 5 | 20,00 | männlic | |
| 16 | | | | | |
| 17 | | | | | |
| 18 | | | | | |
| 19 | | | | | |
| 20 | | | | | |
| 21 | | | | | |
| 22 | | | | | |
| 23 | | | | | |
| 24 | | | | | |
| 25 | | | | | |
| 26 | | | | | |
| 27 | | | | | |
| 28 | | | | | |
| 29 | | | | | |
| 30 | | | | | |
| 31 | | | | | |
| 32 | | | | | |
| 33 | | | | | |
| 34 | | | | | |
| 35 | | | | | |
| 36 | | | | | |
| 37 | | | | | |
| 38 | | | | | |
| 39 | | | | | |
| 40 | | | | | |
| 41 | | | | | |
| 42 | | | | | |
| 43 | | | | | |
| 44 | | | | | |
| 45 | | | | | |
| 46 | | | | | |
| 47 | | | | | |
| 48 | | | | | |
| 49 | | | | | |
| 50 | | | | | |
| 51 | | | | | |
| 52 | | | | | |
| 53 | | | | | |
| 54 | | | | | |
| 55 | | | | | |

Datenansicht Variablenansicht

**Abbildung 4.8:** Ein zu formatierendes SPSS-Datenfile (Datenansicht) von *PERSONEN2.SAV*

In der SPSS-Variablenansicht werden in den Zeilen die einzelnen Variablen angezeigt und in den Spalten Eigenschaften dieser Variablen (siehe Abbildung 4.9).

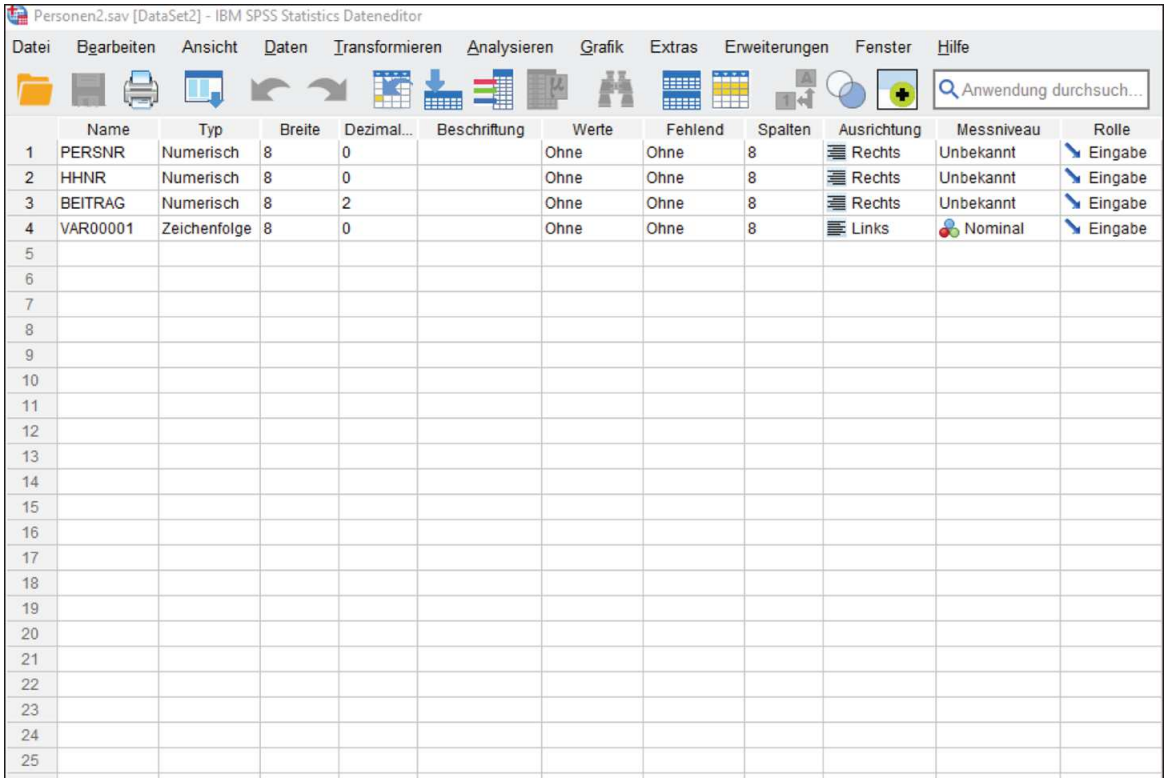
Es empfiehlt sich, für jeden (Ursprungs-)Datensatz zuzüglich zu den Angaben im Variableneditor einen **Codierplan** zur Dokumentation des betreffenden Datensatzes zu erstellen. Dieser Codierplan sollte mindestens den SPSS-Variablennamen, den Variableninhalt und natürlich die Codierung enthalten. Für das Beispiel aus Abbildung 4.9 könnte der Codierplan zum Beispiel wie nachfolgend lauten (wobei ich die Variable *var00001* schon einmal in *geschlecht* umbenannt und zudem ihre Ausprägungen »männlic« und »weiblich« in »1« und »2« umgewandelt habe. Übrigens, wenn Sie sich fragen, wo bei »männlic« das »h« geblieben ist: Das hängt mit der maximalen Anzahl der speicherbaren Zeichen zusammen, mehr dazu weiter unten unter »Variablenbreite und Dezimalstellen«).

In Tabelle 4.1 ist Folgendes angegeben: die Variablenbezeichnung, die Stelle der Variablen im Datensatz, das heißt: wo diese »beginnt«, die Anzahl der im Datensatz »verbrauchten« Stellen durch die betreffende Variable, das Variablenformat und der Variableninhalt im Sinne der Variablencodierung. Die Variable *persnr* beginnt im Datensatz an der ersten Stelle, ist 2-stellig und repräsentiert die Personennummer von Nummer 1 bis Nummer 15. Außerdem ist die Formatierung angegeben: »F2.0« ist angelehnt an die für SPSS relevanten Bezeichnungen aus der Programmiersprache FORTRAN und steht für eine zweistellige numerische Variable ohne Nachkommastellen (FORTRAN = FORmulaTRANslator); ganz genau genommen, steht das »F« für »Fixkommastelle«. Auch die anderen Variablen sind laut Tabelle 4.1 numerisch, wobei für die Variable *beitrag* insgesamt fünf Stellen mit zwei Nachkommastellen vorgesehen sind. In Tabelle 4.1 ist auch *geschlecht* numerisch mit »1« für männliche Personen und mit »2« für weibliche Personen angegeben. Hätten Sie alternativ für

*geschlecht* die ursprünglichen Ausprägungen »maennlic« und »weiblich« verwendet, hätte eine 8-stellige Stringvariable vorgelegen, und die FORTRAN-Formatierung hätte jeweils A8 gelautet (das »A« steht für »alphanumerisch«, was ja nur ein anderer Ausdruck als »String« ist).

| ***Variablenbezeichnung*** | ***Stelle im Datensatz*** | ***Anzahl der Satzstellen*** | ***Format*** | ***Inhalt*** |
|---|---|---|---|---|
| persnr | 1 | 2 | F2.0 | Personennummer 1 bis 15 |
| hhnr | 3 | 1 | F1.0 | Haushaltsnummer 1 bis 5 |
| beitrag | 4 | 5 | F3.2 | Beitragszahlungen in Euro |
| geschlecht | 10 | 1 | F1.0 | Geschlecht: |
| | | | | 1 = männlich |
| | | | | 2 = weiblich |

**Tabelle 4.1:** Beispielhafter Codierplan für die Datei *PERSONEN2.SAV*



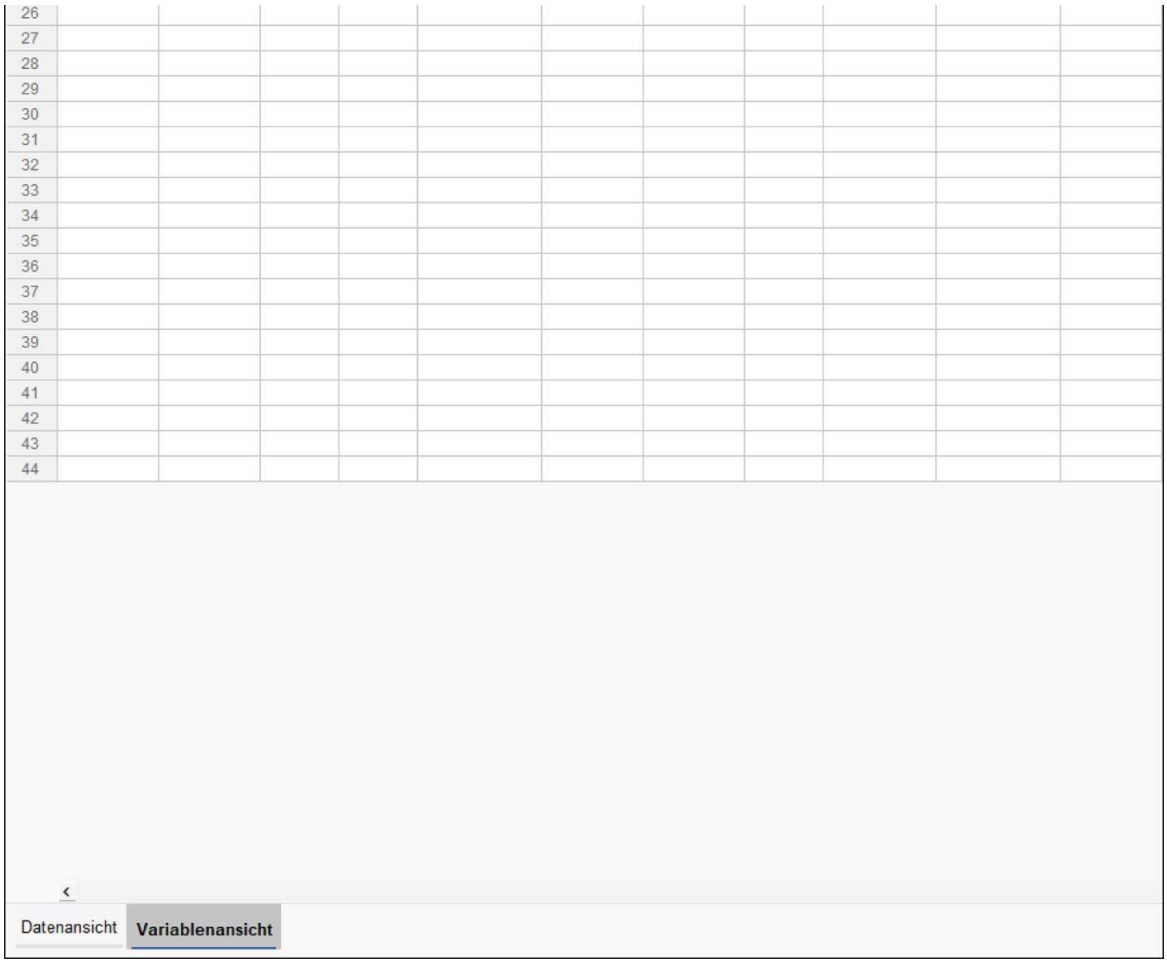



**Abbildung 4.9:** Die SPSS-Variablenansicht von *PERSONEN2.SAV*

Mit einem solchen Codierplan sorgen Sie auch für Klarheit beziehungsweise Nachvollziehbarkeit Ihres Datensatzes bei anderen Personen.

# Zulässige Variablennamen

Konkret wird bei der Variablenansicht in der ersten Spalte der jeweilige **Variablenname** angegeben. In diesem Zusammenhang sind folgende Regeln bei der Namensgebung zu beachten:

- ✔ Der Variablenname muss eindeutig sein, sodass also doppelte Namen nicht erlaubt sind.

- ✔ Ein gültiger Variablenname darf maximal 64 Zeichen beinhalten, und er muss mit einem Buchstaben anfangen.
- ✔ Erlaubt sind folgende Sonderzeichen im Variablennamen: a) der Unterstrich _, also zum Beispiel: *var_1*; b) der Punkt ., also zum Beispiel: *var.1*; c) das Dollarzeichen $, also zum Beispiel: *var$1*; d) die Raute #, also zum Beispiel: *var#1*; e) das At-Zeichen @, also zum Beispiel: *var@1*. Keineswegs erlaubt sind hingegen: das Ausrufezeichen !, das Fragezeichen ?, der Bindestrich -, das Anführungszeichen “ und der Asterisk * sowie das Leerzeichen und führende Ziffern (also Ziffern am Anfang des Variablennamens).
- ✔ Als Variablennamen dürfen Sie keine Schlüsselwörter aus der SPSS-Syntax verwenden, wie ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO oder WITH.

Bei falscher Namensgebung erhalten Sie über ein SPSS-Dialogfeld (via TRANSFORMIEREN| VARIABLE BERECHNEN) eine Fehlermeldung. In Abbildung 4.10 bis Abbildung 4.12 sind drei Beispiele hierfür angegeben.

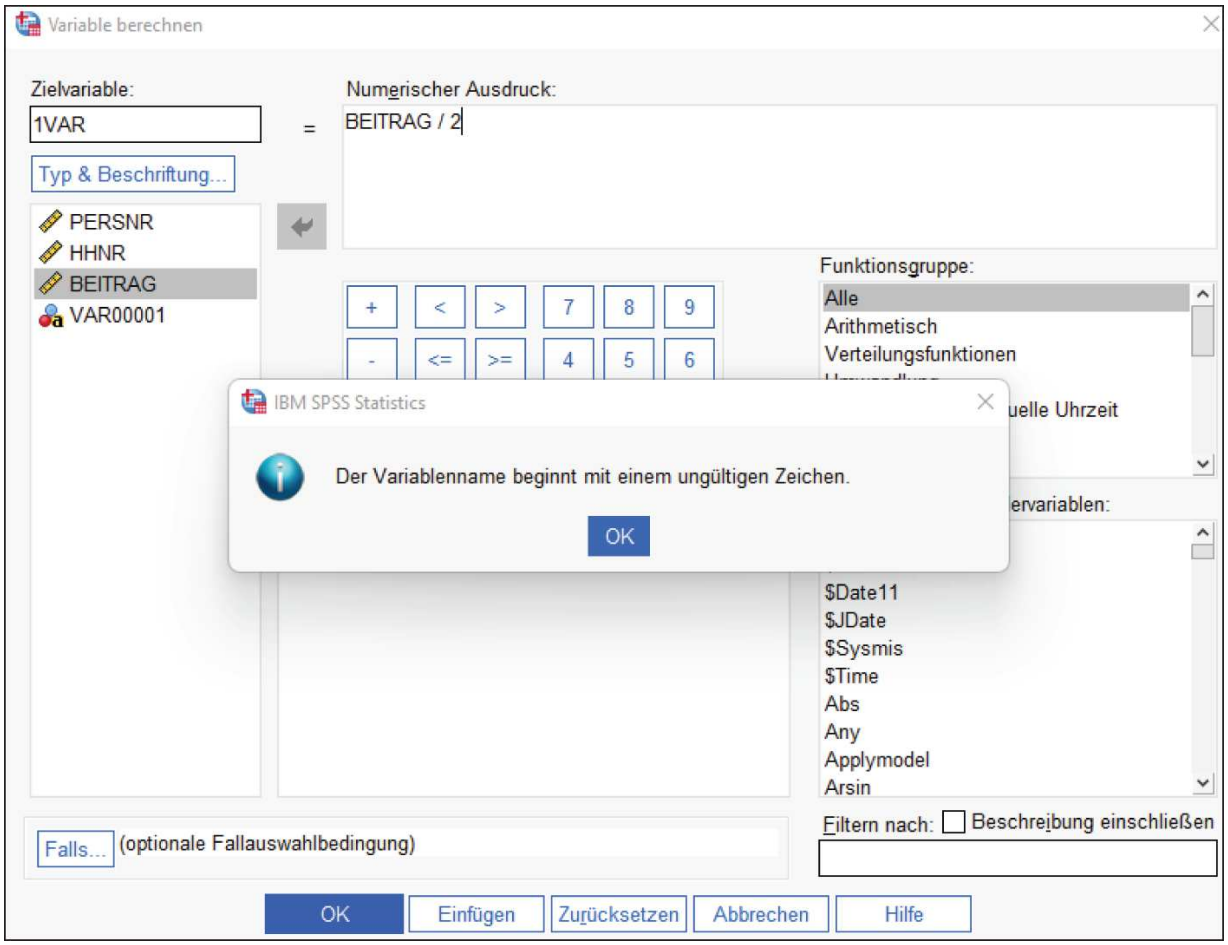



**Abbildung 4.10:** SPSS-Fehlermeldung bei falschem Namensanfang einer Variablen bezüglich der Variablen *1var*

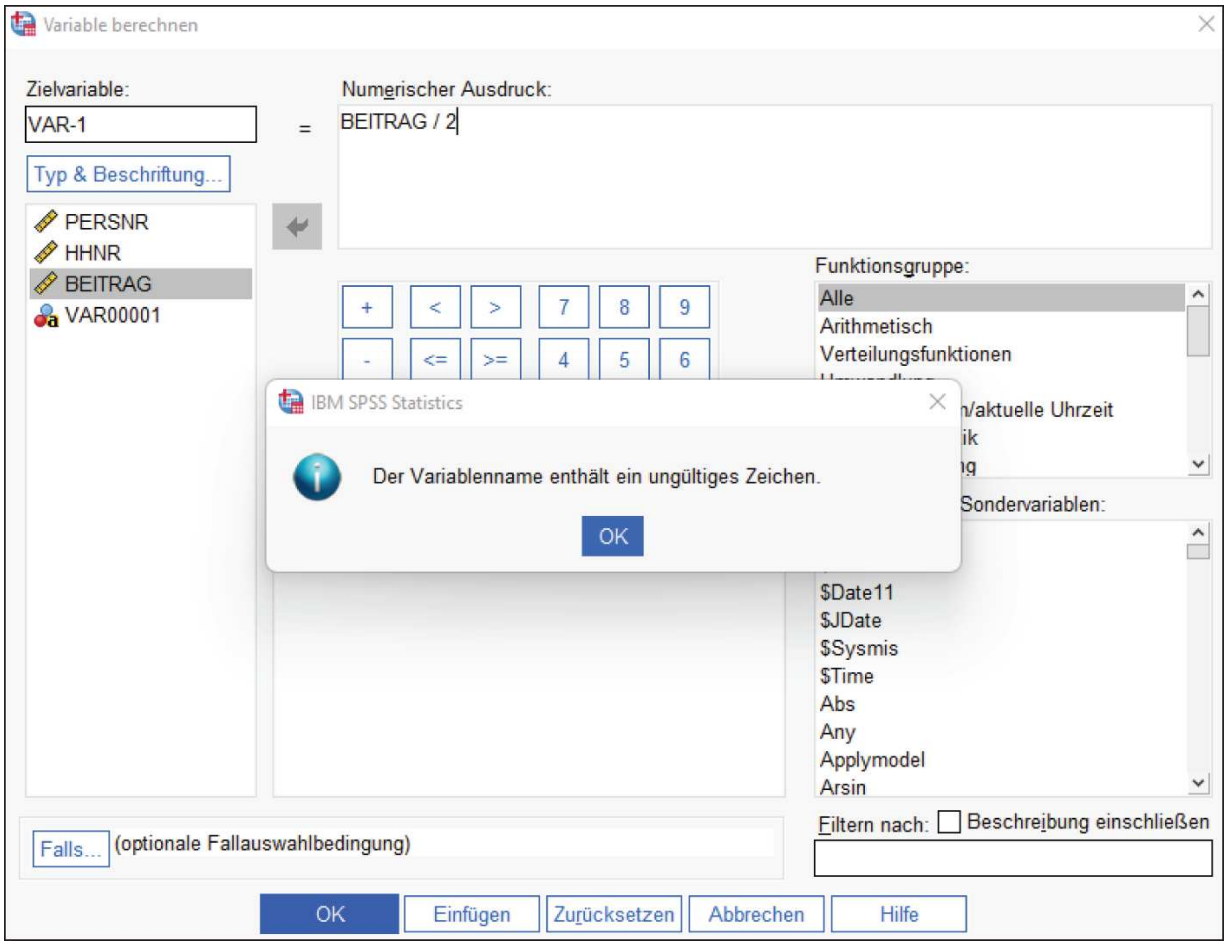



**Abbildung 4.11:** SPSS-Fehlermeldung bei unzulässigem Sonderzeichen in einem Variablennamen bezüglich der Variablen *var-1*

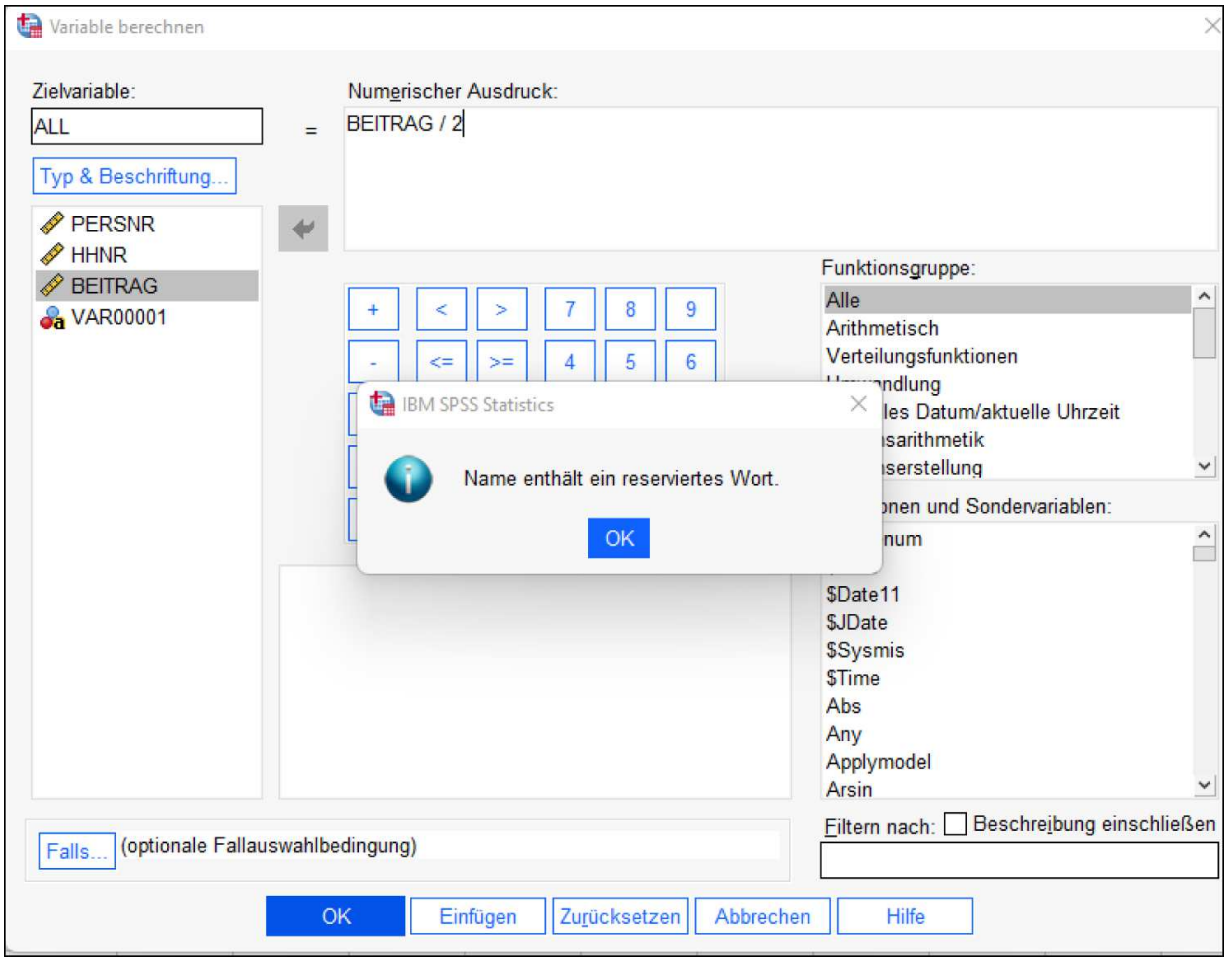



**Abbildung 4.12:** SPSS-Fehlermeldung bei unzulässiger Verwendung eines Schlüsselworts als Variablennamen bezüglich der Variablen *all*

Im Variableneditor können Sie *var0001* beispielsweise in *geschlecht* ändern (wie eben bei der Vorstellung eines Codierplans bereits angedeutet). Dies tun Sie einfach dadurch, dass Sie in der entsprechenden Namenszeile den Eintrag *var0001* per Hand mit *geschlecht* überschreiben. (SPSS unterscheidet dabei nicht zwischen Groß- und Kleinschreibung – in Fachchinesisch spricht man davon, dass SPSS nicht case-sensitiv ist –, sodass Sie auch zum Beispiel *geschlecht*, *Geschlecht* oder *GeScHlEcHt* hätten eingeben können.) Abbildung 4.13 zeigt Ihnen das Ergebnis des Überschreibens von *var0001* durch *geschlecht*.

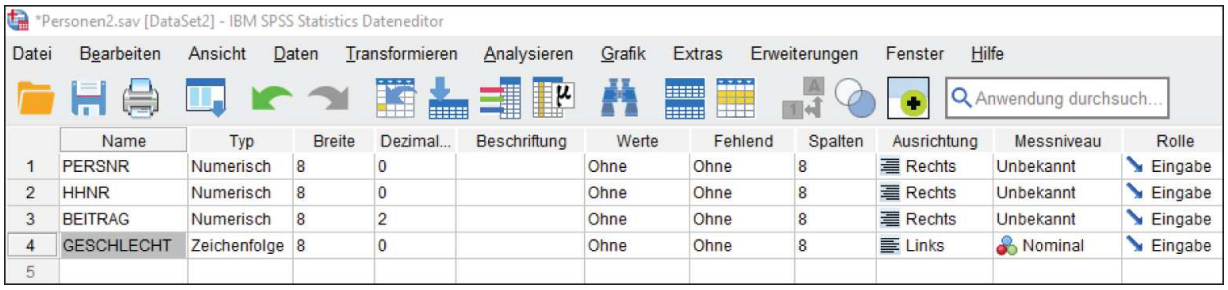



**Abbildung 4.13:** Verändern eines Variablennamens im Variableneditor

# *Variablentyp*

Es folgt in der zweiten Spalte des Variableneditors die wichtige Angabe, welchen **Typs** die betreffende Variable ist. Zur Auswahl stehen – wie Sie Abbildung 4.14 entnehmen können – verschiedene Typen:

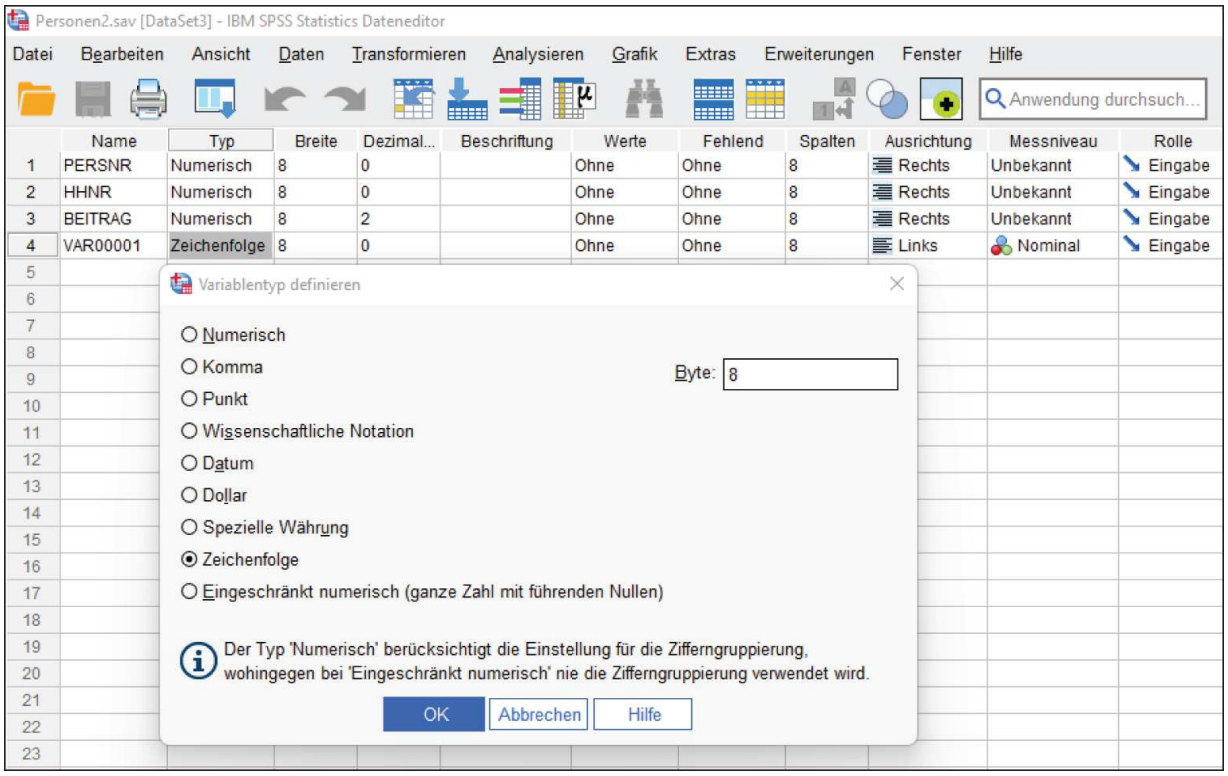
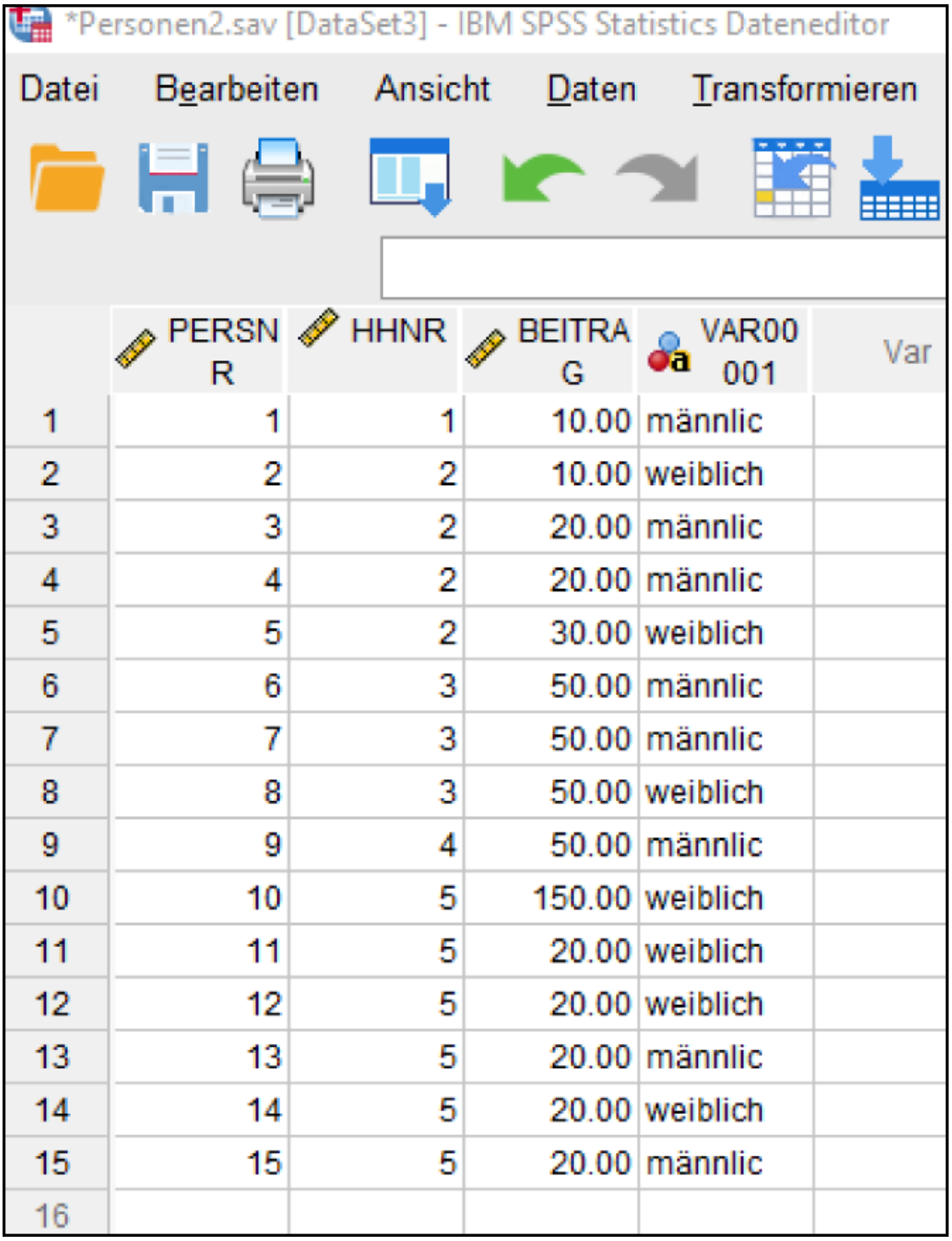



**Abbildung 4.14:** Verändern des Variablentyps im Variableneditor (*PERSONEN2.SAV*)

Im Einzelnen handelt es sich also um folgende Typen (die Beispiele beziehen sich auf den Datensatz *PERSONEN2.SAV*):

- ✔ NUMERISCH: In unserem Beispiel sind *persnr*, *hhnr* und *beitrag* als numerische Variablen – vereinfacht als: »Zahlengrößen« – formatiert. NUMERISCH ist in SPSS der voreingestellte Variablentyp. Möchten Sie Variablen anders formatieren, müssen Sie an dieser Stelle aus NUMERISCH zum Beispiel ZEICHENFOLGE machen.
- ✔ KOMMA: Diese Formatierung ist für numerische Variablen mit Nachkommastellen relevant. In unserem Fall betrifft dies nur *beitrag* (mit zwei Nachkommastellen). Die betreffende Eingabe führt (zugegebenermaßen etwas unerwartet) dazu, dass die Nachkommastellen durch einen Punkt – gemäß der amerikanischen Notation – in der Datenansicht getrennt werden (siehe Abbildung 4.15).

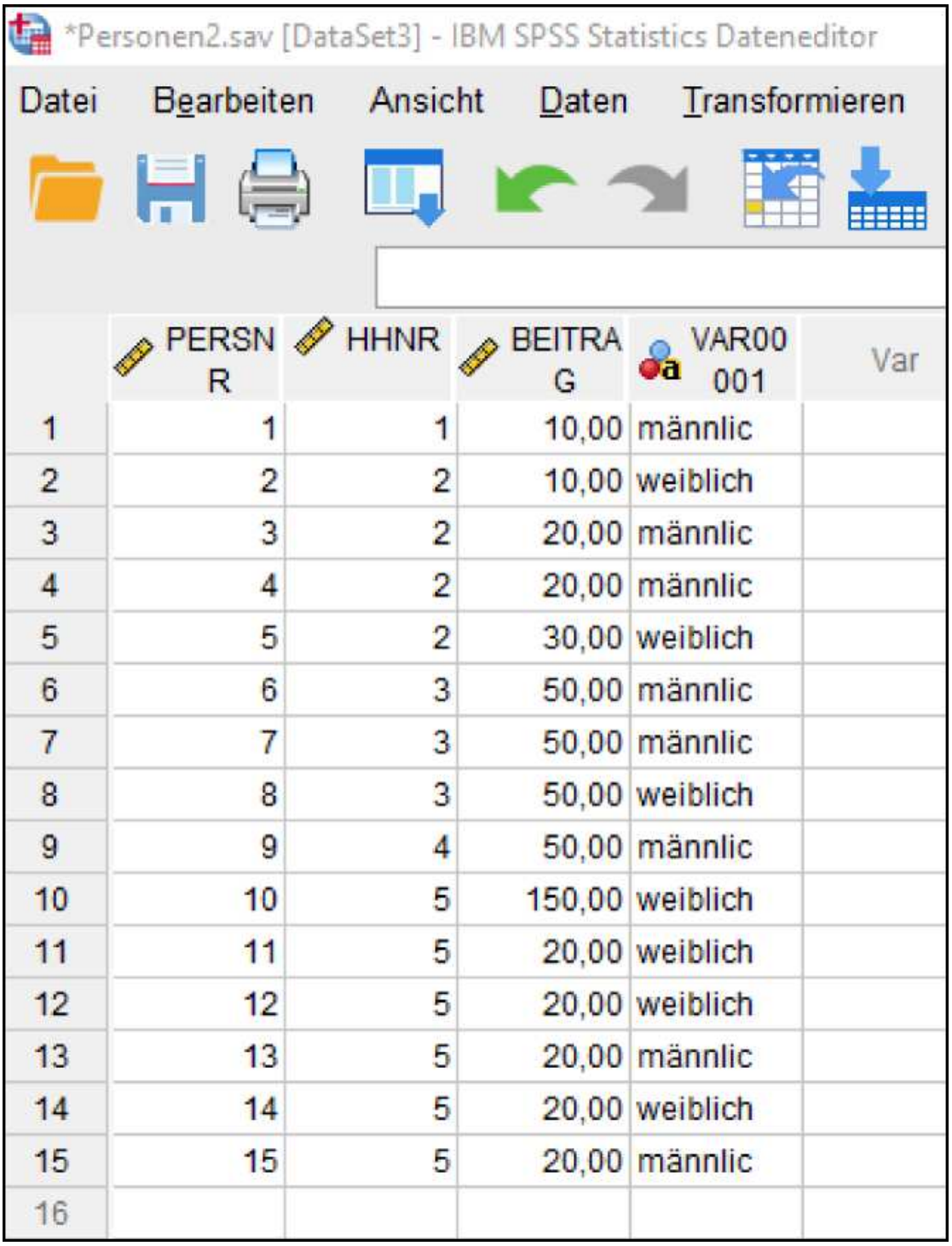



| | PERSNR | HHNR | BEITRAG | VAR00001 | Var |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 10.00 | männlic | |
| 2 | 2 | 2 | 10.00 | weiblich | |
| 3 | 3 | 2 | 20.00 | männlic | |
| 4 | 4 | 2 | 20.00 | männlic | |
| 5 | 5 | 2 | 30.00 | weiblich | |
| 6 | 6 | 3 | 50.00 | männlic | |
| 7 | 7 | 3 | 50.00 | männlic | |
| 8 | 8 | 3 | 50.00 | weiblich | |
| 9 | 9 | 4 | 50.00 | männlic | |
| 10 | 10 | 5 | 150.00 | weiblich | |
| 11 | 11 | 5 | 20.00 | weiblich | |
| 12 | 12 | 5 | 20.00 | weiblich | |
| 13 | 13 | 5 | 20.00 | männlic | |
| 14 | 14 | 5 | 20.00 | weiblich | |
| 15 | 15 | 5 | 20.00 | männlic | |
| 16 | | | | | |

**Abbildung 4.15:** KOMMA-Formatierung der Variablen beitrag (*PERSONEN2.SAV*)

✔ PUNKT: Alternativ können Sie numerische Variablen mit Nachkommastellen auch so formatieren. In unserem Fall wird hierdurch bei *beitrag* die Vorkomma- von der Nachkommastelle

(auch hier zugegebenermaßen recht unerwartet) durch ein Komma getrennt (siehe Abbildung 4.16). Bitte beachten Sie, dass die Angabe von Nachkommastellen in SPSS-Syntaxprogrammen stets der amerikanischen Variante mit dem Punkt folgt (Näheres dazu in Kapitel 5).

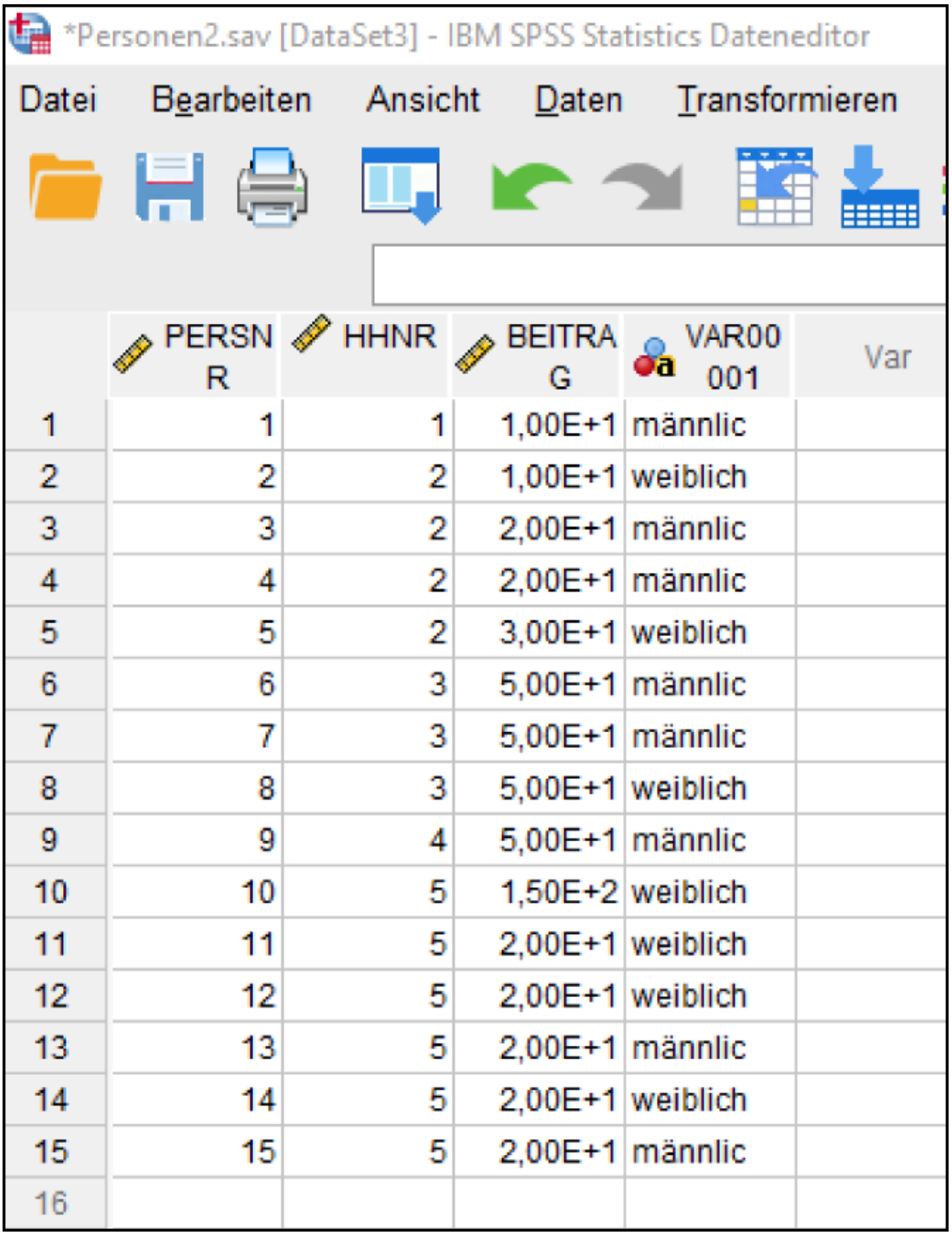



| | PERSNR | HHNR | BEITRAG | VAR00001 | Var |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 10,00 | männlic | |
| 2 | 2 | 2 | 10,00 | weiblich | |
| 3 | 3 | 2 | 20,00 | männlic | |
| 4 | 4 | 2 | 20,00 | männlic | |
| 5 | 5 | 2 | 30,00 | weiblich | |
| 6 | 6 | 3 | 50,00 | männlic | |
| 7 | 7 | 3 | 50,00 | männlic | |
| 8 | 8 | 3 | 50,00 | weiblich | |
| 9 | 9 | 4 | 50,00 | männlic | |
| 10 | 10 | 5 | 150,00 | weiblich | |
| 11 | 11 | 5 | 20,00 | weiblich | |
| 12 | 12 | 5 | 20,00 | weiblich | |
| 13 | 13 | 5 | 20,00 | männlic | |
| 14 | 14 | 5 | 20,00 | weiblich | |
| 15 | 15 | 5 | 20,00 | männlic | |
| 16 | | | | | |

**Abbildung 4.16:** PUNKT-Formatierung der Variablen *beitrag* (*PERSONEN2.SAV*)

✔ WISSENSCHAFTLICHE NOTATION: Hier werden numerische Variablen als Zehnerpotenzen dargestellt. Dies sieht in SPSS dann so aus, dass etwa in *beitrag* aus dem ersten Wert »10« der Ausdruck

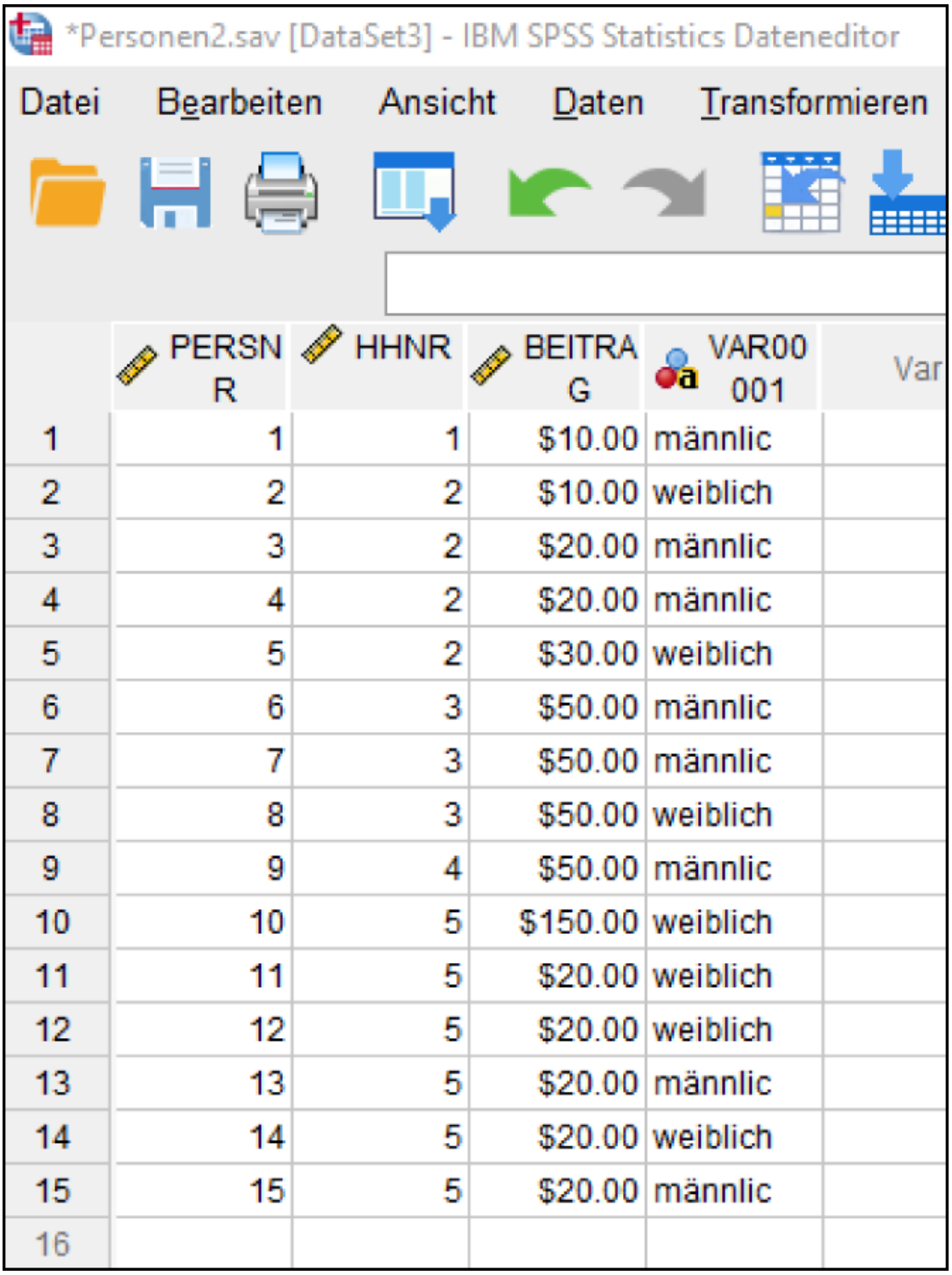
»1,00E+1« (lies: »1 mal 10 hoch 1«, was ja gleich 10 ist) wird (siehe Abbildung 4.17).




| | PERSNR | HHNR | BEITRAG | VAR00001 | Var |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1,00E+1 | männlic | |
| 2 | 2 | 2 | 1,00E+1 | weiblich | |
| 3 | 3 | 2 | 2,00E+1 | männlic | |
| 4 | 4 | 2 | 2,00E+1 | männlic | |
| 5 | 5 | 2 | 3,00E+1 | weiblich | |
| 6 | 6 | 3 | 5,00E+1 | männlic | |
| 7 | 7 | 3 | 5,00E+1 | männlic | |
| 8 | 8 | 3 | 5,00E+1 | weiblich | |
| 9 | 9 | 4 | 5,00E+1 | männlic | |
| 10 | 10 | 5 | 1,50E+2 | weiblich | |
| 11 | 11 | 5 | 2,00E+1 | weiblich | |
| 12 | 12 | 5 | 2,00E+1 | weiblich | |
| 13 | 13 | 5 | 2,00E+1 | männlic | |
| 14 | 14 | 5 | 2,00E+1 | weiblich | |
| 15 | 15 | 5 | 2,00E+1 | männlic | |
| 16 | | | | | |

**Abbildung 4.17:** Formatierung der Variablen *beitrag* mit WISSENSCHAFTLICHE NOTATION (*PERSONEN2.SAV* )

- DATUM: Diese Formatierung ist für Datumsvariablen relevant. SPSS bietet ein paar Möglichkeiten der Datumsdarstellung an.
- DOLLAR: Hier werden aus Ihren numerischen Ausdrücken Dollar-Ausdrücke. Beispielsweise verändert sich der erste Wert bei *beitrag* von »10,00« in »$10,00« (siehe Abbildung 4.18).
- SPEZIELLE WÄHRUNG: Alternativ zur vorstehenden Formatierung können Sie eine andere Währung als den Dollar zur Variablenformatierung verwenden.
- ZEICHENFOLGE: Dies betrifft Stringvariablen, in unserem Beispiel also nur die Variable *geschlecht*. Es empfiehlt sich, auch derartige nominalskalierte Variablen in SPSS numerisch zu formatieren, weil sich dadurch mehr Einsatzmöglichkeiten für solche Variablen ergeben (zum Beispiel im Rahmen der Regressionsanalyse als sogenannte 0/1-Dummies mit den Ausprägungen 0 und 1, also zum Beispiel beim Geschlecht: 0 = männlich und 1 = weiblich). In derartigen Fällen kann daher – beim Vorliegen numerischer Ausprägungen für solche nominalskalierten Merkmale (beispielsweise: 0 = männlich und 1 = weiblich oder alternativ zum Beispiel: 1 = männlich, 2 = weiblich) – die Voreinstellung NUMERISCH einfach beibehalten werden.




| | PERSNR | HHNR | BEITRAG | VAR00001 | Var |
|---|---|---|---|---|---|
| 1 | 1 | 1 | $10.00 | männlic | |
| 2 | 2 | 2 | $10.00 | weiblich | |
| 3 | 3 | 2 | $20.00 | männlic | |
| 4 | 4 | 2 | $20.00 | männlic | |
| 5 | 5 | 2 | $30.00 | weiblich | |
| 6 | 6 | 3 | $50.00 | männlic | |
| 7 | 7 | 3 | $50.00 | männlic | |
| 8 | 8 | 3 | $50.00 | weiblich | |
| 9 | 9 | 4 | $50.00 | männlic | |
| 10 | 10 | 5 | $150.00 | weiblich | |
| 11 | 11 | 5 | $20.00 | weiblich | |
| 12 | 12 | 5 | $20.00 | weiblich | |
| 13 | 13 | 5 | $20.00 | männlic | |
| 14 | 14 | 5 | $20.00 | weiblich | |
| 15 | 15 | 5 | $20.00 | männlic | |
| 16 | | | | | |

**Abbildung 4.18:** DOLLAR-Formatierung der Variablen *beitrag* (*PERSONEN2.SAV*)

✔ EINGESCHRÄNKT NUMERISCH: Hier können Sie eine ganze Zahl mit führenden Nullen formatieren. (Das sieht vielleicht ganz

nett aus, braucht man aber doch eigentlich nicht – oder?)

# *Variablenbreite und Dezimalstellen*

Wie bereits aus Abbildung 4.13 hervorgegangen ist, legen Sie im Rahmen des Variableneditors in der dritten und vierten Spalte die **Variablenbreite** sowie die Anzahl der **Dezimalstellen** fest. Standardmäßig werden bei numerischen Variablen insgesamt acht Stellen vorgegeben, und in unserem Beispiel hat nur *beitrag* ersichtlich zwei vorgegebene Dezimalstellen. Naheliegend ist die Anzahl der Dezimalstellen nur bei numerischen Variablen bedeutsam, und dort auch nur bei der Anzeige im Daten- und Ausgabefenster. Intern rechnet SPSS aber mit einer viel größeren Anzahl von Nachkommastellen.

Auch bei Stringvariablen ist die vorgegebene Variablenbreite »8« (natürlich ohne Dezimalstellen). Im Unterschied zu numerischen Variablen, bei denen die Variablenbreite nur bei der Ausgabe in eine Textdatei bedeutsam ist, ist bei Stringvariablen die Variablenbreite wichtiger: Bei ihnen wird hierdurch die maximale Anzahl der gespeicherten Zeichen festgelegt. Dies führt in unserem Beispiel bei *geschlecht* dazu, dass die intern als »maennlich«, also mit neun Zeichen, dargestellte Ausprägung »männlich« nur als »männlic« im Datenfenster angegeben wird (siehe hierzu etwa Abbildung 4.15 bis Abbildung 4.18).

Wenn Sie jetzt nachträglich bei *geschlecht* die Variablenbreite erhöhen, zum Beispiel von »8« auf »10« (siehe hierzu Datei *PERSONEN4A.SAV* ), können Sie im Dateneditor »männlic« durch »männlich« ersetzen. Dies müssen Sie allerdings in SPSS manuell erledigen. Mit einem Trick sparen Sie sich an dieser Stelle Arbeit: Kopieren Sie aus dem SPSS-Dateneditor die Spalte mit den *geschlecht*-Ausprägungen zum Beispiel nach EXCEL, verändern Sie dort mit SUCHEN UND ERSETZEN die Ausprägung »männlic« in »männlich« und kopieren Sie das Ganze dann wieder zurück nach SPSS. Mit dem SPSS-Befehl RECODE haben Sie im

Übrigen noch eine weitere Alternative innerhalb von SPSS – doch dazu später (in Kapitel 5) mehr!

# *Weitere Einstellungen*

In den dann folgenden Spalten des Variableneditors können Sie Folgendes festlegen:

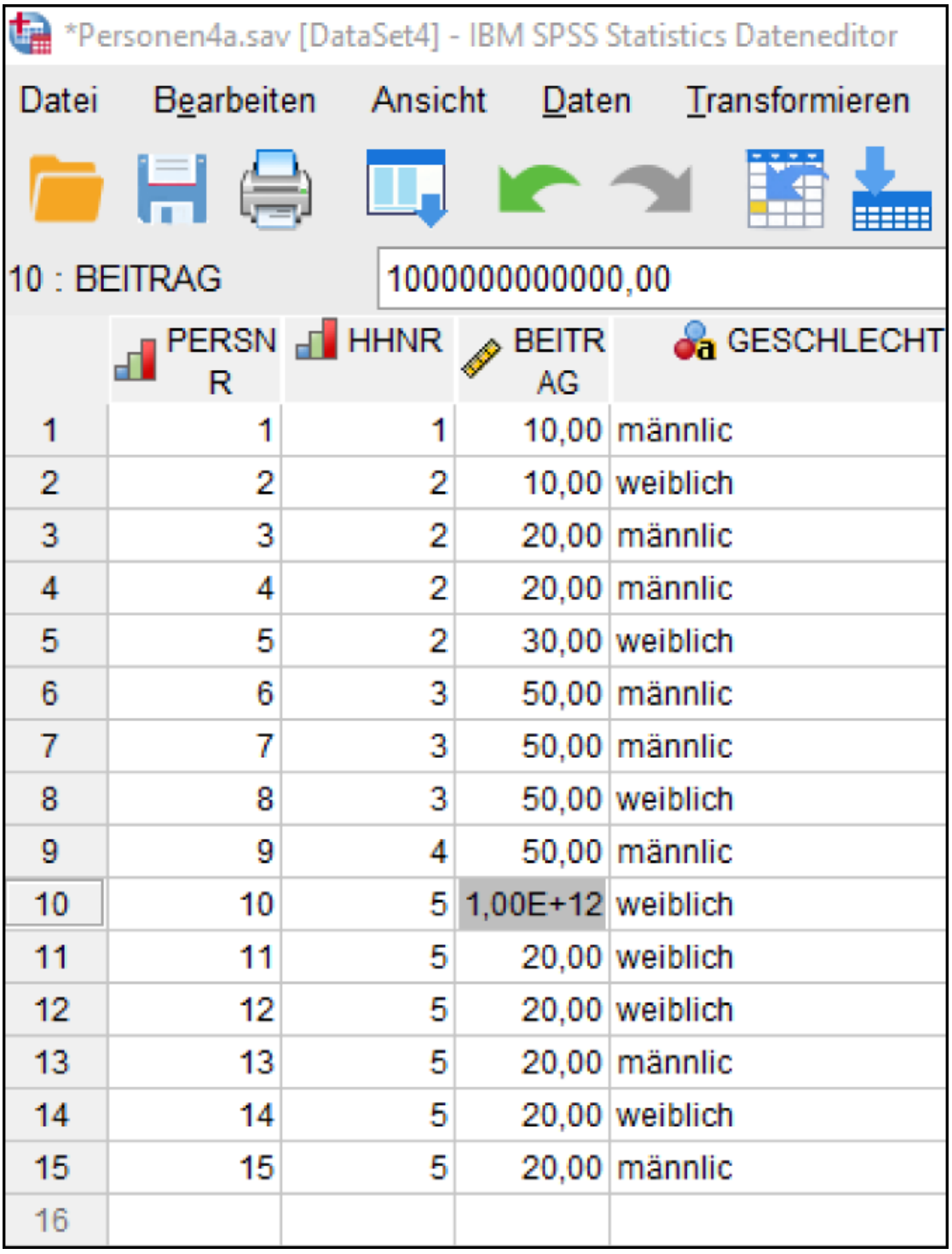
- ✔ BESCHRIFTUNG: Hier können Sie die einzelnen Variablen näher beschreiben, wie dies in Abbildung 4.20 geschehen ist. Für die Variable *persnr* ist dort zum Beispiel das Label »Personennummer« vergeben worden.
- ✔ WERTE: Hier werden für kategoriale Variablen, das heißt für Variablen mit verschiedenen ganzzahligen Ausprägungen (zum Beispiel in Klassen mit den Ausprägungen 1 für die erste Klasse, 2 für die zweite Klasse und so weiter) die inhaltlichen Bedeutungen der einzelnen Kategorien angegeben. Hätten Sie *geschlecht* in Zahlenausdrücke transformiert (zum Beispiel: 1 = männlich, 2 = weiblich), hätte dies hier angegeben werden können; später – im nächsten Abschnitt dieses Kapitels – werden Sie eine solche Transformation noch im Dateneditor durchführen (also bitte noch etwas Geduld an dieser Stelle – danke!).

Trotz der Vereinbarung bestimmter Wertelabels (zum Beispiel: *geschlecht:* 1 = männlich, 2 = weiblich) lässt SPSS im Übrigen auch andere Zahlen im Dateneditor bei der betreffenden Variablen zu, unabhängig davon, ob das inhaltlich sinnvoll ist oder nicht (etwa beim Geschlecht die Eingabe einer »3«).

- ✔ FEHLEND: Werte, die zum Beispiel deshalb im Datensatz fehlen, weil von den Befragten keine Antwort vorliegt, können hier als solche deklariert werden. Es ist üblich, fehlende Werte mit einem ansonsten nicht (zumindest für die betreffende Variable) im Datensatz vorkommenden Wert wie »-9« zu kennzeichnen. Fehlende

Werte können entweder systemdefiniert oder benutzerdefiniert sein. Hier sind die »benutzerdefinierten« fehlenden Werte gemeint. Während systemdefinierte fehlende Werte allgemein dann vorliegen, wenn eine Variablenausprägung fehlt, handelt es sich dann um benutzerdefinierte fehlende Werte, wenn unterschiedliche Arten von fehlenden Werten gekennzeichnet werden sollen. So kann zum Beispiel zwischen bewussten Antwortverweigerungen aus persönlichen Gründen auf eine Frage und fehlenden Antworten aus Desinteresse zumindest an Teilen einer Befragung unterschieden werden.

- SPALTEN: Hiermit ist einfach die Spaltenbreite gemeint, mit der die betreffende Variable im Dateneditor dargestellt wird. Bei zu geringer Spaltenbreite werden numerische Variablen in SPSS in Exponentialschreibweise dargestellt. Reicht auch diese Darstellungsform von der Breite her nicht aus, werden von SPSS nach dem »E« (das Sie hier mit »10 hoch« übersetzen können) Pünktchen gesetzt. Fügen Sie doch einfach einmal im Datensatz *PERSONEN4A.SAV* im Dateneditor bei *beitrag* in der zehnten Zeile den Wert 1 000 000 000 000 (in Worten: eine Billion) ein, und, wie aus Abbildung 4.19 ersichtlich, sofort wird für diese Zelle 1,00E+12 angezeigt. Diese neue Anzeige resultiert daraus, dass Sie eine Zahl mit 13 Stellen in eine Datenzelle mit 8 Stellen (davon 2 Nachkommastellen) geschrieben haben.



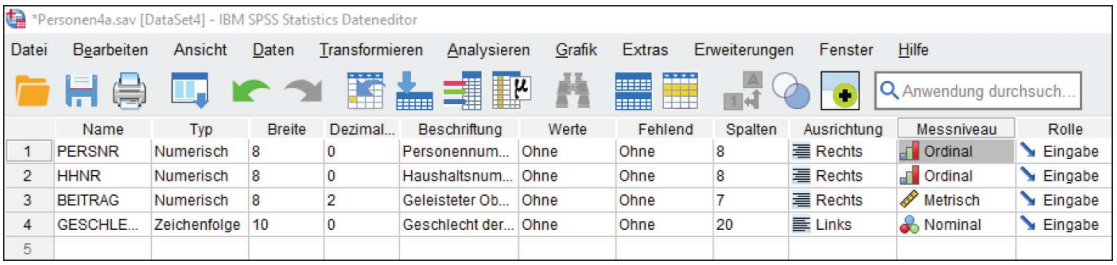

| | PERSNR | HHNR | BEITRAG | GESCHLECHT |
|---|---|---|---|---|
| 1 | 1 | 1 | 10,00 | männlic |
| 2 | 2 | 2 | 10,00 | weiblich |
| 3 | 3 | 2 | 20,00 | männlic |
| 4 | 4 | 2 | 20,00 | männlic |
| 5 | 5 | 2 | 30,00 | weiblich |
| 6 | 6 | 3 | 50,00 | männlic |
| 7 | 7 | 3 | 50,00 | männlic |
| 8 | 8 | 3 | 50,00 | weiblich |
| 9 | 9 | 4 | 50,00 | männlic |
| 10 | 10 | 5 | 1,00E+12 | weiblich |
| 11 | 11 | 5 | 20,00 | weiblich |
| 12 | 12 | 5 | 20,00 | weiblich |
| 13 | 13 | 5 | 20,00 | männlic |
| 14 | 14 | 5 | 20,00 | weiblich |
| 15 | 15 | 5 | 20,00 | männlic |
| 16 | | | | |

**Abbildung 4.19:** Zahlendarstellung im Dateneditor bei zu geringer Spaltenbreite (*PERSONEN4A.SAV*)

✔ AUSRICHTUNG: Diese Rubrik ist ohne größere inhaltliche Bedeutung und meint im Grunde genommen lediglich, ob die

Variablenausprägungen in den einzelnen Zellen links- oder rechtsbündig beziehungsweise mittig ausgerichtet sind. Zahlen werden im Dateneditor standardmäßig rechtsbündig, Stringvariablen linksbündig dargestellt (siehe auch Abbildung 4.14).

- MESSNIVEAU: Zur näheren Kennzeichnung des Skalenniveaus der einzelnen Variablen stehen hier die in der Statistik üblichen Kategorien NOMINAL, ORDINAL und METRISCH zur Verfügung. Im Beispielfall (*PERSONEN4A.SAV* ) sind *persnr* und *hhnr* ordinalskaliert, weil sie zwar numerisch sind, bei ihnen aber die Abstände der Merkmalsausprägungen untereinander keine Rolle spielen. Demgegenüber sind bei der numerischen Variablen *beitrag* auch die Abstände von Bedeutung; sie ist daher metrisch skaliert. Bei *geschlecht* schließlich handelt es sich um eine nominalskalierte Variable, weil die beiden Ausprägungen »männlich« und »weiblich « als gleichwertig gelten. Das Messniveau bei den einzelnen Variablen wird in SPSS in Form von *Piktogrammen* folgendermaßen angezeigt: metrische Variable: Metermaß, ordinale Variable: farbiges, abgestuftes (Häufigkeits-)Diagramm und nominale Variable: dreifarbige, gleich große (im Sinne von: »gleichwertige«) Kreise (siehe Abbildung 4.20).




| | Name | Typ | Breite | Dezimal... | Beschriftung | Werte | Fehlend | Spalten | Ausrichtung | Messniveau | Rolle |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | PERSNR | Numerisch | 8 | 0 | Personennum... | Ohne | Ohne | 8 | Rechts | Ordinal | Eingabe |
| 2 | HHNR | Numerisch | 8 | 0 | Haushaltsnum... | Ohne | Ohne | 8 | Rechts | Ordinal | Eingabe |
| 3 | BEITRAG | Numerisch | 8 | 2 | Geleisteter Ob... | Ohne | Ohne | 7 | Rechts | Metrisch | Eingabe |
| 4 | GESCHLE... | Zeichenfolge | 10 | 0 | Geschlecht der... | Ohne | Ohne | 20 | Links | Nominal | Eingabe |
| 5 | | | | | | | | | | | |

**Abbildung 4.20:** Skalierungsmöglichkeiten in SPSS: NOMINAL, ORDINAL, METRISCH (*PERSONEN4A.SAV* )

- ROLLE: Hier kann einer Variablen eine Rolle zugewiesen werden, der sie in bestimmten Dialogfeldern nachkommt (etwa bei der Regression als unabhängige Variable). Die Rollenzuweisung bezieht sich nur auf Dialogfelder und hat keine Auswirkungen auf die Befehlssyntax. Man unterscheidet folgende Rollen: EINGABE (als Standardeinstellung; zum Beispiel als unabhängige Variable bei einer

Regression), ZIEL (= Ausgabe; zum Beispiel als abhängige Variable bei einer Regression), BEIDES (= Ein- und Ausgabe), OHNE (keine Rolle für die betreffende Variable zugewiesen), PARTITIONIEREN (für separate Stichproben; das heißt, die betreffende Variable dient zur gruppenbezogenen Aufteilung der Stichprobe) und AUFTEILEN (diese Rolle ist nur für ganz spezielle SPSS-Operationen interessant). Wie aus Abbildung 4.21 ersichtlich ist, sind in *PERSONEN4A.SAV* alle Variablen bei ROLLE mit der Standardeinstellung EINGABE versehen.

**Abbildung 4.21:** Zur ROLLE-Formatierung (*PERSONEN4A.SAV*)

Sie können relativ problemlos Attribute – das heißt die Formatierungseinstellungen – einer Variablen auf andere Variablen übertragen. Möchten Sie alle Attribute einer Variablen auf andere Variablen übertragen, markieren Sie im Variableneditor einfach die Zeilennummer (am linken Bildrand) der Quellvariablen, wählen KOPIEREN und markieren anschließend per Mausklick (in Verbindung mit der ⇧ -beziehungsweise der Strg -Taste) die Zielvariable(n). Anschließend müssen Sie nur noch EINFÜGEN auswählen.

Möchten Sie ein einzelnes Attribut einer Variablen auf andere Variablen übertragen, gehen Sie wie folgt vor: Wählen Sie die Quell-Attributszelle und danach KOPIEREN. Schließlich wählen Sie bei den Ziel-Attributszellen EINFÜGEN – fertig!

# *Variablen umcodieren*

Wie eben bereits am Beispiel der Variablen *geschlecht* angedeutet worden ist, ist es manchmal sinnvoll, Variableninhalte zu verändern. Man spricht dann von **Umcodieren**. Ein anderer Ausdruck für das Umcodieren ist: **Transformation**.

## *Umcodieren in derselben Variablen*

Eine Transformations-Möglichkeit ist dabei die **Veränderung von Werten in ein und derselben Variablen**. Im SPSS-Menü können Sie entsprechend über TRANSFORMIEREN|UMCODIEREN IN DIESELBEN VARIABLEN die Merkmalsausprägungen einzelner Variablen verändern. Ich habe nachfolgend – sozusagen endlich, nachdem ich dieses oben ja schon ein paar Mal angekündigt hatte – in der Variablen *geschlecht* (in der an dieser Stelle von *var001* eben auf *geschlecht* veränderten Datendatei *PERSONEN4A.SAV* ) die Ausprägungen »männlic« und »weiblich« in »1« und »2« umcodiert.

In Abbildung 4.22 ist zunächst die Variable ausgewählt worden, für die eine Umcodierung erfolgen soll: *geschlecht*. Nach dem Auswählen von ALTE UND NEUE WERTE… sind dann im sich öffnenden Dialogfenster die gewünschten Umcodierungen vorgenommen worden (siehe Abbildung 4.23).

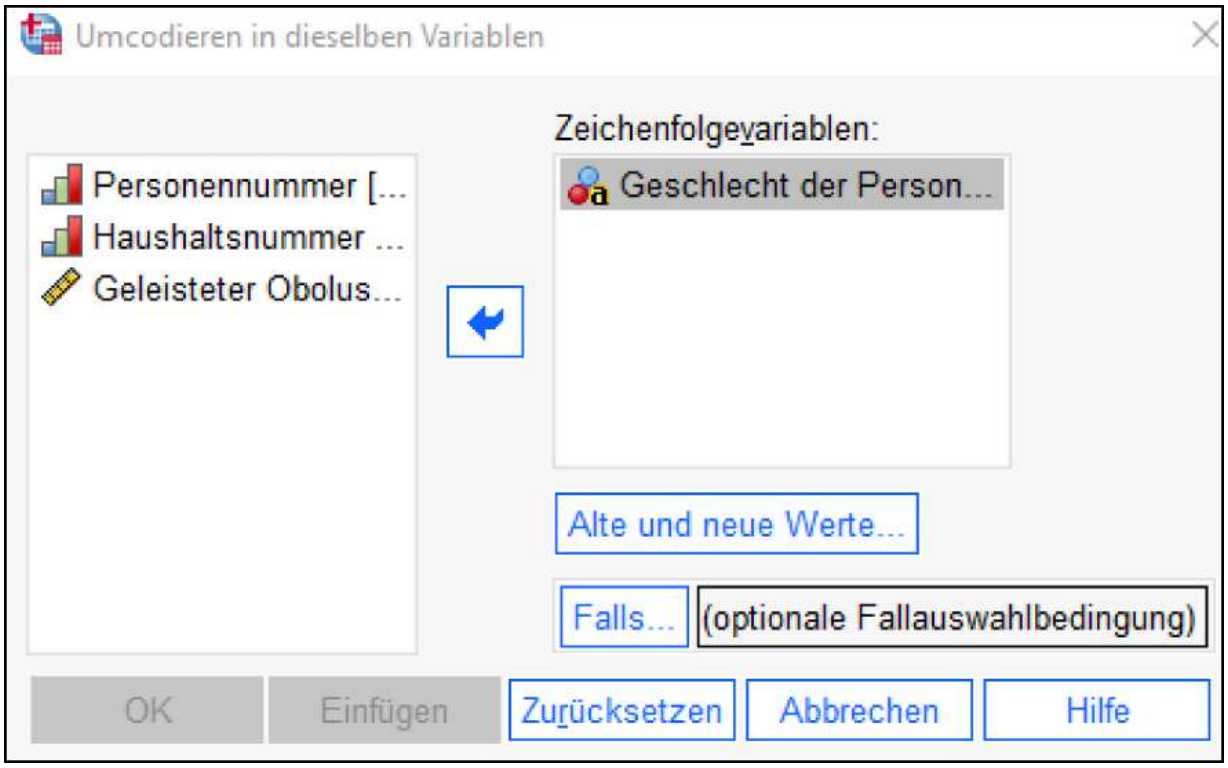
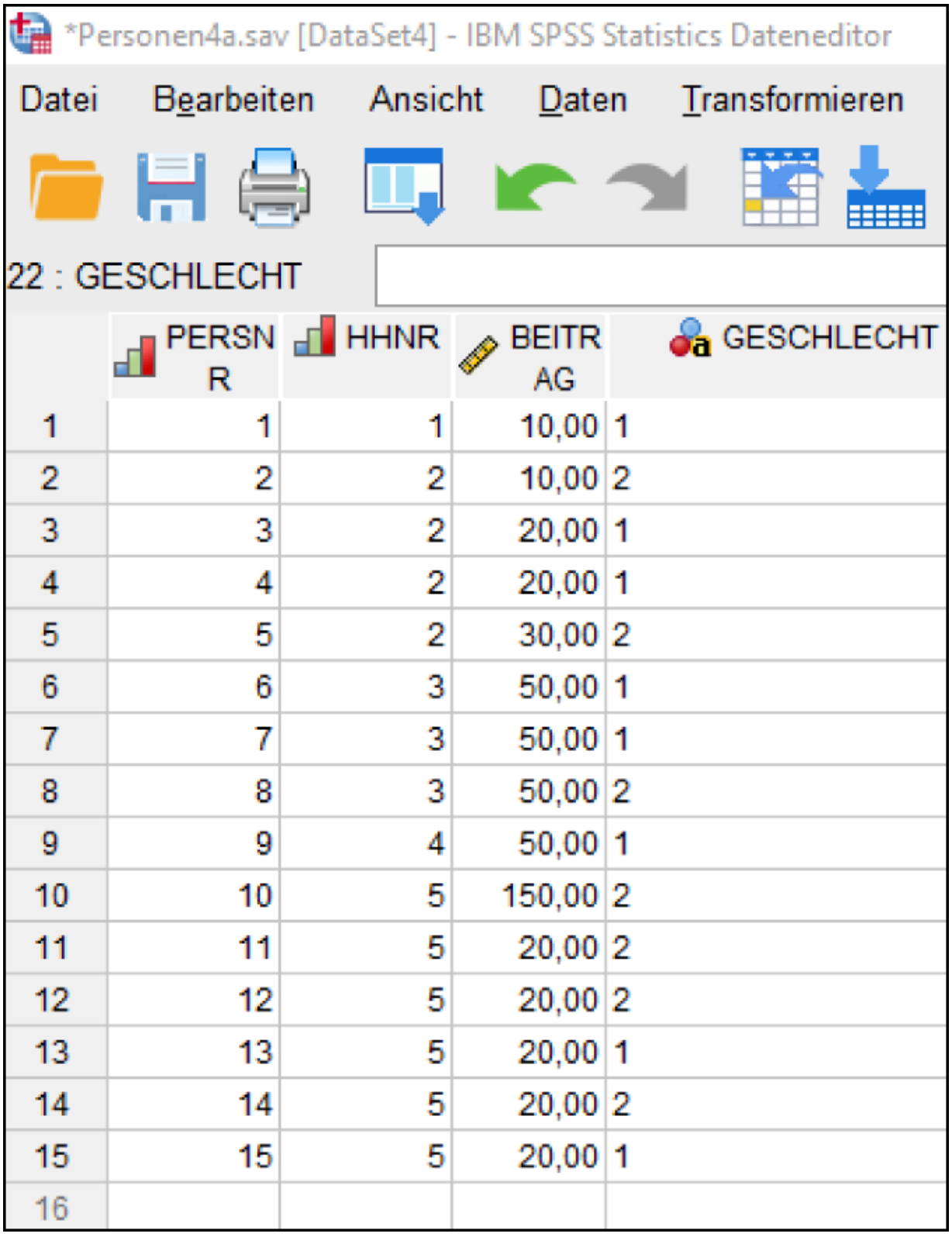



**Abbildung 4.22:** Umcodieren in derselben Variablen am Beispiel der nominalskalierten Variablen *geschlecht* (1)

**Abbildung 4.23:** Umcodieren in derselben Variablen am Beispiel der nominalskalierten Variablen *geschlecht* (2)

Das Ergebnis des Umcodierens in unserem Beispiel mit der Variablen *geschlecht*, wie es sich im Dateneditor darstellt, finden Sie in Abbildung 4.24. Sie sehen, dass sich jetzt überall dort, wo in der vierten Spalte vorher »männlic« stand, eine »1« befindet, und dass überall dort, wo zuvor »weiblich« stand, nunmehr »2« steht. Außerdem sehen Sie anhand von Abbildung 4.24, dass die Variable *geschlecht* – trotz ihrer »Zahlen-Ausprägungen« – immer noch als Stringvariable definiert ist. Das erkennen Sie daran, dass die neuen Ausprägungen von *geschlecht* weiterhin linksbündig angeordnet sind.




| | PERSNR | HHNR | BEITRAG | GESCHLECHT |
|---|---|---|---|---|
| 1 | 1 | 1 | 10,00 | 1 |
| 2 | 2 | 2 | 10,00 | 2 |
| 3 | 3 | 2 | 20,00 | 1 |
| 4 | 4 | 2 | 20,00 | 1 |
| 5 | 5 | 2 | 30,00 | 2 |
| 6 | 6 | 3 | 50,00 | 1 |
| 7 | 7 | 3 | 50,00 | 1 |
| 8 | 8 | 3 | 50,00 | 2 |
| 9 | 9 | 4 | 50,00 | 1 |
| 10 | 10 | 5 | 150,00 | 2 |
| 11 | 11 | 5 | 20,00 | 2 |
| 12 | 12 | 5 | 20,00 | 2 |
| 13 | 13 | 5 | 20,00 | 1 |
| 14 | 14 | 5 | 20,00 | 2 |
| 15 | 15 | 5 | 20,00 | 1 |
| 16 | | | | |

**Abbildung 4.24:** Ergebnis des Umcodierens in derselben Variablen am Beispiel der nominalskalierten Variablen *geschlecht* (in *PERSONEN4A.SAV* )

Naheliegenderweise kommt ein entsprechendes Umcodieren in ein und derselben Variablen nur bei nominal- oder ordinalskalierten Variablen beziehungsweise bei kategorialen Variablen infrage. Bei einer metrisch skalierten Variablen ist diese Form des Umcodierens nur dann sinnvoll, wenn für die betreffende Variable zuvor Klassen gebildet wurden, das heißt, wenn sie in eine kategoriale Variable umgewandelt wurde.

Das Umwandeln einer metrischen in eine kategoriale Variable selbst ist allerdings eine andere Form des Umcodierens: Hier handelt es sich beispielsweise um das Umwandeln der Werte einer Einkommensvariablen mit sehr vielen verschiedenen Ausprägungen (Einkommenswerten) in eine *neue*, eine kategoriale Variable, zum Beispiel in einige Einkommensklassen (etwa: »Bis unter 1.000 Euro/Monat«, »1.000 bis unter 2.000 Euro/Monat«, »2.000 bis unter 3.000 Euro/Monat« und so weiter). Das Umcodieren in eine neue Variable wird im nächsten Abschnitt dieses Kapitels erläutert.

Sie sollten *keinesfalls* bei Ihren **Transformationen** in Bezug auf ein und dieselben Variablen dieselbe Datei als *Quell-* und als *Zieldatei* verwenden. Tun Sie das nämlich, überschreiben Sie Ihre Originaldaten durch die veränderten Daten, und Ihre Originaldaten sind dann sozusagen unwiederbringlich weg. Es kann aber natürlich sein, dass Sie für andere statistische Auswertungszwecke später noch einmal auf Ihre Originaldaten zurückgreifen möchten, was im Falle des Überschreibens der Original- durch die veränderten Daten nicht mehr ginge. Setzen Sie am besten für die Rohdatendatei ein Schreibschutzattribut (beispielsweise im Datei-Browser von Windows unter EIGENSCHAFTEN der betreffenden Datei, zum Beispiel also bei *PERSONEN.SAV*). Wenn Sie ganz besonders vorsichtig veranlagt sind, sollten Sie Ihre Rohdaten in mindestens zwei Dateien speichern (eventuell auch hier noch jeweils mit Schreibschutzattribut versehen). Dann haben Sie immer noch mindestens eine Rohdatendatei in der Hinterhand, wenn mal was mit einer Rohdatendatei schiefgehen sollte. Sollten Sie einen Schreibschutz nicht eingerichtet haben, haben Sie immer noch die

Möglichkeit, im Falle von Datenveränderungen/-transformationen bei der SPSS-Systemanfrage am Ende Ihrer Sitzung, ob Sie Änderungen wirklich durchführen und abspeichern möchten, mit NEIN zu antworten.

## *Umcodieren in einer neuen Variablen*

Neben der Variante des Umcodierens in ein und derselben Variablen gibt es die Variante des **Umcodierens in einer neuen Variablen**. Eine solche Variablentransformation in einer neuen Variablen können Sie im SPSS-Menü über TRANSFORMIEREN|UMCODIEREN IN ANDERE VARIABLEN vornehmen. Dies läuft im Grunde genommen wie beim Umcodieren in ein und derselben Variablen ab, nur mit dem Unterschied, dass Sie die neue Variable anders als die alte bezeichnen.

Möchten Sie eine metrisch skalierte Variable in eine kategoriale umwandeln, bietet sich das Umcodieren in einer neuen Variablen geradezu an. Im Folgenden wird in der Datendatei *PERSONEN4A.SAV* die (metrische) Variable *beitrag* in vier Klassen und in der neuen (kategorialen) Variablen *beitrag_klasse* eingestuft: 1 = Unter 10,00 Euro, 2 = 10,00 bis unter 20,00 Euro, 3 = 20,00 bis unter 30,00 Euro, 4 = 30,00 Euro und mehr. Hierzu wählen Sie im Dialogfeld UMCODIEREN IN ANDERE VARIABLEN auf der linken Seite die Variable *beitrag* aus (siehe Abbildung 4.25), um sie dann im nächsten Schritt in die Rubrik EINGABEVARIABLE → AUSGABEVARIABLE verschieben zu können.

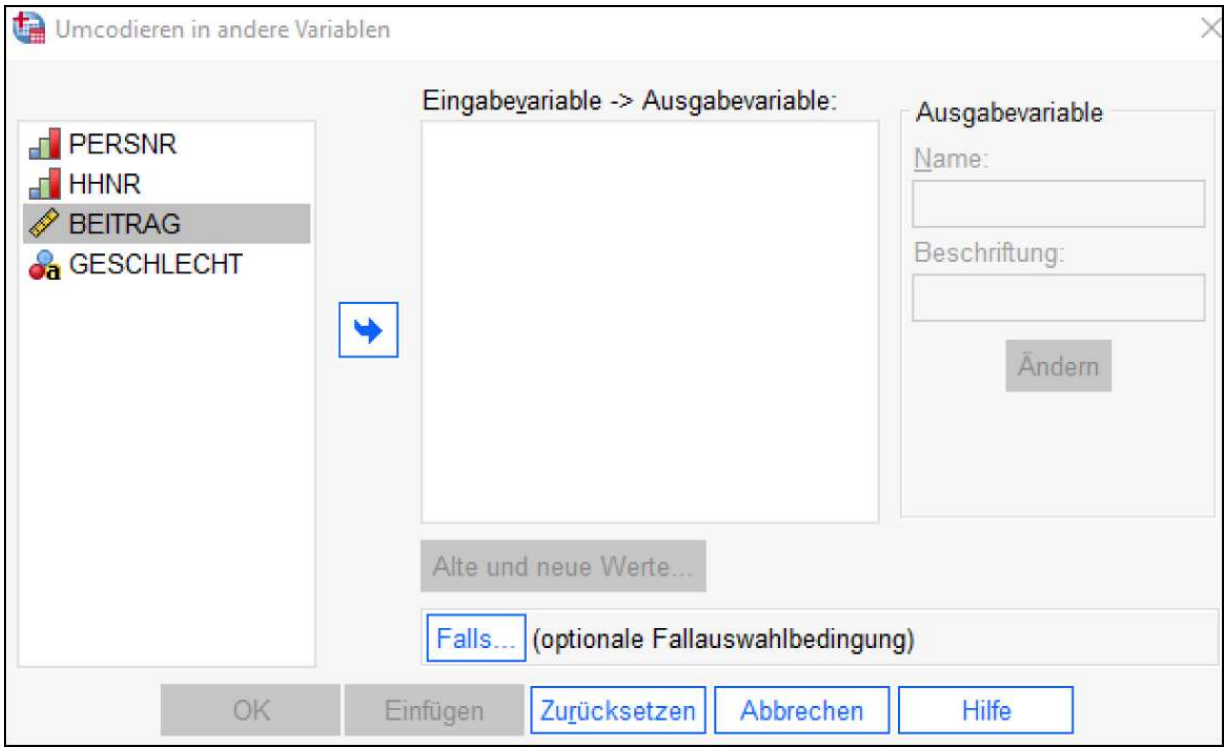



**Abbildung 4.25:** Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (1)

Da es sich bei *beitrag* um eine numerische Variable handelt, ändert sich das Eingabefeld zu NUMERISCHE VARIABLE → AUSGABEVARIABLE (siehe Abbildung 4.26).

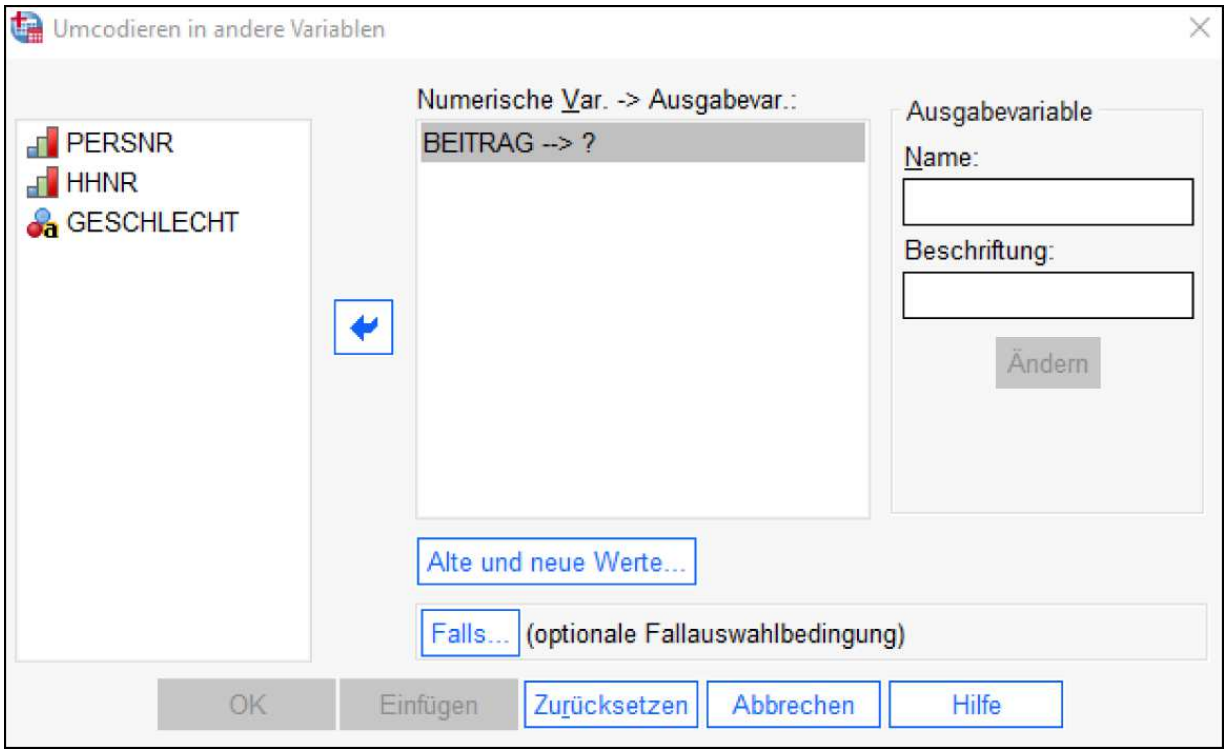



**Abbildung 4.26:** Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (2)

Wie aus Abbildung 4.26 hervorgeht, fehlt noch die Angabe des Namens der neuen Variablen. Das erkennen Sie daran, dass in der Rubrik NUMERISCHE VARIABLE → AUSGABEVARIABLE anstelle des betreffenden Namens noch ein Fragezeichen steht. Diesen unbefriedigenden Zustand ändern Sie bitte jetzt ganz schnell, indem Sie ganz rechts im Dialogfeld unter AUSGABEVARIABLE den gewünschten Namen eingeben, hier: *beitrag_klasse*. Damit Sie später auch noch wissen, welchen Inhalt diese (Ausgabe-)Variable hat, tragen Sie bei BESCHRIFTUNG »Beitragsklassen« ein (siehe Abbildung 4.27). Diese Beschriftung wird dann nämlich von SPSS im Variableneditor in der Beschriftungs-Rubrik übernommen.

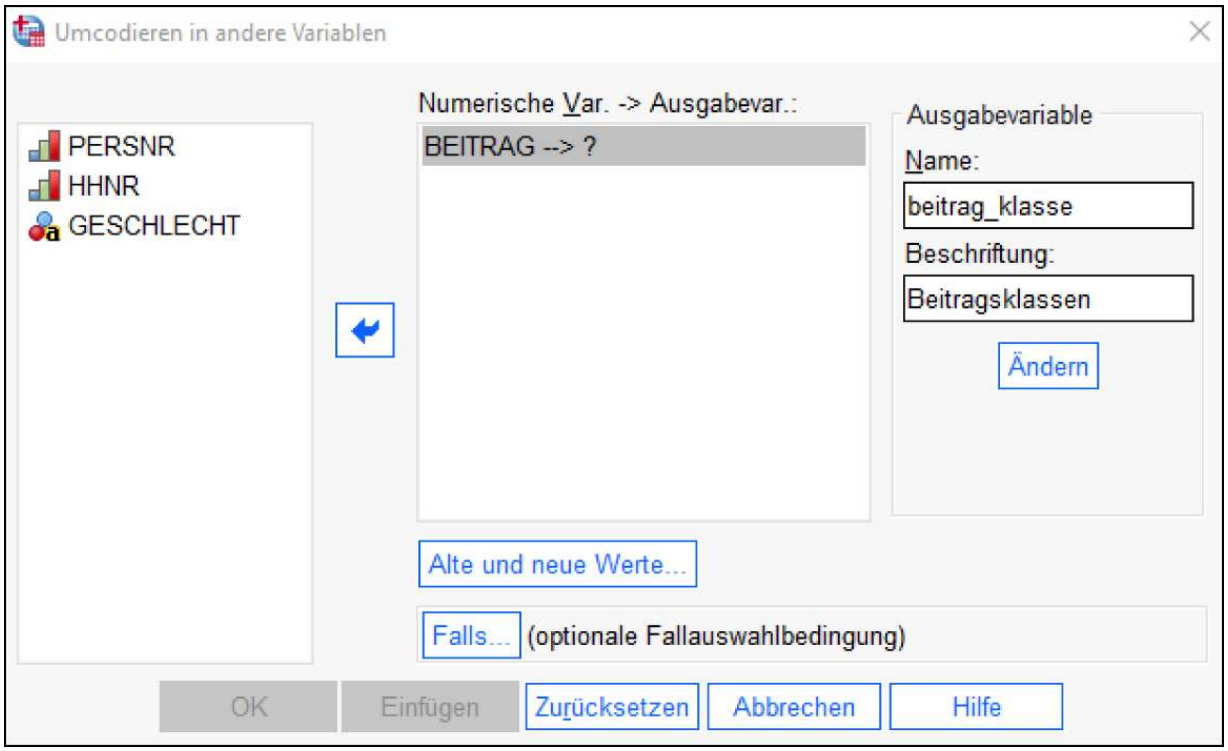



**Abbildung 4.27:** Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (3)

Damit das Umcodieren in einer neuen Variablen überhaupt vollzogen werden kann, müssen Sie im zugehörigen Dialogfeld unbedingt auf der rechten Seite auf die ÄNDERN-Taste klicken. Ansonsten wird der Umcodierungsvorgang nicht in die Wege geleitet! In Abbildung 4.28 ist die ÄNDERN-Taste angeklickt worden, und wie Sie sehen, ist daraufhin das Fragezeichen in der Rubrik NUMERISCHE VARIABLE → AUSGABEVARIABLE verschwunden. Es ist jetzt durch die neue Variable *beitrag_klasse* ersetzt worden. Das war ja unser Ziel!

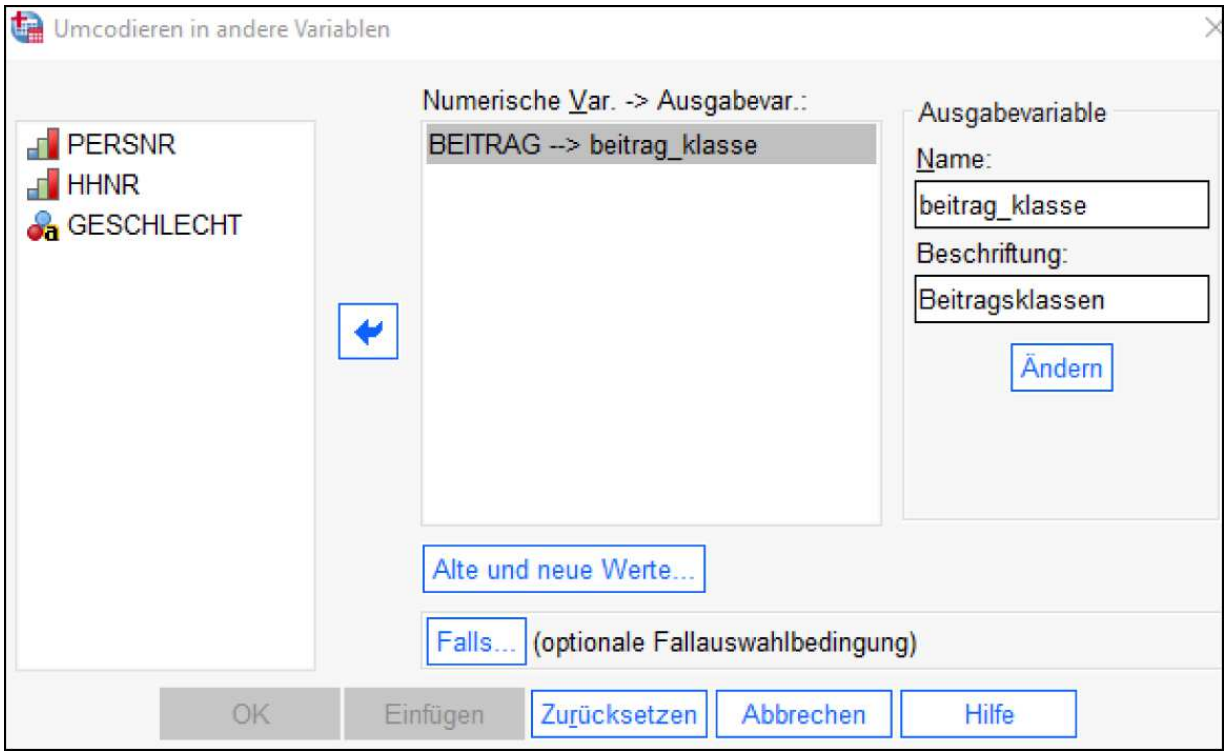



**Abbildung 4.28:** Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (4)

Nach diesen Vorarbeiten kann es jetzt endlich mit dem eigentlichen Umcodieren losgehen. Dazu klicken Sie im Dialogfeld von Abbildung 4.28 die Taste ALTE UND NEUE WERTE an, und es öffnet sich ein neues Dialogfeld, und zwar das in Abbildung 4.29 dargestellte. Sie geben jetzt bitte jeweils unter ALTER WERT den alten Wert ein. In unserem Fall der Umwandlung einer metrischen, nicht-klassifizierten in eine klassifizierte Variable handelt es sich hierbei um Wertebereiche, und zunächst geben Sie für die unterste Klasse unter SPANNWEITE »0 BIS 9,99« ein. Anschließend ordnen Sie diesem Wertebereich unter NEUER WERT die Klassenbezeichnung »1« zu und klicken auf die HINZUFÜGEN-Taste. Dieses machen Sie analog für die Beitragsklassen 2 (»10 BIS 19,99«), 3 (»20 BIS 29,99«) und 4 (»30 und höher«, was Sie dadurch bewerkstelligen, dass Sie bei SPANNWEITE, WERT BIS GRÖSSTER »30« für die Untergrenze dieses Bereichs eingeben). Sie müssten auf diese Art und Weise Abbildung 4.29 erhalten.

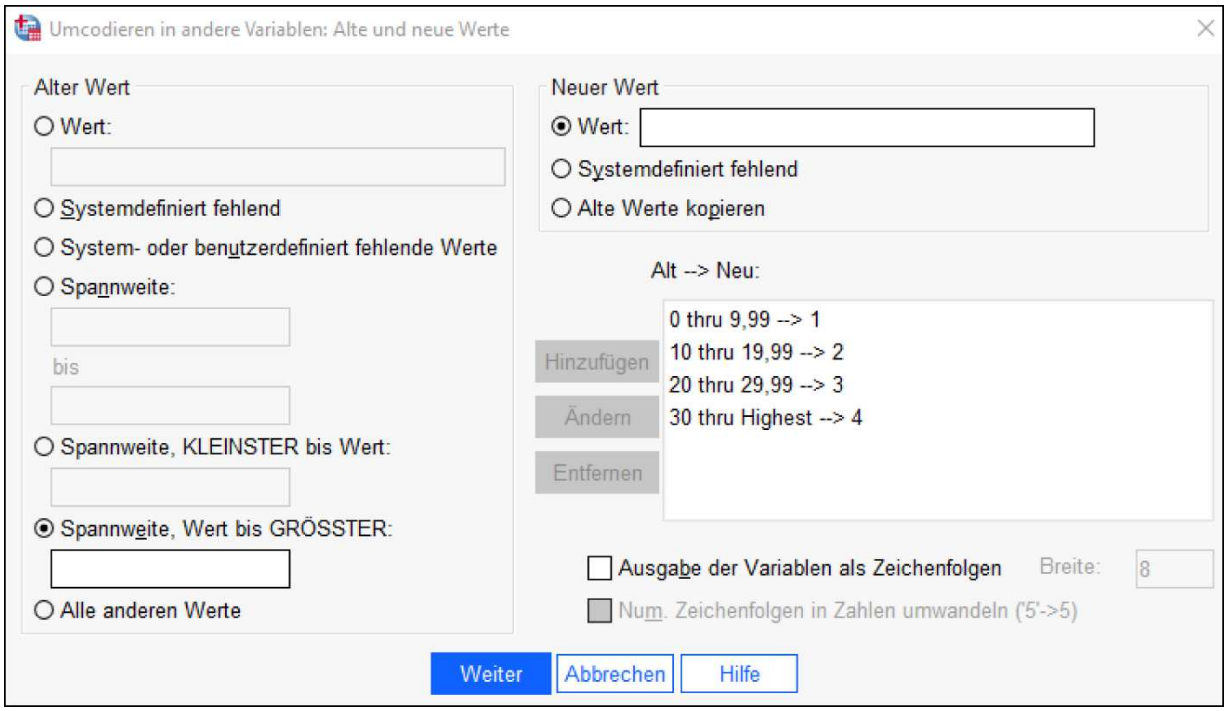
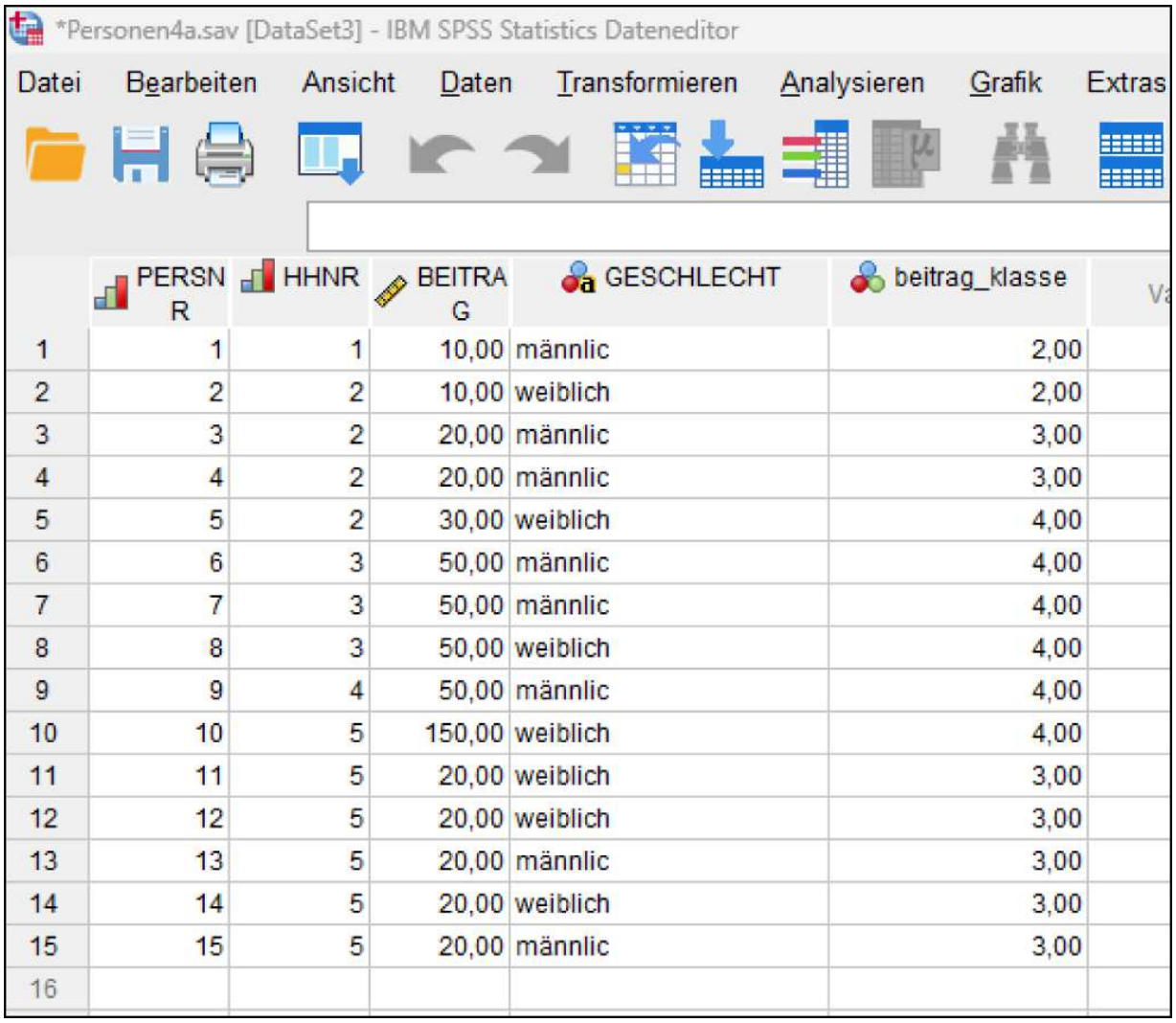



**Abbildung 4.29:** Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (5)

Das Ergebnis des vorstehenden Umcodierens im Dateneditor ist in Abbildung 4.30 dargestellt. Beispielsweise hat die erste Person (in der ersten Datenzeile) einen Beitrag von 10,00 Euro entrichtet. Dieser Betrag fällt nach der vorstehenden Definition in die zweite Beitragsklasse; deshalb steht bei der ersten Person in Abbildung 4.30 bei *beitrag_klasse* auch eine »2«.




| | PERSNR | HHNR | BEITRAG | GESCHLECHT | beitrag_klasse |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 10,00 | männlic | 2,00 |
| 2 | 2 | 2 | 10,00 | weiblich | 2,00 |
| 3 | 3 | 2 | 20,00 | männlic | 3,00 |
| 4 | 4 | 2 | 20,00 | männlic | 3,00 |
| 5 | 5 | 2 | 30,00 | weiblich | 4,00 |
| 6 | 6 | 3 | 50,00 | männlic | 4,00 |
| 7 | 7 | 3 | 50,00 | männlic | 4,00 |
| 8 | 8 | 3 | 50,00 | weiblich | 4,00 |
| 9 | 9 | 4 | 50,00 | männlic | 4,00 |
| 10 | 10 | 5 | 150,00 | weiblich | 4,00 |
| 11 | 11 | 5 | 20,00 | weiblich | 3,00 |
| 12 | 12 | 5 | 20,00 | weiblich | 3,00 |
| 13 | 13 | 5 | 20,00 | männlic | 3,00 |
| 14 | 14 | 5 | 20,00 | weiblich | 3,00 |
| 15 | 15 | 5 | 20,00 | männlic | 3,00 |
| 16 | | | | | |

**Abbildung 4.30:** Ergebnis des Umcodierens in einer anderen Variablen (*PERSONEN4A.SAV*)

Sie können bei Verwendung von TRANSFORMIEREN|UMCODIEREN IN ANDERE VARIABLEN gleichzeitig mehrere Variablen umcodieren. Dies ist nachfolgend anhand des Datensatzes *PERSONEN4A.SAV* zusätzlich für die Variable *hhnr* erfolgt. Deren fünf Ausprägungen 1 bis 5 sollen einfach – zugegebenermaßen etwas sinnfrei, aber es geht hier weniger um Inhalte, sondern um Didaktik! – jeweils um den Wert 10 erhöht und in der neuen Variablen *hhnr2* abgelegt werden. Die entsprechenden Vorarbeiten sind in Abbildung 4.31 und Abbildung 4.32 gezeigt.

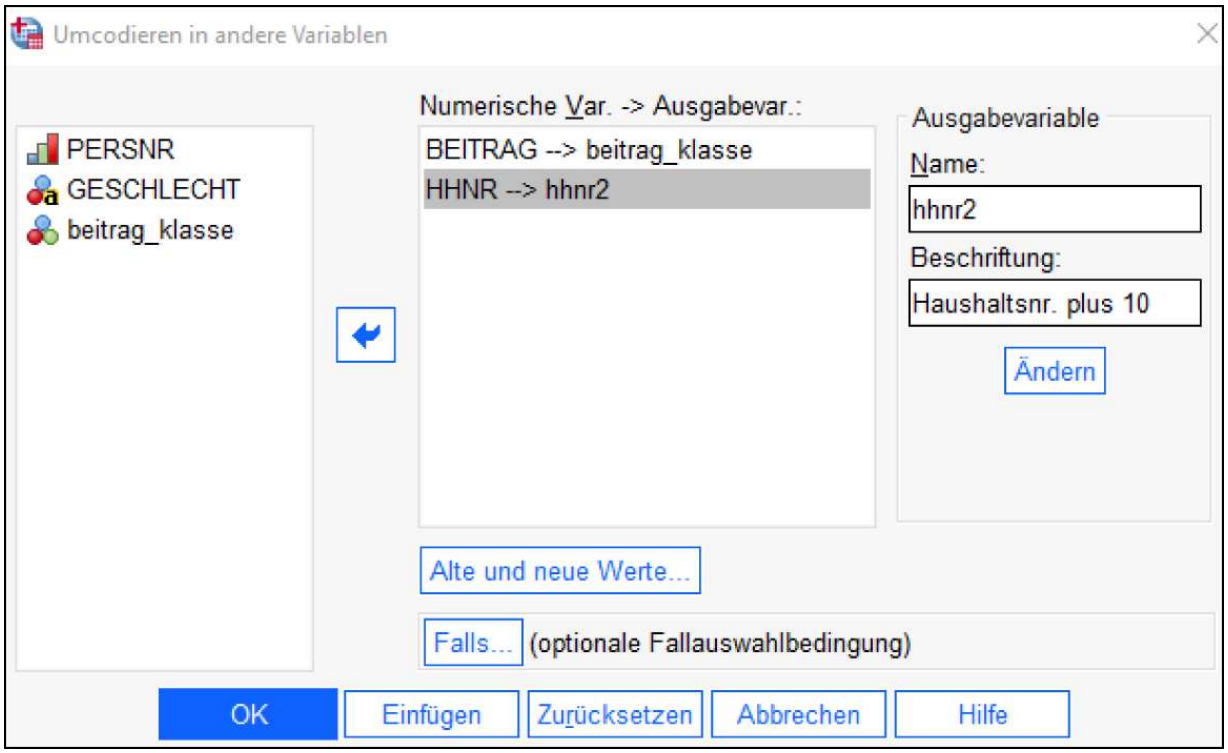



**Abbildung 4.31:** Gleichzeitiges Umcodieren von zwei Variablen in zwei anderen Variablen (*PERSONEN4A.SAV*) (1)

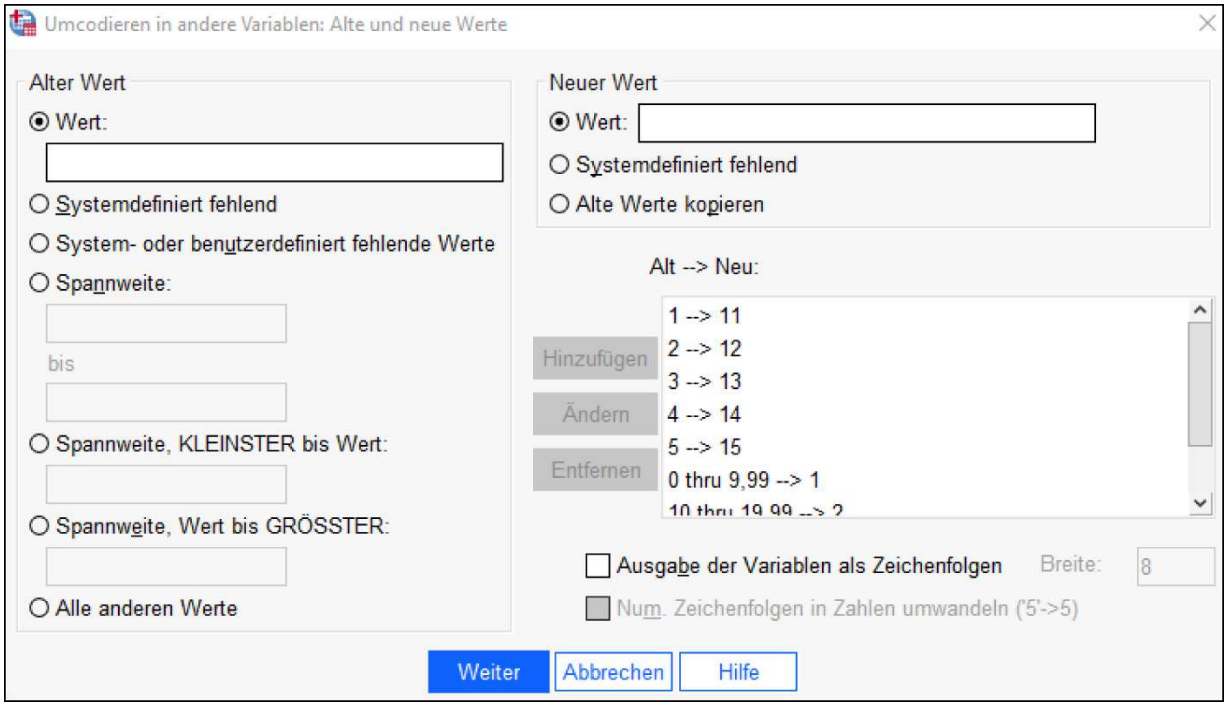
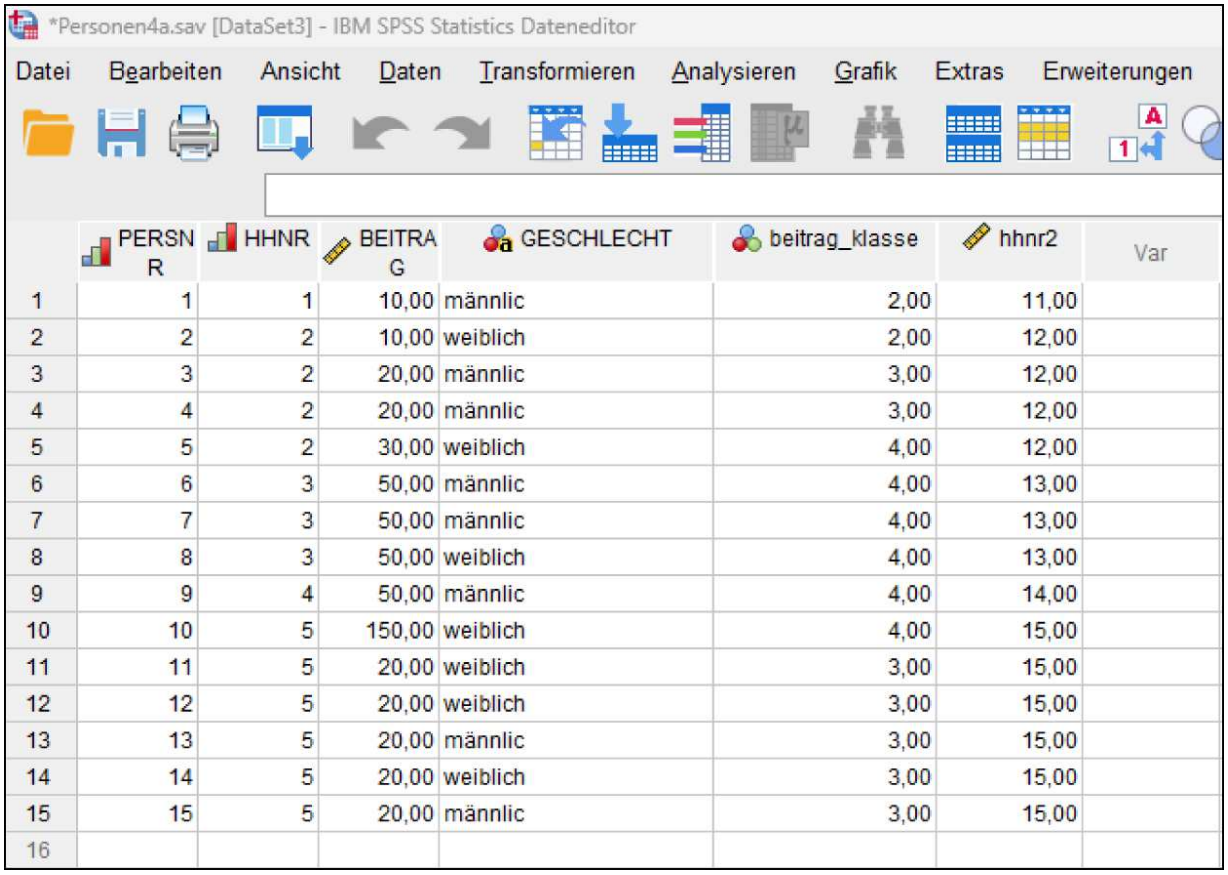



**Abbildung 4.32:** Gleichzeitiges Umcodieren von zwei Variablen in zwei anderen Variablen (*PERSONEN4A.SAV*) (2)

Das erwartete Ergebnis dieses gleichzeitigen Umcodierens zweier Variablen ist in Abbildung 4.33 zu finden. Vom technischen Standpunkt her weist dieses gleichzeitige Umcodieren keinen höheren Schwierigkeitsgrad auf als das Umcodieren einer einzelnen Variablen in einer anderen Variablen.




| | PERSNR | HHNR | BEITRAG | GESCHLECHT | beitrag_klasse | hhnr2 | Var |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 10,00 | männlic | 2,00 | 11,00 | |
| 2 | 2 | 2 | 10,00 | weiblich | 2,00 | 12,00 | |
| 3 | 3 | 2 | 20,00 | männlic | 3,00 | 12,00 | |
| 4 | 4 | 2 | 20,00 | männlic | 3,00 | 12,00 | |
| 5 | 5 | 2 | 30,00 | weiblich | 4,00 | 12,00 | |
| 6 | 6 | 3 | 50,00 | männlic | 4,00 | 13,00 | |
| 7 | 7 | 3 | 50,00 | männlic | 4,00 | 13,00 | |
| 8 | 8 | 3 | 50,00 | weiblich | 4,00 | 13,00 | |
| 9 | 9 | 4 | 50,00 | männlic | 4,00 | 14,00 | |
| 10 | 10 | 5 | 150,00 | weiblich | 4,00 | 15,00 | |
| 11 | 11 | 5 | 20,00 | weiblich | 3,00 | 15,00 | |
| 12 | 12 | 5 | 20,00 | weiblich | 3,00 | 15,00 | |
| 13 | 13 | 5 | 20,00 | männlic | 3,00 | 15,00 | |
| 14 | 14 | 5 | 20,00 | weiblich | 3,00 | 15,00 | |
| 15 | 15 | 5 | 20,00 | männlic | 3,00 | 15,00 | |
| 16 | | | | | | | |

**Abbildung 4.33:** Ergebnis des gleichzeitigen Umcodierens von zwei Variablen in zwei anderen Variablen (*PERSONEN4A.SAV*)

# *Neue Variable berechnen*

Möchten Sie weitergehende mathematische Operationen als zuvor einsetzen, können Sie den Inhalt einer (oder mehrerer) Variablen auch über TRANSFORMIEREN|BERECHNEN in einer neuen Variablen abbilden (beziehungsweise auch in einer bereits bestehenden Variablen – dann überschreiben Sie aber die bisherigen Ausprägungen dieser bestehenden Variablen sozusagen unwiederbringlich!).

Während Sie bislang beim Umcodieren die alten und die neuen Werte jeweils von Hand eingeben mussten, können Sie bei der **Berechnung einer neuen Variablen** in SPSS vorgegebene mathematische Funktionen nutzen und gewissermaßen automatisch Variableninhalte verändern. Um das grundsätzliche Vorgehen zu veranschaulichen, werden in Abbildung 4.34 in der Datei *PERSONEN4A.SAV* die dortigen Euro-Werte der Variablen *beitrag* über den Umrechnungsfaktor 1,95583 (bitte beachten: SPSS erwartet hier die amerikanische Notation; das heißt: anstelle des Kommas einen Punkt!) in DM-Werte in der Variablen *beitrag_dm* umgerechnet.

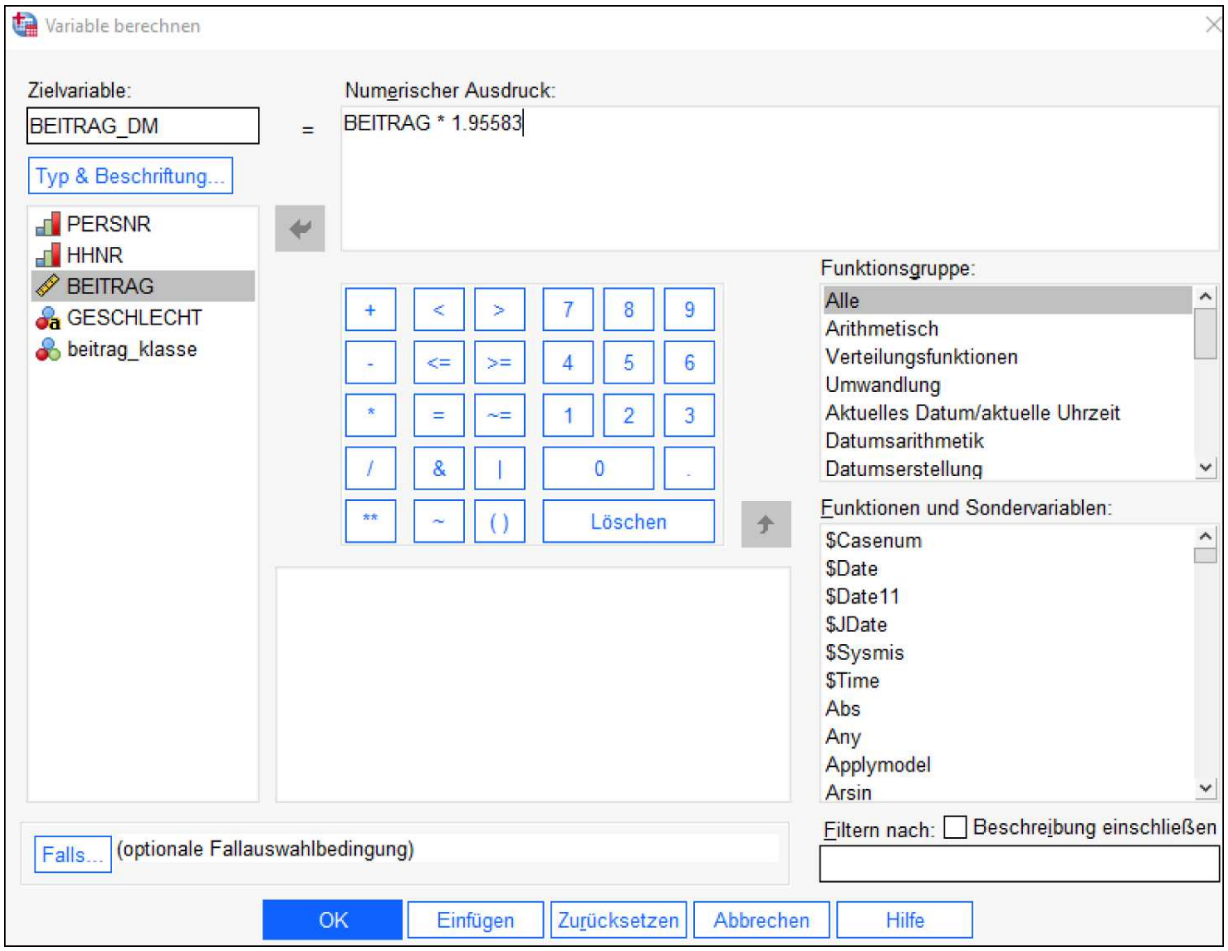
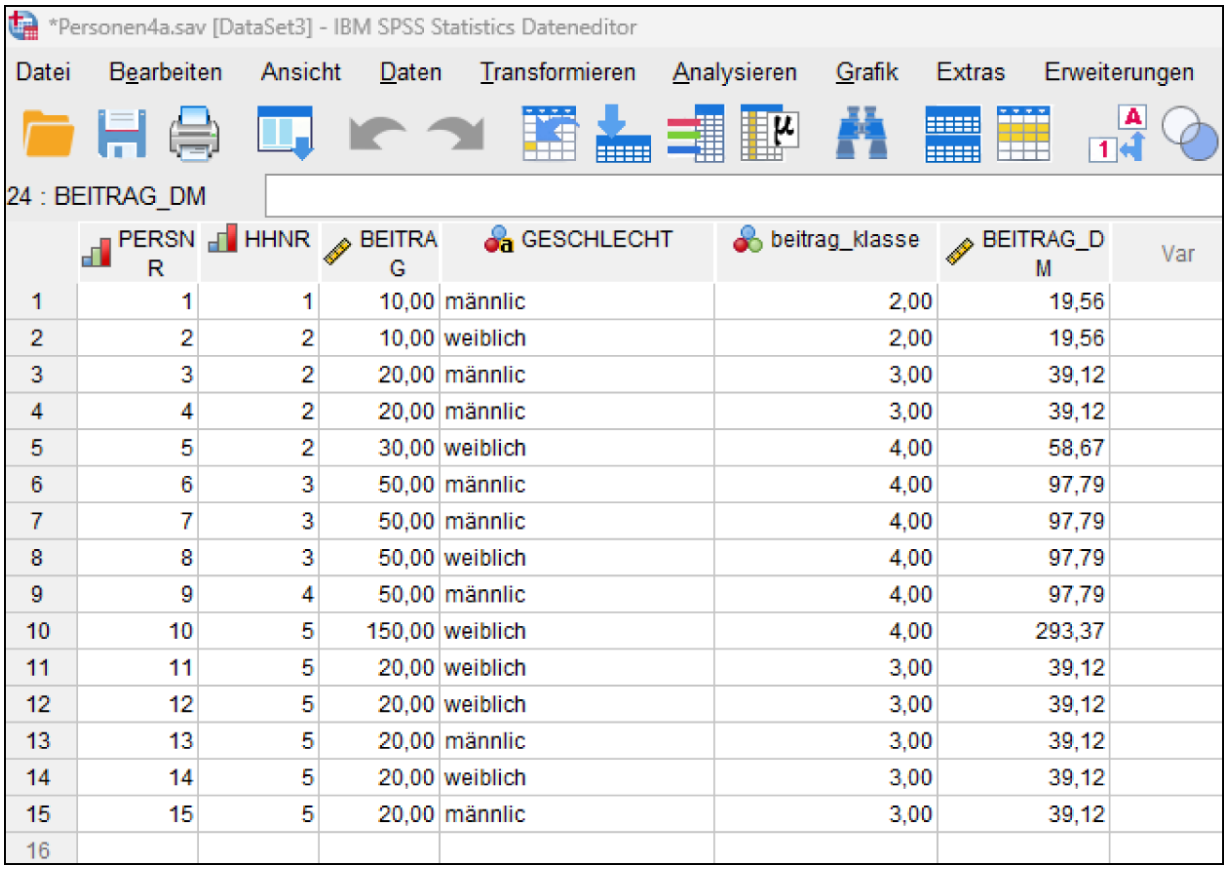



**Abbildung 4.34:** Berechnen einer neuen Variablen (*PERSONEN4A.SAV*)

Als Ergebnis erhalten Sie dann den in Abbildung 4.35 dargestellten Bildschirm in der Datenansicht mit der neuen Variablen *beitrag_dm*. Alle Werte von *beitrag* sind mit dem Wert 1.95583 multipliziert worden,

sodass zum Beispiel für die erste Person (in der ersten Datenzeile) aus den 10,00 Euro nunmehr 19,56 DM geworden sind.




| | PERSNR | HHNR | BEITRAG | GESCHLECHT | beitrag_klasse | BEITRAG_DM | Var |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 10,00 | männlic | 2,00 | 19,56 | |
| 2 | 2 | 2 | 10,00 | weiblich | 2,00 | 19,56 | |
| 3 | 3 | 2 | 20,00 | männlic | 3,00 | 39,12 | |
| 4 | 4 | 2 | 20,00 | männlic | 3,00 | 39,12 | |
| 5 | 5 | 2 | 30,00 | weiblich | 4,00 | 58,67 | |
| 6 | 6 | 3 | 50,00 | männlic | 4,00 | 97,79 | |
| 7 | 7 | 3 | 50,00 | männlic | 4,00 | 97,79 | |
| 8 | 8 | 3 | 50,00 | weiblich | 4,00 | 97,79 | |
| 9 | 9 | 4 | 50,00 | männlic | 4,00 | 97,79 | |
| 10 | 10 | 5 | 150,00 | weiblich | 4,00 | 293,37 | |
| 11 | 11 | 5 | 20,00 | weiblich | 3,00 | 39,12 | |
| 12 | 12 | 5 | 20,00 | weiblich | 3,00 | 39,12 | |
| 13 | 13 | 5 | 20,00 | männlic | 3,00 | 39,12 | |
| 14 | 14 | 5 | 20,00 | weiblich | 3,00 | 39,12 | |
| 15 | 15 | 5 | 20,00 | männlic | 3,00 | 39,12 | |
| 16 | | | | | | | |

**Abbildung 4.35:** Datendatei mit neu berechneter Variablen (*PERSONEN4A.SAV*)

Grundsätzlich können Sie mittels verschiedener **mathematischer Formeln** auf der Basis von TRANSFORMIEREN|VARIABLE BERECHNEN eine neue Variable aus bestehenden Variablen berechnen (beziehungsweise eine bestehende Variable überschreiben, wobei Sie aber immerhin zuvor gefragt werden, ob Sie eine bestehende Variable wirklich überschreiben möchten).

Es stehen unter anderem folgende numerische Ausdrücke zur Verfügung (wie auch in der Mitte beziehungsweise am rechten Bildrand von Abbildung 4.34 (ansatzweise) zu erkennen ist):

- ✔ + (für die Addition)
- ✔ - (für die Subtraktion)

- ✔ * (für die Multiplikation)
- ✔ / (für die Division)
- ✔ ** (zum Potenzieren)
- ✔ ( ) (zum Einklammern mathematischer Ausdrücke)
- ✔ als *arithmetische Funktionen* zum Beispiel: ABS (für die Betragsbildung), EXP (für die Exponentialfunktion mit der Euler'schen Zahl e = 2,71828…, LOG10 (für den dekadischen Logarithmus), LN (für den natürlichen Logarithmus), SQRT (für die Quadratwurzel)
- ✔ als statistische Funktionen zum Beispiel: MEAN (arithmetischer Mittelwert), MAX (Maximalwert), MIN (Minimalwert), SD (Standardabweichung), SUM (Summe)

Dabei gelten folgende **Regeln**:

- ✔ *Auswertungsreihenfolge:* 1. Funktionen; 2. Potenzieren (**); 3. Multiplikation (*), Division (/), Minus-Vorzeichen (-…); 4. Addition (+), Subtraktion (-). Durch Klammersetzung können Sie die vorstehende Auswertungsreihenfolge beeinflussen beziehungsweise verändern: So würden Sie bei dem Beispielsausdruck *var3=2*var1+var2* zunächst den Term 2*var1 ausrechnen und anschließend den Wert von *var2* addieren. Setzen Sie nun eine Klammer dergestalt, dass gilt: *var3=2*(var1+var2)*, berechnen Sie zunächst den Klammerausdruck und multiplizieren Sie ihn anschließend mit 2. Setzen Sie beispielsweise für *var1* den Wert 3 und für *var2* den Wert 5 ein, so erhalten Sie bei der Formel ohne Klammersetzung als Ergebnis den Wert 11 für *var3* und bei der anderen Formel mit Klammersetzung als Ergebnis den Wert 16 für *var3*. Sie sehen, dass die Klammersetzung schon bedeutsam ist!
- ✔ Funktionen mit mehreren Argumenten erfordern das Setzen von genau einem Komma zwischen den Argumenten (zur Übersichtlichkeit eventuell noch durch Leerzeichen ergänzt), also zum Beispiel: SUM (*var1, var2*).

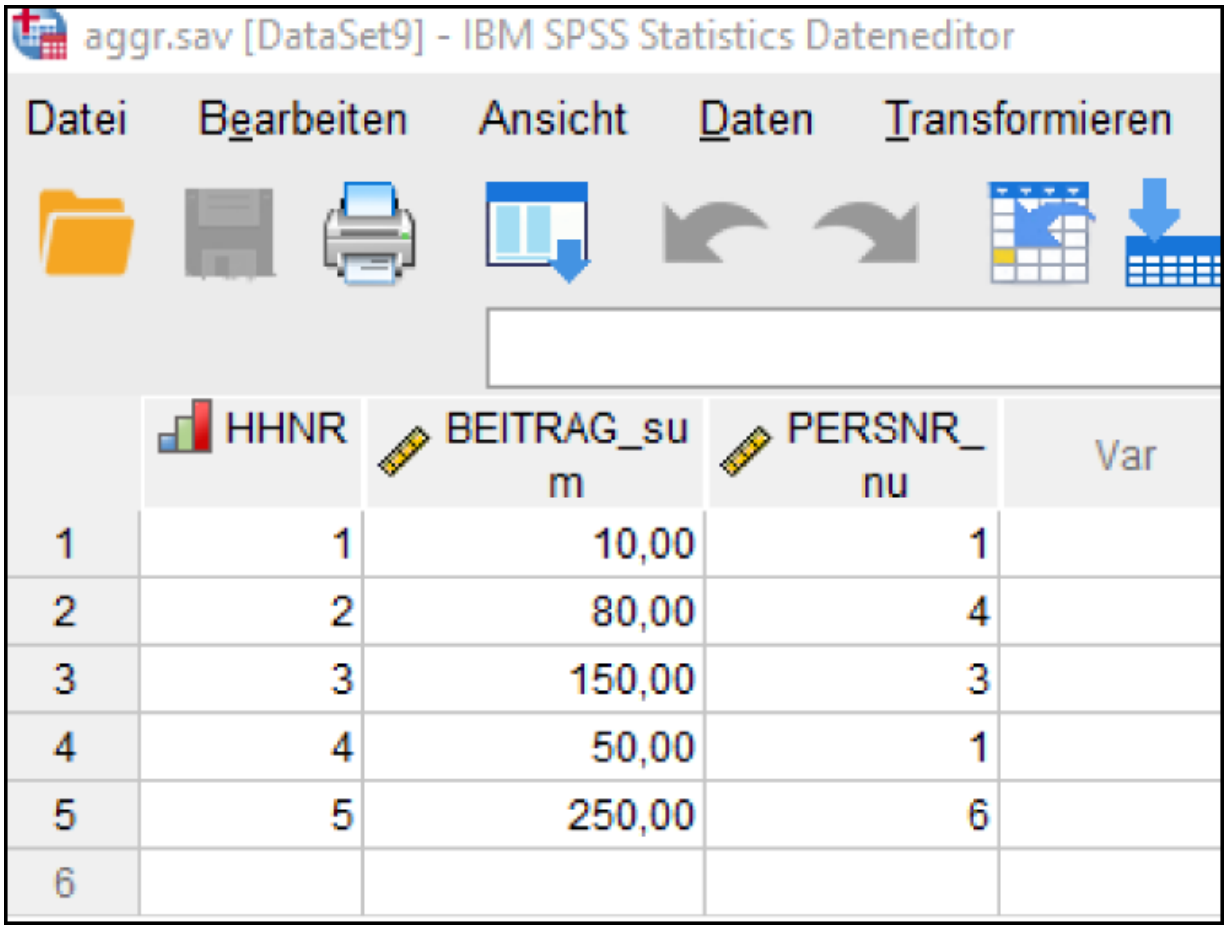
- ✔ In numerischen Ausdrücken ist stets der Dezimalpunkt zu verwenden.

# Daten bearbeiten

## Fälle gewichten

Mitunter möchte man von Stichprobenbefunden auf die (Grund-)Gesamtheit schließen. Zum Beispiel möchte man aus einer Bevölkerungsstichprobe auf die Gesamtbevölkerung schließen. Auch kann es vorkommen, dass Sie von einer Betrachtungsebene auf eine andere schließen möchten, zum Beispiel aus haushaltsbezogenen Informationen auf die Personenebene.

In solchen Fällen können Sie sich in SPSS der **Fallgewichtung** bedienen. Nehmen Sie als Beispiel die in Abbildung 4.36 wiedergegebene Datendatei mit Namen *AGGR.SAV*, die haushaltsbezogene Informationen enthält. Nehmen Sie des Weiteren an, Sie möchten von der Gesamtzahl der Fälle auf der Haushaltsebene (Gesamtzahl der Haushalte; in unserem Beispiel fünf Haushalte) auf die Gesamtzahl der Fälle auf der Personenebene (Gesamtzahl der Personen; in unserem Beispiel 15 Personen) schließen, und zwar hinsichtlich des durchschnittlichen Beitrags (gemäß der Variable *beitrag_sum*).




| | HHNR | BEITRAG_sum | PERSNR_nu | Var |
|---|---|---|---|---|
| 1 | 1 | 10,00 | 1 | |
| 2 | 2 | 80,00 | 4 | |
| 3 | 3 | 150,00 | 3 | |
| 4 | 4 | 50,00 | 1 | |
| 5 | 5 | 250,00 | 6 | |
| 6 | | | | |

**Abbildung 4.36:** Haushaltsbezogene Daten vor Gewichtung (in *AGGR.SAV*)

Letzteres können Sie durch die Gewichtung mit *persnr_nu*, also mit der jeweiligen Haushaltsgröße, bewerkstelligen. Sie müssen hierzu auf DATEN|FÄLLE GEWICHTEN gehen, und dann öffnet sich ein Dialogfeld, in das Sie die betreffende Gewichtungsvariable eingeben dürfen (siehe Abbildung 4.37).

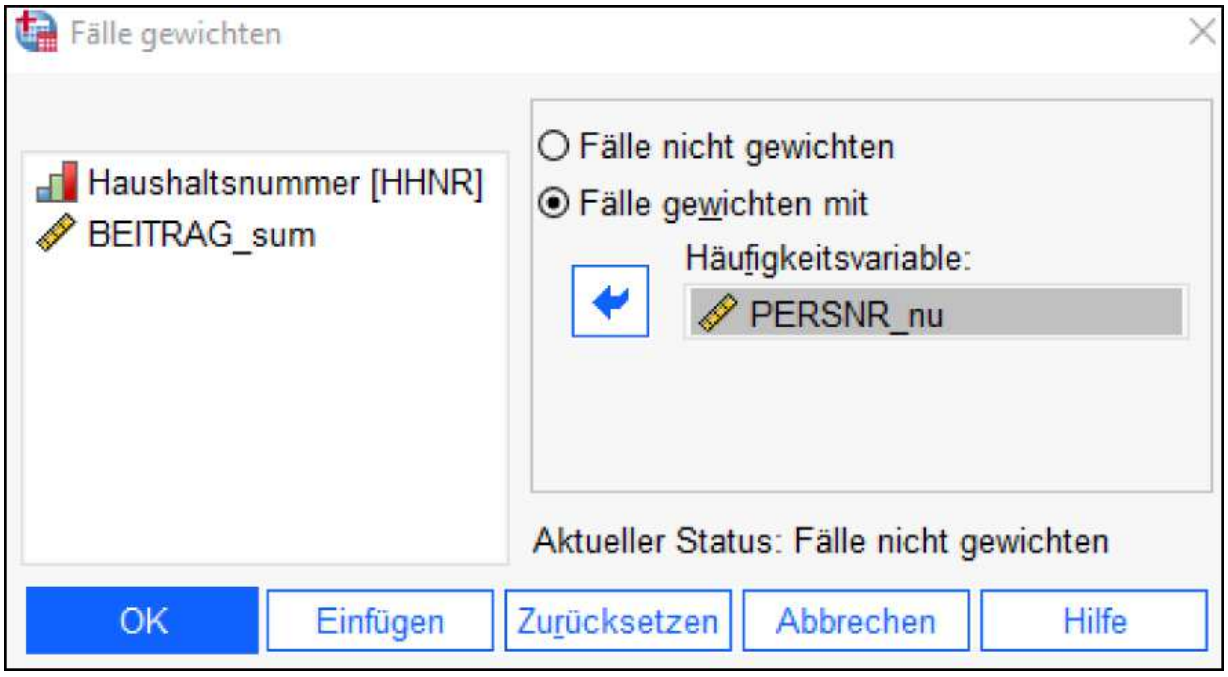



**Abbildung 4.37:** Dialogfeld zur Gewichtung

Mit dieser Gewichtung können Sie nunmehr auf der Personenebene statistische Auswertungen durchführen. Wie bereits geschrieben, soll der durchschnittliche Beitrag berechnet werden. Ohne Gewichtung – das heißt: auf der ursprünglichen Haushaltsebene – ergibt sich folgender Durchschnittsbeitrag:

$$\text{Durchschnittsbeitrag}_{\text{Haushaltsebene}} = \frac{10\,\text{Euro} + 80\,\text{Euro} + 150\,\text{Euro} + 50\,\text{Euro} + 250\,\text{Euro}}{5\,\text{Haushalte}} = 108\,\frac{\text{Euro}}{\text{Haushalt}}.$$

Wenn Sie nun aber die Gewichtung im Sinne von Abbildung 4.37 einschalten, wird der jeweilige Haushaltsbeitrag jedem einzelnen Mitglied des betreffenden Haushalts zugeordnet. Beispielsweise wird die Beitragsleistung des dritten Haushalts in Höhe von 150 Euro drei Personen zugeordnet, also dreimal in der Berechnung berücksichtigt. Es ergibt sich auf der Personenebene, also bei Gewichtung der Haushaltsbeiträge mit der jeweiligen Haushaltsgröße, als Durchschnittsbeitrag:

$$\text{Durchschnittsbeitrag}_{\text{Personenebene}} = \frac{10\,\text{Euro}\cdot 1+80\,\text{Euro}\cdot 4+150\,\text{Euro}\cdot 3+50\,\text{Euro}\cdot 1+250\,\text{Euro}\cdot 6}{15\,\text{Personen}} = 155{,}33\,\frac{\text{Euro}}{\text{Person}}.$$

Sie sehen: Weighting matters! Der personenbezogene Durchschnittsbeitrag unterscheidet sich in diesem Beispiel um fast 50 Euro vom haushaltsbezogenen.

Zur Gewichtung von Fällen kann naheliegenderweise nur eine numerische Variable verwendet werden, wobei negative und fehlende Werte auf 0 gesetzt werden, das heißt bei der Gewichtung nicht berücksichtigt werden. Löschen Sie einfach mal in der Datei *AGGR.SAV* in der zweiten Datenzeile bei der Gewichtungsvariablen *persnr_nu* den dort stehenden Wert (beziehungsweise öffnen Sie gleich die bereits entsprechend präparierte Datei *AGGR_A.SAV* ), und Sie erhalten das Bild in Abbildung 4.38.

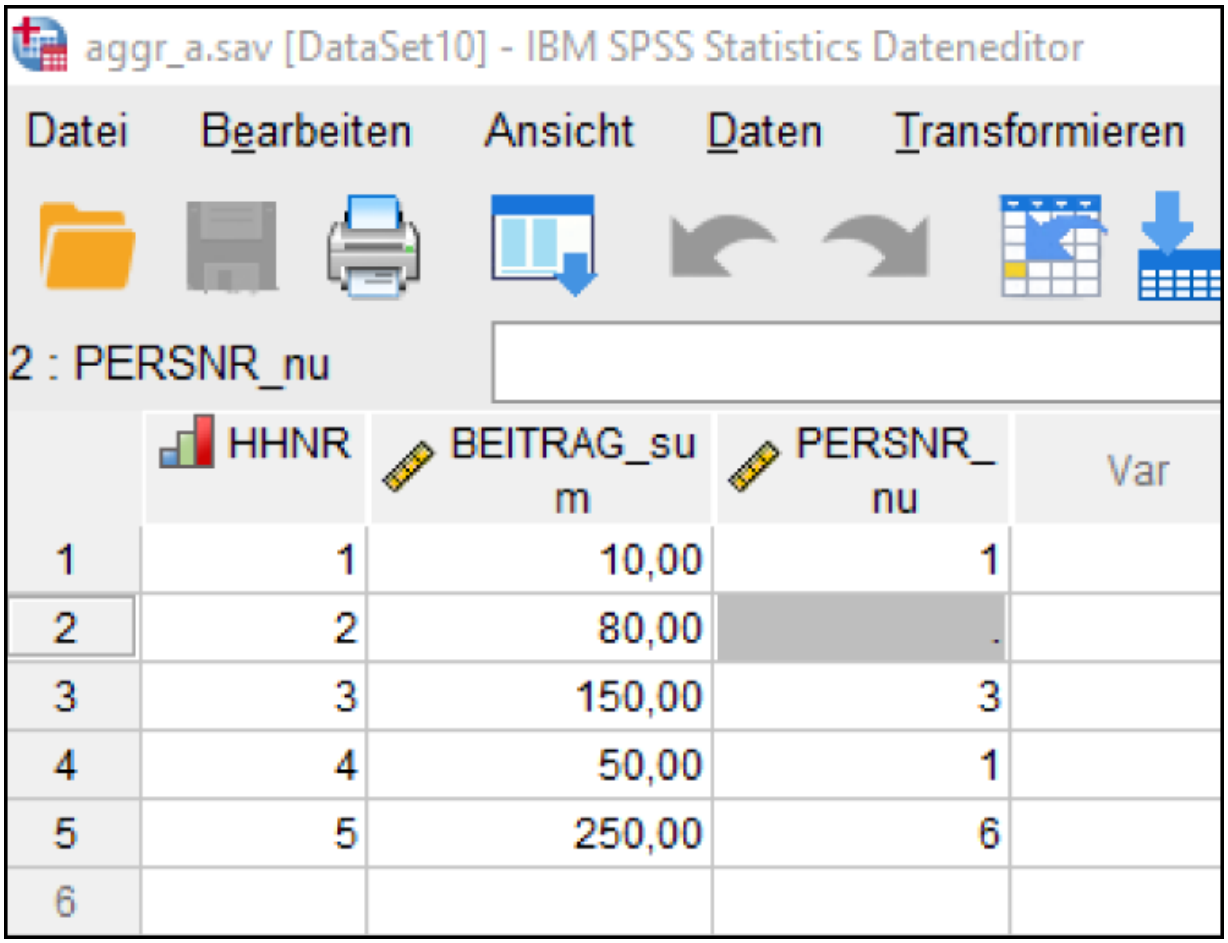



**Abbildung 4.38:** Fehlender Gewichtungswert (*AGGR_A.SAV*)

Wenn Sie nunmehr die Gewichtung mit *persnr_nu* und zudem eine statistische Operation (wie eine Mittelwertberechnung) in SPSS durchführen, gibt SPSS den in Abbildung 4.39 dargestellten Hinweis (natürlich nur, wenn die Gewichtung aktiviert ist).

```
>Warnungsnummer 3211
>Der Wert der Gewichtungsvariablen war für mindestens einen Fall null oder
>negativ oder der Wert fehlte.  Solche Fälle sind für statistische Prozeduren
>und für statistische Grafik, die eine Fallgewichtung mit einem positiven Wert
>erfordern, nicht sichtbar, aber die Fälle verbleiben in der
>Datei und werden von nicht statistischen Funktionen wie z.  B.  LIST und SAVE
>verarbeitet.
```

**Abbildung 4.39:** SPSS-Hinweis bei fehlenden Werten der Gewichtungsvariablen

Bei einer abgespeicherten Datendatei mit aktiver Gewichtung ist die Gewichtung nach deren Öffnen weiter aktiv. Die obige Datei *AGGR.SAV* ist beispielsweise mit aktiver Gewichtung über die Variable *persnr_nu* abgespeichert worden. Wenn Sie diese Datei nun öffnen, ist diese Gewichtung weiterhin aktiv.

Außerdem ist im Zusammenhang mit der Gewichtung unbedingt zu beachten, dass bei Signifikanztests (siehe Teil IV), aber auch zum Beispiel bei einer über eine Gewichtungsvariable gewichteten linearen Regression (siehe Kapitel 9) die Gewichtungsvariable im Durchschnitt stets gleich 1 ist. Der Grund hierfür ist, dass dort die Summe der Gewichte dem Stichprobenumfang gleichen muss. In unserem Beispiel mit *AGGR.SAV* müssen Sie daher die Gewichtungsvariable *persnr_nu* entsprechend anpassen. Dies geht aber in diesem Fall ganz einfach: Sie müssen jeden Wert von *persnr_nu*, der ja die Personenzahl in einem bestimmten Haushalt repräsentiert, durch die Gesamtpersonenzahl (in diesem Fall durch 15) dividieren. Falls Sie vergessen haben sollten, wie das geht, hier zur Erinnerung: Gehen Sie in das Menü TRANSFORMIEREN|BERECHNEN und definieren Sie *persnr_nu = persnr_nu/15*. In diesem Fall würden Sie also die Ausprägungen der bestehenden Variablen *persnr_nu* überschreiben. Alternativ hätten Sie auch eine neue Gewichtungsvariable, zum Beispiel *persnr_nu2*, definieren können als *persnr_nu/15*. Die resultierenden Gewichte für die fünf Haushalte in *AGGR.SAV* sind dann: Haushalt Nummer 1 (1 Person): 1/15, Haushalt Nummer 2 (4 Personen): 4/15, Haushalt Nummer 3 (3 Personen): 3/15, Haushalt Nummer 4 (1 Person): 1/15, Haushalt Nummer 5 (6 Personen): 6/15 – die Gewichtesumme beträgt also, wie gewünscht, 15/15 = 1.

Manchmal erhält man statistische Informationen in einer Häufigkeitstabelle, in der den Merkmalsausprägungen ihre (absoluten) Häufigkeiten zugeordnet sind. Sie können nun die entsprechenden (absoluten) Häufigkeiten als Ausprägungen der betreffenden »Häufigkeits-Variablen« zur Gewichtung der Werte der interessierenden Variablen verwenden und auf diese Weise etwa verschiedene deskriptive statistische Kennziffern wie zum Beispiel Mittelwerte berechnen.

Wenn Sie jetzt neugierig geworden sind und unbedingt für das Vorstehende ein Beispiel suchen, dann blättern Sie doch bitte zu Kapitel 14 und schauen Sie sich das entsprechende Vorgehen an, das dort im Zusammenhang mit den Datendateien *GLEICHVERTEILUNG.SAV* (siehe Abbildung 14.1) und *UNGLEICHVERTEILUNG.SAV* (siehe Abbildung 14.4) erörtert wird.

## *Fälle filtern*

Sie können in SPSS einen Datensatz **filtern**, indem Sie – unter Verwendung logischer Operatoren – einen 0/1-Filter setzen, wie ich Ihnen gleich zeigen werde.

### *Fälle mittels Bedingungen filtern*

Beispielsweise könnte es Ihr Interesse sein, aus dem Datensatz *AGGR.SAV* nur die beiden Einpersonenhaushalte (das heißt Haushalt Nummer 1 und Haushalt Nummer 4) in Ihren Berechnungen zu berücksichtigen. Im SPSS-Menü gehen Sie in diesem Zusammenhang bitte zu DATEN|FÄLLE AUSWÄHLEN, und es öffnet sich die in Abbildung 4.40 gezeigte Dialogbox. Voreingestellt ist offensichtlich ALLE FÄLLE, also kein Filter.

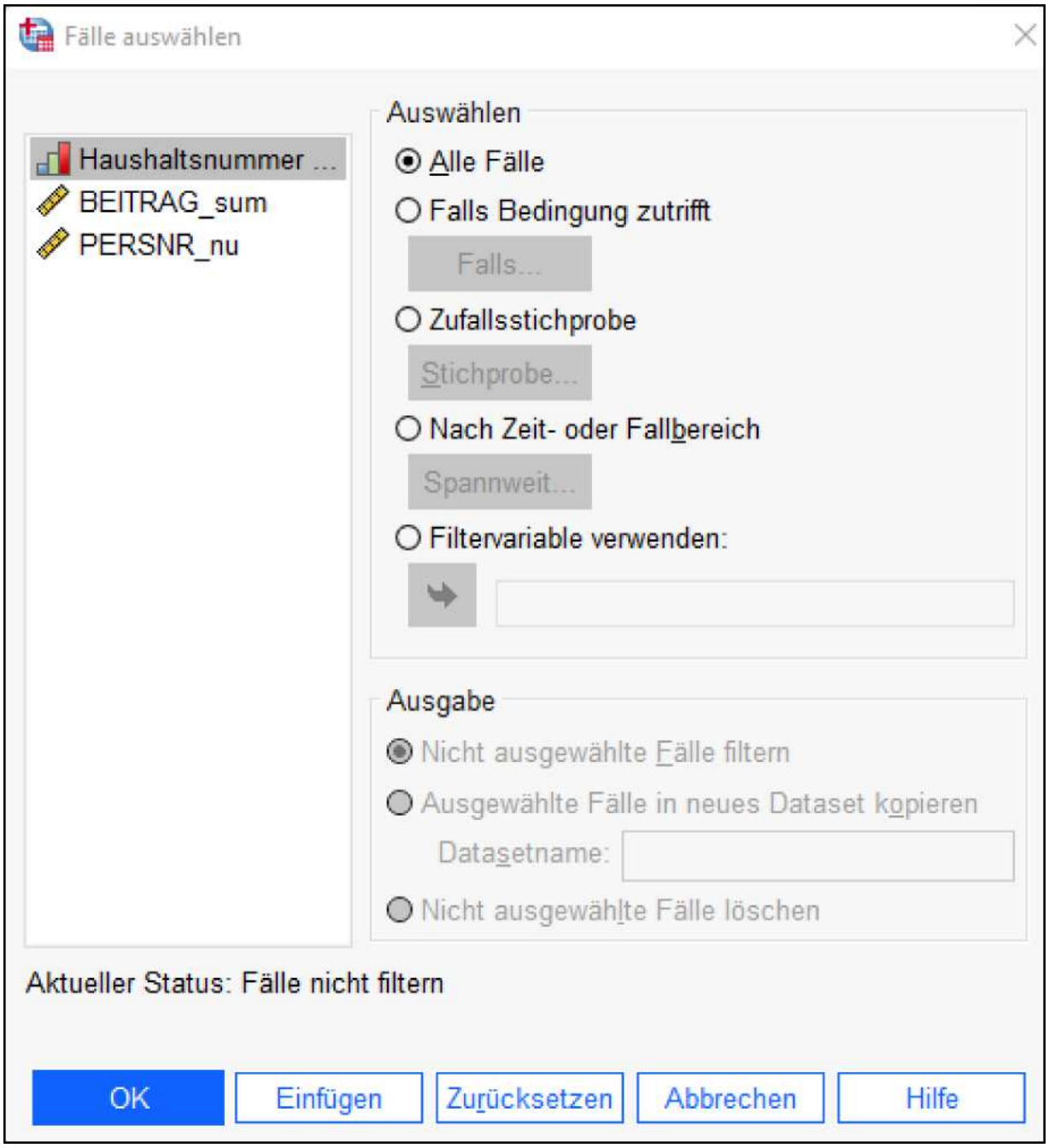



**Abbildung 4.40:** Dialogbox für DATEN|FÄLLE AUSWÄHLEN (*AGGR.SAV*)

Wenn Sie jetzt die Einpersonenhaushalte auswählen möchten, müssen Sie anstelle von ALLE FÄLLE die nächste Rubrik anklicken, also FALLS BEDINGUNG ZUTRIFFT, und dann auch auf den FALLS-Button klicken (siehe Abbildung 4.41).

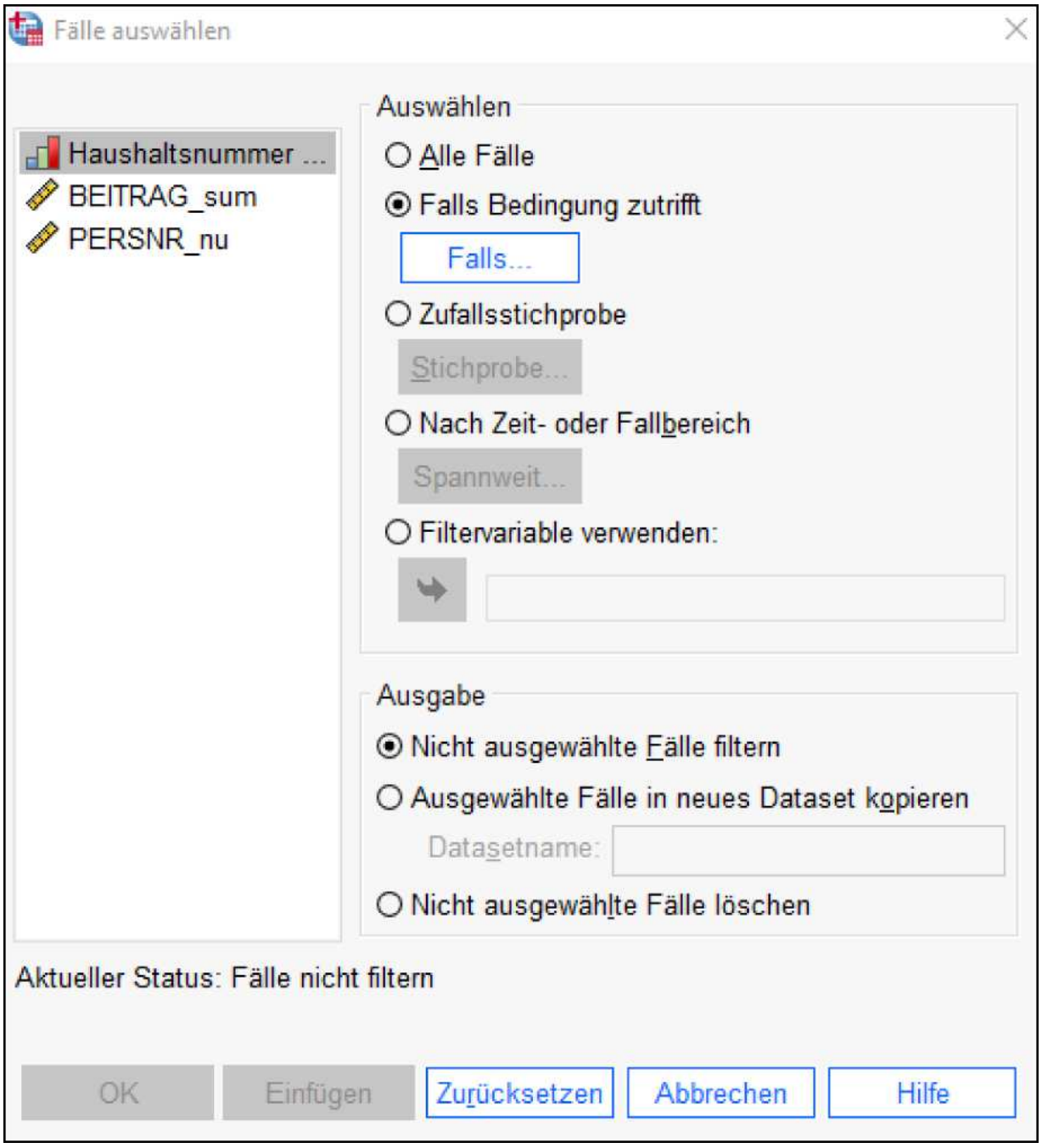



**Abbildung 4.41:** Aktivierung der Fallauswahl in der Dialogbox DATEN|FÄLLE AUSWÄHLEN (*AGGR.SAV*)

Es öffnet sich nun eine neue Dialogbox, in die Sie Ihre Bedingung eingeben, hier: *persnr_nu* = 1. Sie wählen demnach, wie gewünscht, die Fälle aus, in denen die Personenzahl (im Haushalt) gleich 1 ist (siehe Abbildung 4.42).

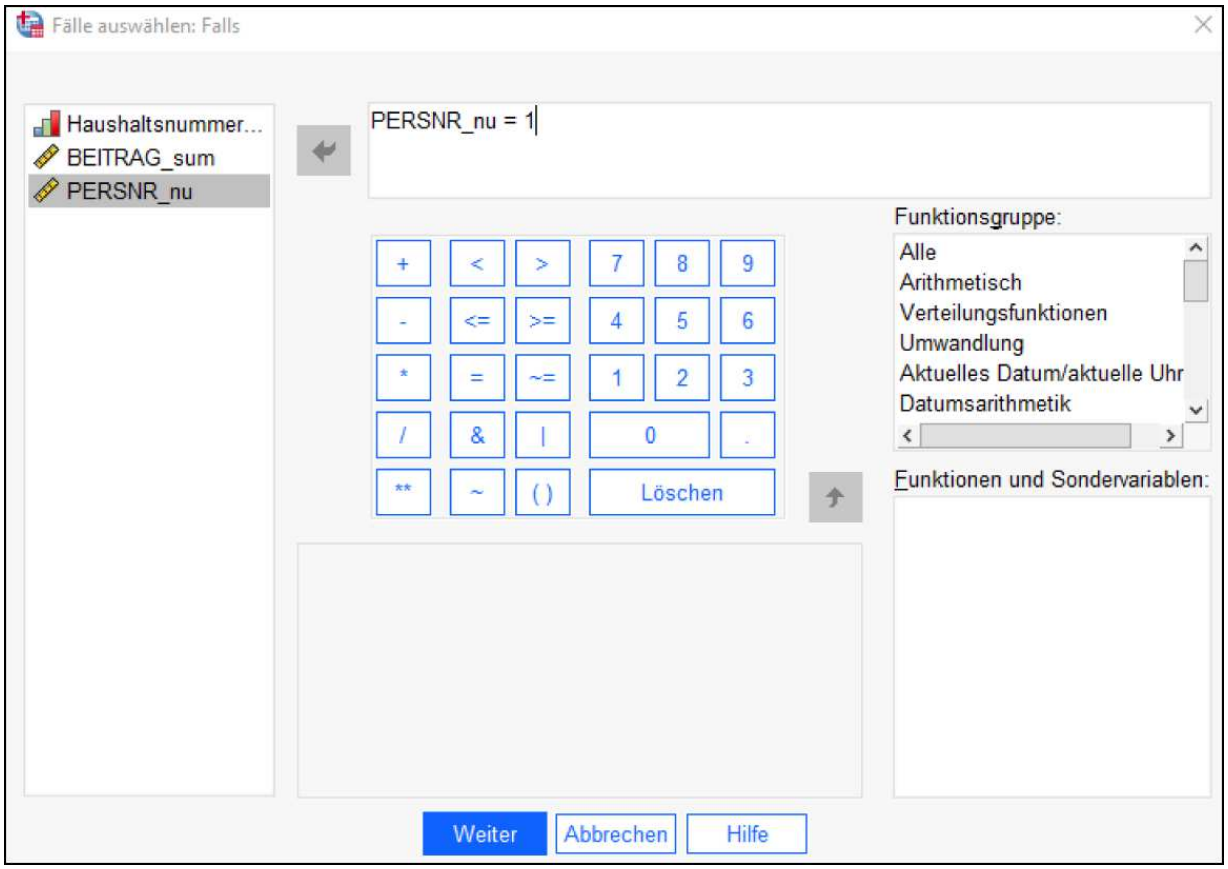
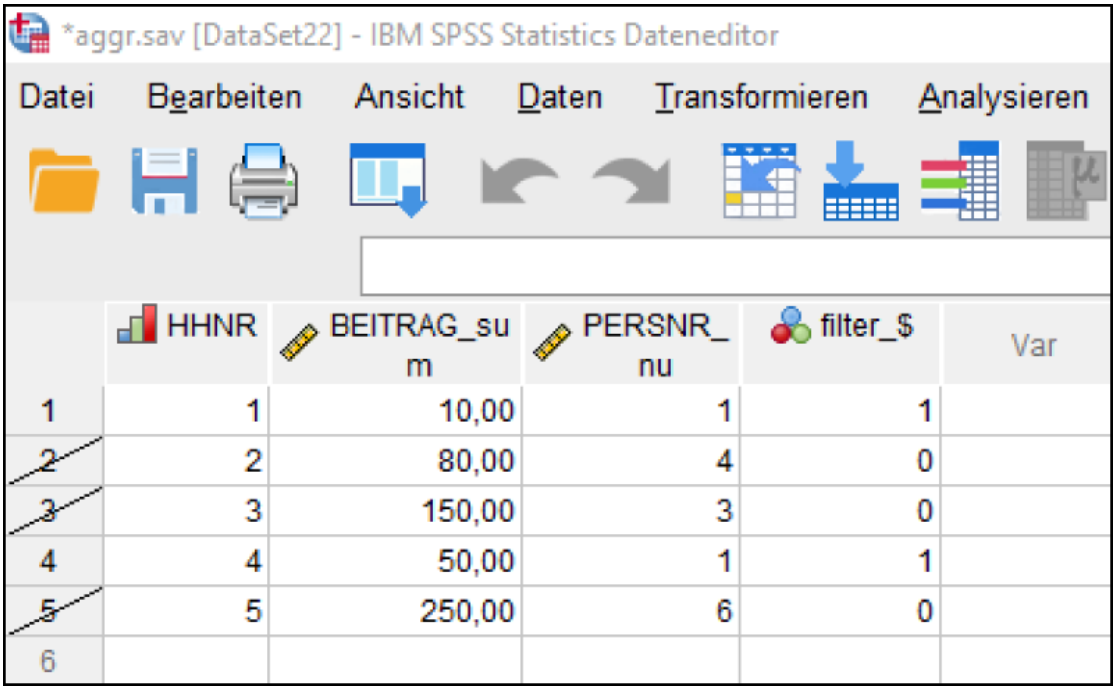



**Abbildung 4.42:** Angabe der Bedingung bei DATEN|FÄLLE AUSWÄHLEN (*AGGR.SAV*)

Wenn Sie nun auf WEITER klicken, gelangen Sie wieder zur vorhergehend genannten Dialogbox, in der Ihnen in der Rubrik AUSGABE für das weitere Arbeiten mit den ausgewählten Fällen drei Wege offenstehen:

- ✔ NICHT AUSGEWÄHLTE FÄLLE FILTERN: Nach Auswahl dieser Filteroption finden Sie im Datensatz eine Filtervariable *filter_$*, die bei den ausgewählten Fällen den Wert 1 und bei den nicht ausgewählten Fällen den Wert 0 annimmt (siehe Abbildung 4.43). Die (Hilfs-)Variable *filter_$* weist folglich den Wert 1 genau dann einem Fall zu, wenn der angegebene logische Ausdruck (im Sinne einer Bedingung; hier: *persnr_nu* = 1) wahr ist, und ansonsten den Wert 0. Sie erkennen die nicht ausgewählten Fälle (*filter_$* = 0) im Dateneditor an den durchgestrichenen Fallnummern. Außerdem steht in der Statuszeile ganz unten rechts: FILTER AKTIV. Sie können den Filter einfach durch Löschen von *filter_$* deaktivieren.

Alternativ können Sie den betreffenden Filter dadurch deaktivieren, dass Sie bei FÄLLE AUSWÄHLEN einfach wieder ALLE FÄLLE anklicken. Wenn Sie eine Datendatei abspeichern, wird *filter_$* mit abgespeichert, ist aber beim nächsten Öffnen der betreffenden Datei nicht mehr aktiv. Sie müssen diese Filtervariable gegebenenfalls erst wieder über DATEN|FÄLLE AUSWÄHLEN aktivieren. Eine in einem Datensatz bereits vorhandene Filtervariable wie *filter_$* können Sie im Übrigen später in der Dialogbox gemäß Abbildung 4.41 unter FILTERVARIABLE VERWENDEN eingeben. Dann können Sie ganz bequem den Filtervorgang starten. Allerdings sollten Sie wissen, was die betreffende Filtervariable genau beinhaltet, das heißt: wonach Sie überhaupt filtern!




| | HHNR | BEITRAG_sum | PERSNR_nu | filter_$ | Var |
|---|---|---|---|---|---|
| 1 | 1 | 10,00 | 1 | 1 | |
| ~~2~~ | 2 | 80,00 | 4 | 0 | |
| ~~3~~ | 3 | 150,00 | 3 | 0 | |
| 4 | 4 | 50,00 | 1 | 1 | |
| ~~5~~ | 5 | 250,00 | 6 | 0 | |
| 6 | | | | | |

**Abbildung 4.43:** Datensatz mit Filtervariable *filter_$* und markierten nicht ausgewählten Fällen (*AGGR.SAV*)

- ✔ AUSGEWÄHLTE FÄLLE IN EIN NEUES DATENBLATT KOPIEREN: Es wird ein neues Datenblatt geöffnet, in das nur die ausgewählten Fälle geschrieben werden. Es ist in Abbildung 4.44 als *AGGR_B.SAV* bezeichnet worden. Sie haben nunmehr also zwei Datenfiles: das ursprüngliche File *AGGR.SAV* mit allen Fällen und

eben dieses neue Datenfile *AGGR_B.SAV*, das in Abbildung 4.45 zu sehen ist.

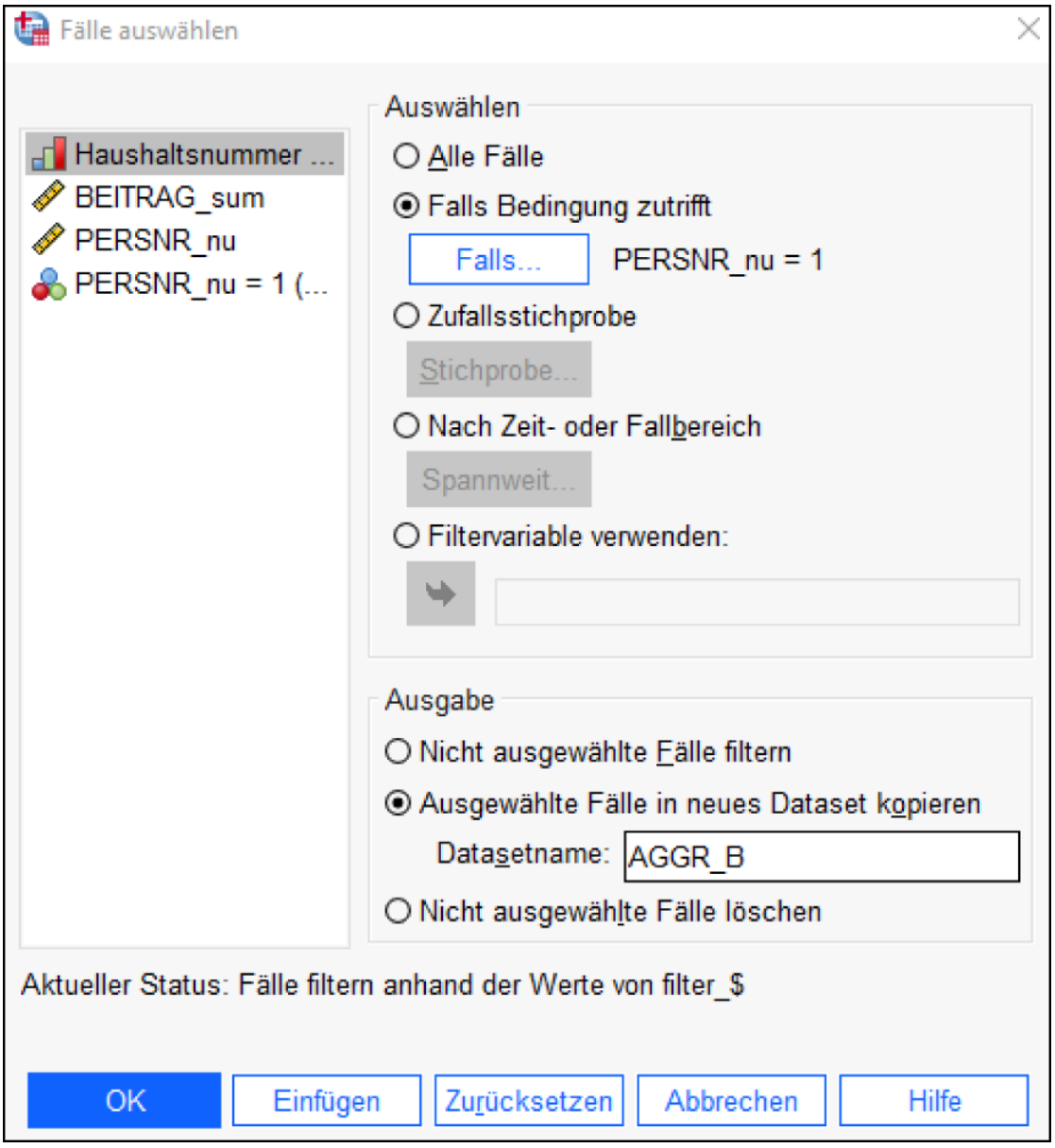



**Abbildung 4.44:** Neuer Datensatz nur mit den ausgewählten Fällen (*AGGR_B.SAV*) (1)

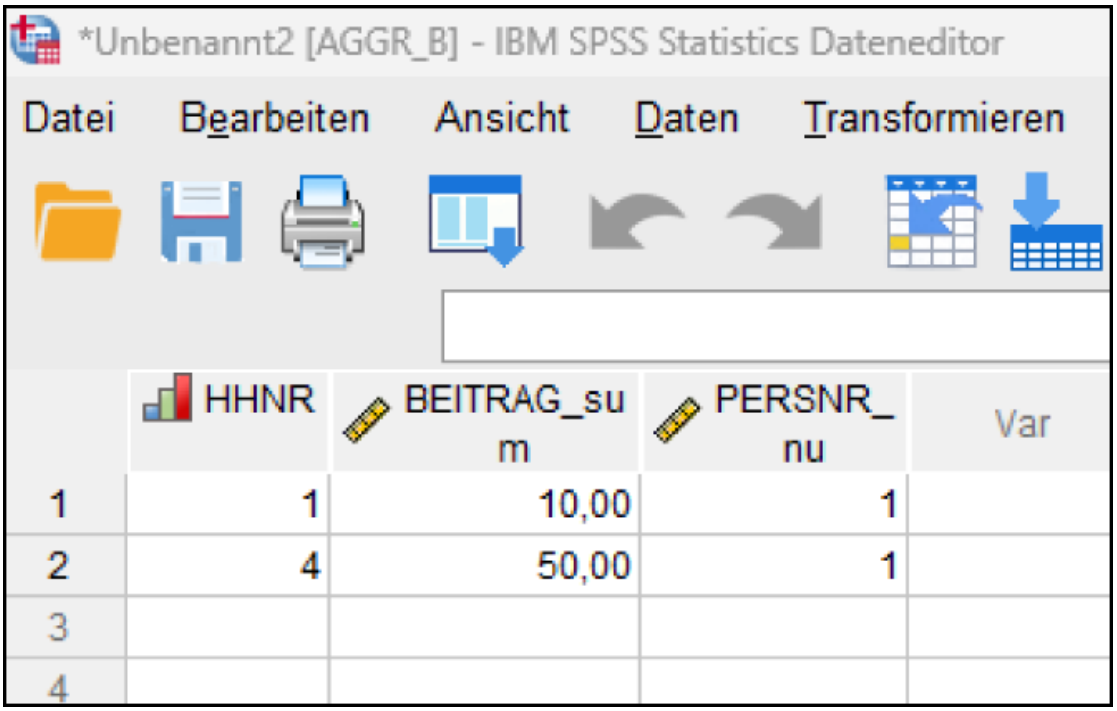



**Abbildung 4.45:** Neuer Datensatz nur mit den ausgewählten Fällen (*AGGR_B.SAV*) (2)

- ✔ NICHT AUSGEWÄHLTE FÄLLE LÖSCHEN: Es werden die nicht ausgewählten Fälle aus *AGGR.SAV* gelöscht. Hier müssen Sie etwas aufpassen: Speichern Sie Ihren Datensatz wieder als *AGGR.SAV* ab, so wird der alte Datensatz überschrieben, und *AGGR.SAV* enthält jetzt nicht mehr fünf, sondern nur noch zwei Fälle (siehe Abbildung 4.46). Bitte beachten Sie, dass sich Abbildung 4.46 im Unterschied zu Abbildung 4.45 nicht auf einen neuen Datensatz, sondern auf den bestehenden Datensatz *AGGR.SAV* bezieht, was Sie ganz oben links in Abbildung 4.46 erkennen können.

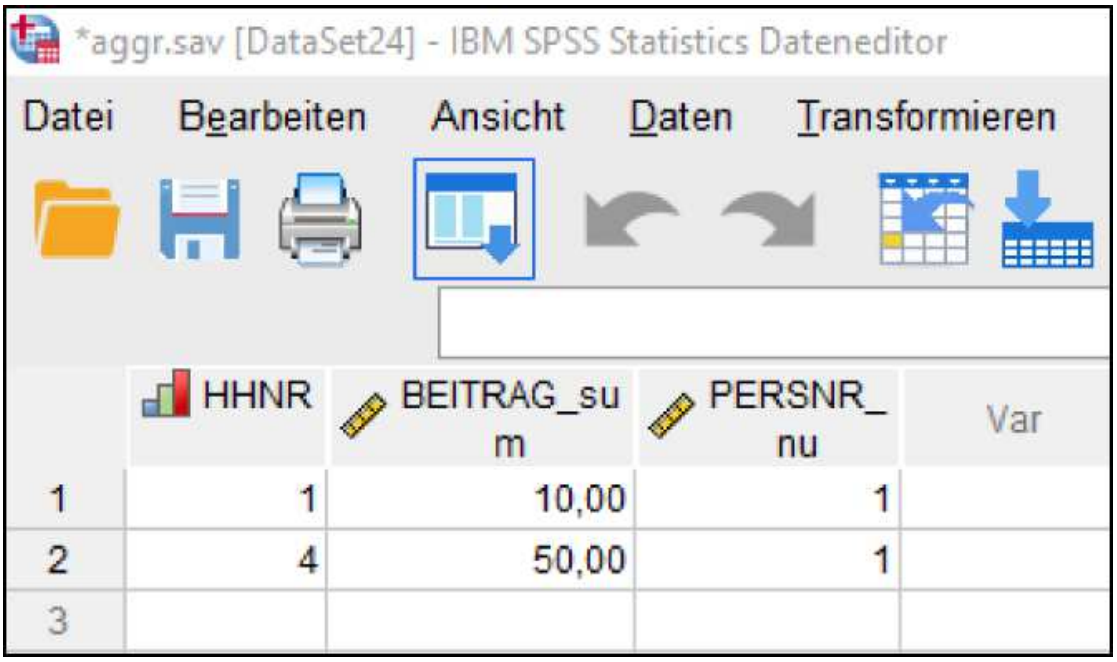



**Abbildung 4.46:** Löschen der nicht ausgewählten Fälle (*AGGR.SAV*)

Die vorstehend genannten Filtervarianten sind bei statistischen Analysen und Grafiken aktiv, nicht hingegen bei einfachen Datentransformationen. Bei Letzteren werden also auch die »ausgefilterten« Fälle berücksichtigt!

Wenn Sie zum Beispiel die Beitragssumme *beitrag_sum* pro Haushalt von den bisher angegebenen Euro-Werten in DM-Werte (in der neuen Variablen *beitrag_dm*) über den Euro/DM-Umrechnungskurs von 1:1,95583 umrechnen möchten, müssen Sie dies zunächst bei TRANSFORMIEREN|VARIABLE BERECHNEN entsprechend eingeben (siehe Abbildung 4.47).

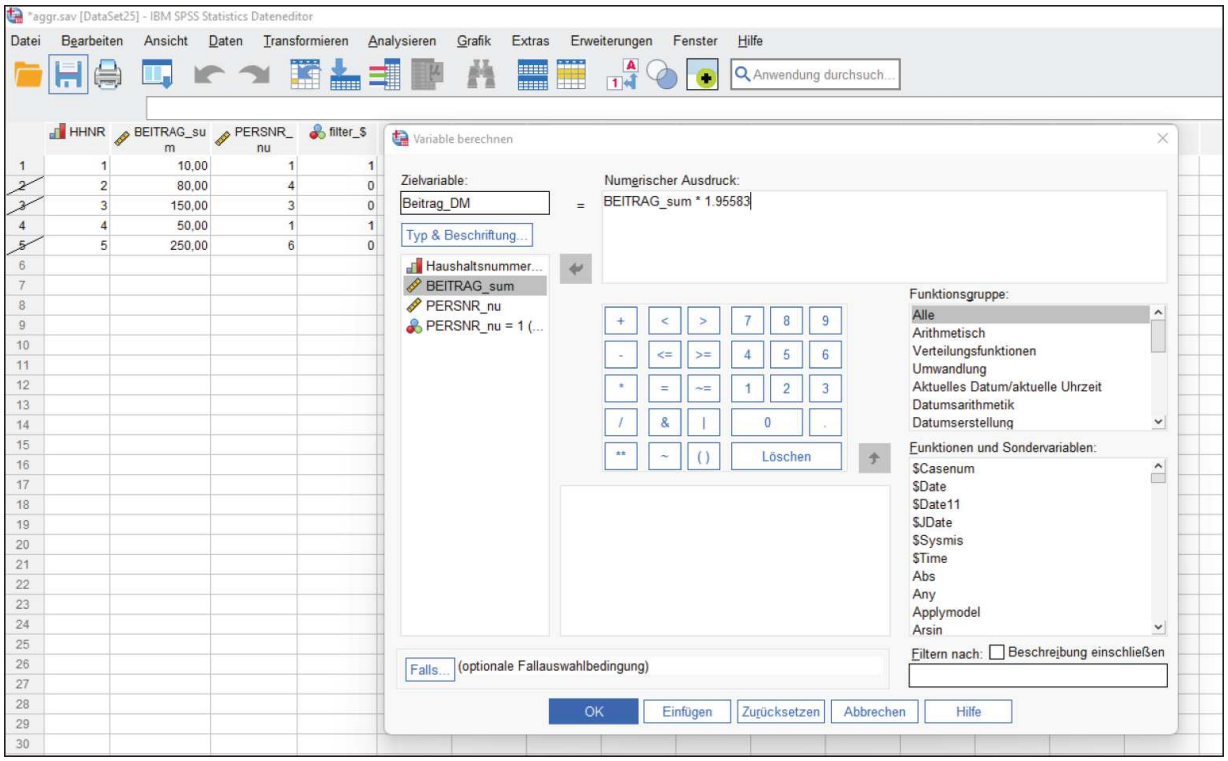
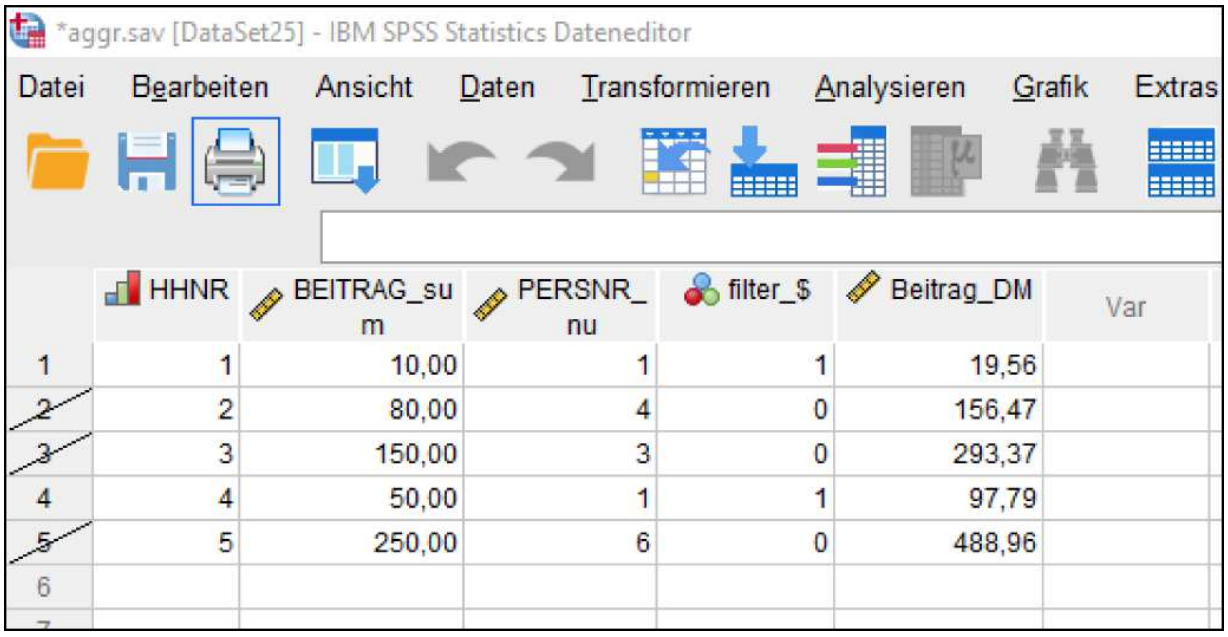



**Abbildung 4.47:** Beispiels-Transformation von Euro- in DM-Werte in einem Beispieldatensatz (*AGGR.SAV*)

Anschließend können Sie im Dateneditor erkennen, dass die vorstehende Variablen-Transformation für *alle* Fälle (also auch für die nicht ausgewählten!) durchgeführt worden ist (siehe Abbildung 4.48).




| | HHNR | BEITRAG_sum | PERSNR_nu | filter_$ | Beitrag_DM | Var |
|---|---|---|---|---|---|---|
| 1 | 1 | 10,00 | 1 | 1 | 19,56 | |
| ~~2~~ | 2 | 80,00 | 4 | 0 | 156,47 | |
| ~~3~~ | 3 | 150,00 | 3 | 0 | 293,37 | |
| 4 | 4 | 50,00 | 1 | 1 | 97,79 | |
| ~~5~~ | 5 | 250,00 | 6 | 0 | 488,96 | |
| 6 | | | | | | |

**Abbildung 4.48:** Variablen-Transformation für einen gefilterten Datensatz (*AGGR.SAV*)

### *Fälle mittels Zufallsstichprobe filtern*

Unter DATEN|FÄLLE AUSWÄHLEN können Sie über die vorgenannten Möglichkeiten hinaus auch noch eine **Zufallsstichprobe** auswählen, wie in Abbildung 4.49 für eine (näherungsweise) 20-Prozent-Stichprobe gezeigt. Die Filterung bezieht sich wieder auf *AGGR.SAV*.

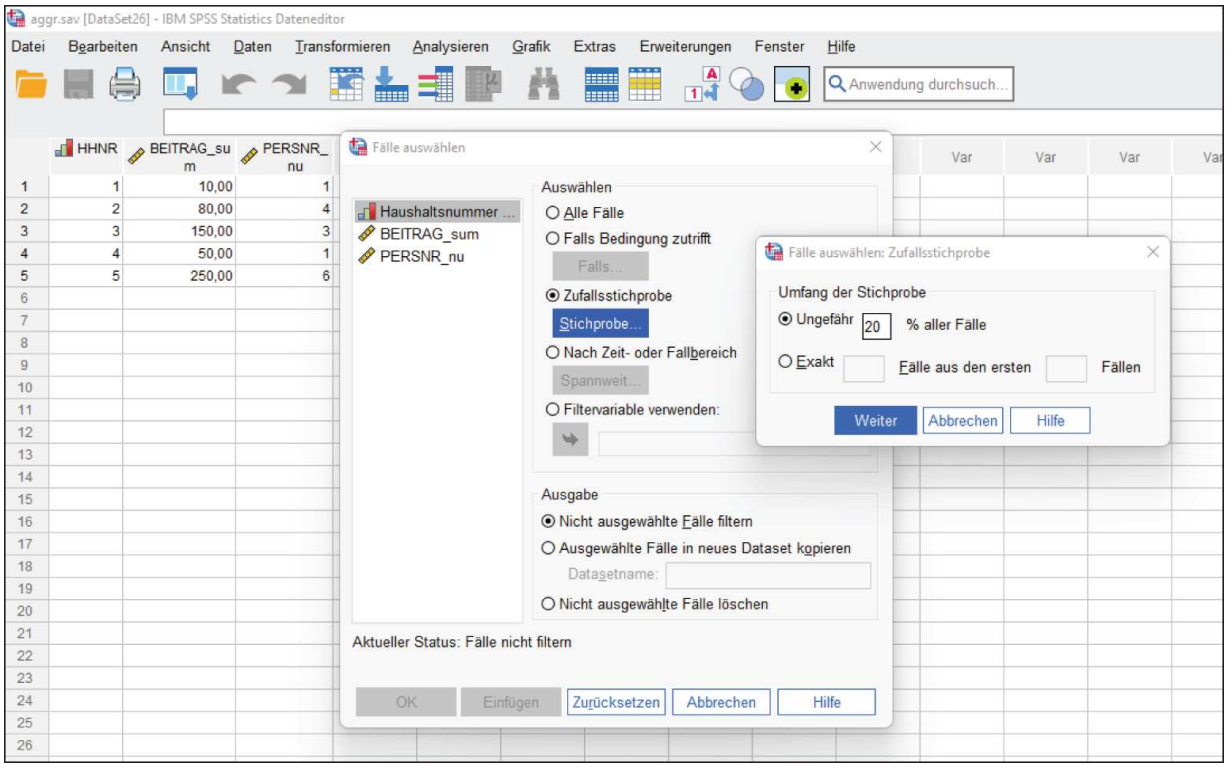
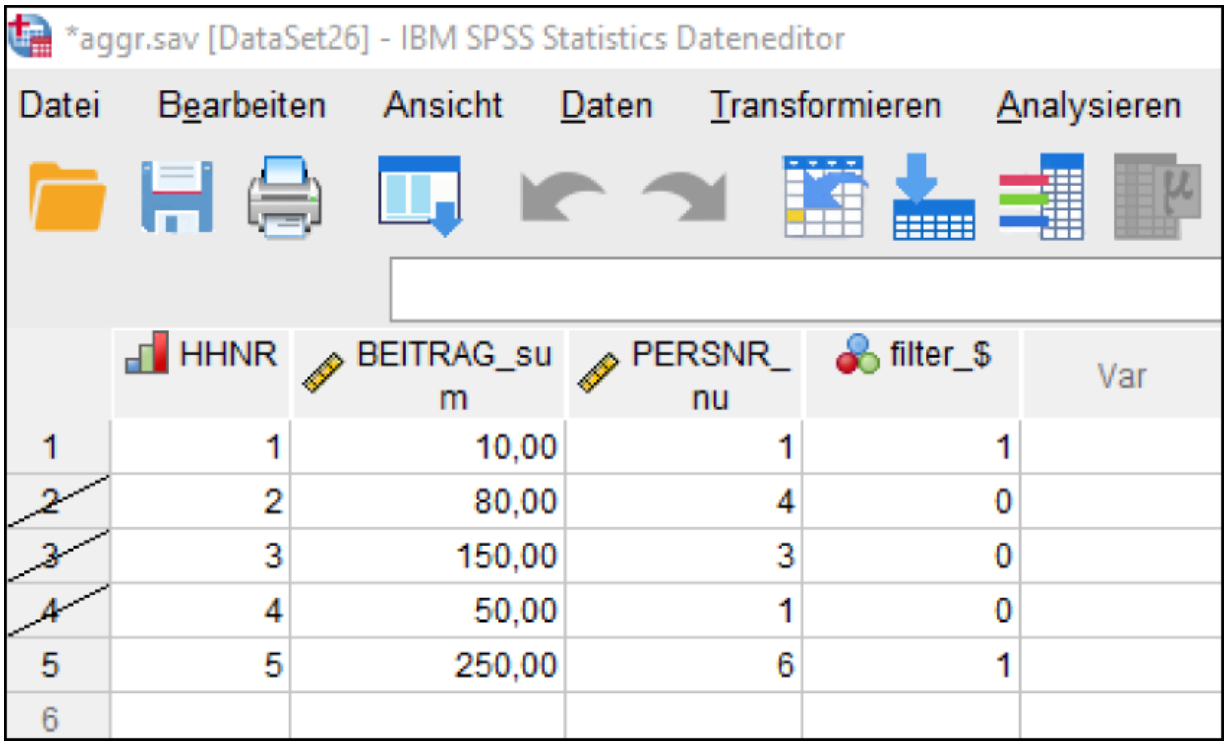



**Abbildung 4.49:** Variante »Zufallsstichprobe« bei DATEN|FÄLLE AUSWÄHLEN

Wichtig ist, dass die Prozentangabe bei der Stichprobenziehung nur ungefähr ist und bei kleinen Stichproben wie hier (*AGGR.SAV* besteht ja nur aus fünf Fällen) zu größeren Abweichungen vom angestrebten Prozentsatz führen kann. In unserem Beispiel werden 2 aus 5 Fällen, das heißt 40 Prozent aller Fälle, zufällig ausgewählt. Wie aus Abbildung 4.50 hervorgeht, sind in unserem Beispiel die Fälle Nummer 1 und Nummer 5 zufällig ausgewählt worden. (Es hätten aber genauso gut zum Beispiel die Fälle Nummer 2 und Nummer 3 oder Nummer 1 und Nummer 4 ausgewählt werden können.)




| | HHNR | BEITRAG_sum | PERSNR_nu | filter_$ | Var |
|---|---|---|---|---|---|
| 1 | 1 | 10,00 | 1 | 1 | |
| ~~2~~ | 2 | 80,00 | 4 | 0 | |
| ~~3~~ | 3 | 150,00 | 3 | 0 | |
| ~~4~~ | 4 | 50,00 | 1 | 0 | |
| 5 | 5 | 250,00 | 6 | 1 | |
| 6 | | | | | |

**Abbildung 4.50:** Ergebnis einer Zufallsauswahl bei DATEN|FÄLLE AUSWÄHLEN (*AGGR.SAV*)

### *Fälle mit Fallangaben filtern*

Eine weitere Filtermöglichkeit, die unter DATEN|FÄLLE AUSWÄHLEN angeboten wird, ist die Angabe von Fällen, die in die Auswahl gelangen sollen. Sie können hierbei einen Fallbereich definieren. In Abbildung 4.51 habe ich für die Datei *AGGR.SAV* offenkundig die beiden ersten Fälle ausgewählt.

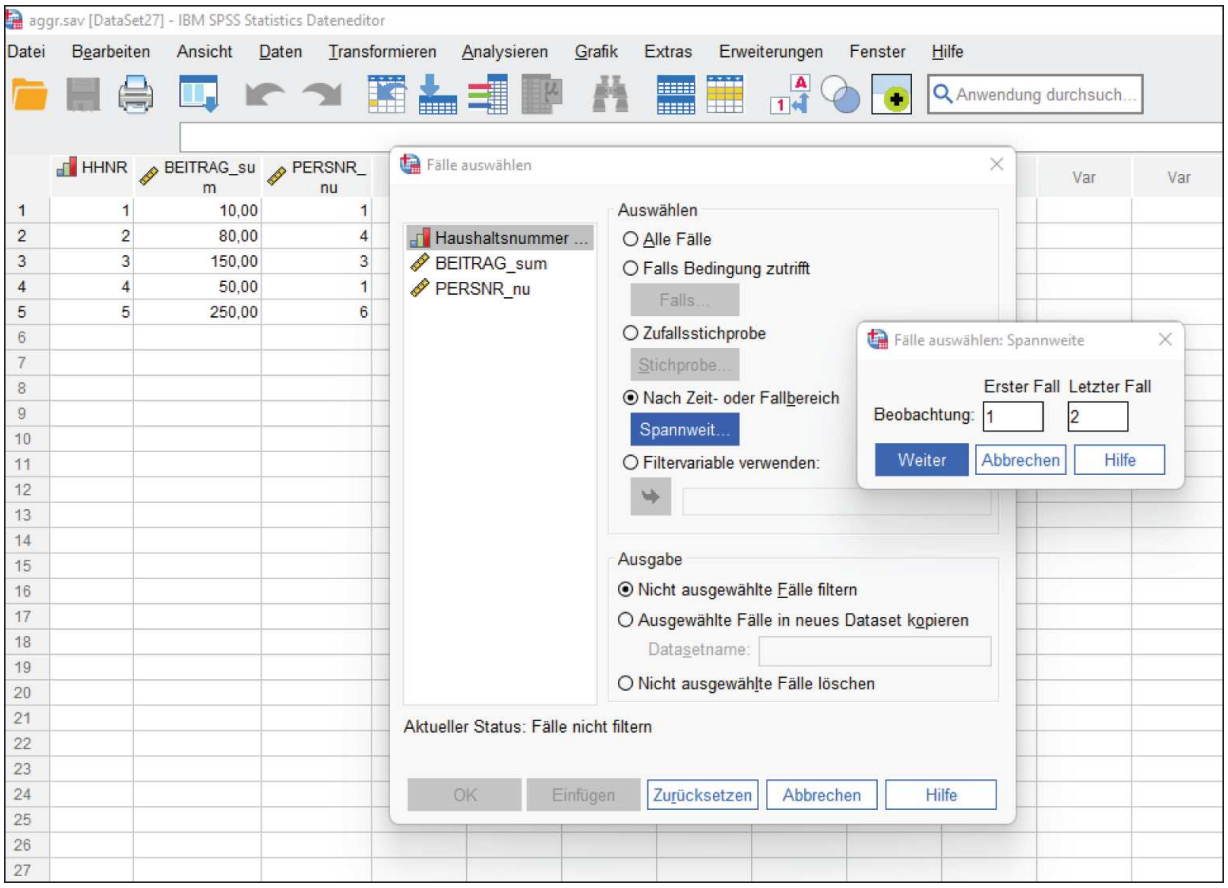
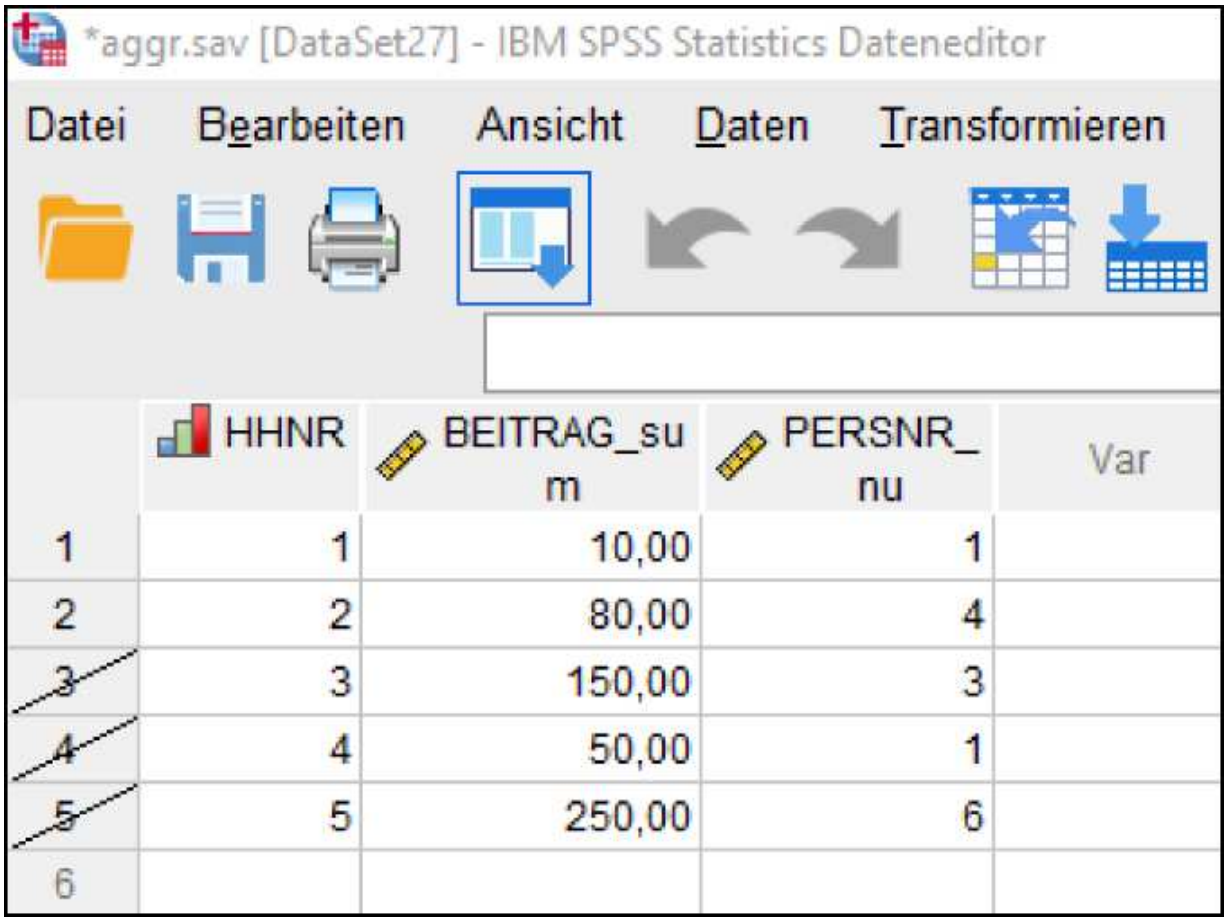



**Abbildung 4.51:** Variante »Fallbereich« bei DATEN|FÄLLE AUSWÄHLEN

Entsprechend werden im Datensatz *AGGR.SAV* nur die beiden ersten Fälle als ausgewählt angezeigt. Bei den anderen Fällen sind die Fallnummern durchgestrichen (siehe Abbildung 4.52).

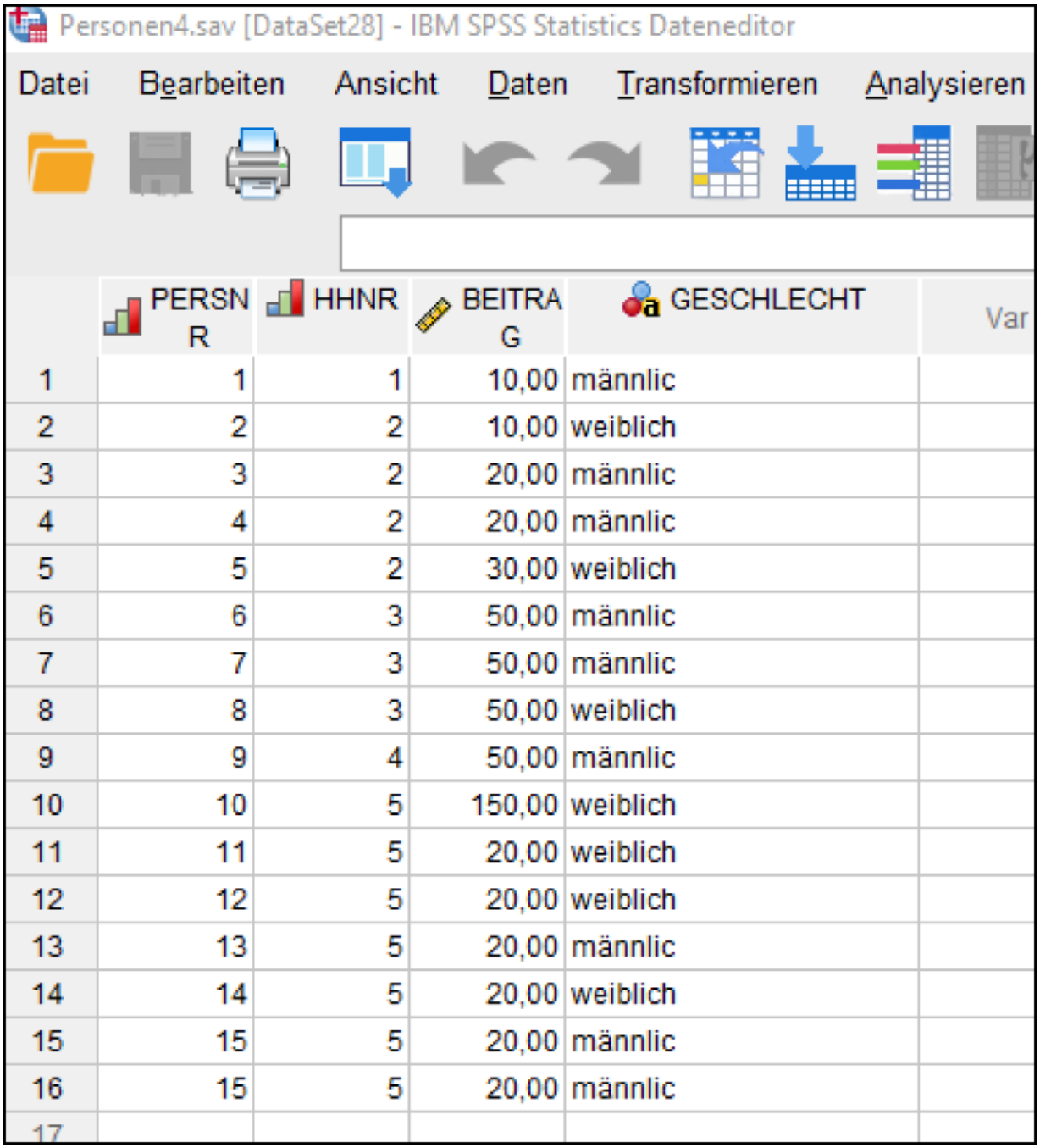



| | HHNR | BEITRAG_sum | PERSNR_nu | Var |
|---|---|---|---|---|
| 1 | 1 | 10,00 | 1 | |
| 2 | 2 | 80,00 | 4 | |
| ~~3~~ | 3 | 150,00 | 3 | |
| ~~4~~ | 4 | 50,00 | 1 | |
| ~~5~~ | 5 | 250,00 | 6 | |
| 6 | | | | |

**Abbildung 4.52:** Ergebnis einer »Fallbereichsauswahl« bei DATEN|FÄLLE AUSWÄHLEN (*AGGR.SAV*)

# *Doppelte Fälle ermitteln*

Um zum Beispiel Übertragungsfehler zu ermitteln, können Sie mit DATEN|DOPPELTE FÄLLE ERMITTELN (versehentlichen) **Doppelungen von Fällen** in Ihrer Datendatei nachgehen. Im folgenden Beispiel (Datei *PERSONEN4.SAV*) sind offensichtlich die Personendatensätze Nummer 15 und Nummer 16 doppelt (siehe Abbildung 4.53).




| | PERSNR | HHNR | BEITRAG | GESCHLECHT | Var |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 10,00 | männlic | |
| 2 | 2 | 2 | 10,00 | weiblich | |
| 3 | 3 | 2 | 20,00 | männlic | |
| 4 | 4 | 2 | 20,00 | männlic | |
| 5 | 5 | 2 | 30,00 | weiblich | |
| 6 | 6 | 3 | 50,00 | männlic | |
| 7 | 7 | 3 | 50,00 | männlic | |
| 8 | 8 | 3 | 50,00 | weiblich | |
| 9 | 9 | 4 | 50,00 | männlic | |
| 10 | 10 | 5 | 150,00 | weiblich | |
| 11 | 11 | 5 | 20,00 | weiblich | |
| 12 | 12 | 5 | 20,00 | weiblich | |
| 13 | 13 | 5 | 20,00 | männlic | |
| 14 | 14 | 5 | 20,00 | weiblich | |
| 15 | 15 | 5 | 20,00 | männlic | |
| 16 | 15 | 5 | 20,00 | männlic | |
| 17 | | | | | |

**Abbildung 4.53:** Doppelter Fall in SPSS-Datendatei *PERSONEN4.SAV*

Mit der eben genannten Menüabfolge DATEN|DOPPELTE FÄLLE ERMITTELN gelangen Sie zum in Abbildung 4.54 dargestellten Dialogfeld, in das Sie bitte einfach die Personennummer *persnr* oben rechts eingeben. Sie geben deshalb hier *persnr* ein, weil Sie untersuchen sollen, ob zwei Personen versehentlich doppelt im Datensatz

vorkommen. In der entsprechenden Dialogbox habe ich ÜBEREINSTIMMENDE FÄLLE AN DEN ANFANG DER DATEI VERSCHIEBEN angekreuzt, sodass unsere beiden doppelten Fälle nunmehr am Anfang der Datendatei *PERSONEN4.SAV* stehen.

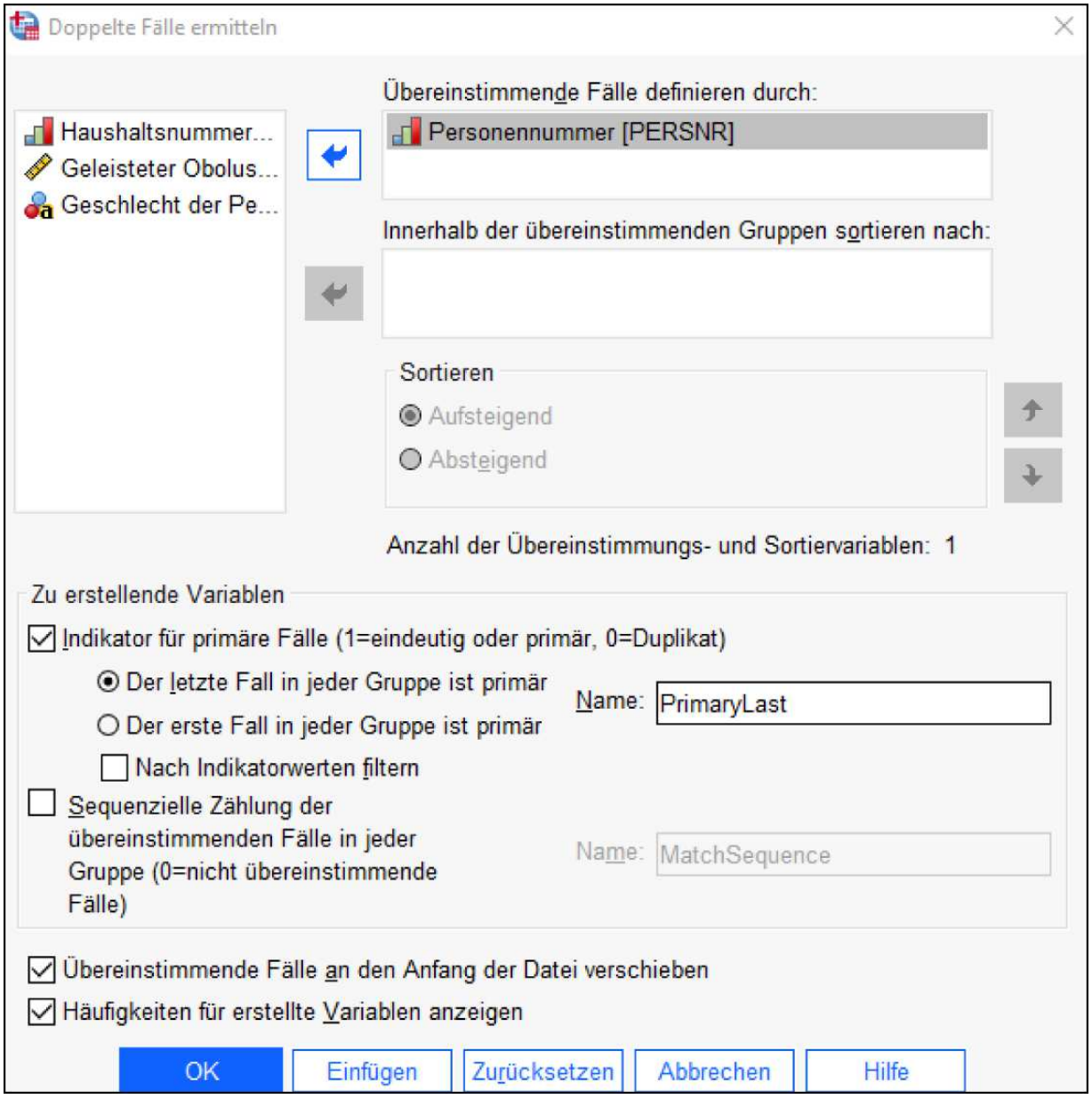



**Abbildung 4.54:** Dialogfeld zur Ermittlung doppelter Fälle

Anschließend ermittelt SPSS für Sie die doppelten Fälle, also die Fälle, die hinsichtlich der Ausprägungen bei allen Variablen gleich sind. Im vorliegenden Beispiel sind offenkundig Fall Nummer 15 und Fall

Nummer 16 identisch, was von SPSS auch korrekt ausgewiesen wird: Es wird ein doppelter Fall ausgewiesen, und als »Primärer Fall« wird Fall Nummer 15 benannt (siehe Abbildung 4.55). Letzteres ist so zu verstehen, dass dieser Fall zweimal im Datensatz ist.

**Statistiken**

Indikator jeder letzten Fallübereinstimmung als Primär

| | | |
|---|---|---|
| N | Gültig | 16 |
| | Fehlend | 0 |

**Indikator jeder letzten Fallübereinstimmung als Primär**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | Doppelter Fall | 1 | 6,3 | 6,3 | 6,3 |
| | Primärer Fall | 15 | 93,8 | 93,8 | 100,0 |
| | Gesamt | 16 | 100,0 | 100,0 | |

**Abbildung 4.55:** SPSS-Ergebnis der Ermittlung doppelter Fälle (*PERSONEN4.SAV*)

In einem nächsten Schritt könnte dieser doppelte Fall (ursprünglich Nummern 15 und 16) einmal aus dem Datensatz gelöscht werden. Dies ist hier ganz einfach: Sie müssen einfach in die erste Datenzeile gehen, diese markieren und dann mittels Tastatur löschen (siehe Abbildung 4.56).

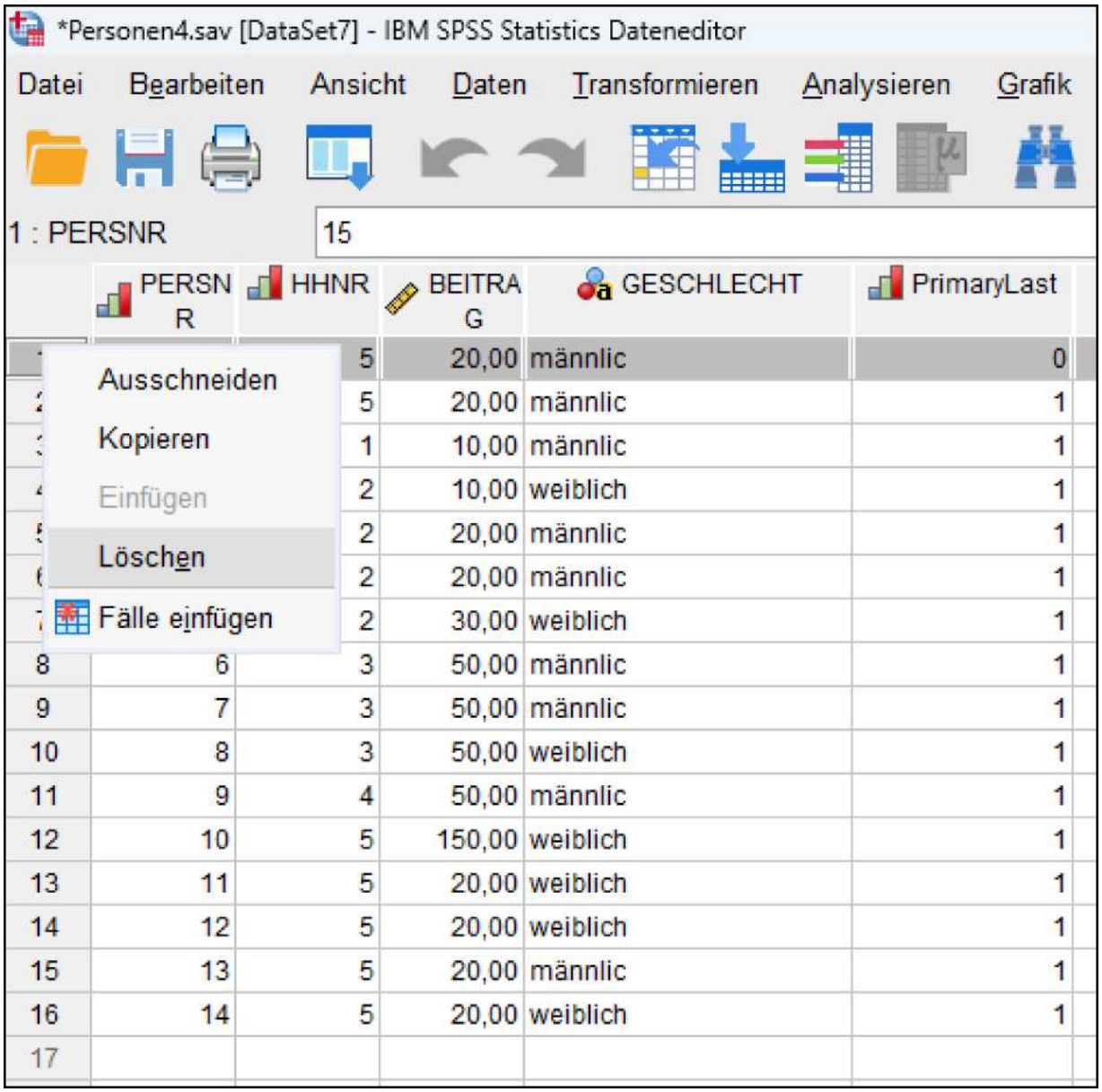



**Abbildung 4.56:** Löschen eines doppelten Falls (*PERSONEN4.SAV*)

# Kapitel 5

# Was Sie an Syntax reinstecken, bekommen Sie an Output raus

## IN DIESEM KAPITEL

Syntaxfiles

Outputfiles

In Kapitel 5 wird zunächst gezeigt, wie Sie in SPSS ohne das Benutzermenü auskommen. Es werden Ihnen in diesem Zusammenhang die Grundlagen der SPSS-Syntaxsprache nähergebracht. Deren Verwendung mündet dann in SPSS-Output.

## *Syntaxfiles*

Bislang haben Sie sich passiv, das heißt **benutzergeführt** innerhalb von SPSS bewegt. SPSS lässt sich aber auch aktiv nutzen, indem Sie über sogenannte **Syntaxprogramme** eine Benutzersteuerung betreiben. Konkret werden mehr oder weniger kleine Programme erstellt, die dann in SPSS abgearbeitet werden und – hoffentlich! – plausible Ergebnisse bringen.

Unter **Syntax** wird hier einfach die SPSS-Befehlssprache verstanden. Während der Nutzung des Programms SPSS läuft sie stets im Hintergrund. Die SPSS-Befehlssprache hat ihren historischen Kern darin, dass SPSS ursprünglich – ohne bequeme Menüsteuerung – auf Großrechnern im »Vor-Windows-Zeitalter« zum Einsatz kam. (Ich möchte Sie ja nicht langweilen, aber ich habe diese »Steinzeit« miterlebt, und ich kann Ihnen sagen, dass das Ganze seinerzeit schon etwas umständlich war. So, jetzt aber wieder in die Gegenwart!)

Über DATEI|NEU|SYNTAX kommen Sie menügesteuert in den Bereich, in den Sie die SPSS-Syntax hineinschreiben können. Diesen Bereich nennt man den **Syntaxeditor**. »Nackt« (das heißt: ohne jegliche Eingaben) sieht er aus, wie in Abbildung 5.1 angegeben. Auf seine konkrete Ausgestaltung wird nachfolgend peu à peu eingegangen.

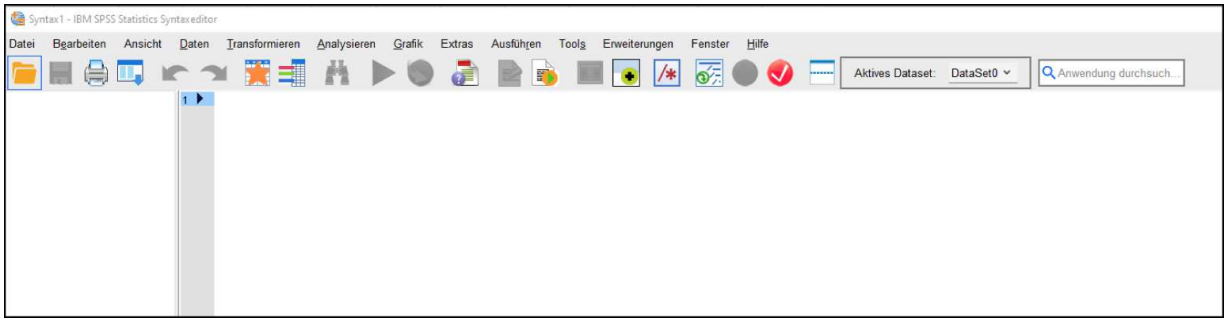



**Abbildung 5.1:** Der SPSS-Syntaxeditor

# *Vorteile von SPSS-Syntaxfiles*

Warum aber sollten Sie sich die Mühe machen und SPSS-Syntaxprogramme überhaupt erstellen? Nun ja, aus meiner Sicht gibt es eine ganze Reihe von Aspekten, die für derartige Programme sprechen:

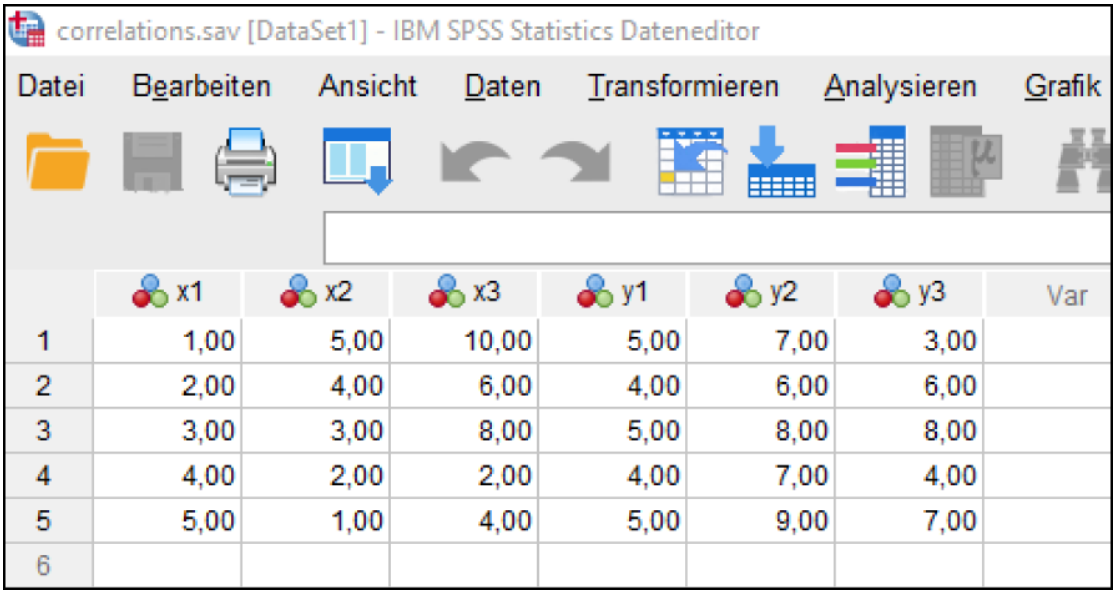
- ✔ Komplexere, insbesondere ineinander verschachtelte statistische Vorgänge lassen sich solcherart *übersichtlich* darstellen (vor allem auch dann, wenn alles gut kommentiert wird). Gerade auch bei längeren Vorgängen kann es schon nerven, wenn Sie im Benutzermenü ständig über die im Grunde genommen gleichen Dialogfelder »gehen« müssen. Einfacher ist es, wenn Sie die entsprechenden **Befehlsabfolgen** hintereinander aufschreiben, und Ihr Computer arbeitet diese Schritt für Schritt ab (während Sie sich in der Zwischenzeit ein gutes Glas Wein oder Ähnliches gönnen …).
- ✔ **Wiederkehrende Befehle,** die sich auf *andere* Gruppen der Untersuchungseinheiten beziehen, können Sie in einem Syntaxprogramm einfach durch Copy-and-Paste vervielfachen. Anschließend können Sie die Bezeichnungen – etwa bezüglich der Variablennamen – einfach durch SUCHEN UND ERSETZEN in Ihrem Syntaxprogramm anpassen. Nehmen Sie zum Beispiel an, Sie

wollten für Männer und Frauen getrennt deren Durchschnittseinkommen in einem Datensatz berechnen. Dann können Sie die betreffenden Befehle für die Männer hinschreiben, anschließend diese Befehle an eine andere Stelle im Syntaxprogramm kopieren und mittels SUCHEN UND ERSETZEN überall dort, wo »männlich« in dem kopierten Befehlsblock steht, einfach »weiblich« einfügen. Gleiche Analysen mit anderen Variablen sind also über SPSS-Syntaxfiles entsprechend leicht umsetzbar, indem der betreffende Syntaxtext einfach (ein wenig) überarbeitet wird.

- ✔ Es ist in SPSS-Syntaxprogrammen vergleichsweise leicht möglich, auch Syntax aus anderen Programmen wie **R** oder **Python** durch entsprechende Befehle zu integrieren (zum Beispiel bei R: BEGIN PROGRAM R…END PROGRAM).
- ✔ Sie können SPSS-Syntaxfiles erstellen, ohne dass Sie überhaupt SPSS auf Ihrem Rechner haben. Angenommen, Sie befinden sich auf Reisen und auf Ihrem Laptop ist kein SPSS installiert, dann können Sie dennoch über einen **Texteditor** (zum Beispiel WORD®) Ihr SPSS-Syntaxprogramm schon einmal schreiben und beispielsweise als *TXT*- oder *DOC(X)*-Datei abspeichern. Wenn Sie später wieder Zugriff auf SPSS haben (etwa auf einem anderen Rechner), können Sie den Inhalt der *TXT*- oder *DOC(X)*-Datei einfach nach SPSS in den Syntaxeditor kopieren und Ihre Auswertungen in SPSS starten.
- ✔ Die Vergabe von **Variablen-** beziehungsweise **Wertelabels**, also die Formatierungen beziehungsweise die Benennung der Variablen und ihrer Ausprägungen, ist mit VARIABLE LABELS beziehungsweise mit VALUE LABELS im Syntaxmodus relativ schnell und problemlos möglich.
- ✔ Bestimmte Operationen sind *nur* über die SPSS-Syntax möglich (beziehungsweise sehr leicht umzusetzen), zum Beispiel das Erstellen von Korrelationsmatrizen, also statistischer Zusammenhangstabellen, für bestimmte Kombinationen aus *mehreren* X- und *mehreren* Y-Variablen. In diesem Beispiel müssen Sie als Syntax lediglich eingeben:

```
CORRELATIONS
/VARIABLES = x1 x2 x3 WITH y1 y2 y3.
```

Für die Beispieldatendatei *CORRELATIONS.SAV* (siehe Abbildung 5.2) habe ich für drei X-Variablen, die in der resultierenden Korrelationsmatrix in den Zeilen stehen, und für drei Y-Variablen, die in den Spalten der Korrelationsmatrix zu finden sind, gemäß der vorstehenden Syntax eine entsprechende Korrelationsanalyse durchgeführt.




| | x1 | x2 | x3 | y1 | y2 | y3 | Var |
|---|---|---|---|---|---|---|---|
| 1 | 1,00 | 5,00 | 10,00 | 5,00 | 7,00 | 3,00 | |
| 2 | 2,00 | 4,00 | 6,00 | 4,00 | 6,00 | 6,00 | |
| 3 | 3,00 | 3,00 | 8,00 | 5,00 | 8,00 | 8,00 | |
| 4 | 4,00 | 2,00 | 2,00 | 4,00 | 7,00 | 4,00 | |
| 5 | 5,00 | 1,00 | 4,00 | 5,00 | 9,00 | 7,00 | |
| 6 | | | | | | | |

**Abbildung 5.2:** Beispieldatendatei *CORRELATIONS.SAV* für eine eingeschränkte Korrelationsmatrix

Wie aus Abbildung 5.3 ersichtlich, werden nur Ergebnis-(Korrelations-)Werte für die Kombinationen *x1/y1, x2/y2* und *x3/y3* ausgewiesen.

**Korrelationen**

| | | y1 | y2 | y3 |
|---|---|---|---|---|
| x1 | Pearson-Korrelation | ,000 | ,693 | ,457 |
| | Sig. (2-seitig) | 1,000 | ,194 | ,439 |
| | N | 5 | 5 | 5 |
| x2 | Pearson-Korrelation | ,000 | -,693 | -,457 |
| | Sig. (2-seitig) | 1,000 | ,194 | ,439 |
| | N | 5 | 5 | 5 |
| x3 | Pearson-Korrelation | ,577 | -,139 | -,076 |
| | Sig. (2-seitig) | ,308 | ,824 | ,903 |
| | N | 5 | 5 | 5 |

**Abbildung 5.3:** Eine eingeschränkte (syntaxgesteuerte) Korrelationsmatrix als Output (*CORRELATIONS.SAV*)

Wenn Sie hingegen über das Menü vorgehen, und zwar über ANALYSIEREN |KORRELATION|BIVARIAT, und im sich öffnenden Dialogfenster sämtliche Variablen unter VARIABLEN eingeben (also *x1, x2, x3, y1, y2* und *y3*), liefert Ihnen SPSS eine ausführlichere und, wie ich finde, unübersichtlichere Korrelationsmatrix. In dieser ausführlicheren Matrix sind alle X/Y-Kombinationen ausgewiesen (siehe den entsprechenden Output in Abbildung 5.4).

**Korrelationen**

| | | x1 | x2 | x3 | y1 | y2 | y3 |
|---|---|---|---|---|---|---|---|
| x1 | Pearson-Korrelation | 1 | -1,000** | -,800 | ,000 | ,693 | ,457 |
| | Sig. (2-seitig) | | <,001 | ,104 | 1,000 | ,194 | ,439 |
| | N | 5 | 5 | 5 | 5 | 5 | 5 |
| x2 | Pearson-Korrelation | -1,000** | 1 | ,800 | ,000 | -,693 | -,457 |
| | Sig. (2-seitig) | <,001 | | ,104 | 1,000 | ,194 | ,439 |
| | N | 5 | 5 | 5 | 5 | 5 | 5 |
| x3 | Pearson-Korrelation | -,800 | ,800 | 1 | ,577 | -,139 | -,076 |
| | Sig. (2-seitig) | ,104 | ,104 | | ,308 | ,824 | ,903 |
| | N | 5 | 5 | 5 | 5 | 5 | 5 |
| y1 | Pearson-Korrelation | ,000 | ,000 | ,577 | 1 | ,721 | ,264 |
| | Sig. (2-seitig) | 1,000 | 1,000 | ,308 | | ,170 | ,668 |
| | N | 5 | 5 | 5 | 5 | 5 | 5 |
| y2 | Pearson-Korrelation | ,693 | -,693 | -,139 | ,721 | 1 | ,508 |
| | Sig. (2-seitig) | ,194 | ,194 | ,824 | ,170 | | ,383 |
| | N | 5 | 5 | 5 | 5 | 5 | 5 |
| y3 | Pearson-Korrelation | ,457 | -,457 | -,076 | ,264 | ,508 | 1 |
| | Sig. (2-seitig) | ,439 | ,439 | ,903 | ,668 | ,383 | |
| | N | 5 | 5 | 5 | 5 | 5 | 5 |

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

**Abbildung 5.4:** Eine umfangreiche (menügesteuerte) Korrelationsmatrix als Output (*CORRELATIONS.SAV*)

Wie kommt nun dieser Unterschied zustande? Um dies herauszufinden, müssen Sie lediglich in die der Abbildung 5.4 zugrunde liegende Syntax schauen und diese Syntax dann mit der Syntax vergleichen, die für Abbildung 5.3 maßgeblich ist. Für Abbildung 5.4 ist (im Wesentlichen) folgende Syntax relevant:

```
CORRELATIONS
/VARIABLES = x1 x2 x3 y1 y2 y3.
```

Es fehlt also das Wörtchen »WITH«, das in der obigen Korrelationsanweisung vorhanden war. Kleine Ursache, große Wirkung also: Dadurch, dass »WITH« fehlt, werden alle X/Y-Kombinationen menügesteuert ausgewiesen!

- ✔ SPSS-Syntaxfiles erlauben es, dass Sie Outputfiles nicht unbedingt abspeichern müssen, sondern lediglich die zugrunde liegenden Befehle, die zu einem bestimmten Output führen. Diesen Output

können Sie mit einem Syntaxfile sozusagen jederzeit wieder erzeugen, ohne ihn unbedingt abspeichern zu müssen.

- ✔ SPSS-Syntaxfiles erleichtern die Kommunikation zwischen verschiedenen Personen (zum Beispiel Forschern) und auch die spätere Nachvollziehbarkeit der Berechnungsvorgänge (ohne das Ursprungsdatenfile verändern zu müssen).
- ✔ Haben Sie ein Syntaxprogramm vorliegen, müssen Sie nur noch auf den AUSFÜHREN-Knopf klicken. Sie kommen also sehr schnell zu einem Ergebnis – zumindest schneller, als wenn Sie sich vergleichsweise umständlich im Menü durch allerlei Dialogfenster beziehungsweise -abfragen quälen müssen.

# *Erforderliche Fähigkeiten und »Fallstricke« im Zusammenhang von SPSS-Syntaxprogrammen*

Bezüglich der Erstellung von SPSS-Syntaxprogrammen sind einige Fähigkeiten vonnöten, und es lauern ein paar »**Fallstricke**« auf Sie:

- ✔ Sie müssen einige **SPSS-Befehle** beherrschen. Das hört sich aber schlimmer an, als es ist, da im Regelfall nur ein paar SPSS-Befehle für Ihre Arbeit von Bedeutung sind. Außerdem gibt es ein umfangreiches Befehls-Handbuch in SPSS (in der Datei *SYNTAXREFERENCE. PDF*) beziehungsweise die Hilfe-Funktion im SPSS-Menü zum Nachschlagen der benötigten Befehle.
- ✔ Des Weiteren können Sie auch, wenn Sie unsicher sind, im SPSS-Benutzermenü auf die Schnelle die von Ihnen gewünschte Operation durchführen. Die betreffende Befehlsabfolge finden Sie gleich darauf im SPSS-**Outputmodus** (Outputfile), das heißt in einem Ausgabefenster, von wo aus Sie die Syntax in den Syntaxeditor kopieren und anschließend nach eigenem Gusto an Ihre Zwecke anpassen können (sofern Sie bei den Viewer-Einstellungen allerdings

BEFEHLE IM PROTOKOLL ANZEIGEN angeklickt haben). Öffnen Sie zum Beispiel einfach mal die Datendatei *PERSONEN4A.SAV*. Dann erscheint in diesem Fall in einem Ausgabefenster – in Abbildung 5.5 in »Ausgabe 1« – der Syntaxtext

```
GET FILE='C:\SPSS-DATEIEN\PERSONEN4A.SAV'.
DATASET NAME DataSet1 WINDOW=FRONT.
```

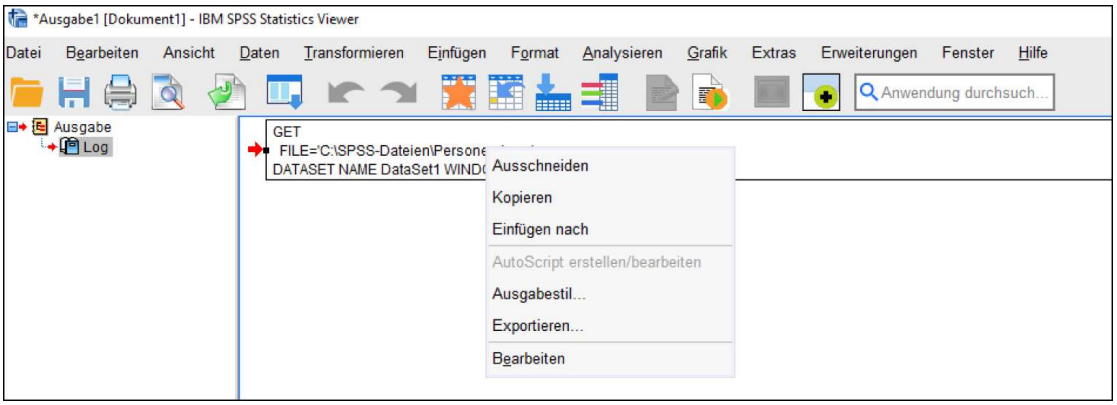



**Abbildung 5.5:** Syntax in einem Ausgabefenster (bezüglich *PERSONEN4A.SAV*)

Wenn Sie mit der rechten Maustaste auf diesen Text klicken, öffnet sich eine Box, in der Sie – wie in Abbildung 5.6 – einfach auf KOPIEREN klicken können. Dadurch kommt dieser Text in die Zwischenablage, und Sie können ihn mittels EINFÜGEN (oder über den Shortcut Strg + V) in den Syntaxeditor in das dort geöffnete Fenster einfügen.

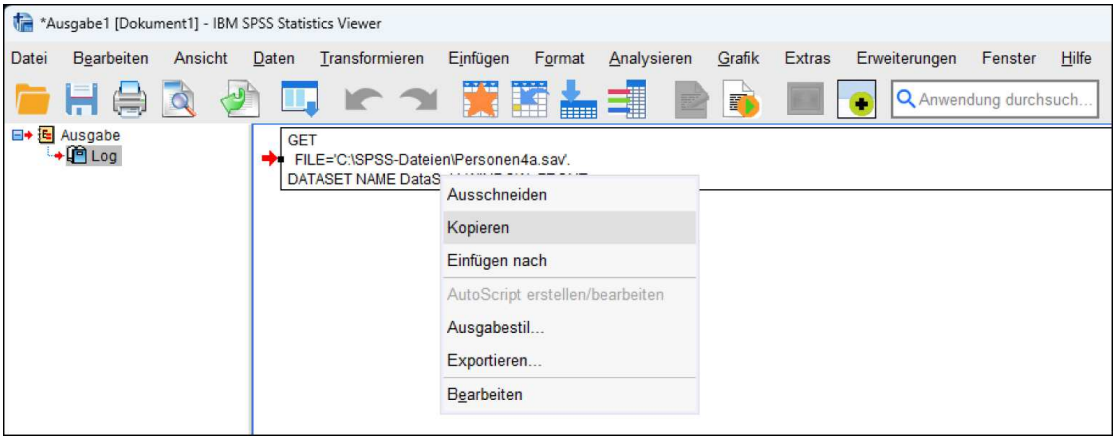



**Abbildung 5.6:** Kopieren von Syntax in einem Ausgabefenster (bezüglich *PERSONEN4A.SAV*)

✔ Alternativ bietet sich Ihnen bei nahezu jedem **Dialogfeld** die Möglichkeit, anstelle von OK vielmehr auf EINFÜGEN zu klicken. Angenommen, Sie möchten in der Datendatei *PERSONEN4A.SAV* nur die Fälle mit einer Personennummer kleiner gleich 10 berücksichtigen, dann können Sie im SPSS-Menü über DATEN|FÄLLE AUSWÄHLEN und die Bedingung *persnr* <= 10 die entsprechende Auswahl vornehmen. Anschließend sollen Sie aber nicht auf OK, sondern auf EINFÜGEN in der betreffenden Dialogbox klicken (das heißt: auf den zweiten Button von links unten in Abbildung 5.7).

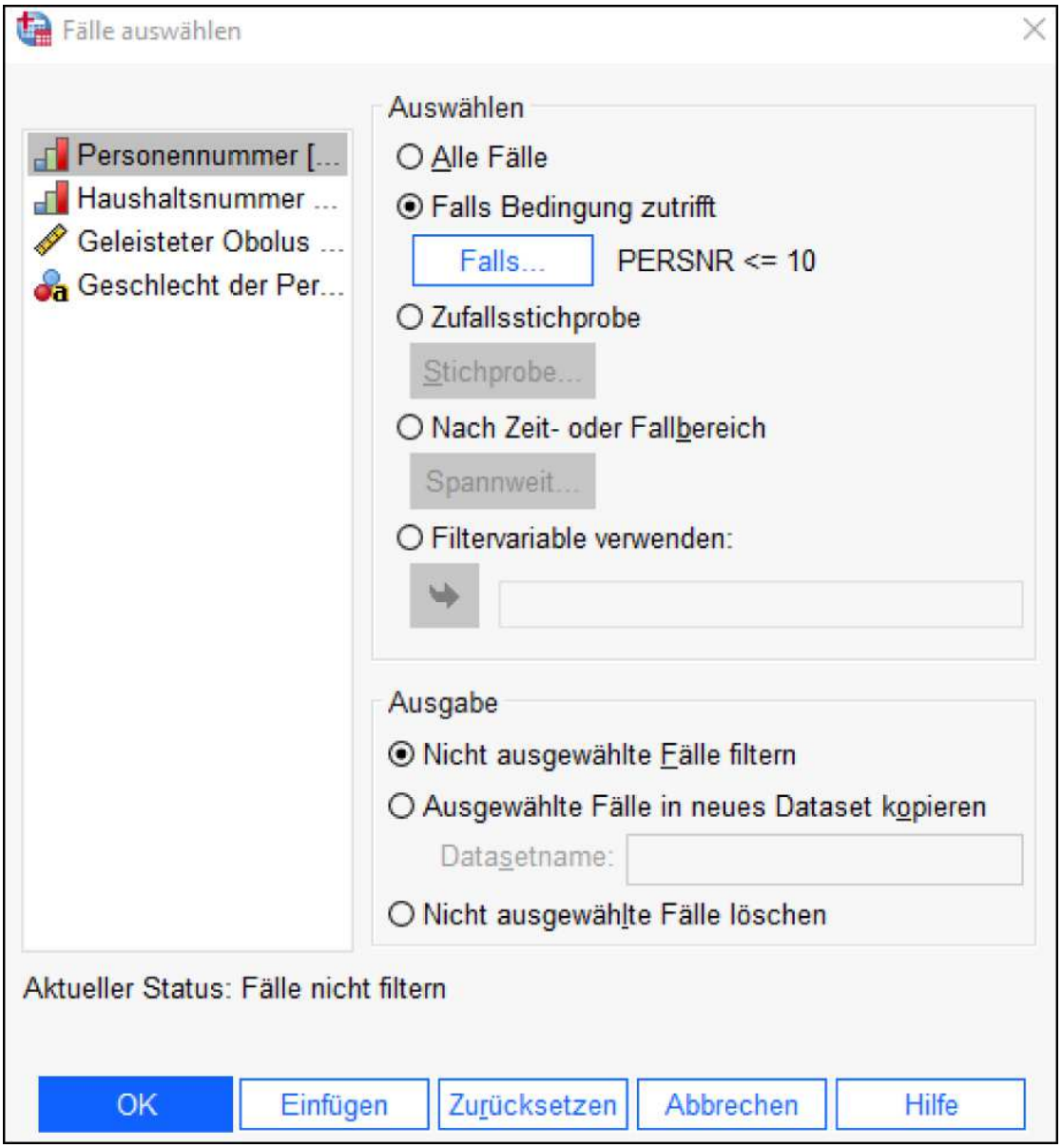


**Abbildung 5.7:** Option EINFÜGEN in einer SPSS-Dialogbox (*PERSONEN4A.SAV*)

Der zugehörige Syntaxtext wird nunmehr quasi automatisch in das geöffnete Syntaxfenster von SPSS geschrieben (siehe Abbildung 5.8). Auf diese Weise können Sie gewissermaßen zumindest halbautomatisch letztlich Ihre eigenen Syntaxprogramme erzeugen.

**Abbildung 5.8:** Im Syntaxeditor automatisch eingefügter Befehlstext (*PERSONEN4A.SAV*)

- Es sind Syntaxfehler möglich, die die (vollständige) Ausführung Ihres Programms blockieren. Im Outputmodus werden in diesem Fall aber **Warnungen** an den betreffenden Syntaxstellen ausgegeben, die Ihnen helfen, den Fehlern auf die Spur zu kommen. Aber auch Schreibfehler oder dergleichen werden direkt im Syntaxeditor angezeigt. Ich habe im Folgenden den SPSS-Befehl COMPUTE extra einmal falsch geschrieben (als KOMPUTE). Schon im Syntaxeditor erkennt man, dass ein Fehler vorliegen muss, da der SPSS-Befehl KOMPUTE nicht fett gedruckt ist (siehe Abbildung 5.9). Dies ist aber bei korrekten Befehlen der Fall.

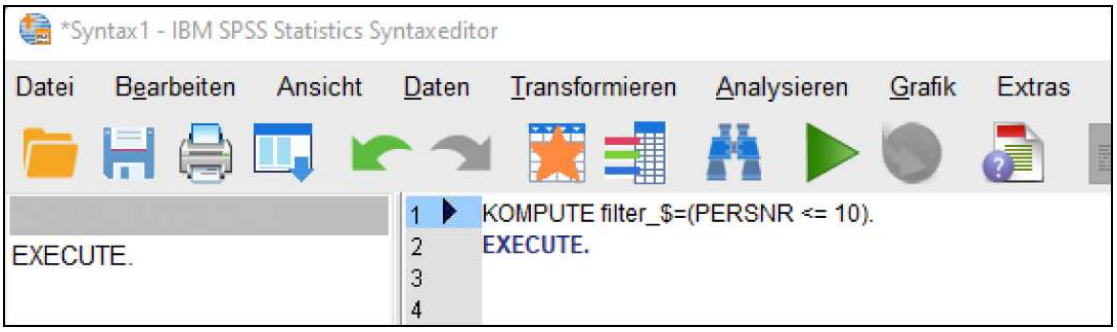



**Abbildung 5.9:** Syntaxfehler im Syntaxeditor (*PERSONEN4A.SAV*)

Die farbigen Syntax-Einstellungen können Sie nach eigenem Gutdünken im SPSS-Menü über BEARBEITEN|OPTIONEN und den dortigen Reiter SYNTAX-EDITOR einstellen. Bei mir sind gemäß Abbildung 5.10 sowohl die Syntaxfarb- als auch die Fehlerfarbcodierung aktiviert. Beispielsweise werden Hauptbefehle in SPSS bei mir fett gedruckt (das »B« steht für das englische Wort »bold«) und in Lila dargestellt.

Was in unserem Fall (mit dem Rechtschreibfehler) genauer los ist, wird in der von SPSS erzeugten Ausgabe deutlich. Dort wird ausgeführt, dass der Befehl KOMPUTE nicht bekannt ist (siehe Abbildung 5.11). Da hat SPSS aber auch so was von recht: Schließlich ist die SPSS-Syntax an die englische Sprache angelehnt, und im Englischen schreibt man schließlich COMPUTE und nicht KOMPUTE!

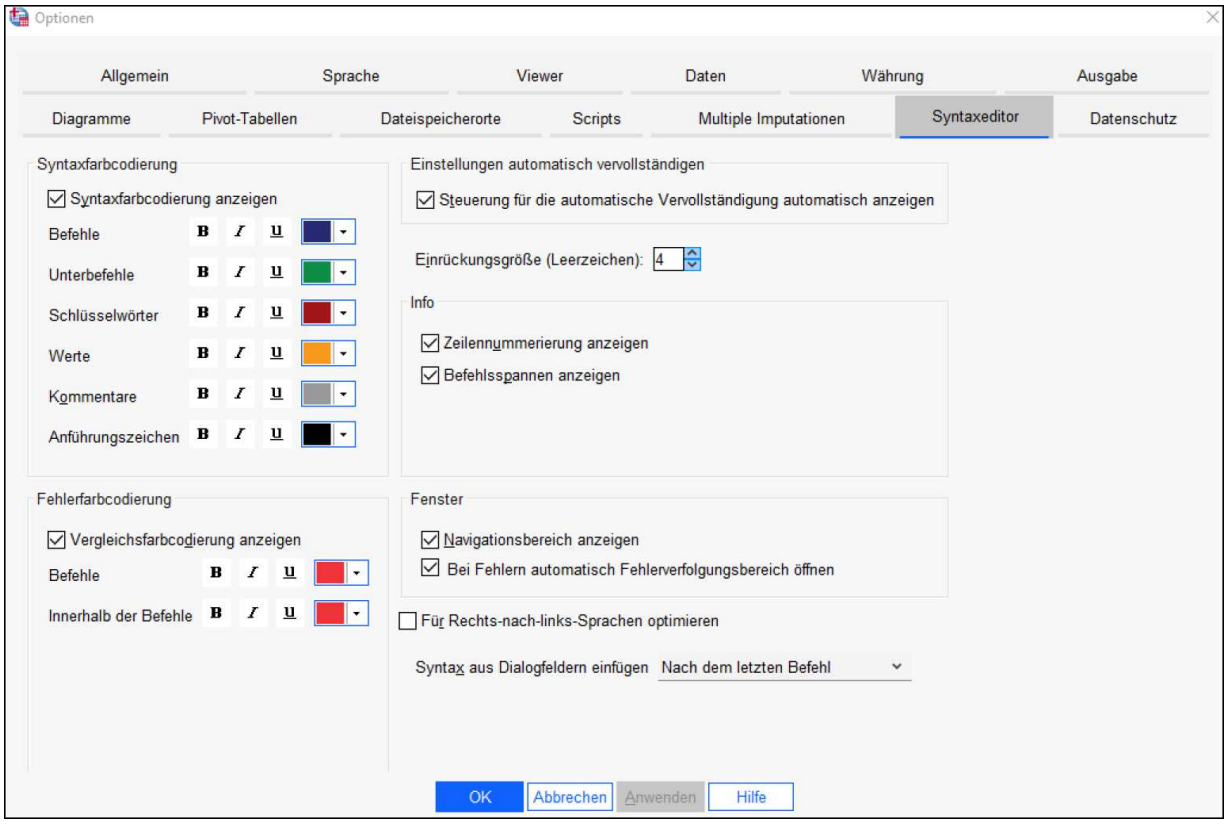



**Abbildung 5.10:** OPTIONEN-Einstellungen für den Syntaxeditor

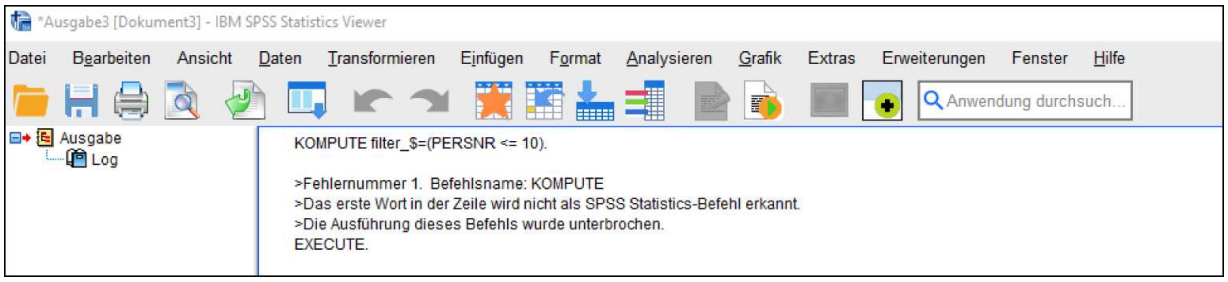



**Abbildung 5.11:** Syntaxfehler-Hinweis im Ausgabefenster (*PERSONEN4A.SAV*)

Ich empfehle bei (sehr) großen Datenfiles die Erstellung kleiner »**Spiel-Datenfiles** «, die vom Aufbau her den Original-Datenfiles gleichen, aber viel weniger Fälle aufweisen. Gerade im Falle von Syntaxfehlern sparen Sie hierdurch nicht nur Rechnerzeit, sondern schonen auch Ihre Nerven! Stellen Sie sich vor, Sie hätten einen Datensatz mit mehreren Tausend Fällen und Hunderten von Variablen und müssten ganz komplizierte Datentransformationen durchführen, dann kann es ungeheuer nerven, wenn zum Beispiel in Befehlszeile Nummer 1.035 ein Syntaxfehler ist und Ihr Syntaxprogramm nach vielleicht einer Stunde Laufzeit abgebrochen wird. Daher sollten Sie hier nach dem Motto »Small is beautiful« handeln und sich ein Datenfile mit einigen wenigen Fällen und einigen wenigen (benötigten) Variablen kreieren, das nur eine Laufzeit von wenigen Minuten (vielleicht sogar nur von wenigen Sekunden) hat. Dann ärgern Sie sich bei einem Programmabbruch vermutlich nicht so sehr, als wenn Sie eine Stunde hoffnungsfroh vor Ihrem PC saßen und am Ende von einem Programmabbruch enttäuscht wurden.

✔ Es ist auf jeden Fall auf eine gute **Kommentierung** Ihrer Syntax zu achten. Ist das nämlich nicht der Fall, können Außenstehende nur schwer nachvollziehen, was Sie da eigentlich gemacht haben. Noch schlimmer: Mit einigem zeitlichen Abstand können Sie selbst nicht mehr nachvollziehen, was genau Sie mit Ihrem Programm ein paar Wochen vorher eigentlich fabrizieren wollten. Ich bin an der Stelle leider auch ein gebranntes Kind … ;-)

# Zum Aufbau von SPSS-Syntaxprogrammen

Eigentlich ist der Aufbau eines **SPSS-Befehls** ganz leicht. Am Anfang einer Befehlszeile kommt der **Befehlsname**, dann die **Variable(n)**, auf die sich das Ganze bezieht, und schließlich die **Operation** in Bezug auf die Variable(n). Den Befehlsnamen nennt man auch *Befehlsschlüsselwort* (»Keyword«) und die Operationen auch *Spezifikationen*, die zum Beispiel arithmetische Operatoren darstellen (zum Beispiel + für die Addition, - für die Subtraktion, / für die Division, * für die Multiplikation und ** für das Potenzieren).

## Ein Beispiel

Ein Beispiel für eine SPSS-Syntax ist die folgende, mit der Sie sich Häufigkeiten für eine Variable *x1* »auswerfen« lassen (inklusive Balkendiagrammen; im inhaltlichen Detail werden solche FREQUENCIES-Berechnungen im Übrigen in Kapitel 6 präsentiert):

```
FREQUENCIES VARIABLES = x1
/BARCHART.
```

- In der ersten Zeile findet sich der Kommando-/Befehlsname (hier: `FREQUENCIES`) zusammen mit einer ersten Spezifikation, nämlich, dass sich dieser Befehl konkret auf die Variable *x1* beziehen soll. Hierdurch wird für diese Variable eine Häufigkeitstabelle ausgegeben (Näheres folgt in inhaltlicher Hinsicht, wie geschrieben, in Kapitel 6). In der zweiten Zeile steht ein Unterbefehl – inhaltlich bedeutet `/BARCHART`, dass für die in der ersten Zeile angegebene Variable *x1* ein sogenanntes Balkendiagramm ausgegeben werden soll. Dass es sich um einen Unterbefehl von `FREQUENCIES` handelt, wird an dem `BARCHART` vorangestellten Schrägstrich deutlich. In SPSS sind bei jedem Befehl die möglichen Unterbefehle genau vorgegeben. Im `FREQUENCIES`-Zusammenhang sind zum Beispiel weitere Unterbefehle:

`/FORMAT = NOTABLE`: Hierdurch wird auf eine Häufigkeitstabelle verzichtet.

`/STATISTICS = ALL`: Hierdurch werden statistische Kennziffern wie Mittelwerte oder Standardabweichungen und dergleichen »ausgeworfen«, wobei Sie – anstelle von `ALL` wie hier – auch nur einzelne Kennziffern wie den Median (MEDIAN) oder das arithmetische Mittel (MEAN) auswählen können:

```
/STATISTICS = MEDIAN MEAN.
```

Wichtig ist, dass die ganze Befehlsabfolge am Ende mit einem Punkt abgeschlossen wird. Dieser Punkt signalisiert SPSS: »Kommando beendet«. Mithilfe einer Befehlsabfolge unterhalten Sie sich sozusagen mit SPSS, wählen ein paar Dinge aus (Variablen, Unterbefehle), um sich dann von SPSS bezogen auf dieses »Gespräch« mit einem »Go!« zu verabschieden. Noch einmal – und darauf werde ich im Folgenden noch mehrfach hinweisen: Ganz am Ende steht bei einer SPSS-Anweisung immer der Punkt!

Bezogen auf den Datensatz *CORRELATIONS.SAV* mit den Variablen *x1, x2, x3, y1, y2* und *y3*, müssen Sie mit der obigen FREQUENCIES-Anweisung unter Auswahl der Variablen *x1* unbedingt beachten, dass Sie bei mehreren geöffneten Datendateien die richtige erwischen. Im Beispiel hier habe ich in der Tat mal mehrere Datendateien geöffnet, und, wie Sie Abbildung 5.12 ganz oben links entnehmen können (dort steht ja »*CORRELATIONS.SAV* [DataSet2] – IBM SPSS Statistics Dateneditor«), ist hier *CORRELATIONS.SAV* das DataSet2.

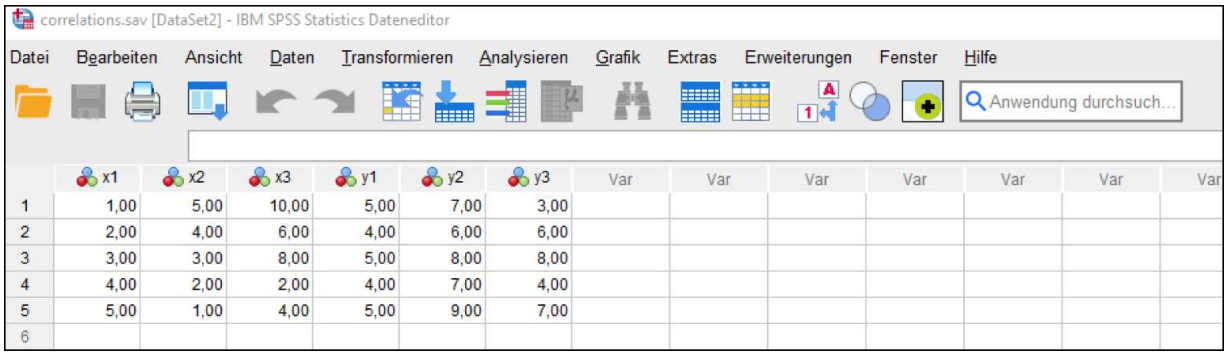



**Abbildung 5.12:** Datendateibezeichnung bei mehreren geöffneten Datendateien

Entsprechend müssen Sie im Syntaxeditor mit Ihrer Syntax dieses DataSet2 auch auswählen (siehe Abbildung 5.13). Sie müssen hier also höllisch aufpassen, dass Sie die für Ihre Zwecke maßgebliche Datendatei auswählen; ansonsten gibt's ein schreckliches Durcheinander (gerade auch noch, wenn die Variablennamen in den geöffneten Datendateien gleich sind)!

**Abbildung 5.13:** Auswahl einer Datendatei (*CORRELATIONS.SAV*) im Syntaxeditor bei mehreren geöffneten Datendateien

Wenn Sie nun im Syntaxeditor oben in der Menüleiste auf AUSFÜHREN|ALLE klicken – dazu im nächsten Unterabschnitt dieses Kapitels gleich mehr! –, wird Ihr Befehl ausgeführt, und in unserem Beispielfall erhalten Sie den gewünschten Output im Ausgabefenster, nämlich eine Häufigkeitstabelle und ein Balkendiagramm für *x1* (beziehungsweise die Häufigkeitsverteilung für *x1*), wie in Abbildung 5.14 dargestellt (Ausgabefenster werden im Übrigen im nächsten Abschnitt dieses Kapitels noch ausführlicher behandelt).

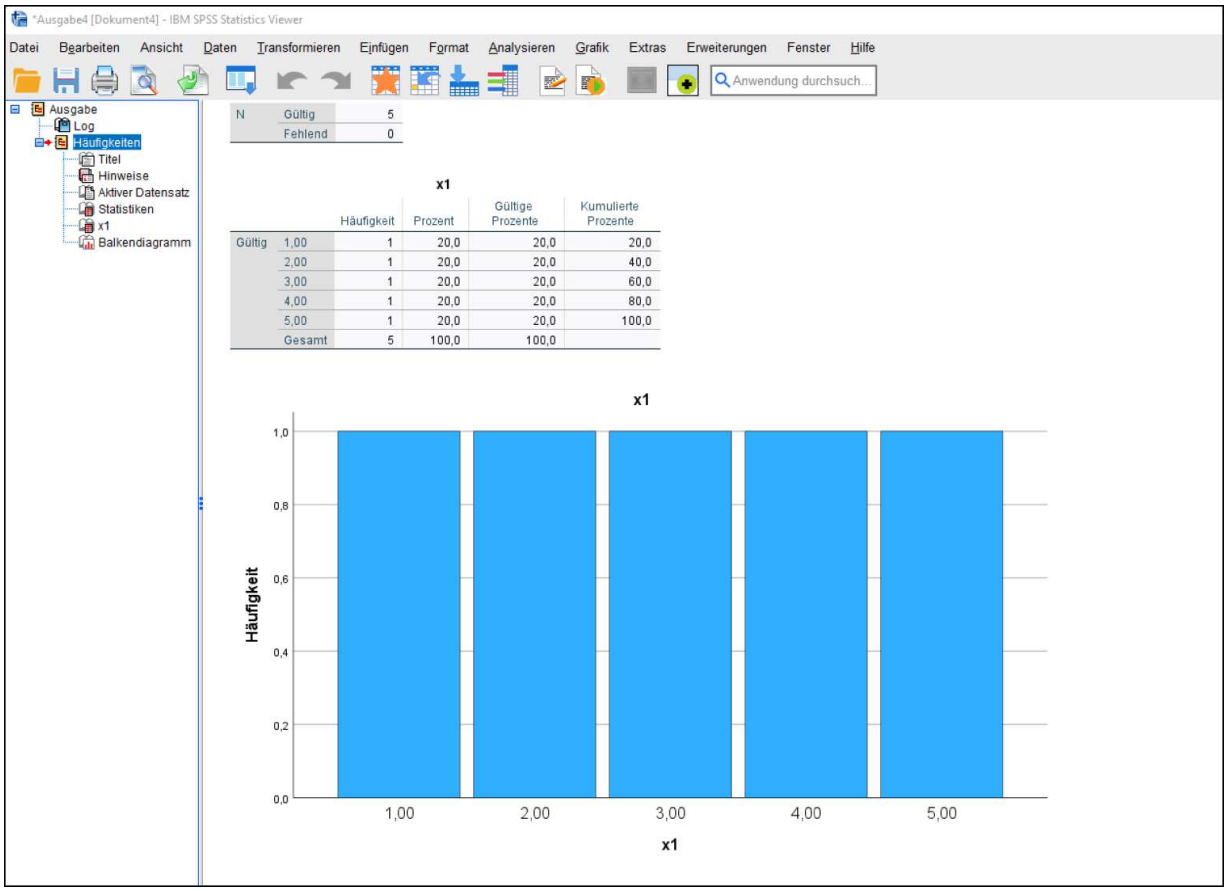



**Abbildung 5.14:** Output eines spezifischen SPSS-Befehls (für *CORRELATIONS.SAV*)

Bitte trennen Sie zwei Elemente der SPSS-Befehlssprache durch mindestens ein Leerzeichen oder durch einen Zeilenwechsel. Wenn Sie sich – um bei der Beispieldatendatei *CORRELATIONS.SAV* zu bleiben – Häufigkeitstabellen für mehrere Variablen ausweisen lassen möchten und Ihre Syntax etwa

```
FREQUENCIES VARIABLES = x1 x2 y1.
```

lautet, müssen Sie sinnvollerweise zwischen den drei hier angegebenen Variablen *x1, x2* und *y1* jeweils ein Leerzeichen setzen. Das ist ja eigentlich auch trivial, handelt es sich doch um drei Variablennamen, also sozusagen um drei Wörter.

Bei Zeichen, die von SPSS eine bestimmte Bedeutung zugewiesen bekommen haben, wie bei den für mathematische Operationen bedeutsamen Zeichen

- »=« (»gleich«)
- »<>« (»ungleich«)
- »<« (»kleiner als«)
- »>« (»größer als«)
- »(« (»geöffnete Klammer«)
- »)« (»geschlossene Klammer«)
- »+« (»plus«, bei der Addition)
- »-« (»minus«, bei der Subtraktion) und so weiter

müssen Sie diese Vorschrift hingegen nicht beachten. Das heißt: Vor und nach derartigen Zeichen brauchen Sie keine Leerzeichen zu setzen; Sie können das aber gefahrlos dennoch tun!

Entsprechend sind zum Beispiel die Befehle

```
FREQUENCIES VARIABLES = x1 x2 y1.
```

und

```
FREQUENCIES VARIABLES=x1 x2 y1.
```

völlig gleichwertig zueinander.

Wie oben ersichtlich, kann ein SPSS-Kommando über mehrere Zeilen hinweg fortgesetzt werden. Dabei sollten Sie aber keine Leerzeile *innerhalb* eines SPSS-Kommandos setzen, da diese analog zum Punkt als Ende des Kommandos interpretiert würde. *Zwischen* zwei Kommandos dürfen Sie hingegen problemlos so viele Leerzeilen setzen, wie Sie lustig sind.

Wenn Sie demnach den – auf die Variable *x1* in der Datendatei *CORRELATIONS.SAV* bezogen – FREQUENCIES-Befehl von seinem Unterbefehl \BARCHART durch eine Leerzeile trennen (siehe Abbildung 5.15), bekommen Sie ein Problem (wie in Abbildung 5.16 dargestellt): SPSS weigert sich, den Unterbefehl auszuführen!

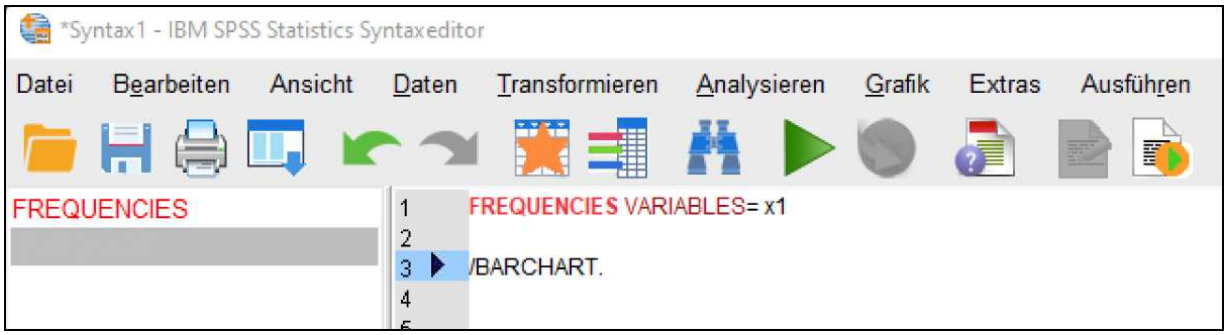



**Abbildung 5.15:** Leerzeile innerhalb eines SPSS-Kommandos

Als Grundregel gilt (wie bereits erwähnt), dass jedes Kommando in einer neuen Zeile beginnt und mit einem Punkt endet.

## *Kommentare*

Der Übersichtlichkeit wegen sollten Sie Ihr Tun im Rahmen von SPSS-Syntaxprogrammen auf jeden Fall mit **Kommentaren** über COMMENT beziehungsweise mit Sternchen * versehen (oder auch mittels: /*…*/). Dadurch wird für Außenstehende sichtbar, was Sie mit Ihrem Programm so alles treiben. Auch ein Kommentar endet mit einem Punkt (siehe Abbildung 5.17). Dadurch wird SPSS angezeigt, dass der Kommentar zu Ende ist.

## Häufigkeiten

**Statistiken**

x1

| | | |
|---|---|---|
| N | Gültig | 5 |
| | Fehlend | 0 |

**x1**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | 1,00 | 1 | 20,0 | 20,0 | 20,0 |
| | 2,00 | 1 | 20,0 | 20,0 | 40,0 |
| | 3,00 | 1 | 20,0 | 20,0 | 60,0 |
| | 4,00 | 1 | 20,0 | 20,0 | 80,0 |
| | 5,00 | 1 | 20,0 | 20,0 | 100,0 |
| | Gesamt | 5 | 100,0 | 100,0 | |

```
/BARCHART.

>Fehlernummer 1.  Befehlsname: /BARCHART
>Das erste Wort in der Zeile wird nicht als SPSS Statistics-Befehl erkannt.
>Die Ausführung dieses Befehls wurde unterbrochen.
```

**Abbildung 5.16:** Unvollständiger Output bei einer Leerzeile innerhalb eines SPSS-Kommandos (für *CORRELATIONS.SAV*)

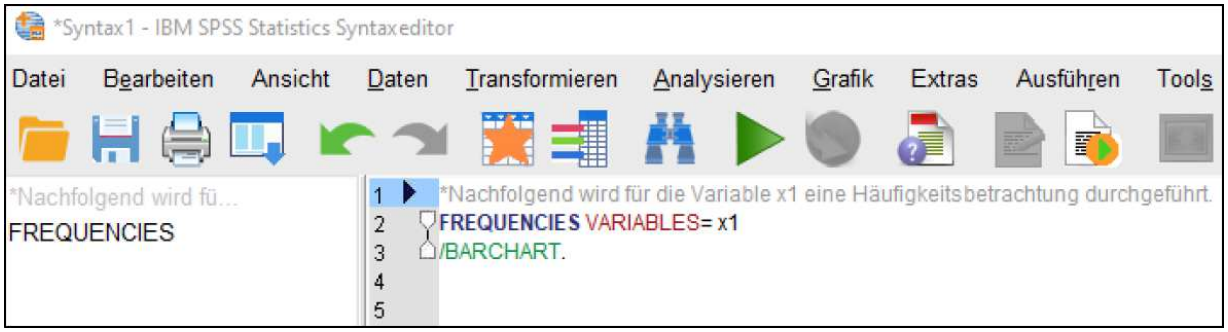



**Abbildung 5.17:** Beendigung eines Kommentars mit einem Punkt

Wenn Ihnen die Beendigung eines Kommentars mit einem Punkt zuwider ist, etwa weil Sie einen Kommentar gerne mit einem anderen Satzzeichen beenden möchten wie einem Doppelpunkt, müssen Sie SPSS das Ende des Kommentars anders anzeigen. Dies tun Sie – wie in Abbildung 5.18 gezeigt – durch eine Leerzeile zwischen Kommentar und nächstem SPSS-Befehl.

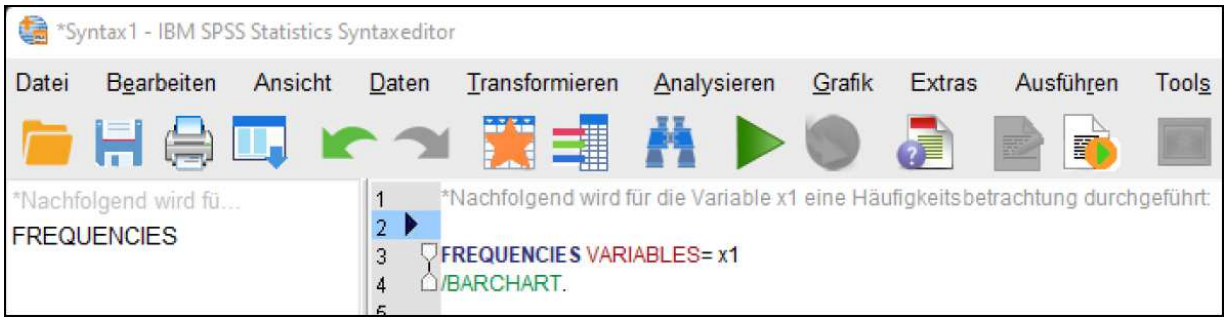



**Abbildung 5.18:** Beendigung eines Kommentars mit einer Leerzeile

Auch ein Kommentar-Kommando darf sich über mehrere Zeilen hinweg erstrecken (siehe Abbildung 5.19).

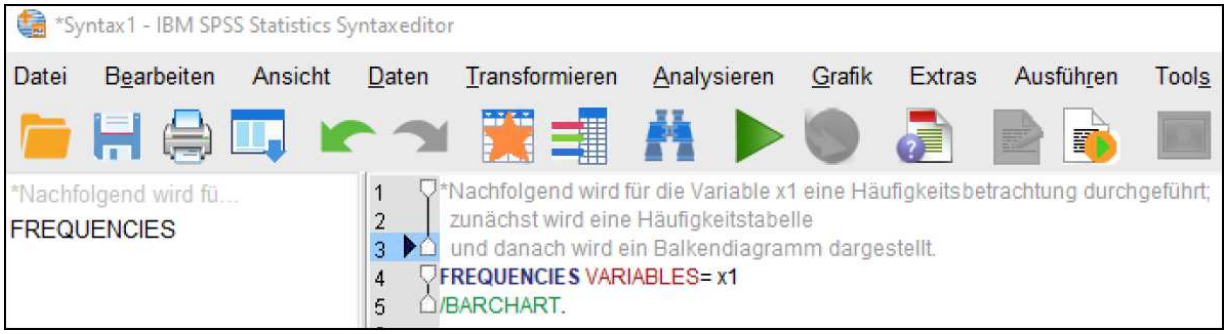



**Abbildung 5.19:** Mehrzeiliger Kommentar

Sie sollten hierbei allerdings keine Leerzeilen innerhalb des Kommentar-Kommandos verwenden, weil hierdurch, wie bereits in Abbildung 5.18 dargestellt, das Ende des Kommentars signalisiert würde. Dieser Fall ist in Abbildung 5.20 dargelegt.

**Abbildung 5.20:** Leerzeile innerhalb eines mehrzeiligen Kommentars

Im vorliegenden Beispiel passiert allerdings nicht viel: SPSS präsentiert Ihnen zwar eine Fehlermeldung, führt aber dennoch »brav« die gewünschten Befehle in Bezug auf die Häufigkeitstabelle und das Balkendiagramm aus (siehe Abbildung 5.21). Dennoch empfehle ich Ihnen, sich grundsätzlich an die Syntaxregeln (auch bei den Kommentaren) zu halten; dann sind Sie auf jeden Fall auf der »sicheren Seite«.

Ein Kommentar-Kommando endet bekanntlich grundsätzlich mit einem Punkt. Ist ein Punkt von Ihnen allerdings als Satzzeichen am Ende einer Kommentarzeile vorgesehen, müssen Sie die nächste Kommentarzeile wieder mit COMMENT beziehungsweise * beginnen. Punkte im Sinne von Satzzeichen innerhalb einer Kommentarzeile (siehe Abbildung 5.22) sind hingegen kein Problem.

Wie Sie nämlich Abbildung 5.23 entnehmen können, wird Ihr Kommentar in diesem Fall in der gewünschten Weise abgebildet.

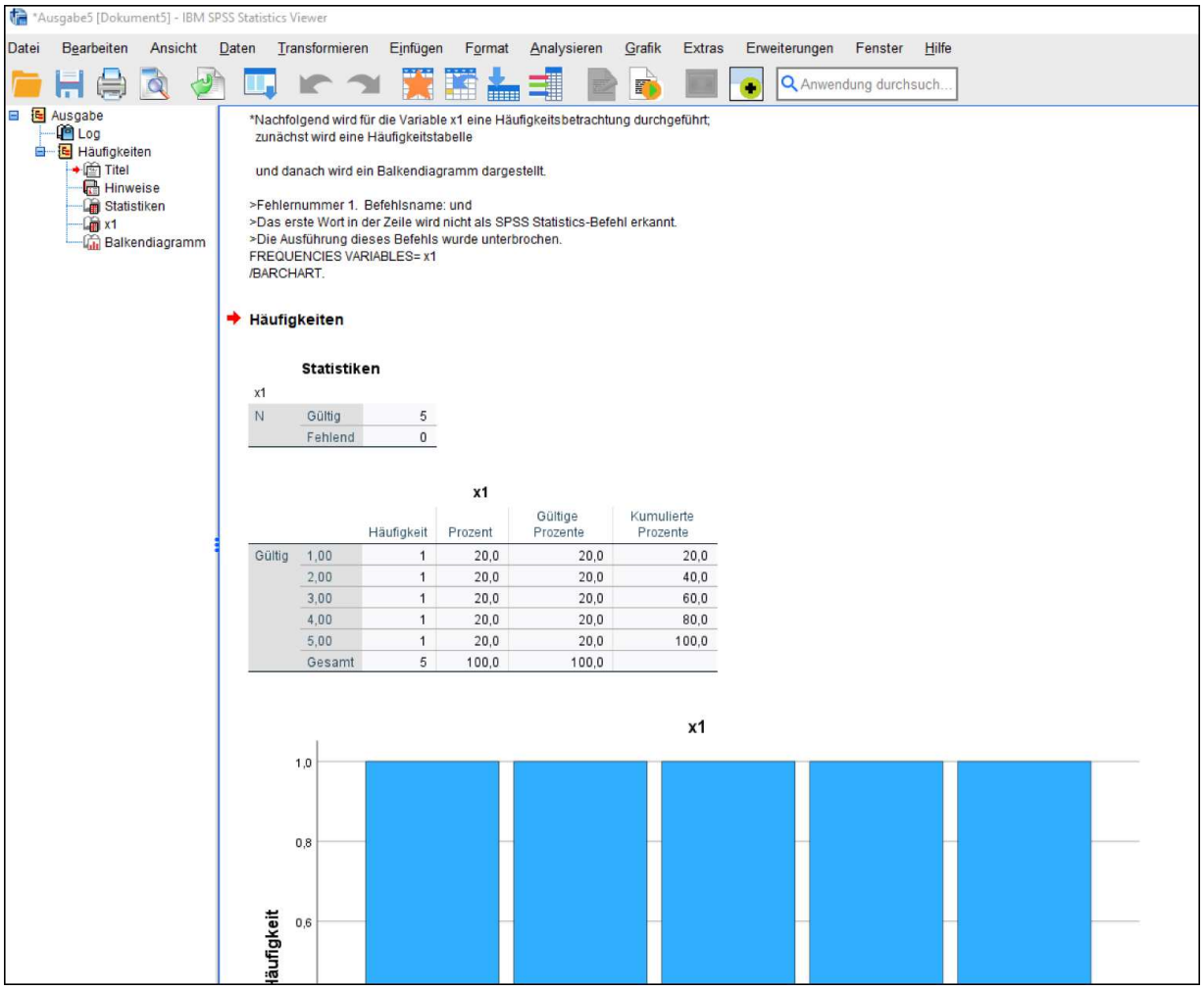



**Abbildung 5.21:** Output bei einer Leerzeile innerhalb eines mehrzeiligen Kommentars (für *CORRELATIONS.SAV*)

**Abbildung 5.22:** Punkt als Satzzeichen innerhalb eines Kommentars

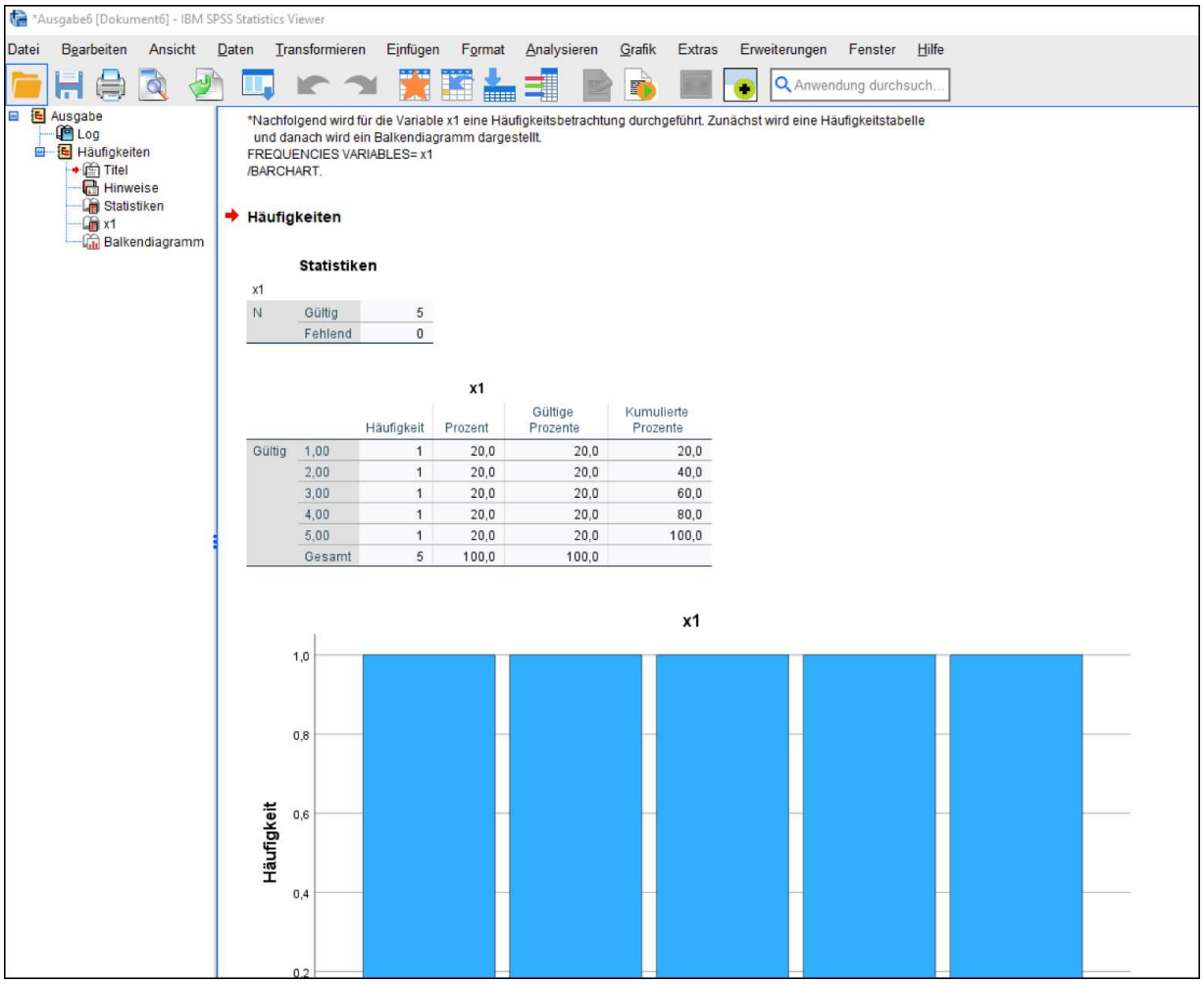



**Abbildung 5.23:** Problemloser Output bei einem Punkt als Satzzeichen innerhalb eines Kommentars (für *CORRELATIONS.SAV*)

# *Zugänge zur SPSS-Syntax*

Sie haben verschiedene Möglichkeiten, mit der SPSS-Syntax zu arbeiten:

- ✔ Zugang über den SPSS-Syntaxeditor
- ✔ Zugang über SPSS-Dialogboxen
- ✔ Copy-and-Paste von Syntax
- ✔ Öffnen einer bestehenden SPSS-Syntaxdatei über DATEI|ÖFFNEN
- ✔ Zugang über die SPSS-Journaldatei *STATISTICS.JNL*

Alle diese Zugangsmöglichkeiten werden im Folgenden präsentiert.

## *Zugang über Syntaxeditor*

Im vorhergehenden Abschnitt dieses Kapitels wurde schon ein paarmal auf die Zugangsmöglichkeit über den **SPSS-Syntaxeditor** eingegangen. Diese Zugangsmöglichkeit erlangen Sie über DATEI|NEU|SYNTAX. Im sich öffnenden Syntaxeditor können Sie jetzt Ihre Befehle in der »SPSS-Sprache« eingeben. SPSS unterbreitet Ihnen hierbei Vorschläge für in SPSS existierende Befehle an, wie dies in Abbildung 5.24 für den GET-FILE-Befehl angedeutet ist. SPSS bietet Ihnen hier drei mögliche Zusätze für GET an: FILE sowie LOCK und PASSWORD. Darüber hinaus werden in einer anderen Farbe vier mögliche Unterbefehle offeriert: DROP, KEEP, MAP und RENAME.

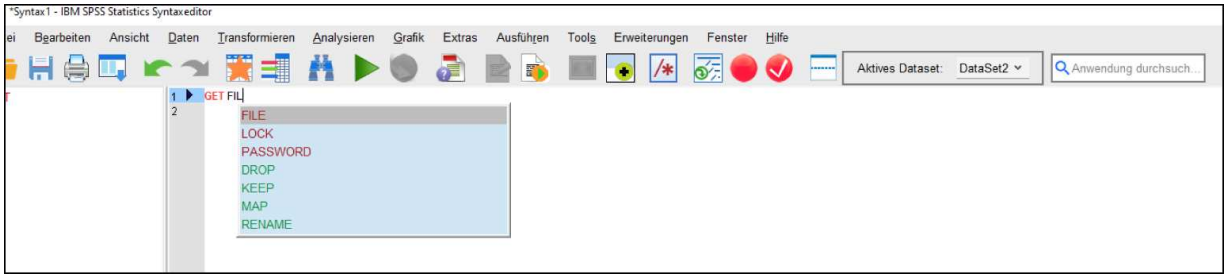



**Abbildung 5.24:** Der SPSS-Syntaxeditor als Zugangsweg

Noch mehr Hilfestellung erhalten Sie, wenn Sie im Menü oben rechts auf den HILFE-Button (HILFE ZUR SYNTAX) klicken (siehe Abbildung 5.25).

**Abbildung 5.25:** Syntaxhilfe über das »Fragezeichen-Symbol« im Syntaxeditor

Im vorliegenden Fall leistet Ihnen SPSS onlinebasiert zum Befehl GET FILE Syntaxhilfe: Es werden (internetbasiert) die grundsätzliche Struktur dieses Befehls beziehungsweise die Programmiermöglichkeiten, die dieser Befehl bietet, gezeigt (siehe Abbildung 5.26).

Alternativ können Sie sich die einzelnen SPSS-Syntaxbefehle und ihren Aufbau auch im mit SPSS mitgelieferten Benutzerhandbuch *SYNTAXREFERENCE.PDF* anschauen.

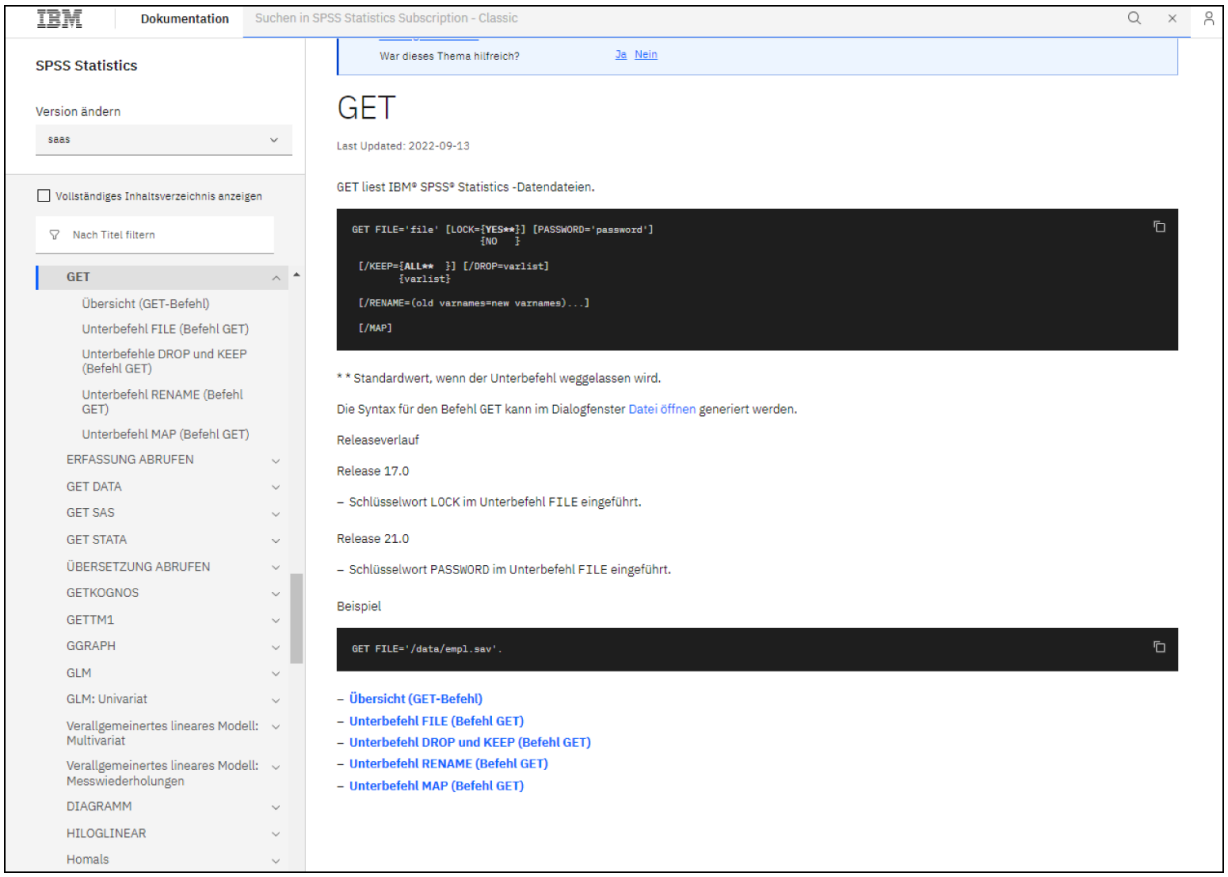



**Abbildung 5.26:** Online-Syntaxhilfe zu einem bestimmten Befehl (hier bezüglich GET FILE)

## *Zugang über Dialogboxen*

Sie finden in einer SPSS-Dialogbox neben dem OK-Button die Variante EINFÜGEN. Klicken Sie hierauf, wird der entsprechende SPSS-Befehl nicht direkt ausgeführt, sondern erst einmal in den Syntaxeditor geschrieben. Auf diese Weise können Sie sich sozusagen per Knopfdruck ein eigenes Syntaxprogramm zusammenstellen. In Abbildung 5.7 und Abbildung 5.8 ist das entsprechende Vorgehen bereits skizziert worden.

## *Zugang über Copy-and-Paste*

Wie schon geschrieben, können Sie Ihre Befehle auch in WORD oder einem anderen Textverarbeitungsprogramm aufschreiben und das Ganze dann später nach SPSS kopieren. Alternativ können Sie über das Menü bestimmte Aktionen durchführen lassen und sich die zugehörige Syntax in ein Outputfile schreiben lassen. Auch diese Syntax können Sie dann

in den Syntaxeditor kopieren. Das entsprechende Vorgehen wurde oben in Abbildung 5.5 und Abbildung 5.6 bereits gezeigt.

## *Zugang über ein bestehendes Syntaxfile*

Existiert bereits ein SPSS-Syntaxfile, so können Sie dieses über DATEI|ÖFFNEN|SYNTAX… einfach öffnen. Ein SPSS-Syntaxfile erkennen Sie an der Extension *.SPS*. In Abbildung 5.27 wird die Syntaxdatei *SYNTAX2.SPS* geöffnet.

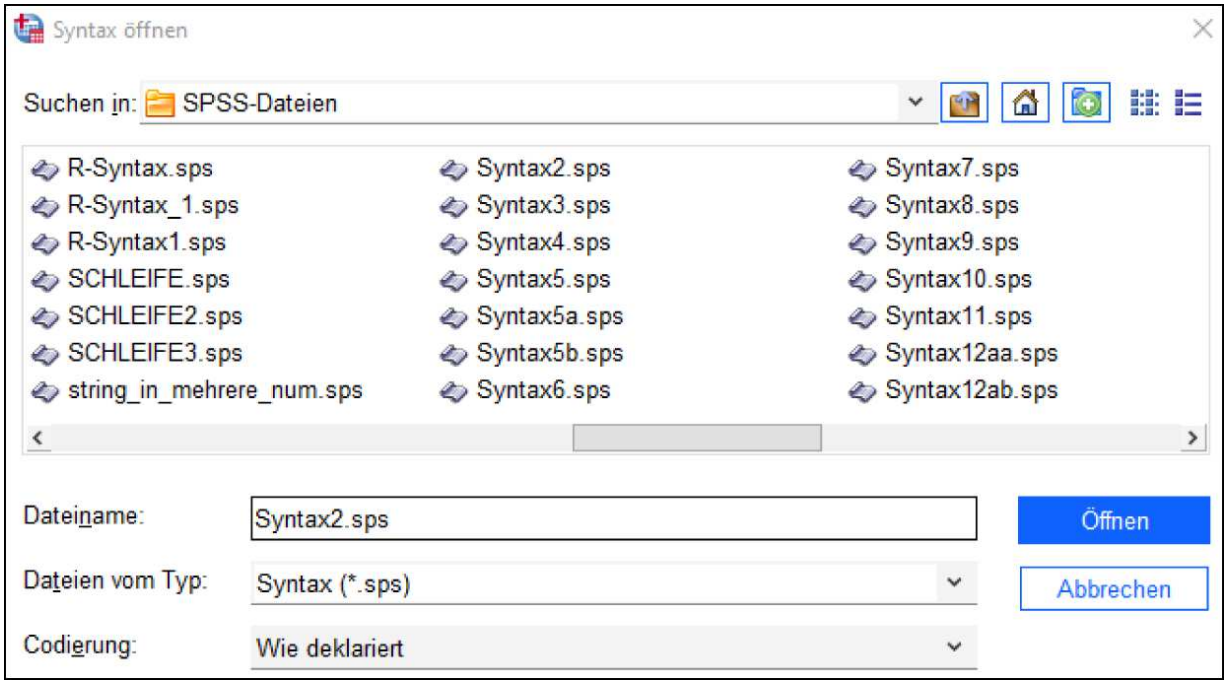



**Abbildung 5.27:** Öffnen eines bereits existierenden SPSS-Syntaxfiles

## *Zugang über ein Journalfile*

In SPSS werden alle Befehle der Menüs TRANSFORMIEREN, ANALYSIEREN und GRAFIKEN in der **SPSS-Journaldatei** *STATISTICS.JNL* protokolliert. Wo sich diese bei Ihnen auf dem Rechner befindet, erfahren Sie über die Menüabfolge BEARBEITEN|OPTIONEN|DATEISPEICHERORTE. Ist dort – wie in Abbildung 5.28 – in der Mitte der betreffenden Dialogbox das Häkchen bei SYNTAX IN JOURNAL AUFZEICHNEN gesetzt, werden die eben genannten Befehle in *STATISTICS.JNL* hineingeschrieben. Allerdings benötigen Sie zum Öffnen Ihrer Journaldatei eine geeignete Applikation,

was nicht so einfach zu bewerkstelligen ist, sodass ich persönlich Ihnen von dieser Zugangsmöglichkeit eher abrate.

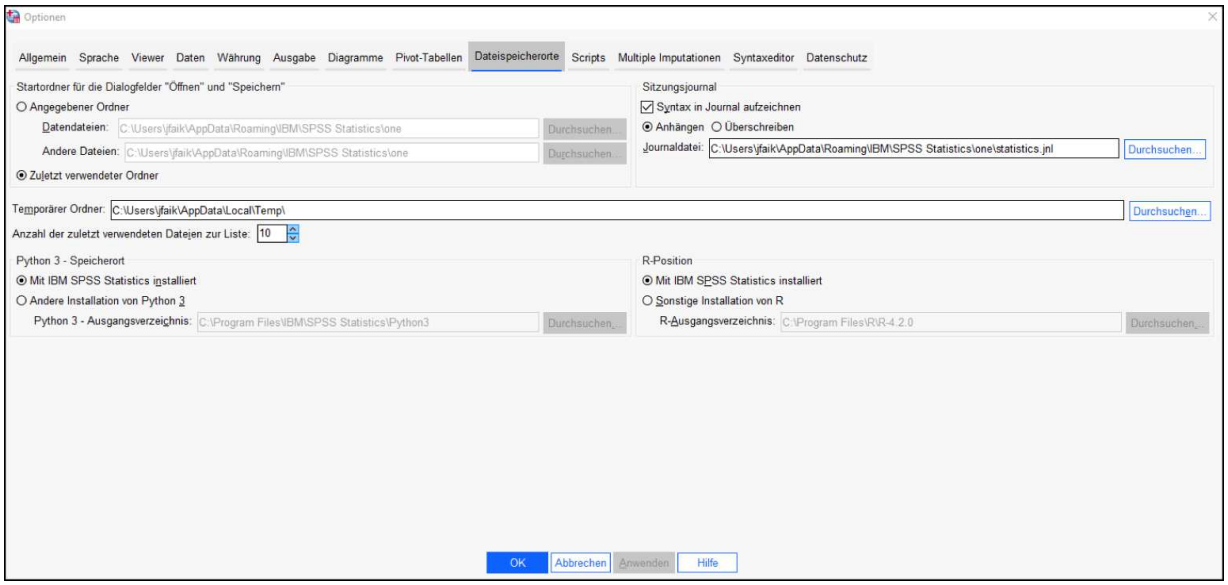



**Abbildung 5.28:** Speicherort für die SPSS-Journaldatei STATISTICS.JNL

# *»SPSS-Grammatik«*

Im Folgenden möchte ich mich mit Ihnen etwas genauer über die »SPSS-Grammatik« unterhalten. Wie eingangs zu diesem Kapitel bereits kurz ausgeführt, ist ein SPSS-Befehl grundsätzlich wie folgt aufgebaut; man spricht auch von einer **Befehlsprozedur**:

```
BEFEHLSSCHLÜSSELWORT VARIABLE(N) OPERATION.
```

Vorstehend stellt der Bestandteil `VARIABLE(N) OPERATION` die bereits eingangs zu diesem Kapitel erwähnten Spezifikationen wie Variablennamen und/oder Variablenveränderungen dar.

Hat die von Ihnen verwendete Befehlsprozedur **Unterbefehle**, werden diese – worauf ebenfalls schon hingewiesen wurde – durch einen Schrägstrich (**Slash**), also durch »/«, voneinander getrennt. Oben in diesem Kapitel wurde ja diesbezüglich schon ein (Analyse-)Beispiel für FREQUENCIES mit den (möglichen) Unterbefehlen /BARCHART, /FORMAT= NOTABLE und zum Beispiel /STATISTICS=ALL gegeben. Allgemein gilt:

```
BEFEHLSSCHLÜSSELWORT VARIABLE(N) OPERATION
= UNTERBEFEHL(E).
```

## *Syntax-Kleinigkeiten mit großen Wirkungen*

Ganz wichtig an der vorstehenden **Befehlssyntax** ist – und ich werde nicht müde, dies zu betonen! – der Punkt am Ende! Erst durch diesen Punkt wird ein Befehl als abgeschlossen angesehen. Also bitte den Punkt am Ende nicht vergessen! **Kommentare** können – wie ebenfalls in diesem Kapitel schon geschrieben wurde – durch einen **Asterisk** (*) symbolisiert werden. Hier ist wichtig, dass dieser Asterisk am Anfang (und eventuell auch noch am Ende) eines Befehls steht; ansonsten würde SPSS den Asterisk als Multiplikationszeichen fehlinterpretieren.

Was ist noch zu beachten?

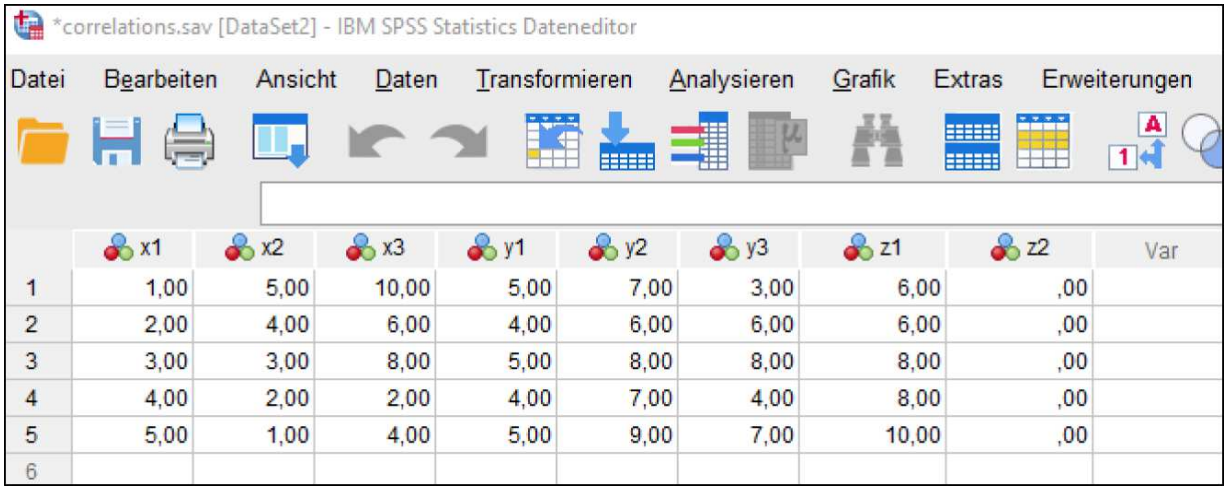
Manchmal ist in SPSS der EXECUTE-Befehl notwendig. So gilt bei **Transformationsbefehlen**, dass diese erst nach EXECUTE ausgeführt werden. Der Grund dafür ist, dass hierdurch unterschiedliche Transformationen auf einen identischen Datenstatus bezogen werden. Wird eine Transformation ohne EXECUTE ausgeführt, wird eine bestimmte SPSS-Operation das Ergebnis einer vorhergehenden Transformation zum Ausgangspunkt nehmen, sich also auf einen anderen Datenstatus beziehen als möglicherweise gewünscht.

Lassen Sie mich dies mit einem Beispiel verdeutlichen: In der Datendatei *CORRELATIONS.SAV* sollen zwei neue Variablen berechnet werden. Die Berechnung neuer Variablen fällt dabei unter den Oberbegriff »Datentransformation«. Konkret sollen die Variablen *z1* = *x1* + *y1* und *z2* = *z1* + *x2* berechnet werden. Hierzu verwenden Sie bitte in SPSS den Befehl COMPUTE.

Schreiben Sie zunächst Folgendes in den Syntaxeditor (Sie finden das Ganze auch in *SYNTAX3. SPS*):

```
COMPUTE z1 = 0.
COMPUTE z2 = 0.
COMPUTE z1 = x1 + y1.
EXECUTE.
COMPUTE z2 = z1 + x2.
```

Wenn Sie nun im Syntaxeditor auf AUSFÜHREN|ALLE klicken, erhalten Sie das Datenfenster gemäß Abbildung 5.29. Offenkundig ist die Variable *z2* noch nicht in den Ausdruck »*z1* + *x2*« verändert worden, während Variable *z1* in den gewünschten Ausdruck »*x1* + *y1*« transformiert wurde. Dieser Unterschied liegt darin, dass nach diesem letztgenannten Ausdruck für *z1* in der Syntax ein EXECUTE gesetzt wurde, während dies nach dem erstgenannten Ausdruck für *z2* unterblieben ist. Würden Sie jetzt die in Abbildung 5.29 gezeigte Datendatei abspeichern und anschließend mit dieser abgespeicherten Datendatei zum Beispiel arithmetische Mittelwerte berechnen, würden Sie in Bezug auf die Variable *z2* Ihr blaues Wunder erleben. Der Ihnen angezeigte arithmetische Mittelwert würde in diesem Fall 0,00 betragen. Eigentlich sollte er aber als Durchschnitt der Summe von *z1* und *x2* gleich 10,6 sein. Hieran sehen Sie, dass die EXECUTE-Anweisung nicht unwichtig ist.




| | x1 | x2 | x3 | y1 | y2 | y3 | z1 | z2 | Var |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1,00 | 5,00 | 10,00 | 5,00 | 7,00 | 3,00 | 6,00 | ,00 | |
| 2 | 2,00 | 4,00 | 6,00 | 4,00 | 6,00 | 6,00 | 6,00 | ,00 | |
| 3 | 3,00 | 3,00 | 8,00 | 5,00 | 8,00 | 8,00 | 8,00 | ,00 | |
| 4 | 4,00 | 2,00 | 2,00 | 4,00 | 7,00 | 4,00 | 8,00 | ,00 | |
| 5 | 5,00 | 1,00 | 4,00 | 5,00 | 9,00 | 7,00 | 10,00 | ,00 | |
| 6 | | | | | | | | | |

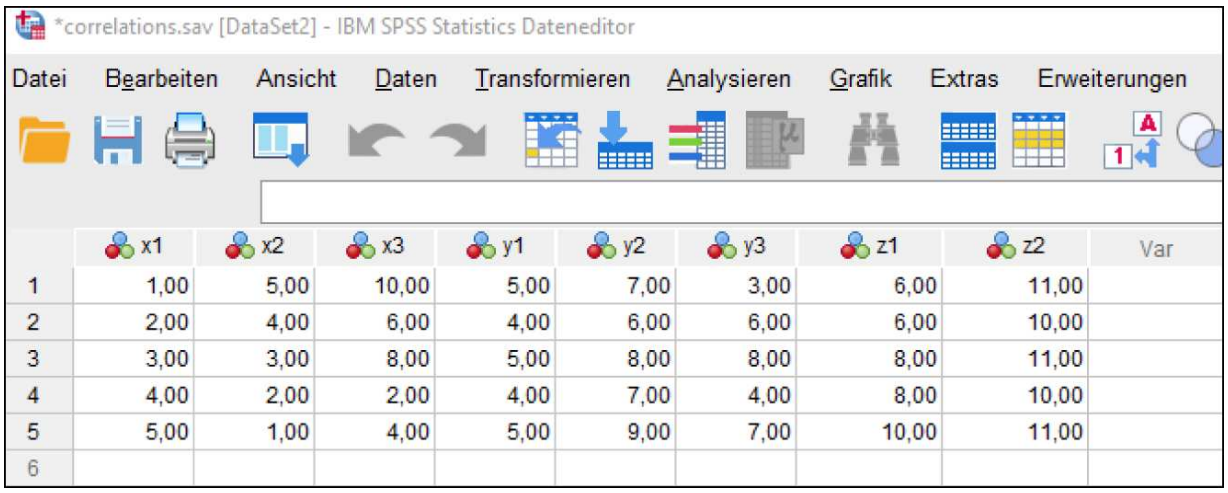
**Abbildung 5.29:** Datenansicht nach Datentransformationen mit EXECUTE-Befehl unter Bezugnahme auf *CORRELATIONS.SAV* (1)

Verschieben Sie nun bitte den EXECUTE-Befehl an das Ende Ihrer Transformationsbefehle (siehe *SYNTAX4.SPS*):

```
COMPUTE z1 = 0.
COMPUTE z2 = 0.
COMPUTE z1 = x1 + y1.
COMPUTE z2 = z1 + x2.
EXECUTE.
```

Jetzt wird auch die Transformation des letzten COMPUTE-Befehls durchgeführt, und die Variable *z2* enthält jetzt nicht mehr ausschließlich Nullen, sondern die Ergebnisse aus der Addition von *z1* und *x2* (siehe Abbildung 5.30).

Folgt auf eine Transformationsanweisung im Übrigen ein statistischer Auswertungs-Befehl – eine sogenannte **Statistikprozedur** – wie FREQUENCIES, werden die vorstehenden Transformationen gewissermaßen automatisch – das heißt: ohne EXECUTE-Befehl – ausgeführt. Die Verwendung von EXECUTE schadet aber nicht. Sie können also im Grunde genommen ohne Bedenken eine Befehlsreihenfolge stets mit EXECUTE abschließen.




| | x1 | x2 | x3 | y1 | y2 | y3 | z1 | z2 | Var |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1,00 | 5,00 | 10,00 | 5,00 | 7,00 | 3,00 | 6,00 | 11,00 | |
| 2 | 2,00 | 4,00 | 6,00 | 4,00 | 6,00 | 6,00 | 6,00 | 10,00 | |
| 3 | 3,00 | 3,00 | 8,00 | 5,00 | 8,00 | 8,00 | 8,00 | 11,00 | |
| 4 | 4,00 | 2,00 | 2,00 | 4,00 | 7,00 | 4,00 | 8,00 | 10,00 | |
| 5 | 5,00 | 1,00 | 4,00 | 5,00 | 9,00 | 7,00 | 10,00 | 11,00 | |
| 6 | | | | | | | | | |

**Abbildung 5.30:** Datenansicht nach Datentransformationen mit EXECUTE-Befehl unter Bezugnahme auf *CORRELATIONS.SAV* (2)

## *Weitere Syntax-Besonderheiten*

Im Zusammenhang mit der vorstehend behandelten SPSS-Syntax sind unter anderem folgende weiteren Besonderheiten zu beachten:

- ✔ Anweisungen können mit den drei ersten Buchstaben abgekürzt werden. Beispielsweise könnten Sie anstelle von `COMPUTE z1 = x1 + y1.` kürzer `COM z1 = x1 + y1.` schreiben. Oder: Sie könnten

`FREQUENCIES VARIABLES = x1` in `FRE VAR = x1.` abkürzen und so weiter.

- ✔ Variablennamen müssen hingegen immer ausgeschrieben werden. Bei ihnen sind führende Nullen im Ziffernteil relevant: Die Variable *var01* ist also nicht mit der Variablen *var1* identisch.
- ✔ Das Dezimaltrennzeichen in einem SPSS-Syntaxprogramm ist immer der Punkt.
- ✔ Sogenannte SPSS-Systemvariablen (zum Beispiel $CASENUM, $DATE, $TIME) beginnen mit dem Dollarzeichen »$« und dürfen nur in Datentransformationen angesprochen werden (also nicht in Statistikprozeduren).
- ✔ Bei arithmetischen Ausdrücken gilt die übliche Reihenfolge: 1. Funktionen; 2. Potenzieren (**), 3. Multiplizieren (*) beziehungsweise Dividieren (/), 4. Addieren (+) beziehungsweise Subtrahieren (-). Durch runde Klammern können Sie diese logischen Ausdrücke steuern, das heißt: die vorgenannte Reihenfolge verändern.
- ✔ In SPSS gibt es als logische Werte: »true« (»richtig«) und »false« (»falsch«), und bei den logischen Operatoren gilt folgende Reihenfolge: 1. Relationale Operatoren (=, <>, <, <=, >, >=; in Worten: gleich, ungleich, kleiner, kleiner gleich, größer, größer gleich); 2. NOT (das heißt: der logische Ausdruck gilt nicht); 3. AND (logisches Und); 4. OR (logisches Oder). Auch hier können Sie diese logischen Ausdrücke durch runde Klammern steuern, das heißt: die vorgenannte Reihenfolge verändern.

# *SPSS-»Grundbefehle«*

Grob kann man die SPSS-Befehle in zwei Gruppen einteilen, nämlich in:

- ✔ Transformations-/Definitionsbefehle und
- ✔ Analysebefehle

Bei den **Transformations-** beziehungsweise **Definitionsbefehlen** geht es im Wesentlichen um das **Management** Ihrer Daten. Dies beinhaltet das Einlesen, Verarbeiten und Abspeichern Ihrer Daten. Demgegenüber beziehen sich die **Analysebefehle** in SPSS, wie der Name andeutet, auf die statistische Auswertung Ihrer Daten.

Ich werde mich in diesem Abschnitt auf die Transformations- beziehungsweise Definitionsbefehle beschränken. Die Analysebefehle und damit die statistischen Auswertungsmöglichkeiten von SPSS werden in den folgenden Teilen dieses Buchs noch ausführlich erörtert.

Für grundlegende **Datenmanagement**-Operationen in SPSS, wie sie bislang schon beschrieben wurden, brauchen Sie einige wenige SPSS-»Grundbefehle«, auf die ich im Folgenden kurz eingehen möchte. Wie Sie auch hier wieder leicht sehen werden, sind die SPSS-Befehle an die englische Sprache angelehnt.

Die nachfolgenden Grundbefehle finden Sie zum Download in verschiedenen (Mini-)Syntaxfiles, deren Name jeweils an der entsprechenden Stelle genannt ist. Diese (Mini-)Syntaxfiles müssen Sie im Syntaxeditor über den Menüpunkt AUSFÜHREN|ALLE einfach nur noch starten. Wenn im Folgenden übrigens Verzeichnisse angegeben sind, so handelt es sich um meine eigenen Bezeichnungen (in »*C:\SPSS-DATEIEN*«). Sie müssten an den betreffenden Stellen natürlich die Verzeichnisse ansprechen, in denen sich die einzelnen Dateien auf Ihrem Rechner befinden.

## *GET-FILE-Befehl*

Mit GET FILE öffnen Sie ein bestehendes SPSS-Datenfile, wobei Sie bitte den Ort (den Ordner), wo dieses Datenfile sich befindet, in Anführungszeichen nach diesem Befehl angeben.

Bitte beachten Sie, dass Sie die nähere Bezeichnung des Speicherorts nicht einfach per Absatzschaltung trennen dürfen. Dann wird der GET-FILE-Befehl nicht ausgeführt! Dies wäre zum Beispiel hier der Fall:

```
GET FILE='C:\SPSS-DATEIEN\
PERSONEN4A.SAV'.
```

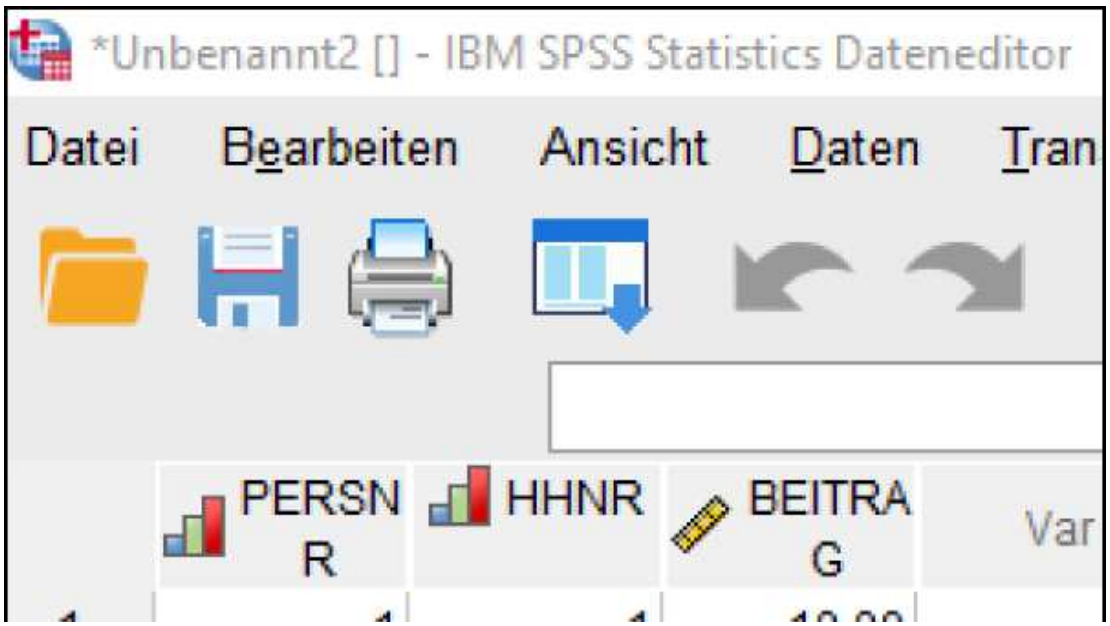
(Wenn nachfolgend Zeilentrennungen beim Speicherort im Buch angezeigt werden, so resultiert dies aus Eigenarten des verwendeten Textverarbeitungsprogramms. Im SPSS-Syntaxeditor jedenfalls dürfen Sie die entsprechende Zeilentrennung nicht durchführen.)

Es gibt zwei wichtige Unterbefehle bei GET FILE: zum einen den Unterbefehl KEEP und zum anderen den Unterbefehl DROP.

Nutzen Sie bitte in diesem Zusammenhang die Datendatei *PERSONEN4A.SAV* als Beispiel. Diese Datendatei besteht aus 15 Fällen für die vier Variablen *persnr, hhnr, beitrag* und *geschlecht*. Wenn Sie aus diesen vier Variablen eine Variablenauswahl treffen möchten, zum Beispiel nur die Variablen *persnr, hhnr* und *beitrag* in Ihrem neuen Datensatz haben möchten, geben Sie die Namen der ausgewählten Variablen in den Unterbefehl KEEP ein:

```
GET FILE='C:\SPSS-DATEIEN\PERSONEN4A.SAV'
/KEEP = persnr hhnr beitrag.
```

Wenn Sie nun diese (in *SYNTAX5A.SPS* enthaltene) Syntax ausführen (mit AUSFÜHREN|ALLE), erhalten Sie in der Tat eine neue Datendatei, die nur noch die gewünschten drei Variablen *persnr, hhnr* und *beitrag* beinhaltet (siehe Abbildung 5.31).



| | | | |
|---|---|---|---|
| 1 | 1 | 1 | 10,00 |
| 2 | 2 | 2 | 10,00 |
| 3 | 3 | 2 | 20,00 |
| 4 | 4 | 2 | 20,00 |
| 5 | 5 | 2 | 30,00 |
| 6 | 6 | 3 | 50,00 |
| 7 | 7 | 3 | 50,00 |
| 8 | 8 | 3 | 50,00 |
| 9 | 9 | 4 | 50,00 |
| 10 | 10 | 5 | 150,00 |
| 11 | 11 | 5 | 20,00 |
| 12 | 12 | 5 | 20,00 |
| 13 | 13 | 5 | 20,00 |
| 14 | 14 | 5 | 20,00 |
| 15 | 15 | 5 | 20,00 |

**Abbildung 5.31:** Neue Datendatei nach Anwendung des GET-FILE-Befehls zusammen mit dem KEEP-Unterbefehl auf Basis von *PERSONEN4A.SAV*

Alternativ zum KEEP-Unterbefehl können Sie beim GET-FILE-Befehl angeben, welche Daten Sie aus dem Ursprungsdatensatz entfernen möchten. Möchten Sie beispielsweise in Ihrem neuen Datensatz die Variable *beitrag* nicht mehr drin haben, geben Sie dies mit dem Unterbefehl DROP an (siehe auch *SYNTAX5B.SPS*):

```
GET FILE='C:\SPSS-DATEIEN\PERSONEN4A.SAV'
/DROP = beitrag.
```

Nachdem Sie auf AUSFÜHREN|ALLE geklickt haben, erhalten Sie im Datenfenster ein neues Datenfile mit den Variablen *persnr, hhnr* und

*geschlecht*; die Variable *beitrag* ist jetzt also, wie gewünscht, nicht mehr im Datensatz enthalten (siehe Abbildung 5.32).

| | PERSNR | HHNR | GESCHLECHT |
|---|---|---|---|
| 1 | 1 | 1 | männlic |
| 2 | 2 | 2 | weiblich |
| 3 | 3 | 2 | männlic |
| 4 | 4 | 2 | männlic |
| 5 | 5 | 2 | weiblich |
| 6 | 6 | 3 | männlic |
| 7 | 7 | 3 | männlic |
| 8 | 8 | 3 | weiblich |
| 9 | 9 | 4 | männlic |
| 10 | 10 | 5 | weiblich |
| 11 | 11 | 5 | weiblich |
| 12 | 12 | 5 | weiblich |
| 13 | 13 | 5 | männlic |

| | | | |
|---|---|---|---|
| 14 | 14 | 5 | weiblich |
| 15 | 15 | 5 | männlic |

**Abbildung 5.32:** Neue Datendatei nach Anwendung des GET-FILE-Befehls zusammen mit dem DROP-Unterbefehl auf Basis von *PERSONEN4A.SAV*

## *SAVE-OUTFILE-Befehl*

Mit SAVE OUTFILE wird ein SPSS-Datenfile gespeichert. Auch hier müssen Sie den Ort (den Ordner), wo das Datenfile »stehen« soll, in Anführungszeichen (nach dem Befehl) schreiben. Haben Sie zum Beispiel das Bedürfnis, das in Abbildung 5.32 zunächst noch temporär erstellte Datenfile dauerhaft zu speichern (unter dem Namen *PERSONEN5.SAV*), geht das ganz einfach mittels

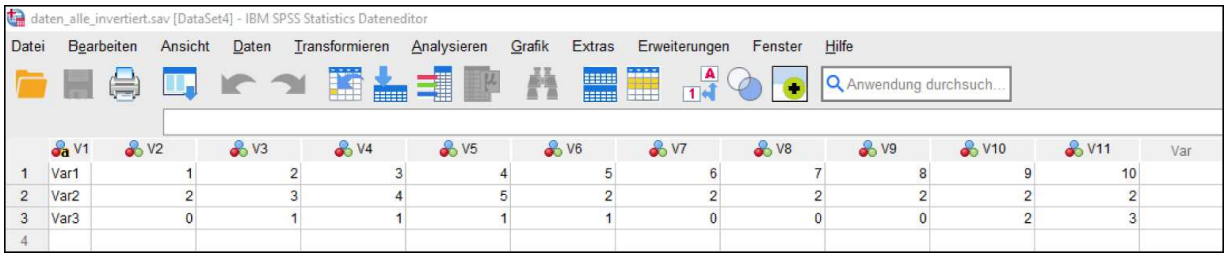
```
SAVE OUTFILE='C:\SPSS-DATEIEN\PERSONEN5.SAV'.
```

Die zu dieser Syntax gehörige Syntaxdatei heißt im Übrigen *SYNTAX6.SPS*.

## *FLIP-Befehl*

Liegt Ihre Datenmatrix nicht in der SPSS-Form mit den Fällen in den Zeilen und den Variablen in den Spalten, sondern genau umgekehrt vor, sieht Ihr Datenfile zum Beispiel wie *DATEN_ALLE_INVERTIERT.SAV* in Abbildung 5.33 aus. Diese »verkehrte« SPSS-Datendatei haben Sie zum Beispiel durch einen Datenimport von EXCEL erhalten (über das Menü DATEN IMPORTIEREN|EXCEL).

Wie aus Abbildung 5.33 hervorgeht, findet sich in der ersten Spalte die nicht weiter benötigte Variable *v1*, deren einziger Sinn und Zweck darin besteht, die drei »eigentlichen« Variablen *var1*, *var2* und *var3* anzuzeigen. Diese Variable *v1* brauchen Sie im Folgenden aber nicht mehr; sie wird daher bei der **Dateninvertierung** nicht mehr berücksichtigt. Unter dem Begriff der Dateninvertierung wird hier einfach verstanden, dass in der Datenmatrix Zeilen und Spalten vertauscht werden, sodass Sie die SPSS-Form einer Datenmatrix mit den Fällen in den Zeilen und den Variablen in den Spalten erhalten.



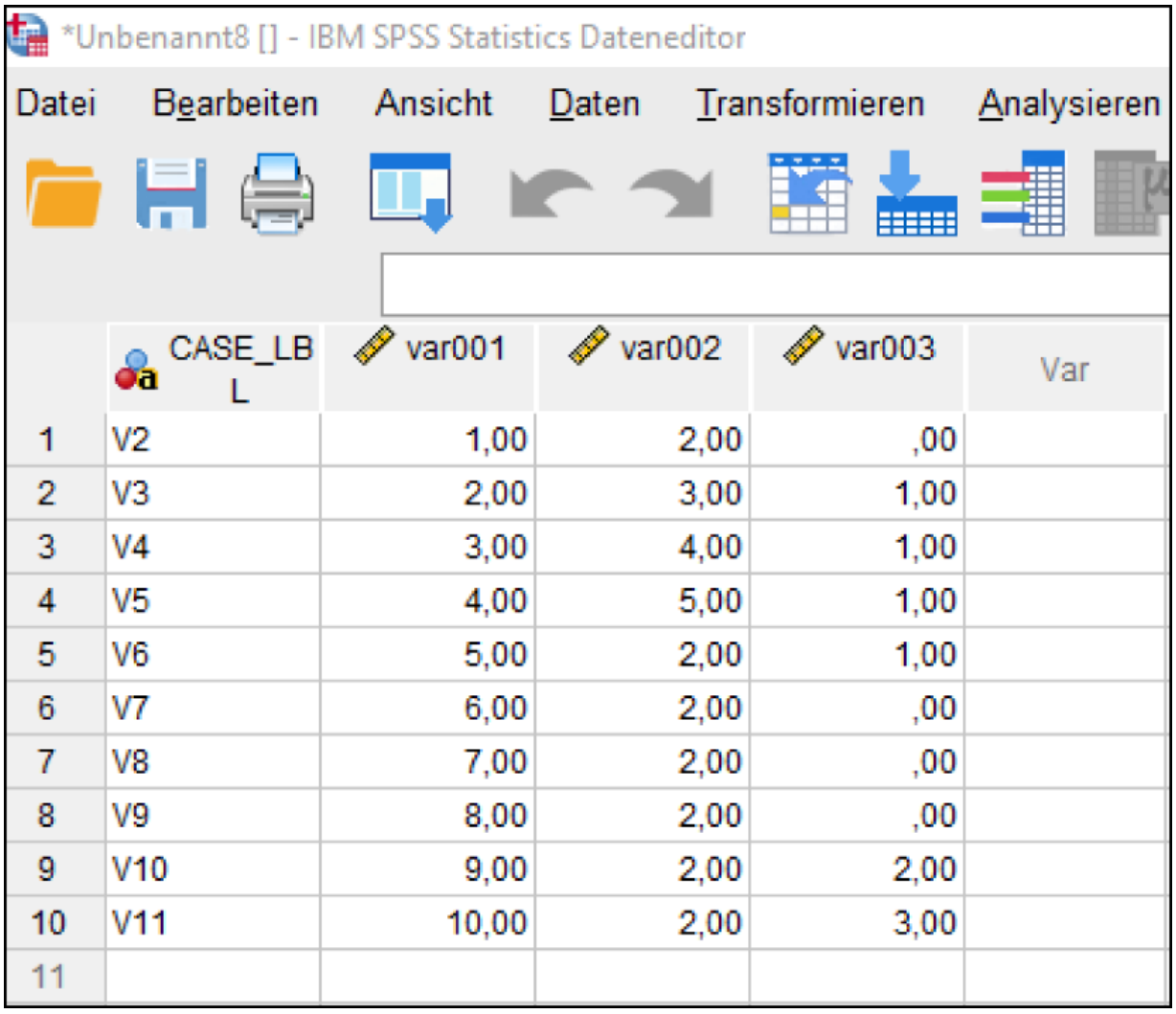

| | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 | V11 | Var |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Var1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| 2 | Var2 | 2 | 3 | 4 | 5 | 2 | 2 | 2 | 2 | 2 | 2 | |
| 3 | Var3 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 2 | 3 | |
| 4 | | | | | | | | | | | | |

**Abbildung 5.33:** Datendatei *DATEN_ALLE_INVERTIERT.SAV* mit den Variablen in den Zeilen und den Fällen in den Spalten

Aus den vorgenannten Gründen werden zunächst – nach dem Öffnen der Datei *DATEN_ALLE_INVERTIERT.SAV* – mit dem FLIP-Befehl nur die Variablen *v2* bis *v11* (»*v2* TO *v11*«) invertiert (die Syntax für dieses Beispiel findet sich am Anfang von *SYNTAX7.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\DATEN_ALLE_INVERTIERT.SAV'.
FLIP VARIABLES= v2 TO v11.
```

Im Ergebnis erhalten Sie eine invertierte Datenmatrix, die auf den ersten Blick schon einmal ganz gut ausschaut (siehe Abbildung 5.34).



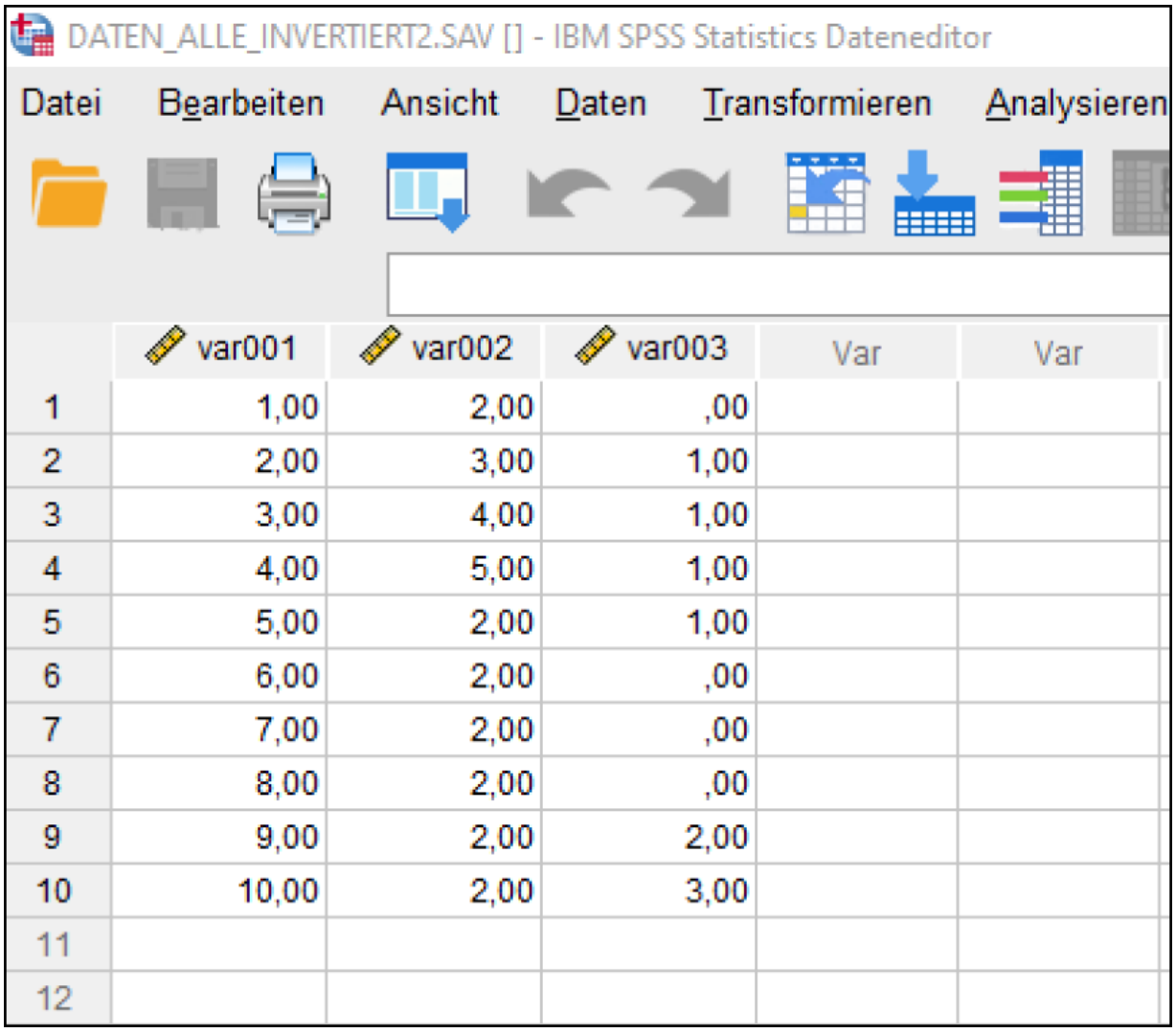

| | CASE_LBL | var001 | var002 | var003 | Var |
|---|---|---|---|---|---|
| 1 | V2 | 1,00 | 2,00 | ,00 | |
| 2 | V3 | 2,00 | 3,00 | 1,00 | |
| 3 | V4 | 3,00 | 4,00 | 1,00 | |
| 4 | V5 | 4,00 | 5,00 | 1,00 | |
| 5 | V6 | 5,00 | 2,00 | 1,00 | |
| 6 | V7 | 6,00 | 2,00 | ,00 | |
| 7 | V8 | 7,00 | 2,00 | ,00 | |
| 8 | V9 | 8,00 | 2,00 | ,00 | |
| 9 | V10 | 9,00 | 2,00 | 2,00 | |
| 10 | V11 | 10,00 | 2,00 | 3,00 | |
| 11 | | | | | |

**Abbildung 5.34:** Invertierte Datendatei *DATEN_ALLE_INVERTIERT.SAV* – FLIP-Befehl (1)

Was jetzt eigentlich nur noch stört, ist die Variable *case_lbl*, die von SPSS eingefügt worden ist und deren Ausprägungen die Variablennamen *v2* bis *v11* sind. Dies ist nicht sonderlich interessant, und daher soll *case_lbl* aus dem Datensatz gelöscht werden. Dazu speichern Sie die invertierte Datendatei aus Abbildung 5.34 zunächst einmal unter *DATEN_ALLE_INVERTIERT2.SAV* ab. Anschließend öffnen Sie diese Datei gleich wieder mit dem GET-FILE-Befehl, allerdings unter Anwendung des DROP-Unterbefehls, der sich jetzt naheliegenderweise auf die auszuschließende Variable *case_lbl* bezieht. Am Ende überschreiben Sie die Datendatei *DATEN_ALLE_INVERTIERT2.SAV* dadurch, dass Sie eine Abspeicherung unter genau diesem Namen vornehmen (siehe weiterhin *SYNTAX7.SPS*):

```
SAVE OUTFILE = 'C:\SPSS-DATEIEN\DATEN_ALLE_INVERTIERT2.SAV'.
GET FILE = 'C:\SPSS-DATEIEN\DATEN_ALLE_INVERTIERT2.SAV'
/DROP = case_lbl.
SAVE OUTFILE = 'C:\SPSS-DATEIEN\DATEN_ALLE_INVERTIERT2.SAV'.
```

Wenn Sie jetzt auch noch diese Syntax »laufen« lassen – im Übrigen ganz einfach über AUSFÜHREN|AUSWAHL, nachdem Sie zuvor den maßgeblichen Syntaxtext mit Ihrer Maus markiert haben – erhalten Sie Ihr »Wunsch-Datenfile« (siehe Abbildung 5.35).




| | var001 | var002 | var003 | Var | Var |
|---|---|---|---|---|---|
| 1 | 1,00 | 2,00 | ,00 | | |
| 2 | 2,00 | 3,00 | 1,00 | | |
| 3 | 3,00 | 4,00 | 1,00 | | |
| 4 | 4,00 | 5,00 | 1,00 | | |
| 5 | 5,00 | 2,00 | 1,00 | | |
| 6 | 6,00 | 2,00 | ,00 | | |
| 7 | 7,00 | 2,00 | ,00 | | |
| 8 | 8,00 | 2,00 | ,00 | | |
| 9 | 9,00 | 2,00 | 2,00 | | |
| 10 | 10,00 | 2,00 | 3,00 | | |
| 11 | | | | | |
| 12 | | | | | |

**Abbildung 5.35:** Invertierte Datendatei *DATEN_ALLE_INVERTIERT2.SAV* – FLIP-Befehl (2)

## *MATCH-FILES-Befehl*

Mit MATCH FILES werden zwei Datenfiles mit unterschiedlichen Variablen zusammengefügt. Obwohl dieser Begriff das Matchen von Daten bezeichnet, ist damit nicht das sogenannte statistische Matchen unterschiedlicher Untersuchungseinheiten gemeint, sondern das Zusammenfügen von Daten für ein und dieselben Untersuchungseinheiten (mit einem statistischen Fachbegriff spricht man in diesem Fall vom Record Linkage).

Gewiefte Statistiker unter Ihnen könnten nämlich angesichts des MATCH-FILES-Befehls etwas irritiert sein, weil sie mit diesem Befehl eventuell das statistische **Matchen** assoziieren. Beim statistischen Matchen geht es darum, anhand von Ähnlichkeiten bei den Fällen zweier Datensätze diese Datensätze fallweise zusammenzubringen.

Beispielsweise könnten Sie – vereinfacht ausgedrückt – in einem Datensatz für einen 58-jährigen Diplom-Volkswirt dessen Einkommenslage und in einem anderen Datensatz für einen ebenfalls 58-jährigen Diplom-Volkswirt dessen Hobbys finden. In einem nächsten Schritt würden Sie in einem gemeinsamen Datensatz diese Informationen zum Einkommen und zu den Hobbys zusammentragen.

Bei diesem Fall muss es sich aber nicht notwendigerweise um ein und dieselbe Person handeln; es könnte sich in diesem Beispiel auch um zwei völlig verschiedene 58-jährige Diplom-Volkswirte handeln.

Anders ist dies beim erwähnten MATCH-FILES-Befehl: Hier wird ein und dieselbe Person anhand einer in beiden Datensätzen identischen Identifizierungsvariablen erfasst. Dies ist das bereits erwähnte **Record Linkage**.

Beim Zusammenfügen beachten Sie bitte unbedingt, dass die Variablen gleich sortiert sind, da ansonsten beim Zusammenfügen den einzelnen Fällen falsche Variablenausprägungen zugeordnet werden!

Wenn Sie zwei Datenfiles mit gleicher Fallzahl, aber unterschiedlichen Variablen zusammenbringen, laden Sie zunächst eines der beiden Datenfiles und geben dann im Rahmen des MATCH-FILES-Befehls die andere Datendatei an. Dabei werden die zusammengeführten Daten erst einmal in eine temporäre, durch ein Sternchen gekennzeichnete Datendatei geschrieben (in der Syntax zu erkennen an »/FILE=*«), ehe der MATCH-FILES-Befehl – das ist wichtig hier! – mit EXECUTE zur Ausführung gebracht wird. Am Ende speichern Sie Ihr neues Datenfile am besten unter einem neuen Namen.

Das vorstehend genannte Prozedere wird nachfolgend am Beispiel der Datendateien *DATEN_VAR1_2.SAV* und *DATEN_VAR1_3.SAV* durchgespielt, das heißt: anhand zweier Dateien mit jeweils gleicher Fallzahl (jeweils 10 Fälle), aber mit (zum Teil) unterschiedlichen Variablen (*DATEN_VAR1_2.SAV: var00001, var00002*; *DATEN_VAR1_3.SAV: var00001, var00003*). Diese beiden Datendateien werden am Ende in *DATEN_VAR1_2_3.SAV* zusammengeführt und gespeichert (siehe auch *SYNTAX8.SPS*):

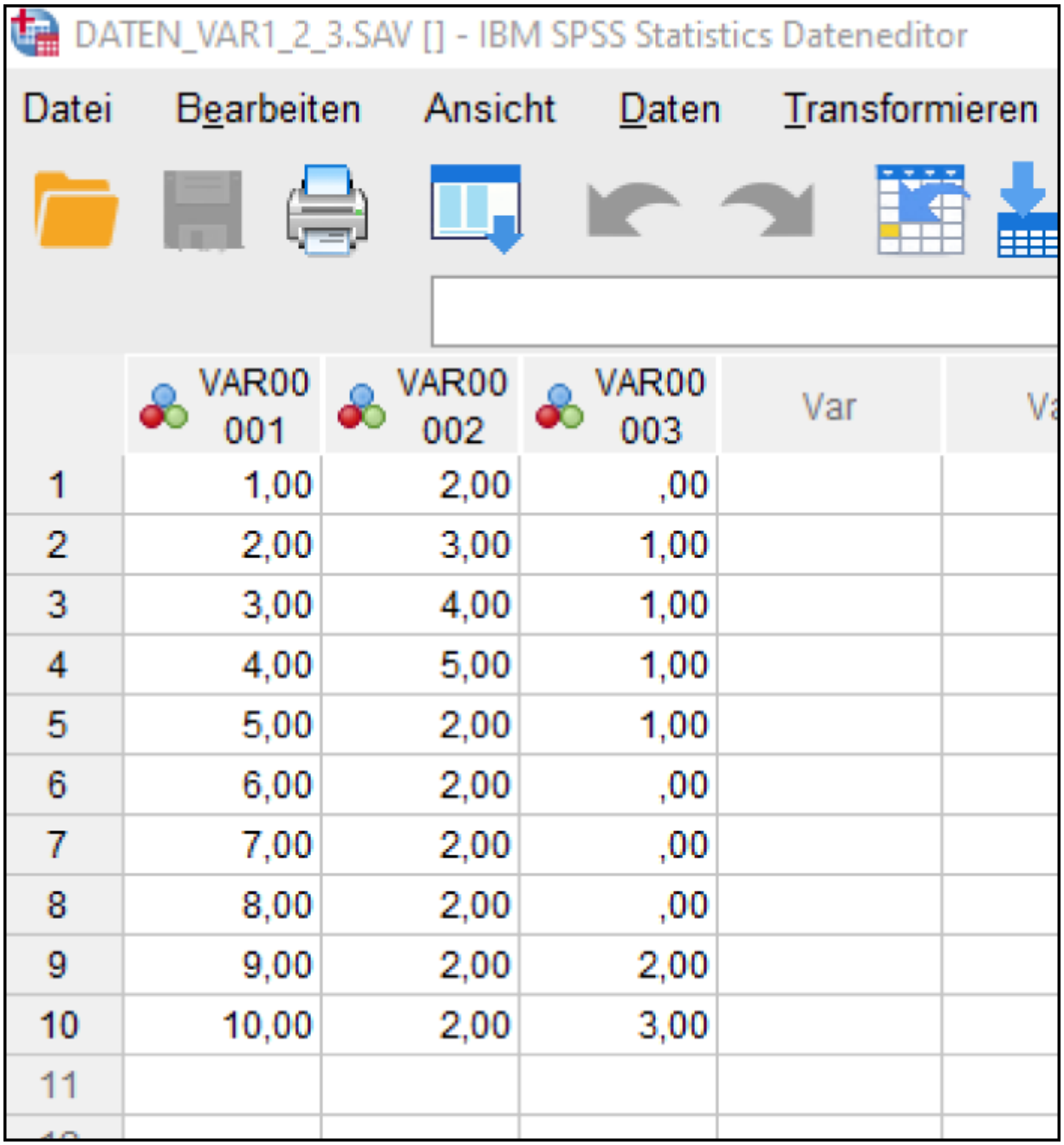
```
GET FILE = 'C:\SPSS-DATEIEN\DATEN_VAR1_2.SAV'.
MATCH FILES /FILE=*
/FILE='C:\SPSS-DATEIEN\DATEN_VAR1_3.SAV'.
EXECUTE.
SAVE OUTFILE='C:\SPSS-DATEIEN\DATEN_VAR1_2_3.SAV'.
```

Das Ergebnis dieser Zusammenführung sieht dann so aus wie in Abbildung 5.36. Hierbei stammen die Variablen *var00001* und *var00002* aus der Datei *DATEN_VAR1_2.SAV* und die Variable *var00003* aus der Datei *DATEN_VAR1_3.SAV*. Selbstverständlich enthält auch *DATEN_VAR1_3.SAV* die Variable *var00001*. Da diese Variable aber doppelt vorkommt, das heißt in beiden »gematchten« Datensätzen, wird sie bei der Zusammenführung einmal gelöscht.

## *ADD-FILES-Befehl*

Mithilfe von ADD FILES fügen Sie zwei Datenfiles zusammen, die gleiche Variablen enthalten, aber unterschiedliche Fälle. Ziehen Sie hierzu bitte die Datendateien *DATEN_1-5.SAV* und *DATEN_6-10.SAV* heran, die jeweils die Variablen *var00001, var00002* und *var00003* beinhalten, aber unterschiedliche Fälle: *DATEN_1-5.SAV* die Fälle

Nummer 1 bis 5 und *DATEN_6-10.SAV* die Fälle Nummer 6 bis 10. Mit ADD FILES werden hier die Fälle der zweitgenannten Datei an die erstgenannte Datei angehängt. Ebenso wie beim MATCH-FILES-Befehl ist es auch beim ADD-FILES-Befehl wichtig, dass ihm ein EXECUTE folgt! Am Ende Ihres kleinen Syntaxprogramms speichern Sie bitte die neue Datendatei unter dem Namen *DATEN_1-10.SAV* ab (zur Syntax siehe auch *SYNTAX9.SPS*):




| | VAR00001 | VAR00002 | VAR00003 | Var | |
|---|---|---|---|---|---|
| 1 | 1,00 | 2,00 | ,00 | | |
| 2 | 2,00 | 3,00 | 1,00 | | |
| 3 | 3,00 | 4,00 | 1,00 | | |
| 4 | 4,00 | 5,00 | 1,00 | | |
| 5 | 5,00 | 2,00 | 1,00 | | |
| 6 | 6,00 | 2,00 | ,00 | | |
| 7 | 7,00 | 2,00 | ,00 | | |
| 8 | 8,00 | 2,00 | ,00 | | |
| 9 | 9,00 | 2,00 | 2,00 | | |
| 10 | 10,00 | 2,00 | 3,00 | | |
| 11 | | | | | |

**Abbildung 5.36:** Nach Fällen gleiche, aber nach Variablen unterschiedliche zusammengeführte Datendateien in *DATEN_VAR1_2_3.SAV* – MATCH-FILES-Befehl

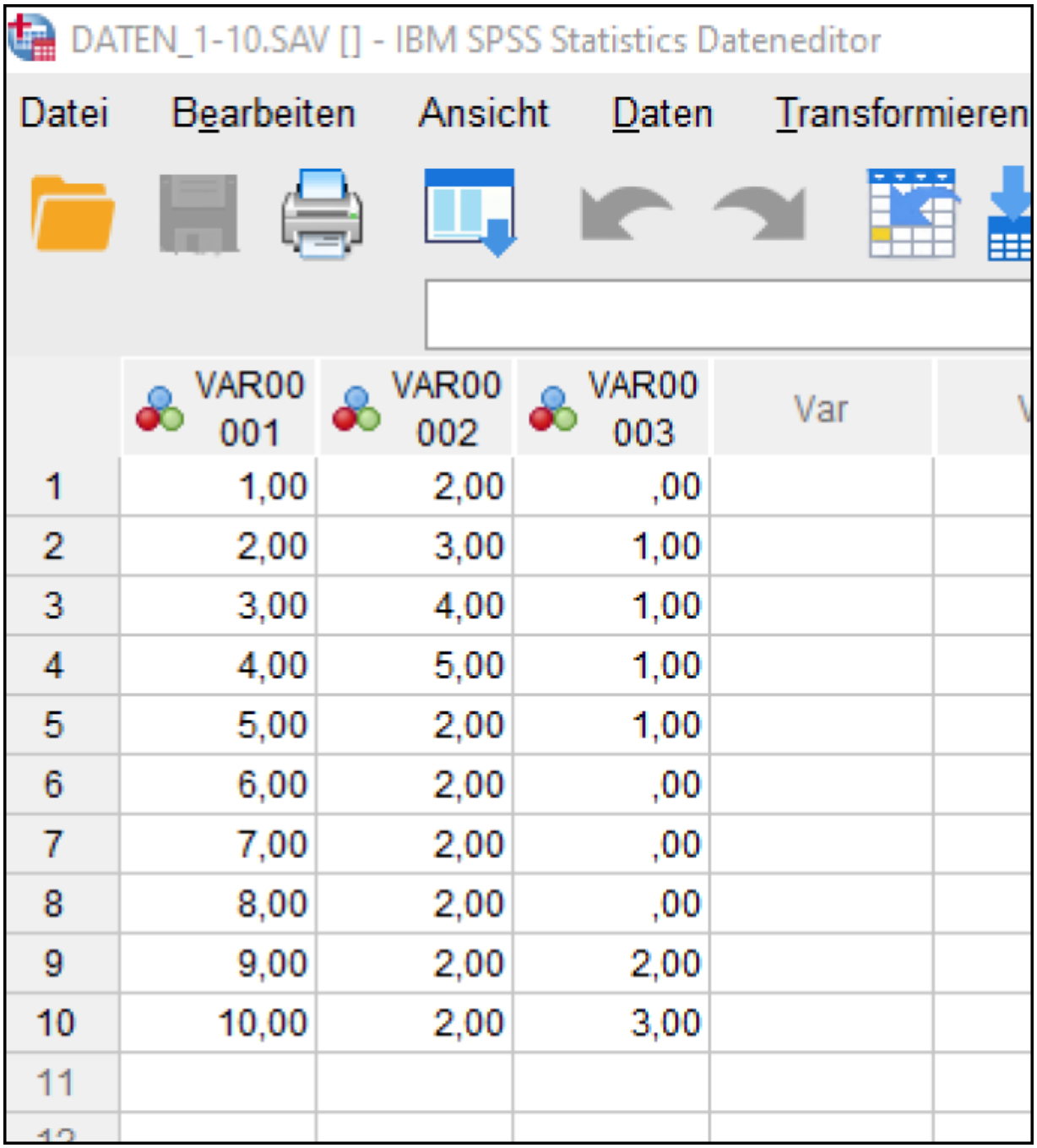
```
GET FILE = 'C:\SPSS-DATEIEN\DATEN_1-5.SAV'.
ADD FILES /FILE=*
/FILE='C:\SPSS-DATEIEN\DATEN_6-10.SAV'.
EXECUTE.
SAVE OUTFILE='C:\SPSS-DATEIEN\DATEN_1-10.SAV'.
```

Die nach den Fällen zusammengefügte Datei *DATEN_1-10.SAV* ist in Abbildung 5.37 dargestellt. Hierbei kommen die Fälle 1–5 aus der Datei *DATEN_1-5.SAV* und die Fälle 6–10 aus der Datei *DATEN_6-10.SAV*.

## *RECODE-, COMPUTE- und IF-Befehl*

Mit RECODE können Sie Variablen insofern umcodieren, als Sie die Ausprägungen einer Variablen verändern und das Ganze dann in dieser oder einer anderen Variablen abspeichern.

Variablentransformationen sind aber auch mittels der COMPUTE-Anweisung möglich. Mit ihr können Sie neue Variablen definieren und Berechnungen für diese neuen Variablen durchführen.




| | VAR00001 | VAR00002 | VAR00003 | Var |
|---|---|---|---|---|
| 1 | 1,00 | 2,00 | ,00 | |
| 2 | 2,00 | 3,00 | 1,00 | |
| 3 | 3,00 | 4,00 | 1,00 | |
| 4 | 4,00 | 5,00 | 1,00 | |
| 5 | 5,00 | 2,00 | 1,00 | |
| 6 | 6,00 | 2,00 | ,00 | |
| 7 | 7,00 | 2,00 | ,00 | |
| 8 | 8,00 | 2,00 | ,00 | |
| 9 | 9,00 | 2,00 | 2,00 | |
| 10 | 10,00 | 2,00 | 3,00 | |
| 11 | | | | |

**Abbildung 5.37:** Nach Variablen gleiche, aber nach Fällen unterschiedliche zusammengeführte Datendateien in *DATEN_1-10.SAV* – ADD-FILES-Befehl

In vielen Fällen sind entsprechende Berechnungen an eine Bedingung gekoppelt, und zwar in Form einer der Berechnung vorangestellten IF-Anweisung. Wenn Sie eine IF-Anweisung entsprechend verwenden, können, nein, müssen Sie bei der nachfolgenden Berechnung den COMPUTE-Ausdruck weglassen.

Im nachfolgenden Beispiel wird auf die Datendatei *PERSONEN4A.SAV* zurückgegriffen. Zunächst sollen die Ausprägungen der Variablen *geschlecht* umcodiert werden: »männlic« in den Wert 0 und »weiblich« in den Wert 1. Da es sich bei *geschlecht* um eine Stringvariable handelt, müssen beim entsprechenden Recodieren in einem Syntaxprogramm die ursprünglichen Ausprägungen »männlic« und »weiblich« ebenso wie die neuen Ausprägungen »0« und »1« jeweils in Hochkommata gesetzt werden.

Neben dieser Recodierung von *geschlecht* soll eine neue Variable *beitrag2* aus der alten Variablen *beitrag* folgendermaßen berechnet werden:

*beitrag2* = 2 * *beitrag*.

Darauf folgend soll in Abhängigkeit vom Geschlecht eine Variable *beitrag3* aus der Variablen *beitrag2* berechnet werden. Für die Männer soll dabei gelten, dass zu der Quadratwurzel aus *beitrag2* der Wert 5 addiert wird (wobei ich die Quadratwurzel durch »**0.5« – in Worten: »hoch 0,5« – operationalisiert habe):

*beitrag3* = 5 + (*beitrag2*)**0.5.

Für die Frauen wird hierbei angenommen, dass zu dem natürlichen Logarithmus von *beitrag2* der Wert 12.26 addiert wird (Sie merken, dass ich Gefahr laufe, mich in einer mathematischen Traumwelt zu verlieren …):

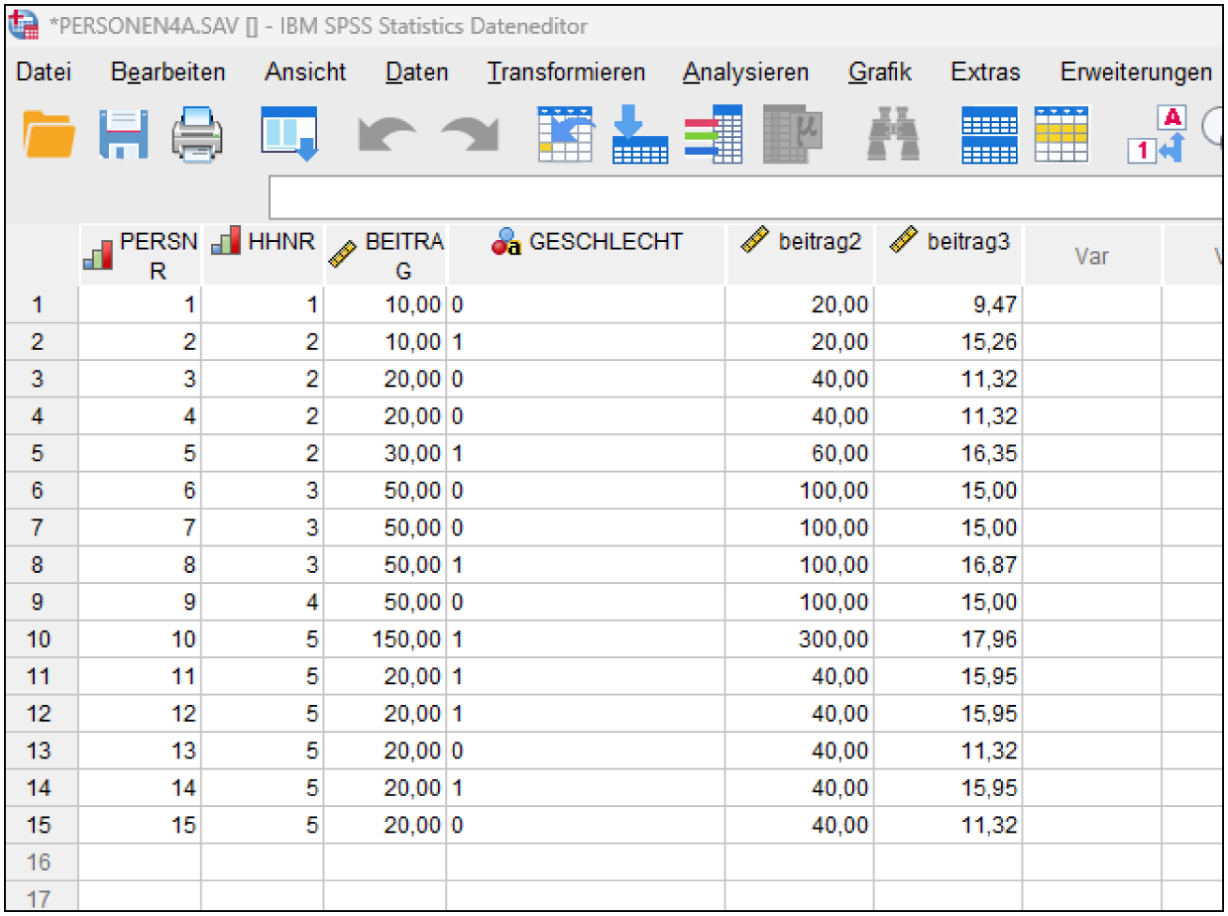
*beitrag3* = 12.26 + ln(*beitrag2*).

Da es sich hier ausschließlich um Datentransformationen handelt, müssen Sie am Ende noch den EXECUTE-Befehl setzen, damit Ihre Veränderungen auch durchgeführt werden. Die entsprechende Syntax sieht dann so aus (siehe auch *SYNTAX10.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN4.SAV'.
RECODE geschlecht ('männlic'='0') ('weiblich'='1').
COMPUTE beitrag2 = 2 * beitrag.
```

```
IF (geschlecht='0') beitrag3 = 5 + (beitrag2)**0.5.
IF (geschlecht='1') beitrag3 = 12.26 + ln(beitrag2).
EXECUTE.
```

Wenn Sie die vorstehende Syntax komplett in den Syntaxeditor eingegeben haben (oder die Datei *SYNTAX10.SPS* aufgerufen und auf AUSFÜHREN|ALLE geklickt haben), erhalten Sie das in Abbildung 5.38 wiedergegebene Datenfenster.




| | PERSNR | HHNR | BEITRAG | GESCHLECHT | beitrag2 | beitrag3 | Var |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 10,00 | 0 | 20,00 | 9,47 | |
| 2 | 2 | 2 | 10,00 | 1 | 20,00 | 15,26 | |
| 3 | 3 | 2 | 20,00 | 0 | 40,00 | 11,32 | |
| 4 | 4 | 2 | 20,00 | 0 | 40,00 | 11,32 | |
| 5 | 5 | 2 | 30,00 | 1 | 60,00 | 16,35 | |
| 6 | 6 | 3 | 50,00 | 0 | 100,00 | 15,00 | |
| 7 | 7 | 3 | 50,00 | 0 | 100,00 | 15,00 | |
| 8 | 8 | 3 | 50,00 | 1 | 100,00 | 16,87 | |
| 9 | 9 | 4 | 50,00 | 0 | 100,00 | 15,00 | |
| 10 | 10 | 5 | 150,00 | 1 | 300,00 | 17,96 | |
| 11 | 11 | 5 | 20,00 | 1 | 40,00 | 15,95 | |
| 12 | 12 | 5 | 20,00 | 1 | 40,00 | 15,95 | |
| 13 | 13 | 5 | 20,00 | 0 | 40,00 | 11,32 | |
| 14 | 14 | 5 | 20,00 | 1 | 40,00 | 15,95 | |
| 15 | 15 | 5 | 20,00 | 0 | 40,00 | 11,32 | |
| 16 | | | | | | | |
| 17 | | | | | | | |

**Abbildung 5.38:** Verschiedene Datentransformationen mit Bezugnahme auf *PERSONEN4A.SAV* – RECODE, COMPUTE, IF

## *VARIABLE-LABELS- und VALUE-LABELS-Befehl*

Mit den Befehlen VARIABLE LABELS und VALUE LABELS vergeben Sie **Variablen-** und **Werteetikette**. Auf diese Weise werden die Bezeichnungen von Variablen ebenso wie die Variablenausprägungen (zum Beispiel bei nominal- oder

ordinalskalierten Variablen) für Dritte bei statistischen Auswertungen der betreffenden Variablen anschaulicher im Sinne von: »sprechend«.

Lassen Sie uns auf die Datendatei *PERSONEN5.SAV* Bezug nehmen. Ziel soll eine Häufigkeitstabelle für die Variable *geschlecht* sein, wobei anstelle des Variablennamens mittels eines VARIABLE-LABELS-Befehls »Geschlecht eines Haushaltsmitglieds« stehen soll. Bei den Ausprägungen dieser Variablen soll über einen VALUE-LABELS-Befehl die Ausprägung »0« durch »männliches Haushaltsmitglied« und die Ausprägung »1« durch »weibliches Haushaltsmitglied« ersetzt werden. Da es sich hier auch sozusagen um Transformationen handelt, wäre jetzt eigentlich ein EXECUTE-Befehl angebracht. Da aber nach dem VALUE-LABELS-Befehl eine statistische Prozedur (in diesem Fall: FREQUENCIES) steht, brauchen Sie hier keine EXECUTE-Anweisung zu geben. Zum Ende hin wird also die erwähnte Häufigkeitsauszählung für *geschlecht* vorgenommen (siehe auch *SYNTAX11.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN5.SAV'.
VARIABLE LABELS geschlecht 'Geschlecht eines Haushaltsmitglieds'.
VALUE LABELS geschlecht '0' 'männliches Haushaltsmitglied' '1' 'weibliches
Haushaltsmitglied'.
FREQUENCIES VARIABLES = geschlecht.
```

Wie Sie anhand von Abbildung 5.39 sehen können, werden bei der Ausgabe der Häufigkeitstabelle anstelle des Variablennamens das Variablenetikett und anstelle der schnöden Merkmalsausprägungen »0« und »1« zwei »sprechende« Werteausprägungen verwendet. Diese Veränderungen erleichtern dem Betrachter dieser Häufigkeitstabelle das Verständnis derselben.

**Geschlecht eines Haushaltsmitglieds**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | männliches Haushaltsmitglied | 8 | 53,3 | 53,3 | 53,3 |
| | weibliches Haushaltsmitglied | 7 | 46,7 | 46,7 | 100,0 |
| | Gesamt | 15 | 100,0 | 100,0 | |

**Abbildung 5.39:** Zuweisung von Variablen- und Werteetiketten – VARIABLE LABELS, VALUE LABELS (auf Basis von *PERSONEN5.SAV*)

## *SELECT-IF-, FILTER-BY- und SPLIT-FILE-Befehl*

Wenn Sie Datensätze nach einzelnen Variablen beziehungsweise deren Ausprägungen **filtern** möchten, steht unter anderem der SELECT-IF-Befehl zur Verfügung.

Sie sollten im Zusammenhang mit diesem Befehl stets die TEMPORARY-Anweisung vor den SELECT-IF-Befehl setzen, damit die Fallauswahl nur temporär wirksam wird und Fälle nicht endgültig gelöscht werden. Setzen Sie nämlich keine TEMPORARY-Anweisung, werden die nicht ausgewählten Fälle sozusagen »physisch«, also auf »Nimmer Wiedersehen«, aus dem Datensatz gelöscht! Seien Sie also in diesen Fällen beim Abspeichern Ihres Datensatzes (sehr) vorsichtig!

Für *PERSONEN5.SAV* wird im Folgenden mittels SELECT IF nach dem Geschlecht gefiltert, und zwar werden für die nachfolgend angestrebte Häufigkeitstabelle der Variablen *beitrag* (mittels FREQUENCY) nur die weiblichen Personen (*geschlecht* = ‘1‘) berücksichtigt (zur Syntax siehe auch *SYNTAX12AA.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN5.SAV'.
TEMPORARY.
SELECT IF (geschlecht='1').
FREQUENCIES VARIABLES = beitrag.
```

Mit der vorstehenden Syntax bleibt wegen TEMPORARY die Datendatei von der Fallzahl her unverändert. Die ausgewiesene Häufigkeitstabelle bezieht sich – wie in Abbildung 5.40 ausgewiesen – hingegen nur auf die insgesamt sieben weiblichen Personen in *PERSONEN5.SAV*.

| Geleisteter Obolus | | | | | |
|---|---|---|---|---|---|
| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
| Gültig | 10,00 | 1 | 14,3 | 14,3 | 14,3 |
| | 20,00 | 3 | 42,9 | 42,9 | 57,1 |
| | 30,00 | 1 | 14,3 | 14,3 | 71,4 |
| | 50,00 | 1 | 14,3 | 14,3 | 85,7 |
| | 150,00 | 1 | 14,3 | 14,3 | 100,0 |
| | Gesamt | 7 | 100,0 | 100,0 | |

**Abbildung 5.40:** Häufigkeitstabelle bei Auswahl einzelner Fälle in *PERSONEN5.SAV* – SELECT IF mit TEMPORARY

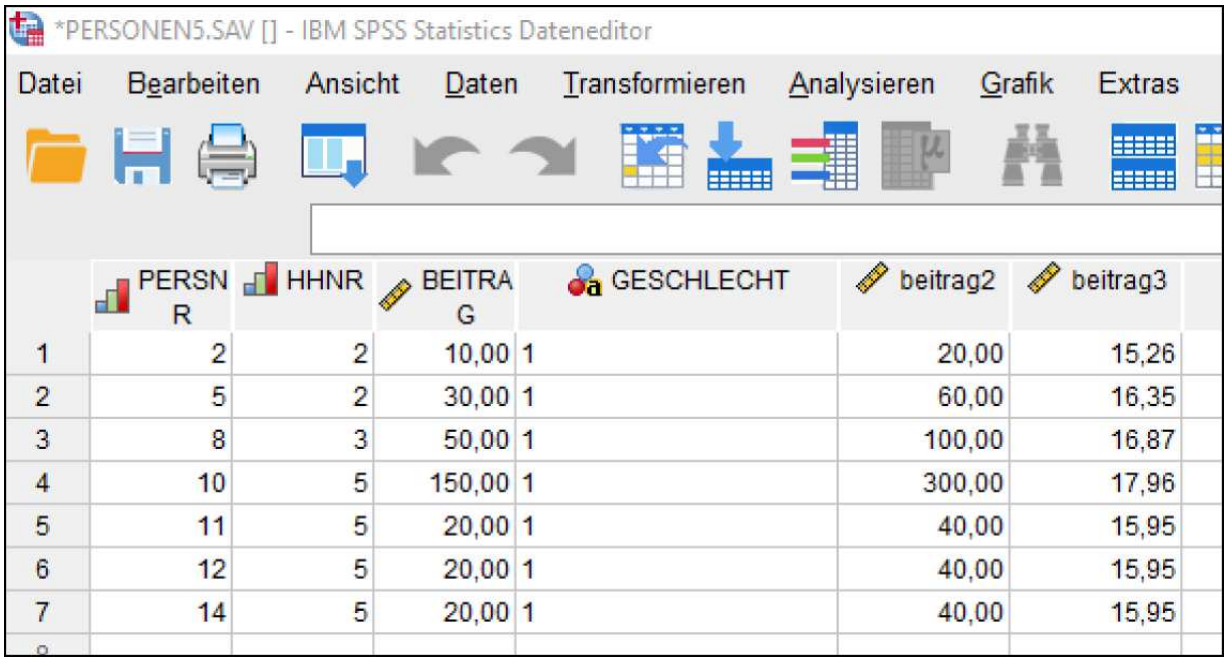
Das gleiche Ergebnis wie in Abbildung 5.40 erhalten Sie auch, wenn Sie den SELECT-IF-Befehl ohne TEMPORARY-Anweisung ausführen (siehe hierzu *SYNTAX12AB.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN5.SAV'.
SELECT IF (geschlecht='1').
FREQUENCIES VARIABLES = beitrag.
```

Allerdings verändert sich der zugrunde liegende Datensatz. Es werden nämlich alle nicht selektierten Fälle (in diesem Beispiel die männlichen Personen) aus dem Datensatz gelöscht (siehe Abbildung 5.41), und wenn Sie jetzt Ihren Datensatz abspeichern, haben Sie nur noch einen verkürzten Datensatz – also aufpassen! Deshalb empfehle ich im Zusammenhang mit dem SELECT-IF-Befehl grundsätzlich die zusätzliche Verwendung von TEMPORARY.

Die vorstehende »Lösch-Problematik« können Sie umgehen, wenn Sie einen Datensatz alternativ über eine **Filtervariable** filtern. In SPSS gibt es hierfür die Variable *filter_$*, die den Wert »1« annimmt, wenn die für die Auswahl maßgebliche Bedingung erfüllt ist, und ansonsten gleich 0 ist. Entsprechend wird *filter_$* über eine COMPUTE-Anweisung zunächst definiert. Anschließend wird über den FILTER-BY-Befehl der Filter gesetzt. Da es sich hier auch um eine Datenveränderung handelt, müssen Sie am Ende wieder eine EXECUTE-Anweisung setzen. Vergessen Sie diese EXECUTE-Anweisung, enthält die resultierende

Datendatei bei *filter_$* lauter fehlende Werte, und alle Fälle werden »ausgefiltert«, das heißt: von den weiteren Analysen ausgeschlossen. Das finden Sie doch sicherlich auch nicht sehr schön – oder?



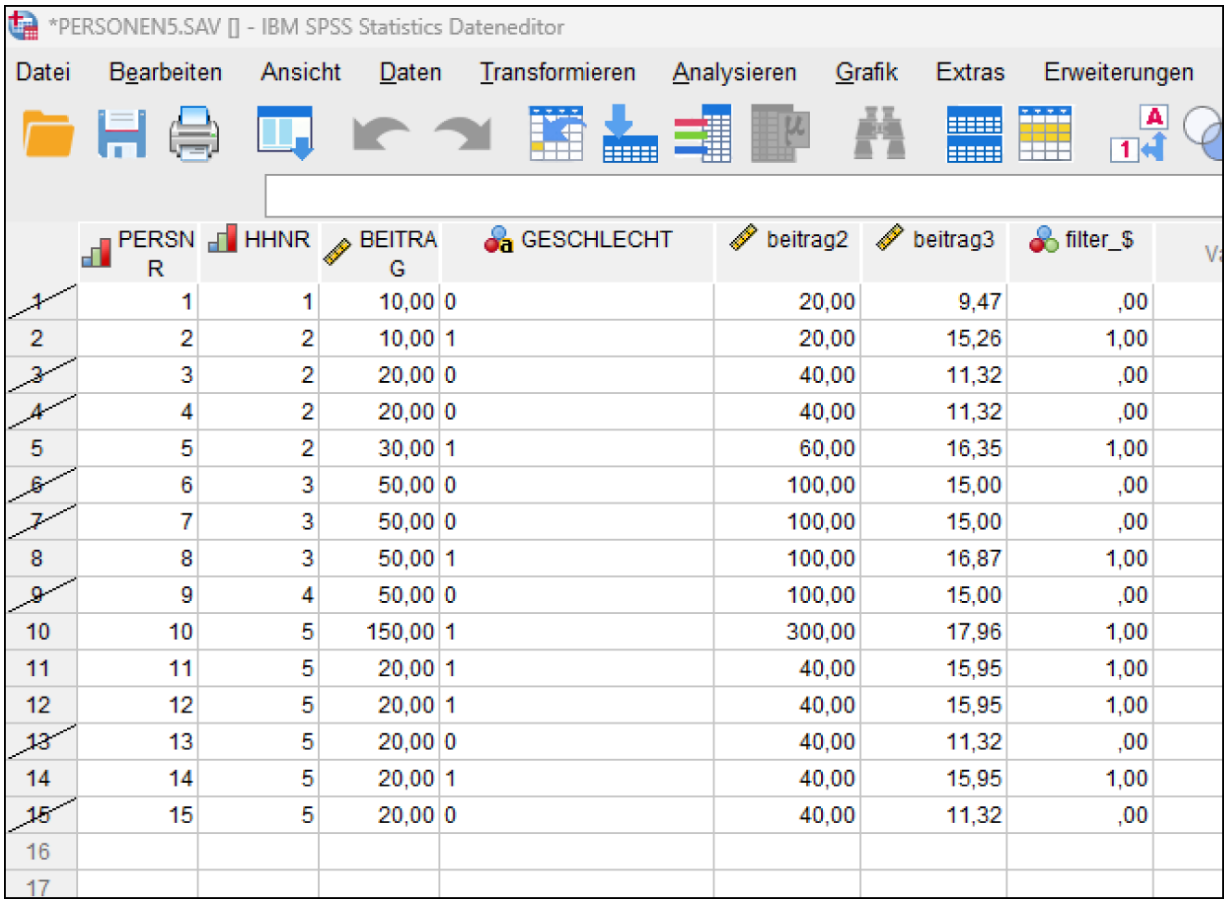

| | PERSNR | HHNR | BEITRAG | GESCHLECHT | beitrag2 | beitrag3 |
|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 10,00 | 1 | 20,00 | 15,26 |
| 2 | 5 | 2 | 30,00 | 1 | 60,00 | 16,35 |
| 3 | 8 | 3 | 50,00 | 1 | 100,00 | 16,87 |
| 4 | 10 | 5 | 150,00 | 1 | 300,00 | 17,96 |
| 5 | 11 | 5 | 20,00 | 1 | 40,00 | 15,95 |
| 6 | 12 | 5 | 20,00 | 1 | 40,00 | 15,95 |
| 7 | 14 | 5 | 20,00 | 1 | 40,00 | 15,95 |

**Abbildung 5.41:** Auswahl von Fällen auf Basis von *PERSONEN5.SAV* – SELECT IF ohne TEMPORARY

Im nachfolgenden Beispiel wird – erneut für die Datendatei *PERSONEN5.SAV* – insofern ein Filter nach dem Geschlecht gesetzt, als nur die Frauen (*geschlecht* = '1') mit ihren Daten ausgewählt werden (zur Syntax siehe auch *SYNTAX12B.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN5.SAV'.
COMPUTE filter_$=(geschlecht = '1').
FILTER BY filter_$.
EXECUTE.
```

Das Ergebnis der vorstehenden Filtersetzung ist aus Abbildung 5.42 zu erkennen. Die nicht ausgewählten Fälle sind hier daran zu erkennen, dass die jeweilige Fallnummer am linken Bildrand durchgestrichen ist. Das ist ein weiterer Unterschied zum SELECT-IF-Befehl, bei dem eine solche Markierung im Datensatz nicht vorgenommen wird.




| | PERSNR | HHNR | BEITRAG | GESCHLECHT | beitrag2 | beitrag3 | filter_$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 10,00 | 0 | 20,00 | 9,47 | ,00 |
| 2 | 2 | 2 | 10,00 | 1 | 20,00 | 15,26 | 1,00 |
| 3 | 3 | 2 | 20,00 | 0 | 40,00 | 11,32 | ,00 |
| 4 | 4 | 2 | 20,00 | 0 | 40,00 | 11,32 | ,00 |
| 5 | 5 | 2 | 30,00 | 1 | 60,00 | 16,35 | 1,00 |
| 6 | 6 | 3 | 50,00 | 0 | 100,00 | 15,00 | ,00 |
| 7 | 7 | 3 | 50,00 | 0 | 100,00 | 15,00 | ,00 |
| 8 | 8 | 3 | 50,00 | 1 | 100,00 | 16,87 | 1,00 |
| 9 | 9 | 4 | 50,00 | 0 | 100,00 | 15,00 | ,00 |
| 10 | 10 | 5 | 150,00 | 1 | 300,00 | 17,96 | 1,00 |
| 11 | 11 | 5 | 20,00 | 1 | 40,00 | 15,95 | 1,00 |
| 12 | 12 | 5 | 20,00 | 1 | 40,00 | 15,95 | 1,00 |
| 13 | 13 | 5 | 20,00 | 0 | 40,00 | 11,32 | ,00 |
| 14 | 14 | 5 | 20,00 | 1 | 40,00 | 15,95 | 1,00 |
| 15 | 15 | 5 | 20,00 | 0 | 40,00 | 11,32 | ,00 |
| 16 | | | | | | | |
| 17 | | | | | | | |

**Abbildung 5.42:** Filtern eines Datensatzes (*PERSONEN5.SAV*) – FILTER BY

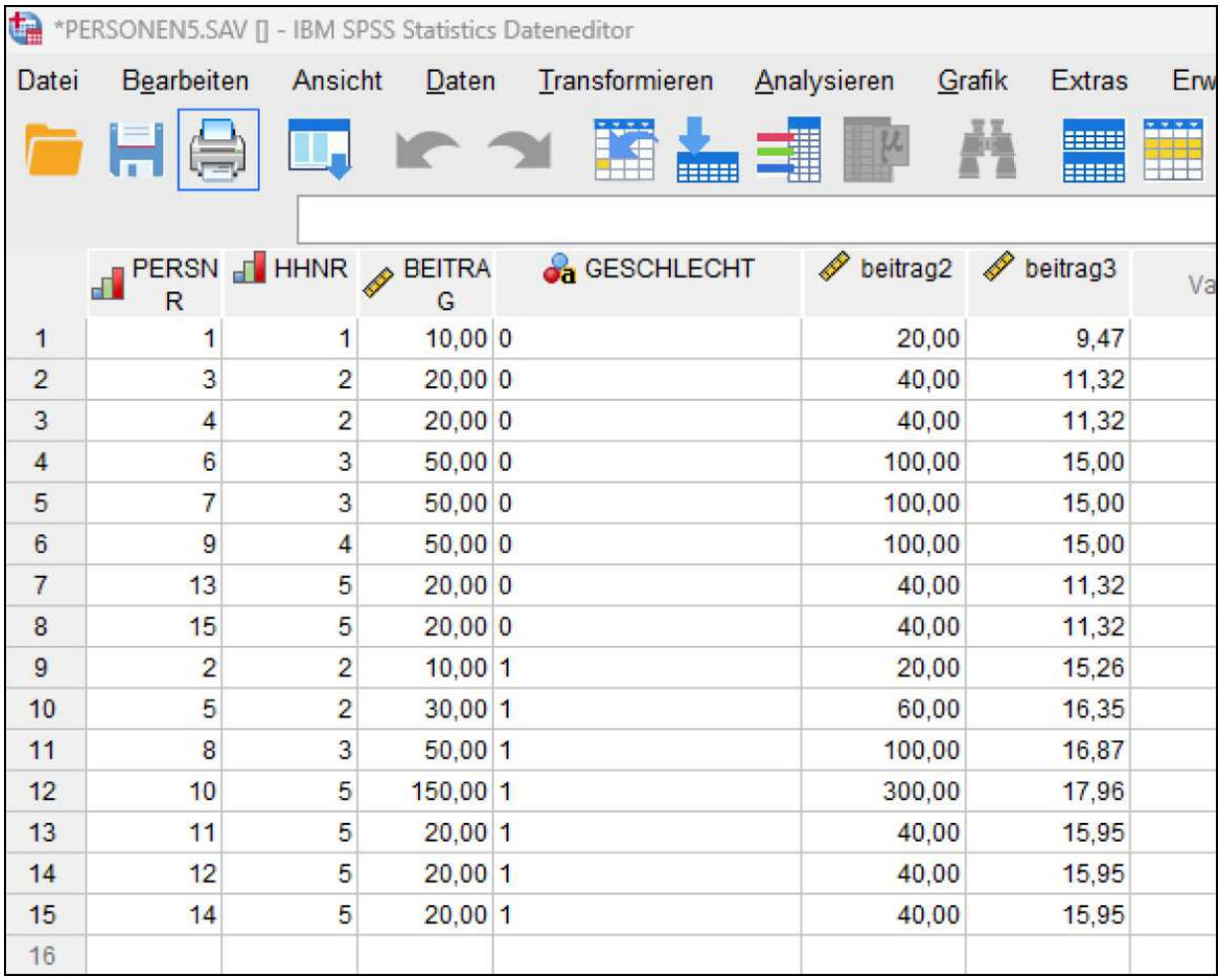
Mit dem einfachen Befehl »FILTER OFF« können Sie das vorstehende Filtern wieder aufheben, und die Fallnummern sind nicht mehr durchgestrichen. Alternativ können Sie auch SPSS den Befehl »USE ALL« erteilen. Auch hierdurch werden alle Fälle wieder aktiviert. Mit anderen Worten: Auch hierdurch wird das Filtern aufgehoben.

Eine Alternative zum Filtern Ihrer Daten (mittels SELECT IF oder FILTER BY) ist der SPLIT-FILE-Befehl. Für sortierte Daten teilt er Ihren Datensatz in Subgruppen auf. Sie müssen dabei zunächst Ihre Daten mit dem SORT-CASES-Befehl nach der Variablen, nach der Sie filtern möchten, sortieren. Möchten Sie zum Beispiel nach Männern und Frauen sortieren, sollten in Ihrem Datensatz zunächst die Fälle für die Gruppe der Männer und dann die Fälle für die Gruppe der Frauen (oder umgekehrt) stehen.

Möchten Sie also für die Beispieldatei *PERSONEN5.SAV* zwischen Männern und Frauen differenzieren, sollten Sie zunächst nach dem Geschlecht sortieren. Anschließend können Sie via SPLIT FILE den Datensatz zwischen Männern und Frauen trennen, ehe Sie schließlich für die von Ihnen unterschiedenen Subgruppen zum Beispiel eine statistische Prozedur wie FREQUENCIES »laufen« lassen. Nachfolgend bezieht sich diese Statistikprozedur auf die Variable *beitrag*. Mit folgendem Syntaxprogramm werden Ihnen also die Häufigkeiten getrennt für Männer und Frauen ausgewiesen (zur Syntax siehe auch *SYNTAX12C.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN5.SAV'.
SORT CASES BY geschlecht.
SPLIT FILE BY geschlecht.
FREQUENCIES VARIABLES = beitrag.
```

Abbildung 5.43 zeigt zunächst, wie die Daten im Dateneditor nach dem Geschlecht sortiert worden sind: In den ersten acht Zeilen der Datenmatrix finden Sie die Fälle für die Gruppe der Männer, danach jene für die Gruppe der Frauen.




| | PERSNR | HHNR | BEITRAG | GESCHLECHT | beitrag2 | beitrag3 | Va |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 10,00 | 0 | 20,00 | 9,47 | |
| 2 | 3 | 2 | 20,00 | 0 | 40,00 | 11,32 | |
| 3 | 4 | 2 | 20,00 | 0 | 40,00 | 11,32 | |
| 4 | 6 | 3 | 50,00 | 0 | 100,00 | 15,00 | |
| 5 | 7 | 3 | 50,00 | 0 | 100,00 | 15,00 | |
| 6 | 9 | 4 | 50,00 | 0 | 100,00 | 15,00 | |
| 7 | 13 | 5 | 20,00 | 0 | 40,00 | 11,32 | |
| 8 | 15 | 5 | 20,00 | 0 | 40,00 | 11,32 | |
| 9 | 2 | 2 | 10,00 | 1 | 20,00 | 15,26 | |
| 10 | 5 | 2 | 30,00 | 1 | 60,00 | 16,35 | |
| 11 | 8 | 3 | 50,00 | 1 | 100,00 | 16,87 | |
| 12 | 10 | 5 | 150,00 | 1 | 300,00 | 17,96 | |
| 13 | 11 | 5 | 20,00 | 1 | 40,00 | 15,95 | |
| 14 | 12 | 5 | 20,00 | 1 | 40,00 | 15,95 | |
| 15 | 14 | 5 | 20,00 | 1 | 40,00 | 15,95 | |
| 16 | | | | | | | |

**Abbildung 5.43:** Sortierung von Daten mithilfe von SORT CASES (*PERSONEN5.SAV*)

Des Weiteren geht aus Abbildung 5.44 der Output in Form getrennter Häufigkeitstabellen für Männer (»0«) und Frauen (»1«) hervor.

**Geleisteter Obolus**

| Geschlecht der Person | | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|---|
| 0 | Gültig | 10,00 | 1 | 12,5 | 12,5 | 12,5 |
| | | 20,00 | 4 | 50,0 | 50,0 | 62,5 |
| | | 50,00 | 3 | 37,5 | 37,5 | 100,0 |
| | | Gesamt | 8 | 100,0 | 100,0 | |
| 1 | Gültig | 10,00 | 1 | 14,3 | 14,3 | 14,3 |
| | | 20,00 | 3 | 42,9 | 42,9 | 57,1 |
| | | 30,00 | 1 | 14,3 | 14,3 | 71,4 |
| | | 50,00 | 1 | 14,3 | 14,3 | 85,7 |
| | | 150,00 | 1 | 14,3 | 14,3 | 100,0 |
| | | Gesamt | 7 | 100,0 | 100,0 | |

**Abbildung 5.44:** Häufigkeitstabellen für Männer und Frauen getrennt (auf Basis von *PERSONEN5.SAV*) – SPLIT FILE

## *AGGREGATE-Befehl*

Möchten Sie von einer unteren Datenebene aus Daten hin zu einer höheren Datenebene zusammenfassen, das heißt mit einem Fachwort: **aggregieren**, sollten Sie den AGGREGATE-Befehl benutzen. Die Elemente der unteren Datenebene sind hierbei Untermengen der oberen Datenebene. Weniger technisch ausgedrückt geht es hier beispielsweise um das Aggregieren von Personendaten auf der Haushaltsebene, zum Beispiel um die Addition der individuellen Einkommen der einzelnen Mitglieder eines privaten Haushalts zum Gesamteinkommen des betreffenden privaten Haushalts.

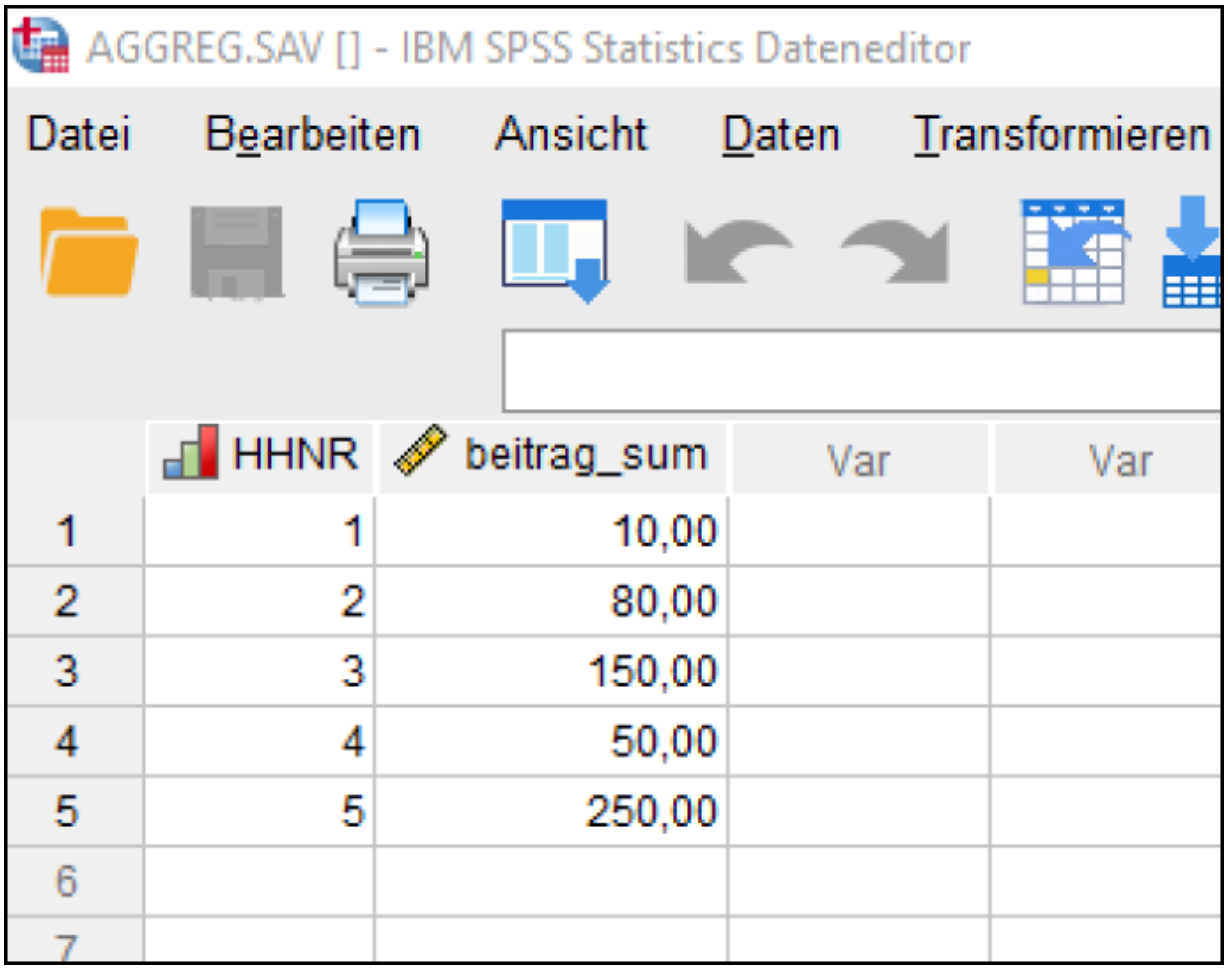
Um den AGGREGATE-Befehl im vorstehenden Sinn zu veranschaulichen, greifen Sie bitte an dieser Stelle auf *PERSONEN4A.SAV* als Beispieldatensatz zurück – das heißt: auf einen Datensatz mit 15 Personen, die sich auf 5 Haushalte verteilen. Ihre Aufgabe ist es jetzt, die personenbezogenen Beitragszahlungen gemäß der Variablen *beitrag* für jeden der Haushalte zu einem haushaltsbezogenen Beitrag in der Variablen *beitrag_sum*

zusammenzufassen. Hierzu dient folgendes Syntaxprogramm (siehe auch *SYNTAX13.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN4A.SAV'.
AGGREGATE
   /OUTFILE='C:\SPSS-DATEIEN\AGGREG.SAV'
   /BREAK=hhnr
   /beitrag_sum=SUM(beitrag).
```

Was genau wurde jetzt mit dem vorstehenden Syntaxprogramm gemacht? Nun, es wurde zunächst angewiesen, dass der neue (dann haushaltsbezogene) Datensatz nach der Aggregation sich in *AGGREG.SAV* wiederfindet. In diese Datei werden also nur die aggregierten Daten von *beitrag_sum* hineingeschrieben.

Danach wurde als **Breakvariable** *hhnr* definiert; das heißt, dass die Haushaltsnummer hier die Variable ist, über die aggregiert wird, was ja auch insofern plausibel ist, als ja hin zur Haushaltsebene aggregiert werden soll. Schließlich wurde hier aus der Vielzahl möglicher Funktionen (wie zum Beispiel auch arithmetischer Mittelwert, Median, Standardabweichung und so weiter) für *beitrag_sum* die Summen-Funktion ausgewählt. Das heißt: Die personenbezogenen Beitragszahlungen, wie sie sich in *beitrag* finden, werden haushaltsbezogen summiert (und als Haushaltswerte in *beitrag_sum* abgelegt). Abbildung 5.45 zeigt den Inhalt von *AGGREG.SAV*.




| | HHNR | beitrag_sum | Var | Var |
|---|---|---|---|---|
| 1 | 1 | 10,00 | | |
| 2 | 2 | 80,00 | | |
| 3 | 3 | 150,00 | | |
| 4 | 4 | 50,00 | | |
| 5 | 5 | 250,00 | | |
| 6 | | | | |
| 7 | | | | |

**Abbildung 5.45:** Haushaltsbezogene Beitragssummen nach der Aggregation von der Personenebene (*AGGREG.SAV*)

Im vorstehenden Beispiel wurde mit *hhnr* eine Breakvariable angegeben. Geben Sie an der entsprechenden Stelle in Ihrem Syntaxprogramm keine Breakvariable an, dann aggregiert SPSS auf der allerhöchsten Datenebene, das heißt: über den kompletten Datensatz. Im Beispiel von *PERSONEN4A.SAV* bedeutet dies, dass die Summenbildung der personellen Beitragszahlungen zum Gesamtbeitrag im gesamten Datensatz führt (siehe auch *SYNTAX14.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN4A.SAV'.
AGGREGATE
   /OUTFILE='C:\SPSS-DATEIEN\AGGREG2.SAV'
   /BREAK=
   /beitrag_sum=SUM(beitrag).
```

Entsprechend besteht die neue (aggregierte) Datei *AGGREG2.SAV* aus lediglich einer Datenzeile, aus der hervorgeht, dass der Gesamtbeitrag über alle Untersuchungseinheiten hinweg 540,00 Euro beträgt (siehe Abbildung 5.46).

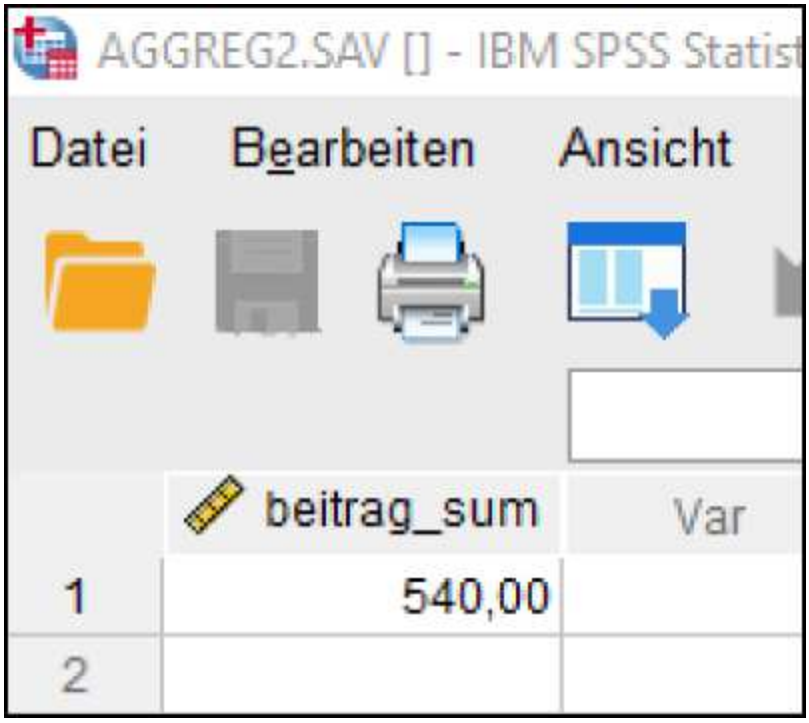



**Abbildung 5.46:** Gesamtbeitragssumme nach der Aggregation von der Personenebene (*AGGREG2.SAV*)

## *Weitere Grundbefehle*

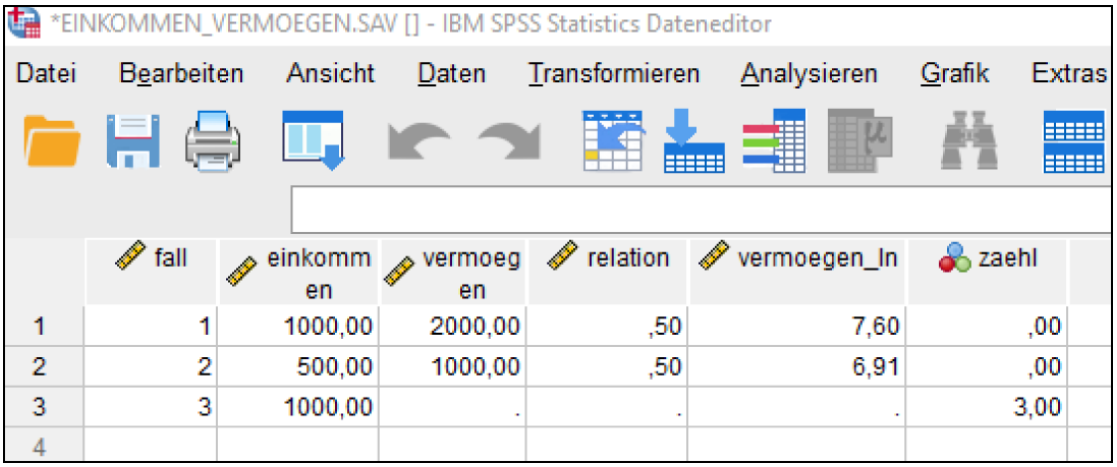
Über die vorstehenden Befehle hinaus gibt es eine Reihe weiterer SPSS-Grundbefehle hinsichtlich des **Datenmanagements**, zum Beispiel (alphabetisch geordnet):

- ✔ COUNT: Der COUNT-Befehl dient dem fallweisen Zählen bestimmter Variablenausprägungen.

  Beispiel: Unter Bezugnahme auf *EINKOMMEN_VERMOEGEN.SAV* sollen mit COUNT in der Zählvariablen *zaehl* für die Variablen *einkommen, vermoegen, relation* und *vermoegen_ln* die systemfehlenden Werte ausgezählt werden (daher steht SYSMIS in Klammern), wobei zur Ausführung dieses Befehls das EXECUTE am Ende wichtig ist (siehe auch *SYNTAX15.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\EINKOMMEN_VERMOEGEN.SAV'.
COUNT zaehl = einkommen vermoegen relation vermoegen_ln (SYSMIS).
EXECUTE.
```

Da nur im dritten Fall bei *vermoegen, relation* und *vermoegen_ln* systemfehlende Werte vorkommen, weist SPSS nur bei diesem dritten Fall einen von Null abweichenden Wert auf, und zwar eben »3« jeweils einmal bei jeder Variablen (siehe Abbildung 5.47).




| | fall | einkommen | vermoegen | relation | vermoegen_ln | zaehl |
|---|---|---|---|---|---|---|
| 1 | 1 | 1000,00 | 2000,00 | ,50 | 7,60 | ,00 |
| 2 | 2 | 500,00 | 1000,00 | ,50 | 6,91 | ,00 |
| 3 | 3 | 1000,00 | . | . | . | 3,00 |
| 4 | | | | | | |

**Abbildung 5.47:** Zur Anwendung des COUNT-Befehls (*EINKOMMEN_VERMOEGEN.SAV*)

Alternativ zu SYSMIS (für das Zählen systemfehlender Werte) stehen beim COUNT-Befehl auch noch die Schlüsselwörter MISSING, LOWEST, HIGHEST und THRU zur Verfügung.

- `LIST`: Der LIST-Befehl dient dem Auflisten einzelner Variablen beziehungsweise Fälle im Ausgabefenster. Sie können dabei einzelne Variablen auswählen, aber mit ALL – wie im nachfolgenden Beispiel – auch alle Variablen eines Datensatzes auflisten lassen. Im nachfolgenden Beispiel sollen unter Bezugnahme auf *PERSONEN4A.SAV* für alle Variablen dieses Datensatzes nur die Fälle 3 bis 11 aufgelistet werden (siehe *SYNTAX16.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN4A.SAV'.
LIST VARIABLES = ALL
/CASES FROM 3 TO 11.
```

Der LIST-Befehl stammt noch aus der guten alten Betriebssystemzeit von DOS. Heutzutage dient er eigentlich nur noch zu einem schnellen

Überblick über ausgewählte Fallbereiche für einige ausgewählte Variablen, was bei größeren Datensätzen hilfreich sein kann – zumindest dann, wenn Sie für die ausgewählten Fallbereiche irgendwelche Auffälligkeiten vermuten.

Wie die Auflistung von Variablen und Fällen in unserem Beispiel aussieht, zeigt Abbildung 5.48.

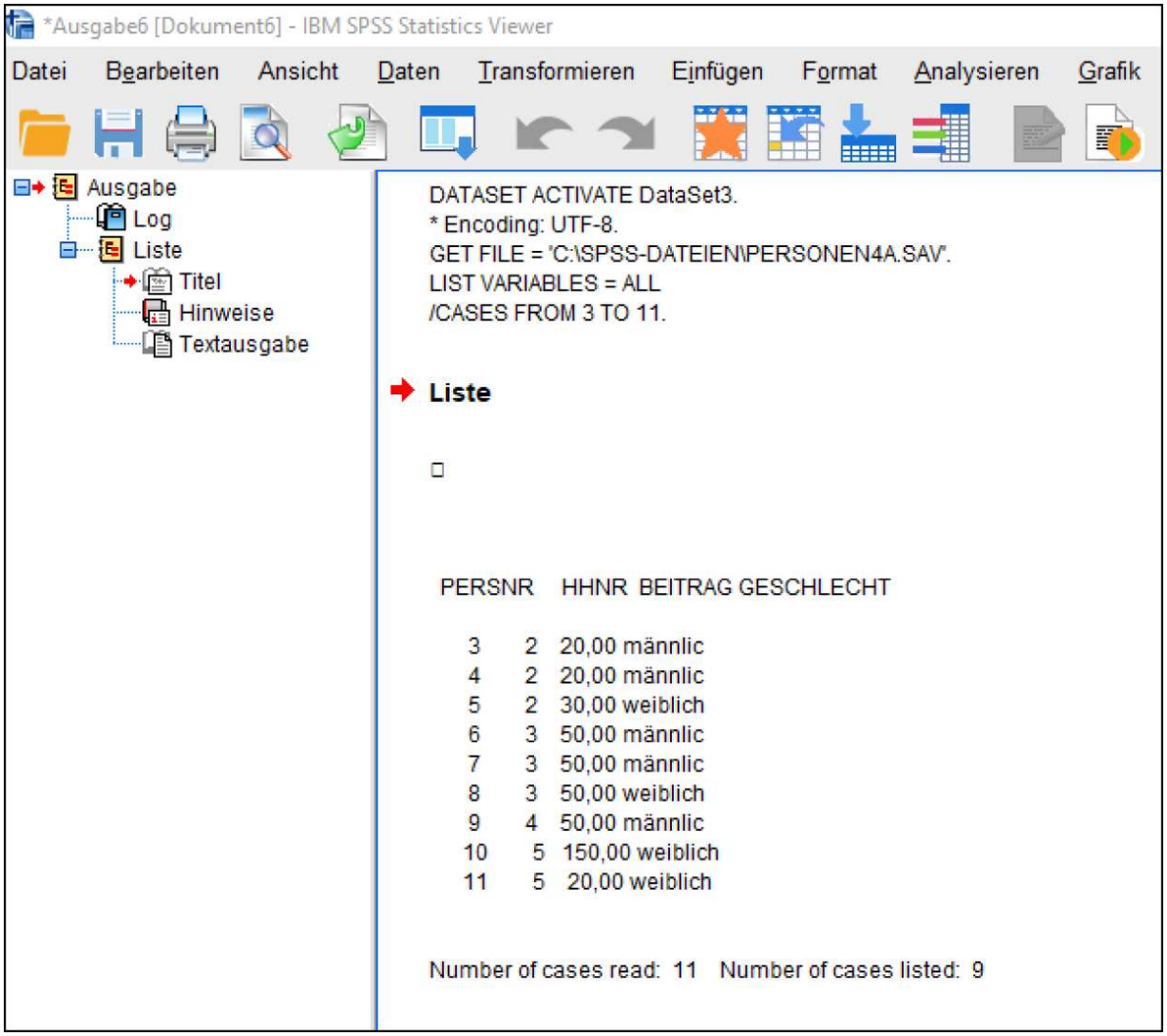



**Abbildung 5.48:** Auflisten ausgewählter Variablen und Fälle – LIST (*PERSONEN4A.SAV*)

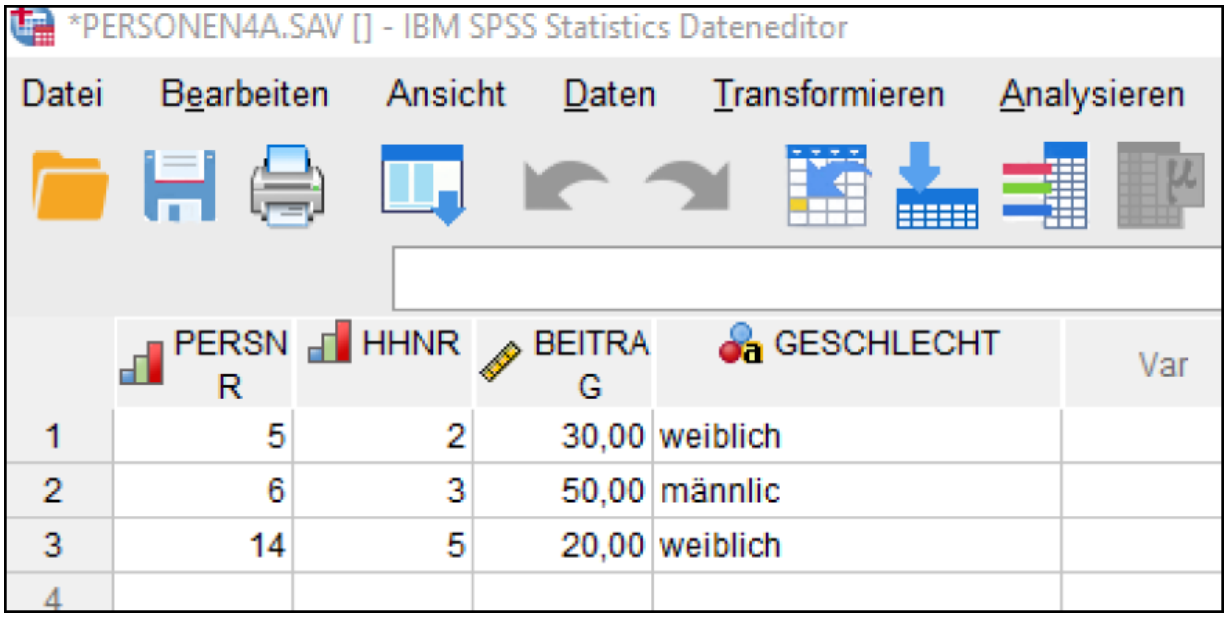
- ✔ `SAMPLE`: Mithilfe des SAMPLE-Befehls können Sie eine zufällige Anzahl von Fällen aus Ihrem Datensatz auswählen. Diese (Zufalls-)Auswahl gilt dann für alle weiteren Prozeduren – es sei denn, Sie setzen vor den SAMPLE-Befehl ein »TEMPORARY.«.

Der SAMPLE-Befehl wird erst wirksam, wenn er mit einem EXECUTE – oder einer (statistischen) Prozedur – abgeschlossen wird. Eine Zufallsauswahl kann zum Beispiel dann sinnvoll sein, wenn Sie eine sehr große Stichprobe (etwa mit mehreren Millionen Datensätzen) zu verarbeiten haben und Ihre komplexen Auswertungen auf eine »verarbeitbare« Fallanzahl reduzieren möchten.

Im folgenden Beispiel wird aus den 15 Fällen der Datendatei *PERSONEN4A.SAV* eine 20-Prozent-Stichprobe gezogen (siehe *SYNTAX17.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN4A.SAV'.
SAMPLE 0.2.
EXECUTE.
```

Bei dieser Anwendung des SAMPLE-Befehls wurden – wie Abbildung 5.49 verdeutlicht – die Fälle mit den Personennummern 5, 6 und 14 gezogen. Wundern Sie sich nicht, wenn bei Ihrer Anwendung des SAMPLE-Befehls auf *PERSONEN4A.SAV* andere Fälle gezogen werden. Es handelt sich hier ja um eine Zufallsauswahl, und es ist das Wesen einer Zufallsauswahl, dass Sie nicht vorhersagen können, welche Fälle genau ausgewählt werden. Auch stimmt – gerade bei kleinen Stichproben – die ausgewählte Fallzahl nicht immer mit dem angegebenen Auswahlsatz überein, da es sich hier in SPSS nur um eine näherungsweise X-Prozent-Zufallsauswahl handelt. In Abbildung 5.49 wurden – sozusagen »zufällig« (bitte entschuldigen Sie diesen kleinen Wortwitz!) – mit drei Fällen genau 20 Prozent der 15 Fälle ausgewählt; es hätten aber auch zum Beispiel zwei oder vier ausgewählte Fälle sein können.




| | PERSNR | HHNR | BEITRAG | GESCHLECHT | Var |
|---|---|---|---|---|---|
| 1 | 5 | 2 | 30,00 | weiblich | |
| 2 | 6 | 3 | 50,00 | männlic | |
| 3 | 14 | 5 | 20,00 | weiblich | |
| 4 | | | | | |

**Abbildung 5.49:** Zufallsauswahl mit dem SAMPLE-Befehl (*PERSONEN4A.SAV*)

- ✔ SYSFILE INFO: Mithilfe des Befehls SYSFILE INFO lassen sich Datei- beziehungsweise Variableninformationen vollständig auflisten. Hierdurch erhalten Sie im Grunde genommen die im Variableneditor enthaltenen Informationen schön übersichtlich im Ausgabefenster.

  Mit der folgenden Syntax (siehe *SYNTAX18.SPS*) werden für die Datendatei PERSONEN4A.SAV die entsprechenden Informationen angefordert:

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN4A.SAV'.
SYSFILE INFO FILE = 'C:\SPSS-DATEIEN\PERSONEN4A.SAV'.
```

Diese im Ausgabefenster dargelegten Informationen sind in Abbildung 5.50 dargestellt. Aus dem oberen Teil von Abbildung 5.50 sind Informationen zur betreffenden Datendatei dargelegt, beispielsweise die Fallzahl (15 Fälle). Im unteren Teil von Abbildung 5.50 finden sich Variableninformationen: Konkret sind dort die Variablenbeschreibungen abgebildet, wie sie von Ihnen im Variableneditor hinterlegt worden sind.

Beispielsweise findet sich dort eine Spalte mit den Formatierungen der einzelnen Variablen: Die numerischen Variablen *persnr* und *hhnr* sind jeweils mit »F8« formatiert. Diese Formatierungsangabe stammt

ursprünglich aus der Programmiersprache FORTRAN, wobei das »F« für »Fixkommastelle« steht. In diesem Fall werden sowohl *persnr* als auch *hhnr* bis zu achtstellig ohne Nachkommastellen im Dateneditor angezeigt. Die numerische Variable *beitrag* ist, wie aus Abbildung 5.50 zu erkennen ist, mit »F8.2« formatiert; das heißt: Sie wird mit bis zu acht Vorkommastellen und mit zwei Nachkommastellen im Dateneditor abgebildet. Bei den genannten numerischen Variablen sind allerdings Berechnungen für Werte mit mehr als acht Stellen vor dem Komma unproblematisch, weil SPSS intern ohnehin mit (viel) mehr als acht Vorkommastellen rechnet. In diesen Fällen würden die Werte der betreffenden numerischen Variablen im Dateneditor in Potenzform dargestellt.

**Informationen zur Datei**

| | | |
|---|---|---|
| Quelle | | C:\SPSS-DATEIEN\PERSONEN4A.SAV |
| Typ | | SPSS Statistics-Datendatei |
| Erstellungsdatum | | 28-AUG-2017 22:48:23 |
| Label | | Keine |
| Zeichenkodierung | | UTF-8 |
| Inhalt der Datei | Datentyp | Fall |
| | Zeilenanzahl des Dokuments | Keine |
| | Variablen-Sets | Keine |
| | Datumsinformation für Forecasting | Keine |
| | Definitionen von Mehrfachantworten | Keine |
| | Information für Data Entry für Windows | Keine |
| | Information für TextSmart | Keine |
| | Information für Modeler | Keine |
| Information über die Daten | Anzahl der Fälle | 15 |
| | Anzahl der Variablenelemente | 5 |
| | Anzahl der benannten Variablen | 4 |
| | Gewichtungsvariable | Keine |
| | Komprimiert | Ja |

**Variablenbeschreibung**

| Name | Position | Label | Messniveau | Format | Spaltenbreite | Ausrichtung |
|---|---|---|---|---|---|---|
| PERSNR | 1 | Personennummer | Ordinal | F8 | 8 | Rechtsbündig |
| HHNR | 2 | Haushaltsnummer | Ordinal | F8 | 8 | Rechtsbündig |
| BEITRAG | 3 | Geleisteter Obolus | Skalar | F8.2 | 8 | Rechtsbündig |
| GESCHLECHT | 4 | Geschlecht der Person | Nominal | A10 | 20 | Linksbündig |

**Abbildung 5.50:** Variablen- und Dateninformationen – SYSFILE INFO (*PERSONEN4A.SAV*)

Die Stringvariable *geschlecht* schließlich ist als »A10« formatiert. Dabei steht das »A« für »alphanumerisch«. Diese Variable ist im vorliegenden Beispiel in ihrer Darstellung im Dateneditor, aber auch – und das ist ein Unterschied zu den numerischen Variablen – hinsichtlich ihrer statistischen Weiterverarbeitung auf zehn Stellen, das heißt Zeichen, beschränkt.

- `TITLE`: Mit dem TITLE-Befehl versehen Sie Ihren Output an der entsprechenden Stelle mit einem Titel. Insofern ist dieser Befehl eine Alternative zu Kommentaren, die bei entsprechender Einstellung ja auch im Ausgabefenster angezeigt werden. Allerdings ist der TITLE-Befehl als Tabellen- oder Diagrammüberschrift konzipiert und sollte sinnvollerweise auch dort eingesetzt werden. Nach TITLE ist der Text der Überschrift (= eine Zeichenkette) in Hochkommata oder Anführungszeichen zu setzen.

  Nehmen Sie einfach mal an, dass Sie mit der folgenden Syntax (siehe auch *SYNTAX19A.SPS*) für den Datensatz *PERSONEN4A.SAV* eine Häufigkeitsauszählung der Variablen *beitrag* getrennt nach Geschlecht durchführen möchten. In diesem Fall können Sie vor dem die betreffende Häufigkeitstabelle erzeugenden FREQUENCIES-Befehl als Überschrift der Tabelle – zugegebenermaßen recht langatmig und inhaltlich zumindest leicht fragwürdig – formulieren: »Häufigkeitstabelle für die Beitragszahlungen – nach Geschlecht getrennt, Faik-Beispieldatensatz, erstellt am 26.09.2022 von Jürgen Faik, der gerne lange Titel kreiert und Spaß an Statistik hat, gerade an Häufigkeitstabellen, und auch an ökonomischen Zusammenhängen interessiert ist«:

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN4A.SAV'.
SORT CASES BY geschlecht.
SPLIT FILE BY geschlecht.
TITLE 'Häufigkeitstabelle für die Beitragszahlungen – nach Geschlecht
getrennt, Faik-Beispieldatensatz, erstellt am 26.09.2022 von Jürgen Faik,
der gerne lange Titel kreiert und Spaß an Statistik hat, gerade an
Häufigkeitstabellen, und auch an ökonomischen Zusammenhängen interessiert
ist'.
FREQUENCIES VARIABLES = beitrag.
```

Dann erhalten Sie den in Abbildung 5.51 dargestellten Output.

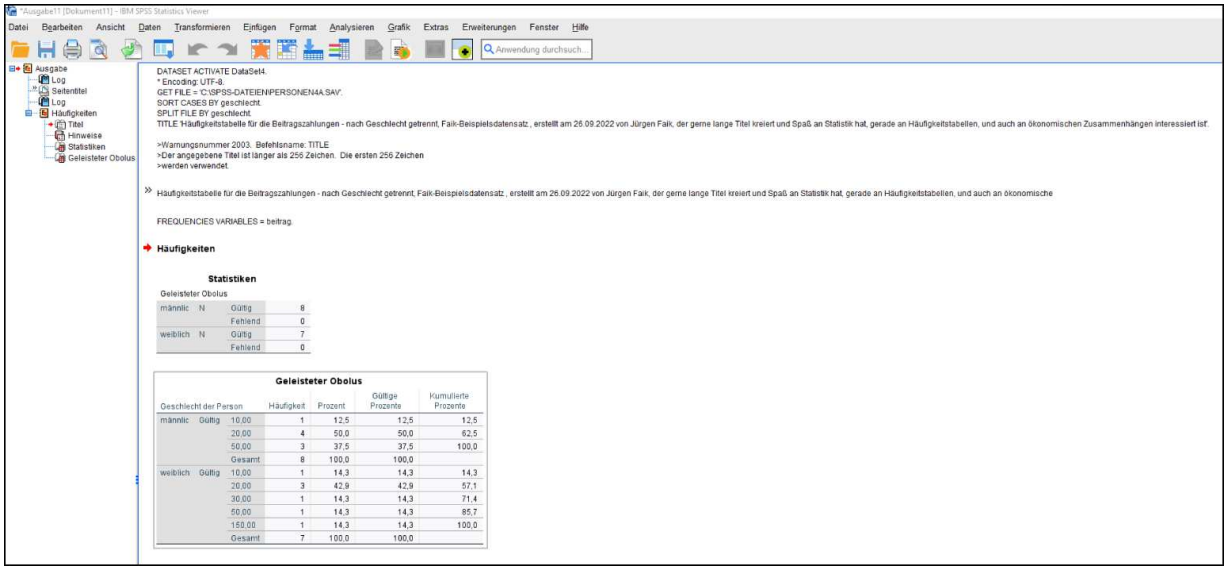



**Abbildung 5.51:** Zu lange Titelzeile – TITLE (*PERSONEN4A.SAV*)

Offensichtlich ist der gewählte Titel zu lang. Titel sind in SPSS 29 nämlich auf 256 Zeichen beschränkt. Sie könnten jetzt Ihren Titel auf 256 Zeichen kürzen. Um aber dennoch eine Tabellenüberschrift mit mehr als 256 Zeichen zu kreieren, können Sie den TITLE-Befehl über mehr als eine Zeile verteilen, zum Beispiel so (siehe *SYNTAX19B.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\PERSONEN4A.SAV'.
SORT CASES BY geschlecht.
SPLIT FILE BY geschlecht.
TITLE 'Häufigkeitstabelle für die Beitragszahlungen – nach Geschlecht
getrennt, Faik-Beispieldatensatz, erstellt am 26.09.2022 von Jürgen Faik,'.
TITLE 'der gerne lange Titel kreiert und Spaβ an Statistik hat, gerade an
Häufigkeitstabellen, und auch an ökonomischen Zusammenhängen interessiert
ist'.
FREQUENCIES VARIABLES = beitrag.
```

Jetzt gibt es von SPSS keine Fehlermeldung mehr (siehe Abbildung 5.52). Allerdings ist ein derartig langer Titel wie eben ja keineswegs die Regel, und 256 Zeichen sind für einen Titel eine ganze Menge! An dieser Stelle gibt es im Übrigen gegenüber den Vorgängerversionen von SPSS 29 einen echten Fortschritt, denn in

diesen Vorgängerversionen lag bezüglich der Titel eine Zeichenbeschränkung auf maximal 60 Zeichen vor.

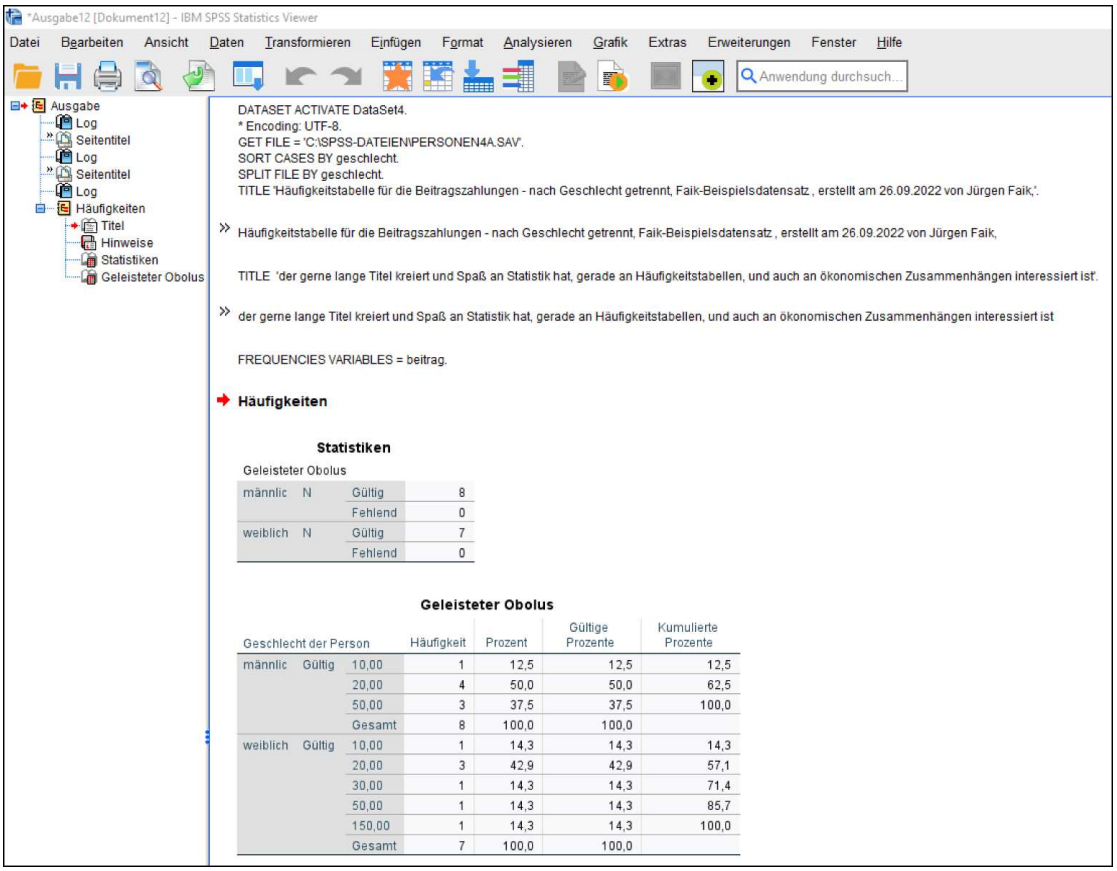



**Abbildung 5.52:** Auf zwei Zeilen verteilter Titel – TITLE (*PERSONEN4A.SAV*)

WEIGHT BY: Mit WEIGHT BY werden Fälle gewichtet. Wenn Sie diese Gewichtung wieder aufheben möchten, geben Sie einfach WEIGHT OFF ein.

Im Folgenden werden mithilfe des haushaltsbezogenen Datensatzes *AGGR.SAV* mit den Variablen *hhnr* (Haushaltsnummer), *beitrag_sum* (Haushaltsgesamtbeitrag) und *persnr_nu* (Anzahl der Haushaltsmitglieder) zum einen mit der Haushaltsgröße gewichtete Statistiken für *beitrag_sum* und zum anderen ungewichtete Statistiken für dieselbe Variable berechnet. Im erstgenannten Fall wird der Haushaltsbeitrag jedem Haushaltsmitglied (fiktiv) zugewiesen, und Sie erhalten Statistiken für alle (15) Personen. Demgegenüber werden im zweitgenannten Fall – ohne Gewichtung – Statistiken für alle (5) Haushalte berechnet.

Zu diesem Zweck wird in der Datendatei *AGGR.SAV* vor der ersten FREQUENCIES-Anweisung die Gewichtung mit der Gewichtungsvariablen »Haushaltsgröße« über den WEIGHT-BY-Befehl aktiviert, später dann aber – vor der zweiten FREQUENCIES-Anweisung – mittels der WEIGHT OFF-Anweisung wieder deaktiviert (siehe auch *SYNTAX20.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\AGGR.SAV'.
WEIGHT BY persnr_nu.
FREQUENCIES VARIABLES = beitrag_sum
/FORMAT = NOTABLE
/STATISTICS=ALL.
WEIGHT OFF.
FREQUENCIES VARIABLES = beitrag_sum
/FORMAT = NOTABLE
/STATISTICS=ALL.
```

Wie aus Abbildung 5.53 und Abbildung 5.54 ersichtlich ist, unterscheiden sich die ausgewiesenen Statistiken im Fall mit Gewichtung von denen im Fall ohne Gewichtung. Beispielsweise beträgt das arithmetische Mittel für den Gesamtbeitrag im Fall mit Gewichtung 155,33 Euro pro Person, während es sich im Fall ohne Gewichtung auf 108,00 Euro pro Haushalt beläuft.

**Statistiken**

BEITRAG_sum

| | | |
|---|---|---|
| N | Gültig | 15 |
| | Fehlend | 0 |
| Mittelwert | | 155,3333 |
| Standardfehler des Mittelwerts | | 22,75892 |
| Median | | 150,0000 |
| Modus | | 250,00 |
| Std.-Abweichung | | 88,14490 |
| Varianz | | 7769,524 |
| Schiefe | | -,103 |
| Standardfehler der Schiefe | | ,580 |
| Kurtosis | | -1,621 |
| Standardfehler der Kurtosis | | 1,121 |
| Spannweite | | 240,00 |
| Minimum | | 10,00 |
| Maximum | | 250,00 |
| Summe | | 2330,00 |

**Abbildung 5.53:** Statistiken im Fall mit Gewichtung – WEIGHT BY (*AGGR.SAV*)

**Statistiken**

**Statistiken**

BEITRAG_sum

| | | |
|---|---|---|
| N | Gültig | 5 |
| | Fehlend | 0 |
| Mittelwert | | 108,0000 |
| Standardfehler des Mittelwerts | | 42,23742 |
| Median | | 80,0000 |
| Modus | | 10,00[a] |
| Std.-Abweichung | | 94,44575 |
| Varianz | | 8920,000 |
| Schiefe | | ,880 |
| Standardfehler der Schiefe | | ,913 |
| Kurtosis | | ,073 |
| Standardfehler der Kurtosis | | 2,000 |
| Spannweite | | 240,00 |
| Minimum | | 10,00 |
| Maximum | | 250,00 |
| Summe | | 540,00 |

a. Mehrere Modi vorhanden. Der kleinste Wert wird angezeigt.

**Abbildung 5.54:** Statistiken im Fall ohne Gewichtung – WEIGHT OFF (*AGGR.SAV*)

# *Ein kleines SPSS-Beispiel-Syntaxprogramm*

Vorstehend haben Sie eine Reihe von SPSS-Grundbefehlen kennengelernt. Sozusagen im Sinne einer Übung werden im Folgenden die meisten dieser Befehle in einem zusammenhängenden Beispiel angewendet.

## *RENAME-, ERASE-Befehl und ADDVARIABLES-Modus*

In dem folgenden Syntaxprogramm finden sich allerdings am Anfang und am Ende des Programms noch zwei weitere Befehle. Am Anfang des Programms wird der RENAME-Befehl genutzt. Dies ist dort notwendig, weil die kurz zuvor invertierte Datendatei von SPSS die Variablennamen *var001* bis *var005* zugewiesen bekommen hat, diese Namen aber durch die gewünschten und für die Zusammenführung einzelner Datensätze notwendigen Variablennamen *persnr, hhnr, einkommen, vermoegen* und *verbrauch* ersetzt werden müssen.

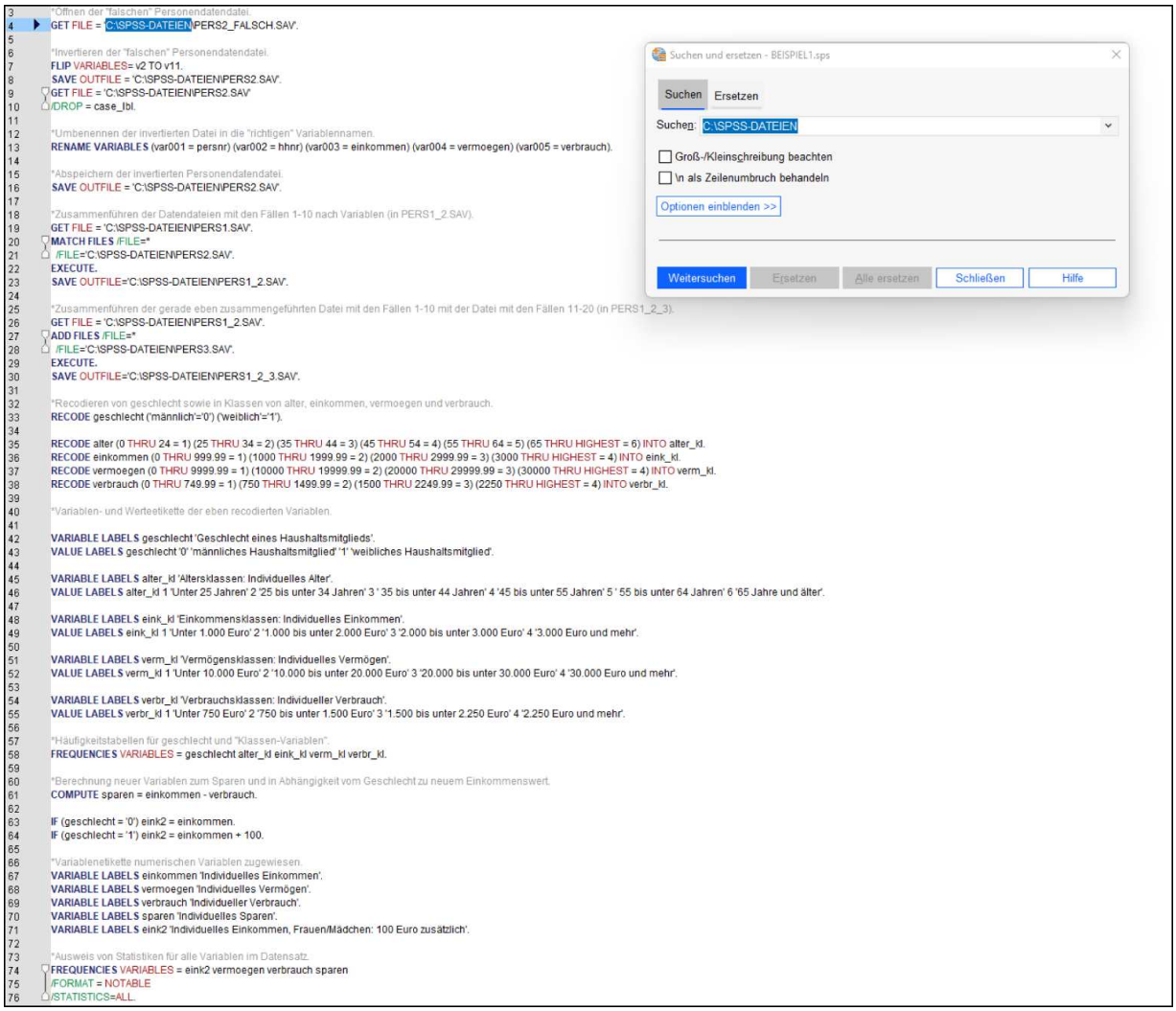
Ganz zum Schluss des Programms wird der ERASE-FILE-Befehl angewendet, um dort eine **Arbeitsdatei** zu löschen. Eine Arbeitsdatei ist eine temporär angelegte *SAV*-Datei, in die in der Regel Zwischenergebnisse von SPSS-Syntaxprogrammen als Daten in der üblichen SPSS-Form mit den Fällen in den Zeilen und den Variablen in den Spalten abgelegt werden. Diese Daten stehen für die spätere Weiterverarbeitung in SPSS zur Verfügung. Da Arbeitsdateien die gleiche Dateikennzeichnung wie endgültige Datendateien, aber auch Rohdatendateien haben (*SAV*), können Sie, müssen aber nicht, Arbeitsdateien anders als die anderen Datendateien bezeichnen (zum Beispiel durch den Namenszusatz *_TEMP* vor der Extension *SAV*). Dann behalten Sie einen besseren Überblick über Ihre verschiedenen Datendateien!

Die eben erwähnte Löschung, und zwar der Datei *HH_ALLE.SAV*, ist notwendig, weil ansonsten bei einem erneuten Aufruf des Syntaxprogramms ein Namenskonflikt bezüglich dieser Datei auftreten könnte und dadurch das Syntaxprogramm gestoppt würde. Gemäß Ihrer SPSS-Grundeinstellungen könnte (beziehungsweise bei mir: wird) zum Beispiel bei einem Neuaufruf Ihres Syntaxprogramms aus der bereits abgespeicherten Datei *HH_ALLE.SAV* von SPSS an dieser Stelle: *HH_ALLE2.SAV*. Wenn Sie nun in Ihrem Programm Ihre Arbeitsdatei über den Namen *HH_ALLE.SAV* zur Weiterverwendung aufrufen, erkennt SPSS diese nicht (weil es sich jetzt auf den Namen *HH_ALLE2.SAV* eingestellt hat), und die Ausführung Ihres Syntaxprogramms wird abgebrochen.

Außerdem findet sich im hinteren Teil des Programms beim letzten AGGREGATE-Befehl und dem ersten Unterbefehl die Spezifikation MODE = ADDVARIABLES. Diese Spezifikation ist erforderlich, damit das dort über alle Haushalte hinweg berechnete Durchschnittseinkommen jedem Haushalt (fiktiv) zugewiesen wird und nicht etwa in einer einzeiligen Datendatei »verschwindet«.

Die nachfolgende Syntax finden Sie in *BEISPIEL1.SPS*. Durch die umfassende Kommentierung dieses Programms sind die einzelnen inhaltlichen Schritte quasi selbsterklärend. Das Syntaxprogramm bringen Sie mit AUSFÜHREN|ALLE »zum Laufen«.

An vielen Stellen des Syntaxprogramms ist meine Verzeichnisstruktur angegeben. Diese sollten Sie mittels SUCHEN UND ERSETZEN durch Ihre Verzeichnisstruktur ersetzen (an allen Stellen des Programms), wie in Abbildung 5.55 angedeutet.



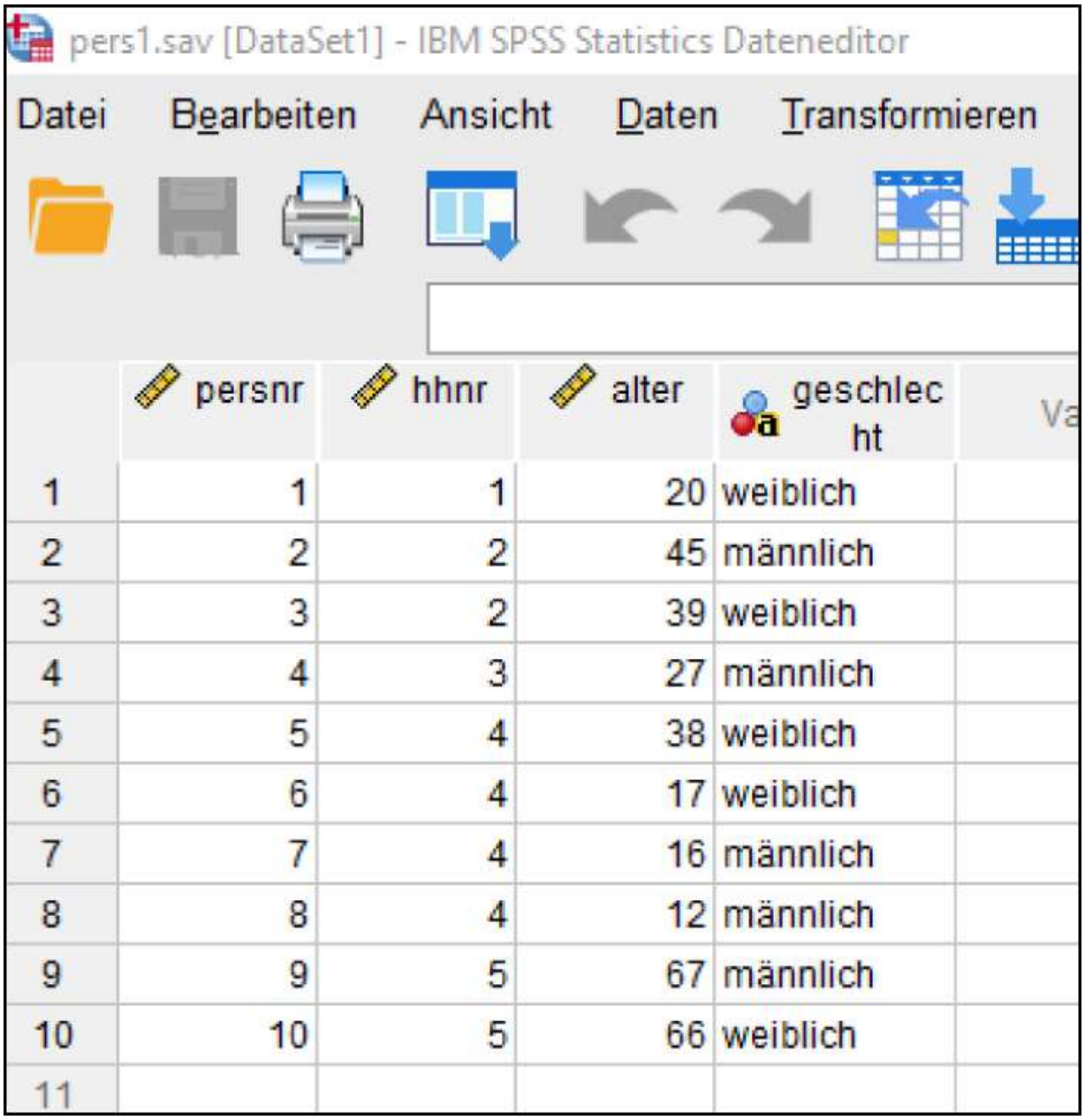

```
*Öffnen der "falschen" Personendatendatei.
GET FILE = 'C:\SPSS-DATEIEN\PERS2_FALSCH.SAV'.

*Invertieren der "falschen" Personendatendatei.
FLIP VARIABLES= v2 TO v11.
SAVE OUTFILE = 'C:\SPSS-DATEIEN\PERS2.SAV'.
GET FILE = 'C:\SPSS-DATEIEN\PERS2.SAV'
/DROP = case_lbl.

*Umbenennen der invertierten Datei in die "richtigen" Variablennamen.
RENAME VARIABLES (var001 = persnr) (var002 = hhnr) (var003 = einkommen) (var004 = vermoegen) (var005 = verbrauch).

*Abspeichern der invertierten Personendatendatei.
SAVE OUTFILE = 'C:\SPSS-DATEIEN\PERS2.SAV'.

*Zusammenführen der Datendateien mit den Fällen 1-10 nach Variablen (in PERS1_2.SAV).
GET FILE = 'C:\SPSS-DATEIEN\PERS1.SAV'.
MATCH FILES /FILE=*
 /FILE='C:\SPSS-DATEIEN\PERS2.SAV'.
EXECUTE.
SAVE OUTFILE='C:\SPSS-DATEIEN\PERS1_2.SAV'.

*Zusammenführen der gerade eben zusammengeführten Datei mit den Fällen 1-10 mit der Datei mit den Fällen 11-20 (in PERS1_2_3).
GET FILE = 'C:\SPSS-DATEIEN\PERS1_2.SAV'.
ADD FILES /FILE=*
 /FILE='C:\SPSS-DATEIEN\PERS3.SAV'.
EXECUTE.
SAVE OUTFILE='C:\SPSS-DATEIEN\PERS1_2_3.SAV'.

*Recodieren von geschlecht sowie in Klassen von alter, einkommen, vermoegen und verbrauch.
RECODE geschlecht ('männlich'='0') ('weiblich'='1').

RECODE alter (0 THRU 24 = 1) (25 THRU 34 = 2) (35 THRU 44 = 3) (45 THRU 54 = 4) (55 THRU 64 = 5) (65 THRU HIGHEST = 6) INTO alter_kl.
RECODE einkommen (0 THRU 999.99 = 1) (1000 THRU 1999.99 = 2) (2000 THRU 2999.99 = 3) (3000 THRU HIGHEST = 4) INTO eink_kl.
RECODE vermoegen (0 THRU 9999.99 = 1) (10000 THRU 19999.99 = 2) (20000 THRU 29999.99 = 3) (30000 THRU HIGHEST = 4) INTO verm_kl.
RECODE verbrauch (0 THRU 749.99 = 1) (750 THRU 1499.99 = 2) (1500 THRU 2249.99 = 3) (2250 THRU HIGHEST = 4) INTO verbr_kl.

*Variablen- und Werteetikette der eben recodierten Variablen.

VARIABLE LABELS geschlecht 'Geschlecht eines Haushaltsmitglieds'.
VALUE LABELS geschlecht '0' 'männliches Haushaltsmitglied' '1' 'weibliches Haushaltsmitglied'.

VARIABLE LABELS alter_kl 'Altersklassen: Individuelles Alter'.
VALUE LABELS alter_kl 1 'Unter 25 Jahren' 2 '25 bis unter 34 Jahren' 3 ' 35 bis unter 44 Jahren' 4 '45 bis unter 55 Jahren' 5 ' 55 bis unter 64 Jahren' 6 '65 Jahre und älter'.

VARIABLE LABELS eink_kl 'Einkommensklassen: Individuelles Einkommen'.
VALUE LABELS eink_kl 1 'Unter 1.000 Euro' 2 '1.000 bis unter 2.000 Euro' 3 '2.000 bis unter 3.000 Euro' 4 '3.000 Euro und mehr'.

VARIABLE LABELS verm_kl 'Vermögensklassen: Individuelles Vermögen'.
VALUE LABELS verm_kl 1 'Unter 10.000 Euro' 2 '10.000 bis unter 20.000 Euro' 3 '20.000 bis unter 30.000 Euro' 4 '30.000 Euro und mehr'.

VARIABLE LABELS verbr_kl 'Verbrauchsklassen: Individueller Verbrauch'.
VALUE LABELS verbr_kl 1 'Unter 750 Euro' 2 '750 bis unter 1.500 Euro' 3 '1.500 bis unter 2.250 Euro' 4 '2.250 Euro und mehr'.

*Häufigkeitstabellen für geschlecht und "Klassen-Variablen".
FREQUENCIES VARIABLES = geschlecht alter_kl eink_kl verm_kl verbr_kl.

*Berechnung neuer Variablen zum Sparen und in Abhängigkeit vom Geschlecht zu neuem Einkommenswert.
COMPUTE sparen = einkommen - verbrauch.

IF (geschlecht = '0') eink2 = einkommen.
IF (geschlecht = '1') eink2 = einkommen + 100.

*Variablenetikette numerischen Variablen zugewiesen.
VARIABLE LABELS einkommen 'Individuelles Einkommen'.
VARIABLE LABELS vermoegen 'Individuelles Vermögen'.
VARIABLE LABELS verbrauch 'Individueller Verbrauch'.
VARIABLE LABELS sparen 'Individuelles Sparen'.
VARIABLE LABELS eink2 'Individuelles Einkommen, Frauen/Mädchen: 100 Euro zusätzlich'.

*Ausweis von Statistiken für alle Variablen im Datensatz.
FREQUENCIES VARIABLES = eink2 vermoegen verbrauch sparen
 /FORMAT = NOTABLE
 /STATISTICS=ALL.
```

**Abbildung 5.55:** Ersetzen der Verzeichnisstruktur des Autors durch Ihre eigene in *BEISPIEL1.SPS*

## *Genutzte Rohdatendateien und ihre Zusammenführung*

Zu Beginn des nachfolgenden Syntaxprogramms werden drei Rohdatendateien verwendet und kurz darauf zusammengeführt (nach Variablen und nach Fällen). Im Einzelnen handelt es sich um die drei in Abbildung 5.56 bis Abbildung 5.58 dargestellten Datendateien: *PERS1.SAV*, *PERS2_FALSCH.SAV* und *PERS3.SAV*.

Die Datendatei *PERS1.SAV* enthält für 10 Fälle Informationen in Form der Variablen *persnr, hhnr, alter* und *geschlecht* (siehe Abbildung 5.56).

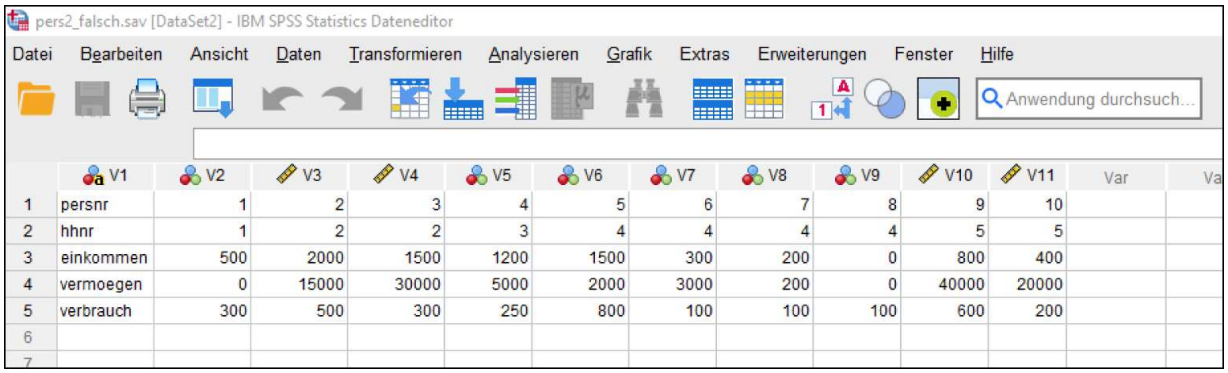



| | persnr | hhnr | alter | geschlecht | Va |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 20 | weiblich | |
| 2 | 2 | 2 | 45 | männlich | |
| 3 | 3 | 2 | 39 | weiblich | |
| 4 | 4 | 3 | 27 | männlich | |
| 5 | 5 | 4 | 38 | weiblich | |
| 6 | 6 | 4 | 17 | weiblich | |
| 7 | 7 | 4 | 16 | männlich | |
| 8 | 8 | 4 | 12 | männlich | |
| 9 | 9 | 5 | 67 | männlich | |
| 10 | 10 | 5 | 66 | weiblich | |
| 11 | | | | | |

**Abbildung 5.56:** Beispiel-Rohdatendatei *PERS1.SAV* (vier Variablen, Fälle 1–10)

Die Rohdatendatei *PERS2_FALSCH.SAV* ist – wie anhand von Abbildung 5.57 zu erkennen ist – in der für SPSS »falschen« Struktur angeordnet, und zwar mit ihren 10 Fällen in den Spalten und den Variablen *persnr, hhnr, einkommen, vermoegen* und *verbrauch* in den Zeilen. Sie muss also zunächst invertiert, das heißt, in die für SPSS »richtige« Struktur gebracht werden. Anschließend wird die invertierte Datendatei mit *PERS1.SAV* »zusammengebracht«, da sie

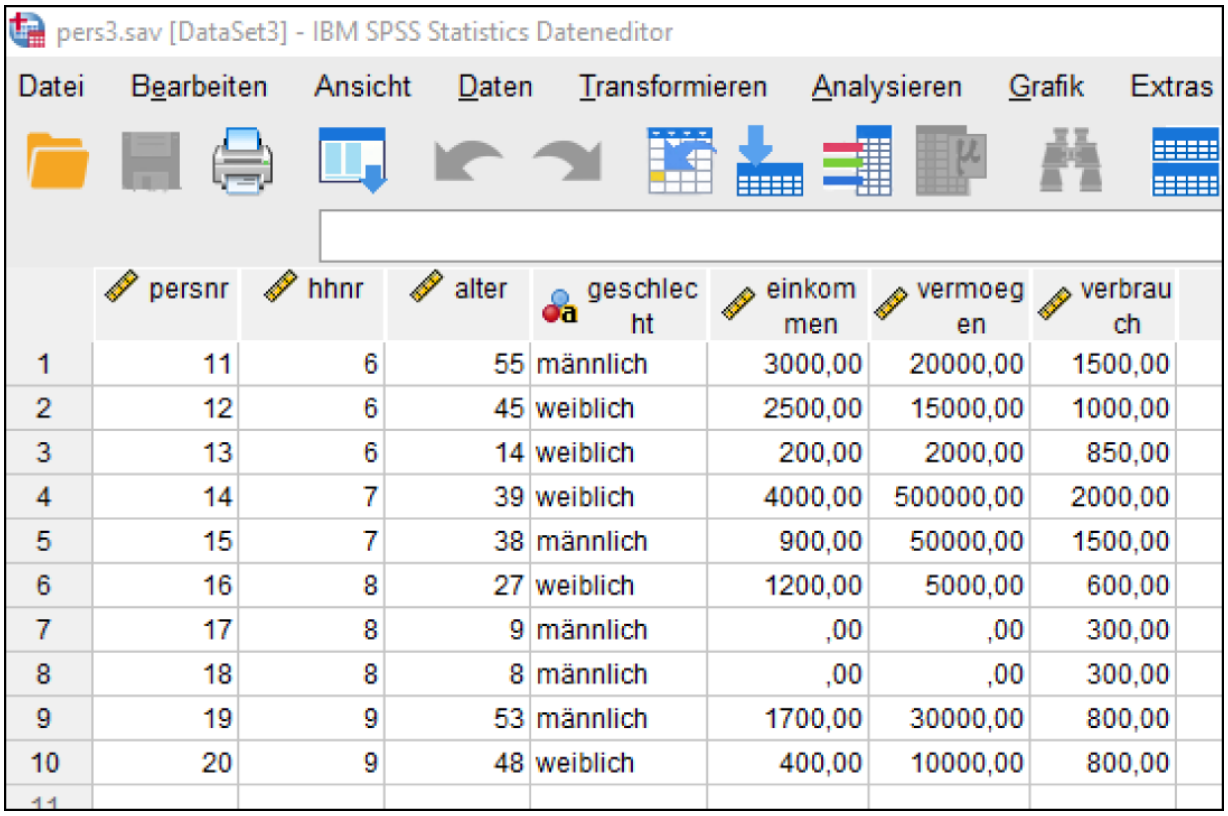
erkennbarerweise für dieselben Fälle wie bei *PERS1.SAV* andere Variablen enthält. Die Schlüsselvariable für diese Zusammenführung ist naheliegenderweise *persnr*, da diese Variable die einzelnen Fälle kennzeichnet. Durch die vorgenannte Zusammenführung erhalten Sie folglich für eine gegebene Anzahl von Fällen (hier: für 10 Personen) mehr Informationen, als durch die zugrunde liegenden Rohdatensätze jeweils für sich allein genommen bereitgestellt werden.




| | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 | V11 | Var | Va |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | persnr | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
| 2 | hhnr | 1 | 2 | 2 | 3 | 4 | 4 | 4 | 4 | 5 | 5 | | |
| 3 | einkommen | 500 | 2000 | 1500 | 1200 | 1500 | 300 | 200 | 0 | 800 | 400 | | |
| 4 | vermoegen | 0 | 15000 | 30000 | 5000 | 2000 | 3000 | 200 | 0 | 40000 | 20000 | | |
| 5 | verbrauch | 300 | 500 | 300 | 250 | 800 | 100 | 100 | 100 | 600 | 200 | | |
| 6 | | | | | | | | | | | | | |

**Abbildung 5.57:** Beispiel-Rohdatendatei *PERS2_FALSCH.SAV* (alle elf Variablen, Fälle 1–10, invertiert)

Dieselben Variablen wie in der aus *PERS1.SAV* und *PERS2_FALSCH.SAV* zusammengeführten Gesamtdatei enthält auch die aus Abbildung 5.58 ersichtliche Rohdatendatei *PERS3.SAV*. Die zusammengeführte Gesamtdatei und *PERS3.SAV* unterscheiden sich allerdings hinsichtlich der Fallnummern, wie Sie leicht anhand der unterschiedlichen Personennummern *persnr* erkennen können. In der genannten zusammengeführten Gesamtdatei befinden sich die Fälle 1 bis 10, während *PERS3.SAV* die Fälle 11 bis 20 enthält. In einem nächsten Schritt werden daher diese beiden Datendateien fallbezogen zusammengeführt.




| | persnr | hhnr | alter | geschlecht | einkommen | vermoegen | verbrauch |
|---|---|---|---|---|---|---|---|
| 1 | 11 | 6 | 55 | männlich | 3000,00 | 20000,00 | 1500,00 |
| 2 | 12 | 6 | 45 | weiblich | 2500,00 | 15000,00 | 1000,00 |
| 3 | 13 | 6 | 14 | weiblich | 200,00 | 2000,00 | 850,00 |
| 4 | 14 | 7 | 39 | weiblich | 4000,00 | 500000,00 | 2000,00 |
| 5 | 15 | 7 | 38 | männlich | 900,00 | 50000,00 | 1500,00 |
| 6 | 16 | 8 | 27 | weiblich | 1200,00 | 5000,00 | 600,00 |
| 7 | 17 | 8 | 9 | männlich | ,00 | ,00 | 300,00 |
| 8 | 18 | 8 | 8 | männlich | ,00 | ,00 | 300,00 |
| 9 | 19 | 9 | 53 | männlich | 1700,00 | 30000,00 | 800,00 |
| 10 | 20 | 9 | 48 | weiblich | 400,00 | 10000,00 | 800,00 |

**Abbildung 5.58:** Beispiel-Rohdatendatei *PERS3.SAV* (alle Variablen, Fälle 11–20)

Für die dadurch entstehende neue Gesamtdatei werden in dem folgenden Syntaxprogramm einige Variablentransformationen und dann ein paar kleinere statistische Auswertungen durchgeführt (konkret: Häufigkeitstabellen nach Alters-, Einkommens-, Vermögens- und Verbrauchsklassen sowie statistische Kennziffern für Einkommens-, Vermögens- und Verbrauchsgrößen beziehungsweise für Kombinationen aus ihnen). Da die betreffende neue Gesamtdatei sich auf personenbezogene Angaben (für insgesamt 20 Personen) bezieht, werden die entsprechenden statistischen Berechnungen auf der Personenebene durchgeführt. Die zugehörigen SPSS-Befehle sind im ersten Teil des nachfolgenden Syntaxprogramms enthalten.

## *Aggregationen*

Im zweiten Abschnitt des folgenden Syntaxprogramms werden die vorstehend erwähnten personenbezogenen Angaben auf die Haushaltsebene bezogen. Dazu werden verschiedene Variablen über den

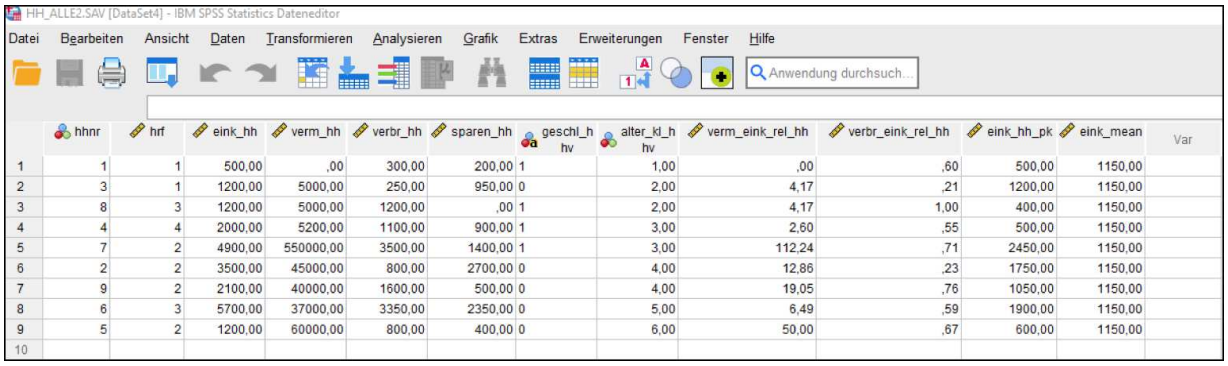
AGGREGATE-Befehl von der Personen- auf die Haushaltsebene »verschoben«, besser ausgedrückt: aggregiert.

In diesem Zusammenhang ist in Abbildung 5.59 die haushaltsbezogene Datendatei *HH_ALLE2.SAV* abgebildet, die Sie nach der letzten Aggregation erhalten. Dieser aggregierte Datensatz beinhaltet mit *hrf* eine Gewichtungsvariable. Inhaltlich repräsentiert *hrf* die Anzahl der jeweiligen Personen in einem Haushalt. Über *hrf* ist es möglich, aus dem haushaltsbezogenen Datensatz *HH_ALLE2.SAV* auch Aussagen in Bezug auf die Gesamtzahl der Personen abzuleiten. Das heißt: Die jeweiligen haushaltsbezogenen Variablenausprägungen werden jedem einzelnen Haushaltsmitglied sozusagen fiktiv zugeordnet. Aus 9 haushaltsbezogenen Fällen werden solcherart 20 personenbezogene Fälle bei Ihren jeweiligen statistischen Berechnungen. Sie blähen also sozusagen über die Gewichtung Ihre Fallzahl bei Ihren Berechnungen etwas auf. Das ist jetzt gar nicht angeberisch gemeint, sondern kann inhaltlich durchaus sinnvoll sein, zum Beispiel, wenn Sie Aussagen über die Wohlstandslage der einzelnen Haushaltsmitglieder treffen möchten.

Des Weiteren enthält *HH_ALLE2.SAV* haushaltsbezogene Einkommens-, Vermögens-, Verbrauchs- und Sparsummen (*eink_hh, verm_hh, verbr_hh* und *spar_hh*) in dem Sinne, dass die individuellen Werte zu einem haushaltsbezogenen Wert summiert worden sind. Das Haushaltseinkommen (*eink_hh*) ist dabei auch noch in eine Pro-Kopf-Größe umgerechnet worden (in *eink_hh_pk*). Zudem sind Vermögens-Einkommens- und Verbrauchs-Einkommensrelationen gebildet worden (*verm_eink_rel_hh* und *verbr_eink_rel_hh*).

Der letzten Spalte von *HH_ALLE2.SAV* können Sie entnehmen, dass die den Gesamteinkommensdurchschnitt widerspiegelnde (und durch eine Aggregation über den gesamten haushaltsbezogenen Datensatz – AGGREGATE-Befehl ohne Breakvariable – hervorgegangene) Variable *eink_mean* jedem der neun Haushalte zugeordnet ist (technisch wird dies über einen bestimmten Modus – MODE=ADDVARIABLES – beim AGGREGATE-Befehl erreicht). Das ist hier inhaltlich deshalb erforderlich, weil die Pro-Kopf-Haushaltseinkommen der einzelnen Haushalte jeweils mit diesem Durchschnittswert im Programm

verglichen werden, und zwar an der Stelle, an der die (temporäre) SELECT-IF-Anweisung zur Anwendung kommt. Diese (temporäre) SELECT-IF-Anweisung wählt hier nämlich die Haushalte aus, deren Pro-Kopf-Haushaltseinkommen geringer ist als die Hälfte des durchschnittlichen Pro-Kopf-Haushaltseinkommens (das heißt: 0,5 * *eink_mean*).




| | hhnr | hrf | eink_hh | verm_hh | verbr_hh | sparen_hh | geschl_hhv | alter_kl_hhv | verm_eink_rel_hh | verbr_eink_rel_hh | eink_hh_pk | eink_mean | Var |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 500,00 | ,00 | 300,00 | 200,00 | 1 | 1,00 | ,00 | ,60 | 500,00 | 1150,00 | |
| 2 | 3 | 1 | 1200,00 | 5000,00 | 250,00 | 950,00 | 0 | 2,00 | 4,17 | ,21 | 1200,00 | 1150,00 | |
| 3 | 8 | 3 | 1200,00 | 5000,00 | 1200,00 | ,00 | 1 | 2,00 | 4,17 | 1,00 | 400,00 | 1150,00 | |
| 4 | 4 | 4 | 2000,00 | 5200,00 | 1100,00 | 900,00 | 1 | 3,00 | 2,60 | ,55 | 500,00 | 1150,00 | |
| 5 | 7 | 2 | 4900,00 | 550000,00 | 3500,00 | 1400,00 | 1 | 3,00 | 112,24 | ,71 | 2450,00 | 1150,00 | |
| 6 | 2 | 2 | 3500,00 | 45000,00 | 800,00 | 2700,00 | 0 | 4,00 | 12,86 | ,23 | 1750,00 | 1150,00 | |
| 7 | 9 | 2 | 2100,00 | 40000,00 | 1600,00 | 500,00 | 0 | 4,00 | 19,05 | ,76 | 1050,00 | 1150,00 | |
| 8 | 6 | 3 | 5700,00 | 37000,00 | 3350,00 | 2350,00 | 0 | 5,00 | 6,49 | ,59 | 1900,00 | 1150,00 | |
| 9 | 5 | 2 | 1200,00 | 60000,00 | 800,00 | 400,00 | 0 | 6,00 | 50,00 | ,67 | 600,00 | 1150,00 | |
| 10 | | | | | | | | | | | | | |

**Abbildung 5.59:** Haushaltsbezogene Datendatei *HH_ALLE2.SAV* nach Durchführen aller Aggregationen am Ende des Syntaxprogramms

## *Das Beispiel-Syntaxprogramm in Gänze*

Insgesamt hat das Beispiel-Syntaxfile *BEISPIEL1.SPS* folgendes Aussehen:

```
GET FILE = 'C:\SPSS-DATEIEN\PERS2_FALSCH.SAV'.
* Invertieren der "falschen" Personendatendatei.
FLIP VARIABLES= v2 TO v11.
SAVE OUTFILE = 'C:\SPSS-DATEIEN\PERS2.SAV'.
GET FILE = 'C:\SPSS-DATEIEN\PERS2.SAV'
/DROP = case_lbl.

*Umbenennen der invertierten Datei in die "richtigen" Variablennamen.
RENAME VARIABLES (var001 = persnr) (var002 = hhnr) (var003 = einkommen)
(var004 = vermoegen) (var005 = verbrauch).

*Abspeichern der invertierten Personendatendatei.
SAVE OUTFILE = 'C:\SPSS-DATEIEN\PERS2.SAV'.

*Zusammenführen der Datendateien mit den Fällen 1-10 nach Variablen (in
PERS1_2.SAV).
GET FILE = 'C:\SPSS-DATEIEN\PERS1.SAV'.
MATCH FILES /FILE=*
      /FILE='C:\SPSS-DATEIEN\PERS2.SAV'.
EXECUTE.
```

```
  SAVE OUTFILE='C:\SPSS-DATEIEN\PERS1_2.SAV'.

  *Zusammenführen der gerade eben zusammengeführten Datei mit den Fällen 1-10
mit der Datei mit den Fällen 11-20 (in PERS1_2_3).
  GET FILE = 'C:\SPSS-DATEIEN\PERS1_2.SAV'.
  ADD FILES /FILE=*
        /FILE='C:\SPSS-DATEIEN\PERS3.SAV'.
  EXECUTE.
  SAVE OUTFILE='C:\SPSS-DATEIEN\PERS1_2_3.SAV'.

  *Recodieren von geschlecht sowie in Klassen von alter, einkommen, vermoegen
und verbrauch.
  RECODE geschlecht ('männlich'='0') ('weiblich'='1').
  RECODE alter (0 THRU 24 = 1) (25 THRU 34 = 2) (35 THRU 44 = 3)
  (45 THRU 54 = 4) (55 THRU 64 = 5) (65 THRU HIGHEST = 6) INTO alter_kl.
  RECODE einkommen (0 THRU 999.99 = 1) (1000 THRU 1999.99 = 2)
  (2000 THRU 2999.99 = 3) (3000 THRU HIGHEST = 4) INTO eink_kl.
  RECODE vermoegen (0 THRU 9999.99 = 1) (10000 THRU 19999.99 = 2)
  (20000 THRU 29999.99 = 3) (30000 THRU HIGHEST = 4) INTO verm_kl.
  RECODE verbrauch (0 THRU 749.99 = 1) (750 THRU 1499.99 = 2)
  (1500 THRU 2249.99 = 3) (2250 THRU HIGHEST = 4) INTO verbr_kl.

  *Variablen- und Werteetikette der eben recodierten Variablen.
  VARIABLE LABELS geschlecht 'Geschlecht eines Haushaltsmitglieds'.
  VALUE LABELS geschlecht '0' 'männliches Haushaltsmitglied' '1' 'weibliches
  Haushaltsmitglied'.
  VARIABLE LABELS alter_kl 'Altersklassen: Individuelles Alter'.
  VALUE LABELS alter_kl 1 'Unter 25 Jahren' 2 '25 bis unter 34 Jahren' 3 ' 35
  bis unter 44 Jahren' 4 '45 bis unter 55 Jahren' 5 ' 55 bis unter 64 Jahren' 6
  '65 Jahre und älter'.
  VARIABLE LABELS eink_kl 'Einkommensklassen: Individuelles Einkommen'.
  VALUE LABELS eink_kl 1 'Unter 1.000 Euro' 2 '1.000 bis unter 2.000 Euro' 3
  '2.000 bis unter 3.000 Euro' 4 '3.000 Euro und mehr'.
  VARIABLE LABELS verm_kl 'Vermögensklassen: Individuelles Vermögen'.
  VALUE LABELS verm_kl 1 'Unter 10.000 Euro' 2 '10.000 bis unter 20.000 Euro'
  3 '20.000 bis unter 30.000 Euro' 4 '30.000 Euro und mehr'.
  VARIABLE LABELS verbr_kl 'Verbrauchsklassen: Individueller Verbrauch'.
  VALUE LABELS verbr_kl 1 'Unter 750 Euro' 2 '750 bis unter 1.500 Euro' 3 '
  1.500 bis unter 2.250 Euro' 4 '2.250 Euro und mehr'.
  *Häufigkeitstabellen für geschlecht und "Klassen-Variablen".
  FREQUENCIES VARIABLES = geschlecht alter_kl eink_kl verm_kl verbr_kl.

  *Berechnung neuer Variable zum Sparen und in Abhängigkeit vom Geschlecht zu
neuem Einkommenswert.
  COMPUTE sparen = einkommen - verbrauch.
  IF (geschlecht = '0') eink2 = einkommen.
  IF (geschlecht = '1') eink2 = einkommen + 100.
```

```
  *Variablenetikette numerischen Variablen zugewiesen.
  VARIABLE LABELS einkommen 'Individuelles Einkommen'.
  VARIABLE LABELS vermoegen 'Individuelles Vermögen'.
  VARIABLE LABELS verbrauch 'Individueller Verbrauch'.
  VARIABLE LABELS sparen 'Individuelles Sparen'.
  VARIABLE LABELS eink2 'Individuelles Einkommen, Frauen/Mädchen: 100 Euro
zusätzlich'.

  *Ausweis von Statistiken für alle Variablen im Datensatz.
  FREQUENCIES VARIABLES = eink2 vermoegen verbrauch sparen
  /FORMAT = NOTABLE
  /STATISTICS=ALL.

  *Abspeichern der veränderten personenbezogenen Gesamtdatei unter neuem Namen
(PERS_ALLE).
  SAVE OUTFILE = 'C:\SPSS-DATEIEN\PERS_ALLE.SAV'.

  *Aggregation dieser personenbezogenen Gesamtdatei zu einer
  haushaltsbezogenen Gesamtdatei HH_ALLE.SAV für bestimmte Variablen.
  *Hierbei: hrf als Gewichtungsvariable i. S. v. Personenanzahl in Haushalt,
  Variablen _hh bezogen auf Haushalt, _hhv bezogen auf 1. Person im Haushalt.
  GET FILE = 'C:\SPSS-DATEIEN\PERS_ALLE.SAV'.
  AGGREGATE
     /OUTFILE='C:\SPSS-DATEIEN\HH_ALLE.SAV'
     /BREAK=hhnr
     /hrf = NU(persnr)
     /eink_hh = SUM(einkommen)
     /verm_hh = SUM(vermoegen)
     /verbr_hh = SUM(verbrauch)
     /sparen_hh = SUM (sparen)
     /geschl_hhv= FIRST(geschlecht)
     /alter_kl_hhv = FIRST (alter_kl).

  *Variablen- bzw. Werteetikette im neuen haushaltsbezogenen Datensatz
HH_ALLE.SAV.
  GET FILE = 'C:\SPSS-DATEIEN\HH_ALLE.SAV'.
  VARIABLE LABELS hrf 'Gewichtungsfaktor: Anzahl der Haushaltsmitglieder'.
  VARIABLE LABELS eink_hh 'Haushaltseinkommen'.
  VARIABLE LABELS verm_hh 'Haushaltsvermögen'.
  VARIABLE LABELS verbr_hh 'Haushaltsverbrauch'.
  VARIABLE LABELS sparen_hh 'Haushaltsersparnis'.
  VARIABLE LABELS geschl_hhv 'Geschlecht des Haushaltsvorstands'.
  VALUE LABELS geschl_hhv '0' 'männlicher Haushaltsvorstand' '1' 'weiblicher
  Haushaltsvorstand'.
  VARIABLE LABELS alter_kl_hhv 'Altersklasse des Haushaltsvorstands (HHV)'.
  VALUE LABELS alter_kl_hhv 1 'Unter 25 Jahren (HHV)' 2 '25 bis unter 34
  Jahren (HHV)' 3 ' 35 bis unter 44 Jahren (HHV)' 4 '45 bis unter 55 Jahren
  (HHV)' 5 ' 55 bis unter 64 Jahren (HHV)' 6 '65 Jahre und älter (HHV)'.
```

```
*Berechnung neuer Variablen: Vermögens-Einkommens-, Verbrauchs-
Einkommensrelation sowie Pro-Kopf-Einkommen.
COMPUTE verm_eink_rel_hh = 0.
IF (eink_hh > 0) verm_eink_rel_hh = verm_hh/eink_hh.
VARIABLE LABELS verm_eink_rel_hh 'Vermögens-Einkommensrelation'.
COMPUTE verbr_eink_rel_hh = 0.
IF (eink_hh > 0) verbr_eink_rel_hh = verbr_hh/eink_hh.
VARIABLE LABELS verbr_eink_rel_hh 'Verbrauchs-Einkommensrelation'.
COMPUTE eink_hh_pk = eink_hh/hrf.
VARIABLE LABELS eink_hh_pk 'Haushalts-Pro-Kopf-Einkommen'.

*Berechnung des durchschnittlichen Haushalts-Pro-Kopf-Einkommens über alle
Haushalte hinweg, wobei MODE-Einstellung
*dazu führt, dass der Durchschnittswert allen Haushalten zugeordnet wird.
AGGREGATE
   /OUTFILE=* MODE=ADDVARIABLES
   /BREAK=
   /eink_mean = MEAN(eink_hh_pk).

*Temporäres Filtern für die Gruppe derjenigen Haushalte, deren Pro-Kopf-
Einkommen niedriger als die Hälfte des Durchschnitts-Pro-Kopf-Einkommens
liegt.
TEMPORARY.
SELECT IF (eink_hh_pk < 0.5 * eink_mean).

*Berechnung ausgewählter Statistiken (Median, Durchschnitt) für die
ausgefilterte Gruppe, einmal gewichtet (mit hrf) und einmal ungewichtet.
WEIGHT BY hrf.
TITLE 'Einkommensschwache Haushalte, gewichtet: Statistiken'.
FREQUENCIES VARIABLES = eink_hh_pk verm_hh verbr_hh verm_eink_rel_hh
verbr_eink_rel_hh
/FORMAT= NOTABLE
/STATISTICS = MEDIAN MEAN.
TEMPORARY.
SELECT IF (eink_hh_pk < 0.5 * eink_mean).
WEIGHT OFF.
TITLE 'Einkommensschwache Haushalte, ungewichtet: Statistiken'.
FREQUENCIES VARIABLES = eink_hh_pk verm_hh verbr_hh verm_eink_rel_hh
verbr_eink_rel_hh
/FORMAT= NOTABLE
/STATISTICS = MEDIAN MEAN.

*Aufteilen des Datensatzes nach dem Alter des Haushaltsvorstands.
SORT CASES BY alter_kl_hhv.
SPLIT FILE BY alter_kl_hhv.

*Berechnung ausgewählter Statistiken (Median, Durchschnitt) für die
```

```
  aufgeteilten Altersgruppen (Haushaltsvorstand), einmal gewichtet (mit hrf) und
einmal ungewichtet.
  WEIGHT BY hrf.
  TITLE 'Median, Durchschnitt nach Altersgruppen, gewichtet'.
  FREQUENCIES VARIABLES = eink_hh_pk verm_eink_rel_hh verbr_eink_rel_hh
  /FORMAT= NOTABLE
  /STATISTICS = MEDIAN MEAN.
  WEIGHT OFF.
  TITLE 'Median, Durchschnitt nach Altersgruppen, ungewichtet'.
  FREQUENCIES VARIABLES = eink_hh_pk verm_eink_rel_hh verbr_eink_rel_hh
  /FORMAT= NOTABLE
  /STATISTICS = MEDIAN MEAN.

  *Abspeichern der haushaltsbezogenen Gesamtdatei unter neuem Namen.
  SAVE OUTFILE = 'C:\SPSS-DATEIEN\HH_ALLE2.SAV'.

  *Löschen der Arbeitsdatei HH_ALLE.SAV, um beim erneuten Durchführen des
  Syntaxfiles keine Komplikationen mit dem Dateinamen zu erhalten.
  ERASE FILE = 'C:\SPSS-DATEIEN\HH_ALLE.SAV'.
```

# *Auswertungsergebnisse*

Damit Sie Ihre Auswertungsergebnisse mit meinen vergleichen können (aber gewiss dieselben Ergebnisse wie ich haben werden), habe ich in Abbildung 5.60 bis Abbildung 5.63 sämtliche Berechnungsergebnisse aus dem Syntaxprogramm *BEISPIEL1.SPS* zusammengestellt, die Sie nach der Ausführung des Syntaxfiles über AUSFÜHREN|ALLE erhalten.

**Geschlecht eines Haushaltsmitglieds**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | männliches Haushaltsmitglied | 10 | 50,0 | 50,0 | 50,0 |
| | weibliches Haushaltsmitglied | 10 | 50,0 | 50,0 | 100,0 |
| | Gesamt | 20 | 100,0 | 100,0 | |

**Altersklassen: Individuelles Alter**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | Unter 25 Jahren | 7 | 35,0 | 35,0 | 35,0 |
| | 25 bis unter 34 Jahren | 2 | 10,0 | 10,0 | 45,0 |
| | 35 bis unter 44 Jahren | 4 | 20,0 | 20,0 | 65,0 |
| | 45 bis unter 55 Jahren | 4 | 20,0 | 20,0 | 85,0 |
| | 55 bis unter 64 Jahren | 1 | 5,0 | 5,0 | 90,0 |

| | | | | | |
|---|---|---|---|---|---|
| | 65 Jahre und älter | 2 | 10,0 | 10,0 | 100,0 |
| | Gesamt | 20 | 100,0 | 100,0 | |

**Einkommensklassen: Individuelles Einkommen**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | Unter 1.000 Euro | 11 | 55,0 | 55,0 | 55,0 |
| | 1.000 bis unter 2.000 Euro | 5 | 25,0 | 25,0 | 80,0 |
| | 2.000 bis unter 3.000 Euro | 2 | 10,0 | 10,0 | 90,0 |
| | 3.000 Euro und mehr | 2 | 10,0 | 10,0 | 100,0 |
| | Gesamt | 20 | 100,0 | 100,0 | |

**Vermögensklassen: Individuelles Vermögen**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | Unter 10.000 Euro | 10 | 50,0 | 50,0 | 50,0 |
| | 10.000 bis unter 20.000 Euro | 3 | 15,0 | 15,0 | 65,0 |
| | 20.000 bis unter 30.000 Euro | 2 | 10,0 | 10,0 | 75,0 |
| | 30.000 Euro und mehr | 5 | 25,0 | 25,0 | 100,0 |
| | Gesamt | 20 | 100,0 | 100,0 | |

**Verbrauchsklassen: Individueller Verbrauch**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | Unter 750 Euro | 12 | 60,0 | 60,0 | 60,0 |
| | 750 bis unter 1.500 Euro | 5 | 25,0 | 25,0 | 85,0 |
| | 1.500 bis unter 2.250 Euro | 3 | 15,0 | 15,0 | 100,0 |
| | Gesamt | 20 | 100,0 | 100,0 | |

**Abbildung 5.60:** Personenbezogene FREQUENCIES-Ergebnisse in Form von Häufigkeitstabellen im vorderen Teil von *BEISPIEL1.SPS* (für *PERS_ALLE.SAV*)

Abbildung 5.60 beinhaltet Häufigkeitstabellen für bestimmte Variablen, wie Sie sich aus den FREQUENCIES-Anweisungen im vorderen Teil des Programms auf der Personenebene ergeben. Bezüglich der Variablen »Geschlecht eines Haushaltsmitglieds« sehen Sie, dass in der personenbezogenen Gesamtdatei genau die Hälfte aller Personen jeweils männlich beziehungsweise weiblich ist. Die in der personenbezogenen Gesamtdatei abgebildete Alters(klassen)verteilung zeigt für das Beispiel den höchsten Häufigkeitswert bei den unter 25-Jährigen mit 35,0 Prozent auf. Bezüglich der Einkommens- und Vermögensverteilung sind im Beispiel mehr als die Hälfte (55,0 Prozent; beim Einkommen)

beziehungsweise genau die Hälfte der Personen (beim Vermögen) mit ihrem individuellen Einkommen beziehungsweise mit ihrem individuellen Vermögen in der jeweils untersten Klasse (beim Einkommen: unter 1.000 Euro; beim Vermögen: unter 10.000 Euro) »verortet«.

Wichtig ist zu betonen, dass bei den folgenden Häufigkeitstabellen keine Gewichtungsvariable im Sinne einer Hochrechnungsvariablen verwendet wurde. Theoretisch denkbar wäre zum Beispiel eine Gewichtungsvariable, die die Stichprobengröße der Gesamtdatendatei (20 Personen) auf die deutsche Gesamtbevölkerung (gut 80 Millionen Personen) hochrechnet.

Ebenfalls im vorderen Teil des Syntaxprogramms befinden sich die statistischen (FREQUENCIES-)Anweisungen zur Berechnung bestimmter statistischer Kennziffern. Sie beziehen sich demnach auf die personenbezogene Gesamtdatendatei *PERS_ALLE.SAV*. Abbildung 5.61 können Sie zum Beispiel die Durchschnittswerte verschiedener Variablen entnehmen: Beispielsweise beträgt das durchschnittliche individuelle Vermögen 37.360,00 Euro. Auch hier ist die Personenfallzahl nicht in irgendeiner Weise hochgerechnet worden (zum Beispiel auf die Bevölkerungsgröße Deutschlands).

**Statistiken**

| | | Individuelles Einkommen, Frauen/Mädchen: 100 Euro zusätzlich | Individuelles Vermögen | Individueller Verbrauch | Individuelles Sparen |
|---|---|---|---|---|---|
| N | Gültig | 20 | 20 | 20 | 20 |
| | Fehlend | 0 | 0 | 0 | 0 |
| Mittelwert | | 1165,0000 | 37360,0000 | 645,0000 | 470,0000 |
| Standardfehler des Mittelwerts | | 246,90452 | 24570,07422 | 117,92616 | 174,99624 |
| Median | | 850,0000 | 7500,0000 | 550,0000 | 200,0000 |
| Modus | | ,00 | ,00 | 300,00 | 200,00 |
| Std.-Abweichung | | 1104,19058 | 109880,71235 | 527,38182 | 782,60698 |
| Varianz | | 1219236,842 | 12073770947 | 278131,579 | 612473,684 |
| Schiefe | | 1,205 | 4,342 | 1,196 | ,358 |
| Standardfehler der Schiefe | | ,512 | ,512 | ,512 | ,512 |
| Kurtosis | | 1,235 | 19,171 | ,985 | -,932 |
| Standardfehler der Kurtosis | | ,992 | ,992 | ,992 | ,992 |
| Spannweite | | 4100,00 | 500000,00 | 1900,00 | 2650,00 |
| Minimum | | ,00 | ,00 | 100,00 | -650,00 |
| Maximum | | 4100,00 | 500000,00 | 2000,00 | 2000,00 |
| Summe | | 23300,00 | 747200,00 | 12900,00 | 9400,00 |

**Abbildung 5.61:** Personenbezogene FREQUENCIES-Ergebnisse in Form statistischer Kennziffern im vorderen Teil von *BEISPIEL1.SPS* (für *PERS_ALLE.SAV*)

Während sich die in Abbildung 5.60 und Abbildung 5.61 dargestellten Ergebnisse auf die (ungewichtete) Personenebene und damit auf die personenbezogene Gesamtdatendatei *PERS_ALLE.SAV* beziehen, sind die Berechnungsergebnisse der beiden folgenden Abbildungen aus der aggregierten haushaltsbezogenen Gesamtdatendatei *HH_ALLE2.SAV* hervorgegangen. Sie finden sich im zweiten Teil des Syntaxprogramms.

In diesem Zusammenhang sind in Abbildung 5.62 die Fälle berücksichtigt, deren Pro-Kopf-Haushaltseinkommen weniger als die Hälfte des durchschnittlichen Pro-Kopf-Haushaltseinkommens insgesamt beträgt (im Beispiel: 0,5 · 1.150 Euro = 575 Euro). Unter Anwendung des Gewichtungsfaktors *hrf* (also der Haushaltsgröße) werden – wie aus dem oberen Teil von Abbildung 5.62 hervorgeht – acht Personen berücksichtigt (»herausgefiltert«), deren durchschnittliches

Pro-Kopf-Haushaltseinkommen – um einfach mal einen Wert aus Abbildung 5.62 auszuwählen – 462,50 Euro/Person beträgt. Demgegenüber finden sich im unteren Teil die ungewichteten Angaben auf der Haushaltsebene für insgesamt drei Haushalte. Das hier berechnete durchschnittliche Pro-Kopf-Haushaltseinkommen für diese drei Haushalte beläuft sich, wie ersichtlich, auf 466,67 Euro/Haushalt.

**Statistiken**

| | | Haushalts-Pro-Kopf-Einkommen | Haushaltsvermögen | Haushaltsverbrauch | Vermögens-Einkommensrelation | Verbrauchs-Einkommensrelation |
|---|---|---|---|---|---|---|
| N | Gültig | 8 | 8 | 8 | 8 | 8 |
| | Fehlend | 0 | 0 | 0 | 0 | 0 |
| Mittelwert | | 462,5000 | 4475,0000 | 1037,5000 | 2,8625 | ,7250 |
| Median | | 500,0000 | 5100,0000 | 1100,0000 | 2,6000 | ,5750 |

**Statistiken**

| | | Haushalts-Pro-Kopf-Einkommen | Haushaltsvermögen | Haushaltsverbrauch | Vermögens-Einkommensrelation | Verbrauchs-Einkommensrelation |
|---|---|---|---|---|---|---|
| N | Gültig | 3 | 3 | 3 | 3 | 3 |
| | Fehlend | 0 | 0 | 0 | 0 | 0 |
| Mittelwert | | 466,6667 | 3400,0000 | 866,6667 | 2,2556 | ,7167 |
| Median | | 500,0000 | 5000,0000 | 1100,0000 | 2,6000 | ,6000 |

**Abbildung 5.62:** Mit SELECT IF selektierte »einkommensschwache Haushalte« mit statistischen Kennziffern (FREQUENCIES) im hinteren Teil von *BEISPIEL1.SPS*, gewichtet (oben) und ungewichtet (unten) (für *HH_ALLE2.SAV*)

Abbildung 5.63 schließlich zeigt Ihnen für alle Personen (linker Teil der Abbildung) beziehungsweise für alle Haushalte (rechter Teil der Abbildung) Zusammenhänge zwischen dem Alter des Haushaltsvorstands und bestimmten Wohlstandsvariablen auf. Der linke Teil der Abbildung ist demnach über die Gewichtungsvariable *hrf* erstellt worden und bezieht sich auf 20 Personen. Im Unterschied hierzu ist beim rechten Teil der Abbildung keine (Personen-)Gewichtung durchgeführt worden. Folglich liegen den Ergebnissen im rechten Teil der Abbildung neun Haushalte zugrunde.

Um Ihnen mal kurz darzulegen, wie Abbildung 5.63 zu interpretieren ist, hier zunächst zwei Beispiele in Bezug auf den linken Abbildungsteil: In

Haushalten mit einem unter 25-jährigen Haushaltsvorstand lebt insgesamt eine Person, deren durchschnittliches Pro-Kopf-Haushaltseinkommen 500,00 Euro/Person beträgt; in Haushalten mit einem 25- bis 34-jährigen Haushaltsvorstand gibt es insgesamt vier Personen mit einem durchschnittlichen Pro-Kopf-Haushaltseinkommen in Höhe von 600,00 Euro/Person.

**Statistiken**

| Altersklasse des Haushaltsvorstands (HHV) | | | Haushalts-Pro-Kopf-Einkommen | Vermögens-Einkommensrelation | Verbrauchs-Einkommensrelation |
|---|---|---|---|---|---|
| Unter 25 Jahren (HHV) | N | Gültig | 1 | 1 | 1 |
| | | Fehlend | 0 | 0 | 0 |
| | Mittelwert | | 500,0000 | ,0000 | ,6000 |
| | Median | | 500,0000 | ,0000 | ,6000 |
| 25 bis unter 34 Jahren (HHV) | N | Gültig | 4 | 4 | 4 |
| | | Fehlend | 0 | 0 | 0 |
| | Mittelwert | | 600,0000 | 4,1667 | ,8021 |
| | Median | | 400,0000 | 4,1667 | 1,0000 |
| 35 bis unter 44 Jahren (HHV) | N | Gültig | 6 | 6 | 6 |
| | | Fehlend | 0 | 0 | 0 |
| | Mittelwert | | 1150,0000 | 39,1483 | ,6048 |
| | Median | | 500,0000 | 2,6000 | ,5500 |
| 45 bis unter 55 Jahren (HHV) | N | Gültig | 4 | 4 | 4 |
| | | Fehlend | 0 | 0 | 0 |
| | Mittelwert | | 1400,0000 | 15,9524 | ,4952 |
| | Median | | 1400,0000 | 15,9524 | ,4952 |
| 55 bis unter 64 Jahren (HHV) | N | Gültig | 3 | 3 | 3 |
| | | Fehlend | 0 | 0 | 0 |
| | Mittelwert | | 1900,0000 | 6,4912 | ,5877 |
| | Median | | 1900,0000 | 6,4912 | ,5877 |
| 65 Jahre und älter (HHV) | N | Gültig | 2 | 2 | 2 |
| | | Fehlend | 0 | 0 | 0 |
| | Mittelwert | | 600,0000 | 50,0000 | ,6667 |
| | Median | | 600,0000 | 50,0000 | ,6667 |

**Statistiken**

| Altersklasse des Haushaltsvorstands (HHV) | | | Haushalts-Pro-Kopf-Einkommen | Vermögens-Einkommensrelation | Verbrauchs-Einkommensrelation |
|---|---|---|---|---|---|
| Unter 25 Jahren (HHV) | N | Gültig | 1 | 1 | 1 |
| | | Fehlend | 0 | 0 | 0 |
| | Mittelwert | | 500,0000 | ,0000 | ,6000 |
| | Median | | 500,0000 | ,0000 | ,6000 |
| 25 bis unter 34 Jahren (HHV) | N | Gültig | 2 | 2 | 2 |
| | | Fehlend | 0 | 0 | 0 |
| | Mittelwert | | 800,0000 | 4,1667 | ,6042 |
| | Median | | 800,0000 | 4,1667 | ,6042 |
| 35 bis unter 44 Jahren (HHV) | N | Gültig | 2 | 2 | 2 |
| | | Fehlend | 0 | 0 | 0 |
| | Mittelwert | | 1475,0000 | 57,4224 | ,6321 |
| | Median | | 1475,0000 | 57,4224 | ,6321 |
| 45 bis unter 55 Jahren (HHV) | N | Gültig | 2 | 2 | 2 |
| | | Fehlend | 0 | 0 | 0 |
| | Mittelwert | | 1400,0000 | 15,9524 | ,4952 |
| | Median | | 1400,0000 | 15,9524 | ,4952 |
| 55 bis unter 64 Jahren (HHV) | N | Gültig | 1 | 1 | 1 |
| | | Fehlend | 0 | 0 | 0 |
| | Mittelwert | | 1900,0000 | 6,4912 | ,5877 |
| | Median | | 1900,0000 | 6,4912 | ,5877 |
| 65 Jahre und älter (HHV) | N | Gültig | 1 | 1 | 1 |
| | | Fehlend | 0 | 0 | 0 |
| | Mittelwert | | 600,0000 | 50,0000 | ,6667 |
| | Median | | 600,0000 | 50,0000 | ,6667 |

**Abbildung 5.63:** Mit SPLIT BY aufgeteilte Datei nach Altersgruppen, das heißt: nach dem Alter des Haushaltsvorstands mit statistischen Kennziffern auf der Haushaltsebene (FREQUENCIES) im hinteren Teil von *BEISPIEL1.SPS*, gewichtet (links) und ungewichtet (rechts) (für *HH_ALLE2.SAV*)

So, das waren Interpretationsbeispiele für den linken Abbildungsteil – hier kommen jetzt auch noch ungefragt entsprechende Interpretationsbeispiele für den rechten Abbildungsteil, der sich auf die Haushaltsanzahl bezieht und dem keine Variablengewichtung zugrunde liegt: Bei dem einen Haushalt mit einem unter 25-jährigen Haushaltsvorstand (einem Einpersonenhaushalt) beläuft sich das durchschnittliche Pro-Kopf-Haushaltseinkommen auf 500,00 Euro/Haushalt, und bei den zwei Haushalten mit einem 25- bis 34-jährigen Haushaltsvorstand beträgt dieser Durchschnitt 800,00 Euro/Haushalt.

# *Ausführen eines SPSS-Syntaxprogramms*

Wenn Sie ein Syntaxprogramm geschrieben haben, müssen Sie es nur noch zum »Laufen« bringen. Wie oben schon recht oft vorweggenommen (damit Sie schon einmal ein paar Ergebnisse gesehen und hier nicht nur Theorie betrieben haben!), geschieht dies im SPSS-Syntaxeditor einfach dadurch, dass Sie in der Befehlsleiste auf AUSFÜHREN gehen und dann auf ALLE klicken. Alternativ können Sie auch die »Wiedergabetaste« drücken oder den Shortcut Strg + R auswählen.

Wenn Sie nicht nur solche Mini-Datensätze wie hier, sondern Datensätze mit zig Tausend Fällen und mehreren Hundert Variablen zu verarbeiten haben, können die Laufzeiten Ihrer entsprechenden Programme mehrere Minuten betragen. Eventuell dauert Ihnen das aber zu lange (weil Sie vielleicht ein wichtiges privates Date vor der Brust haben), oder Sie haben während der Ausführung Ihres Programms einen Syntaxfehler festgestellt (Nobody is perfect!) und sich für die schnelle Beendigung der Programmausführung entschieden. In diesen Fällen können Sie im Menü über DATEI|PROZESSOR ANHALTEN die weitere Ausführung Ihres Programms stoppen und stattdessen zu Ihrem Date gehen oder Ihren Syntaxfehler beseitigen (und dann das Programm später nochmals starten).

Wenn Sie nur Teile Ihres Syntaxprogramms ausführen lassen möchten, müssen Sie diese Teile Ihres Syntaxfiles markieren (eventuell ab einer bestimmten, via Cursor markierten Stelle) und anschließend auf AUSFÜHREN|AUSWAHL klicken.

Greifen Sie doch einfach mal von unseren zahlreichen Beispiel-Syntaxfiles *SYNTAX20.SPS* heraus:

```
GET FILE = 'C:\SPSS-DATEIEN\AGGR.SAV'.
WEIGHT BY persnr_nu.
FREQUENCIES VARIABLES = beitrag_sum
/FORMAT = NOTABLE
/STATISTICS=ALL.
WEIGHT OFF.
FREQUENCIES VARIABLES = beitrag_sum
/FORMAT = NOTABLE
/STATISTICS=ALL.
```

Wenn Sie auf AUSFÜHREN|ALLE klicken (siehe Abbildung 5.64), wird das gesamte Syntaxprogramm durchlaufen, und in diesem Fall werden zwei statistische Übersichten als Output ausgewiesen – einmal gewichtet und einmal ungewichtet (siehe Abbildung 5.65).

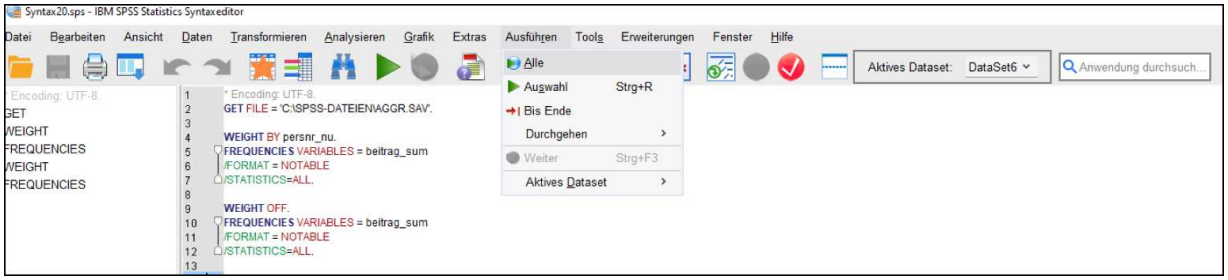



**Abbildung 5.64:** Syntax-Beispiel für AUSFÜHREN|ALLE (in Bezug auf *AGGR.SAV*)

**Statistiken**

BEITRAG_sum

| | | |
|---|---|---|
| N | Gültig | 15 |
| | Fehlend | 0 |
| Mittelwert | | 155,3333 |
| Standardfehler des Mittelwerts | | 22,75892 |
| Median | | 150,0000 |
| Modus | | 250,00 |
| Std.-Abweichung | | 88,14490 |
| Varianz | | 7769,524 |
| Schiefe | | -,103 |
| Standardfehler der Schiefe | | ,580 |
| Kurtosis | | -1,621 |
| Standardfehler der Kurtosis | | 1,121 |
| Spannweite | | 240,00 |
| Minimum | | 10,00 |
| Maximum | | 250,00 |
| Summe | | 2330,00 |

**Statistiken**

BEITRAG_sum

| | | |
|---|---|---|
| N | Gültig | 5 |
| | Fehlend | 0 |
| Mittelwert | | 108,0000 |
| Standardfehler des Mittelwerts | | 42,23742 |
| Median | | 80,0000 |
| Modus | | 10,00[a] |
| Std.-Abweichung | | 94,44575 |
| Varianz | | 8920,000 |
| Schiefe | | ,880 |
| Standardfehler der Schiefe | | ,913 |
| Kurtosis | | ,073 |
| Standardfehler der Kurtosis | | 2,000 |
| Spannweite | | 240,00 |
| Minimum | | 10,00 |
| Maximum | | 250,00 |
| Summe | | 540,00 |

a. Mehrere Modi vorhanden. Der kleinste Wert wird angezeigt.

**Abbildung 5.65:** Ausführlicher Output im Syntax-Beispiel bei AUSFÜHREN|ALLE (mit Gewichtung: linke Seite; ohne Gewichtung: rechte Seite) (in Bezug auf *AGGR.SAV*)

Markieren Sie hingegen nur ausgewählte Teile Ihres Syntaxprogramms und klicken anschließend auf AUSFÜHREN|AUSWAHL (siehe Abbildung 5.66), werden von SPSS auch nur die markierten Teile des Syntaxprogramms ausgeführt. Im vorliegenden Beispielfall mit *SYNTAX20.SPS* führt dies dazu, dass nur die erste im Programm enthaltene statistische Anweisung ausgeführt, das heißt, nur die erste statistische Übersicht als Output ausgewiesen wird (siehe Abbildung 5.67).

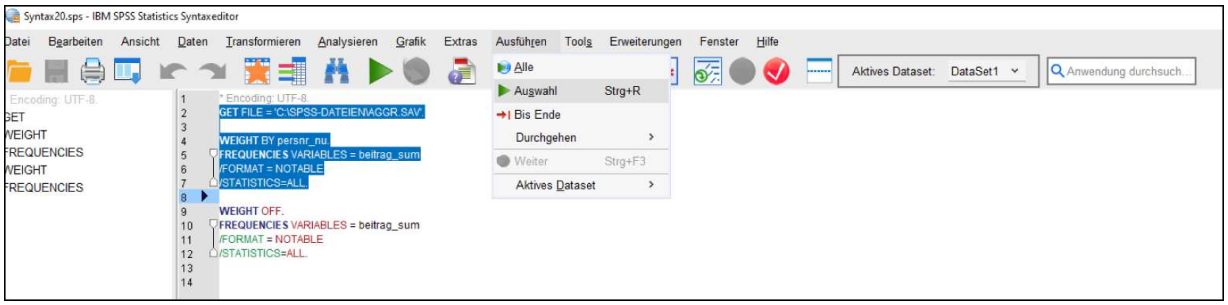



**Abbildung 5.66:** Syntax-Beispiel für AUSFÜHREN|AUSWAHL (in Bezug auf *AGGR.SAV*)

**Statistiken**

BEITRAG_sum

| | | |
|---|---|---|
| N | Gültig | 15 |
| | Fehlend | 0 |
| Mittelwert | | 155,3333 |
| Standardfehler des Mittelwerts | | 22,75892 |
| Median | | 150,0000 |
| Modus | | 250,00 |
| Std.-Abweichung | | 88,14490 |
| Varianz | | 7769,524 |
| Schiefe | | -,103 |
| Standardfehler der Schiefe | | ,580 |
| Kurtosis | | -1,621 |
| Standardfehler der Kurtosis | | 1,121 |
| Spannweite | | 240,00 |
| Minimum | | 10,00 |
| Maximum | | 250,00 |
| Summe | | 2330,00 |

**Abbildung 5.67:** Reduzierter Output im Syntax-Beispiel bei AUSFÜHREN|AUSWAHL (in Bezug auf *AGGR.SAV*)

# *Outputfiles*

In jedem Fall erhalten Sie in den vorstehenden Fällen Output, und damit sind Sie auch schon beim nächsten großen Thema: bei den Outputfiles. Die Ergebnisse von Syntaxfiles werden nämlich, sofern es sich um Endergebnisse handelt, in **Outputfiles** protokolliert.

An verschiedenen Stellen wurde Ihnen – gerade im vergangenen Abschnitt dieses Kapitels – bereits Output präsentiert. Im Folgenden möchte ich gerne die Outputthematik etwas umfassender vorstellen.

# *Viewer-Einstellungen*

Lassen Sie uns zunächst mit einer Bezeichnung beginnen: In SPSS werden die Ergebnisse in Tabellen- und Diagrammform (je nach SPSS-Grundeinstellung: mit oder ohne zugrunde liegender Syntax) in einem **Ausgabefenster** dargestellt. Dieses Ausgabefenster nennt man mit einem englischen Wort auch **Viewer**. Ein Ausgabefenster öffnet sich in SPSS erst, wenn Sie irgendeine datenbezogene Aktion durchgeführt haben. In der folgenden Abbildung 5.68 hat sich das Ausgabefenster deshalb geöffnet, weil eine Datendatei namens *123.SAV* geöffnet worden ist. (Der in diesem Fall übrigens auftretende Warnhinweis hat hier keine größere Bedeutung, und Sie können ihn einfach ignorieren.)

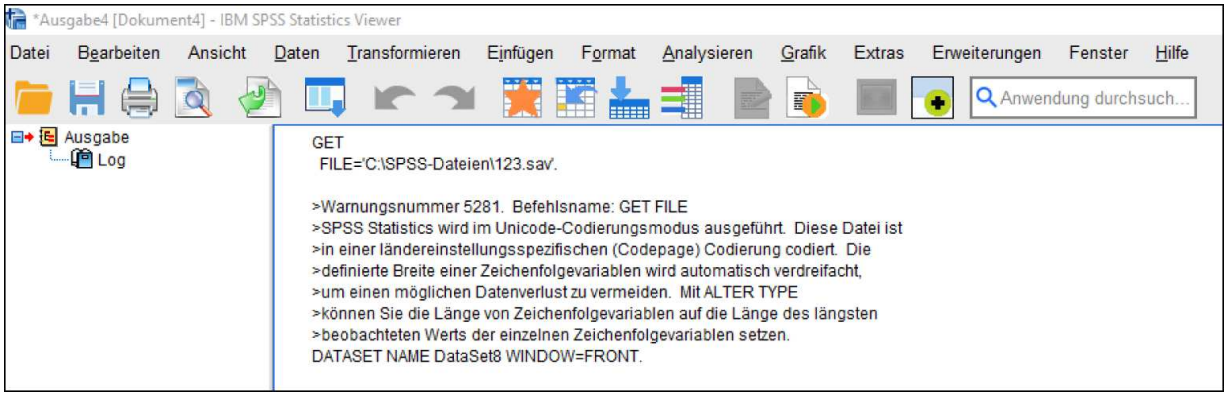



**Abbildung 5.68:** Öffnen eines Ausgabefensters nach Öffnen einer Datendatei (*123.SAV*)

Welcher Art die Anzeige im Viewer ist, können Sie im SPSS-Menü über BEARBEITEN| OPTIONEN|VIEWER steuern (siehe Abbildung 5.69).

Neben Formatierungsangaben zur Schriftart, -größe und so weiter auf der rechten Seite dieses Reiters in Bezug auf den Titel, den ausgegebenen Text und dergleichen finden Sie auf der linken Seite dieses Dialogfelds Möglichkeiten, bestimmte Anzeigen im Viewer an- und auszuschalten. Sollte Sie zum Beispiel das Protokollieren der einzelnen Syntaxbefehle im Viewer stören, können Sie dies dort abschalten.

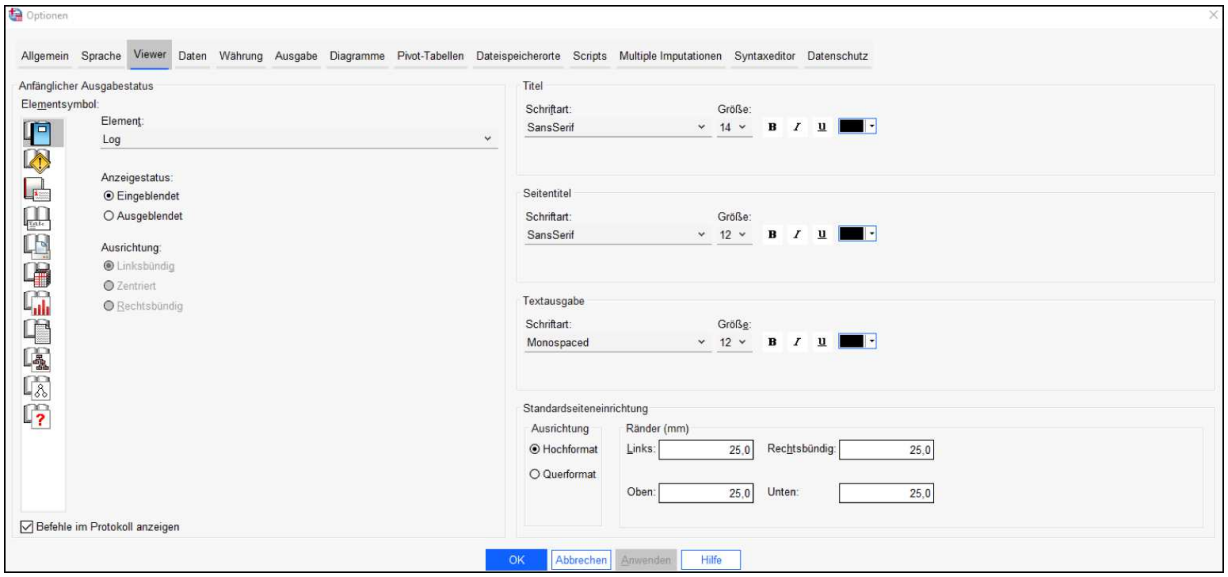



**Abbildung 5.69:** Viewer-Einstellungen

Um die Anzeige der Syntax zu unterdrücken, wählen Sie sie auf der linken Seite des Reiters BEARBEITEN|OPTIONEN|VIEWER unter ELEMENT »Log« (wie »Logbuch« im Sinne des Syntax-Protokolls) aus und klicken anschließend in der darunterliegenden Rubrik ANZEIGESTATUS auf AUSGEBLENDET (siehe Abbildung 5.70). Unter BEARBEITEN|OPTIONEN|VIEWER können Sie auch noch weitere Viewer-Elemente ein- beziehungsweise ausblenden, zum Beispiel die SPSS-Warnungen oder die SPSS-Hinweise.

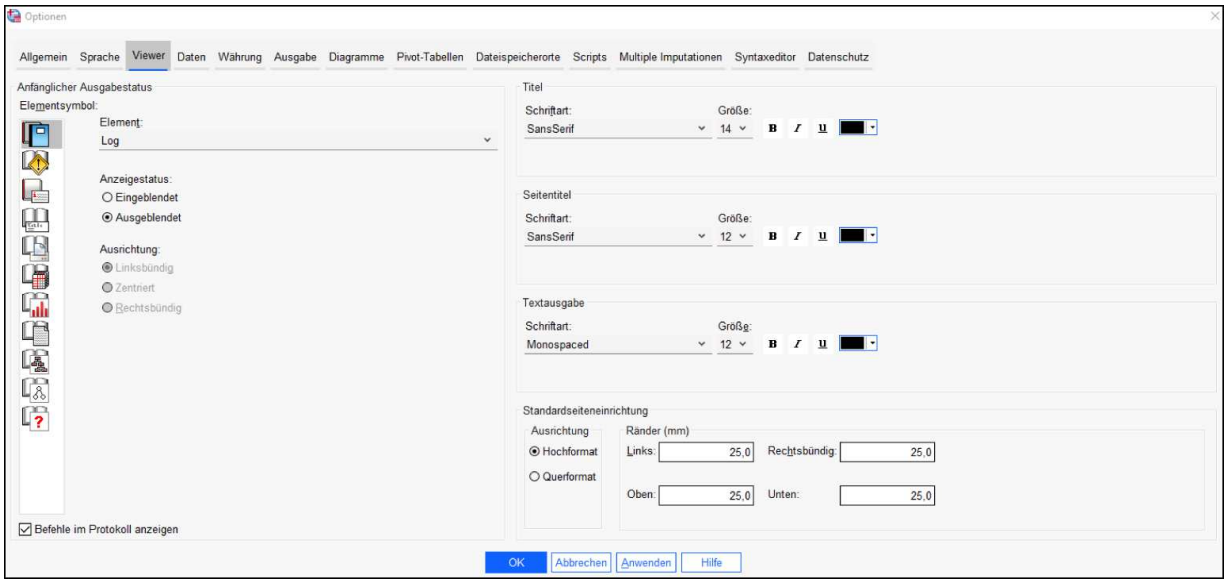



**Abbildung 5.70:** Ausschalten des Syntax-Protokolls

Als Folge der vorstehend erwähnten Festlegung »Ausblenden des Syntax-Protokolls« wird die Syntax im Outputmodus nicht mehr mit angegeben. Lassen Sie zum Beispiel von unseren zahlreichen Syntaxfiles *SYNTAX20.SPS* »laufen«, dann erhalten Sie mit der vorstehenden Festlegung einen syntaxfreien Output. Dies ist anhand des Beispiels mit *SYNTAX20.SPS* in Abbildung 5.71 veranschaulicht. Wenn Sie Ihren Output zu Präsentationszwecken verwenden möchten (etwa für ein Meeting), erscheint die Unterdrückung des Syntax-Protokolls durchaus sinnvoll. Möchten Sie hingegen anhand des Outputs etwaige Fehlerquellen in Ihrem Syntaxprogramm aufdecken, kann das Syntax-Protokoll hilfreich sein, weil Sie durch mögliche Warn-/Fehlerhinweise von SPSS auf derartige Fehler in Ihrer Syntax direkt hingewiesen werden.

Der Viewer ist links in den Navigationsbereich im Sinne einer Gliederungsansicht und rechts in den entscheidenden Inhaltsbereich aufgeteilt.

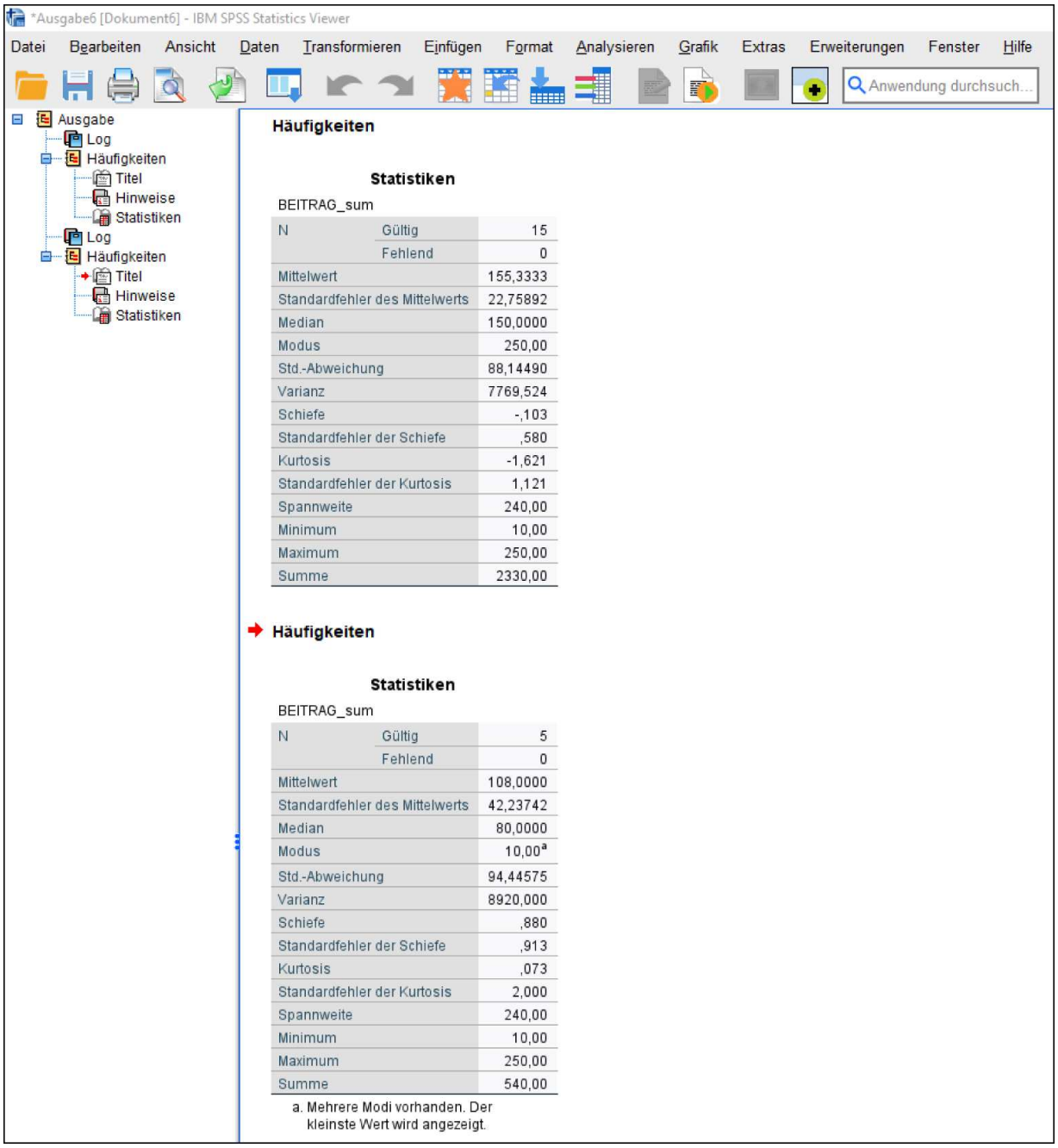



**Abbildung 5.71:** SPSS-Output nach Ausschalten des Syntax-Protokolls (in Bezug auf *AGGR.SAV*)

# *Navigationsbereich*

Wie bereits erwähnt, befindet sich der **Navigationsbereich** auf der li nken Seite eines Ausgabefensters. Um seine Elemente gut zu verstehen,

öffnen Sie bitte das Syntaxfile *SYNTAX21.SPS* und lassen es anschließend »laufen«:

```
GET FILE = 'C:\SPSS-DATEIEN\AGGR.SAV'.
FREQUENCIES VARIABLES = beitrag_sum
/FORMAT = NOTABLE
/STATISTICS=ALL.
DESCRIPTIVES VARIABLES = beitrag_sum
/STATISTICS=ALL.
```

Mittels zweier statistischer Prozeduren (FREQUENCIES und DESCRIPTIVES), auf die in den Kapiteln 6 und 7 noch ausführlicher eingegangen wird, werden in diesem Beispiel-Syntaxprogramm für die Variable *beitrag_sum* statistische Kennziffern ausgegeben. Sie haben also zwei **Ausgabeblöcke**: zum einen den FREQUENCIES- und zum anderen den DESCRIPTIVES-Ausgabeblock. In Abbildung 5.72 sind diese beiden Ausgabeblöcke auf der linken Seite des Ausgabefensters, im Navigationsbereich, mit dem deutschen Begriff »Häufigkeiten« überschrieben. Beiden Ausgabeblöcken ist im Navigationsbereich jeweils das zugehörige Syntaxprotokoll (Überschrift: »Log«) vorangestellt. Unterteilt sind beide Ausgabeblöcke jeweils in drei Bereiche:

- ✔ Titel,
- ✔ Hinweise und
- ✔ Statistiken.

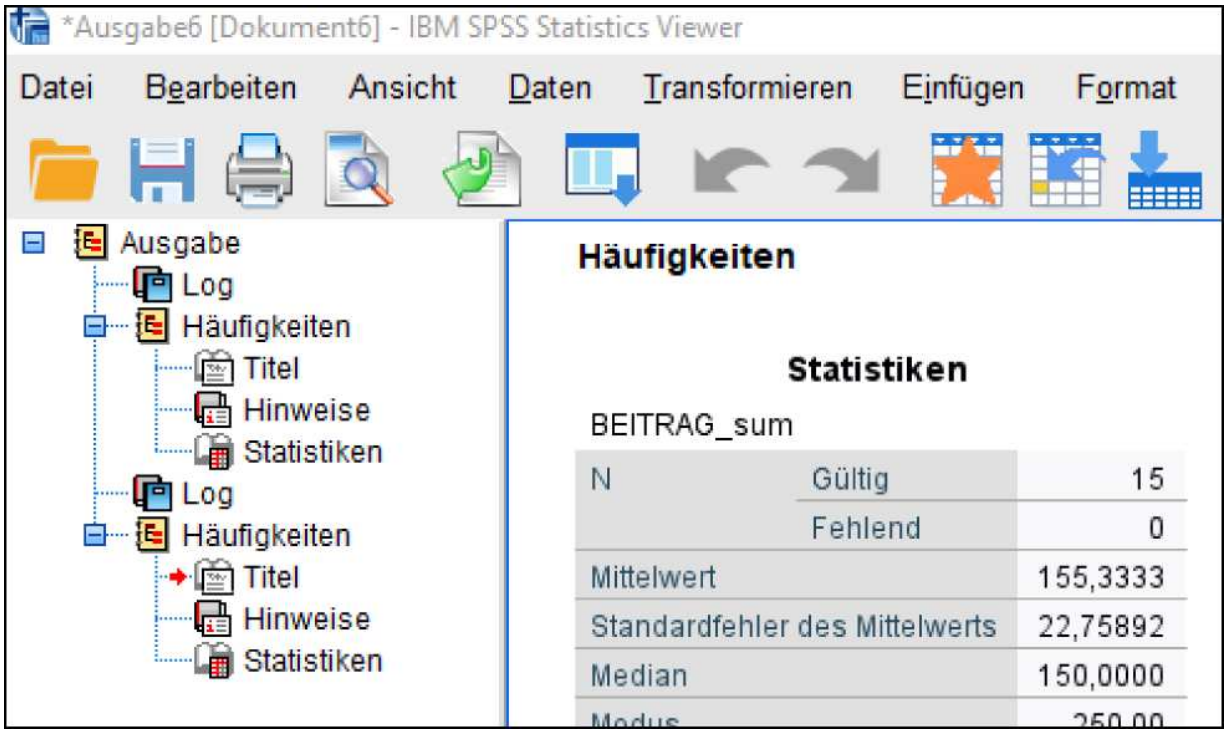



**Abbildung 5.72:** Der Navigationsbereich des Viewers (in Bezug auf *AGGR.SAV*)

Lassen Sie mich bitte mit Ihnen diese drei Bereiche für das vorstehende Beispiel, konkret für die FREQUENCIES-Anweisung (also für den ersten der Häufigkeiten-Ausgabeblöcke), kurz durchgehen. Klicken Sie deshalb im Navigationsbereich zunächst einmal auf »Titel« im Häufigkeiten-Ausgabeblock. Wie Abbildung 5.73 veranschaulicht, ist hiermit die Überschrift des gesamten Ausgabeblocks gemeint (= »Häufigkeiten«). Dies erkennen Sie daran, dass SPSS rechts im Ausgabefenster (im Inhaltsbereich) diese Überschrift umrahmt. Sie könnten sie nun im Inhaltsbereich manuell verändern.

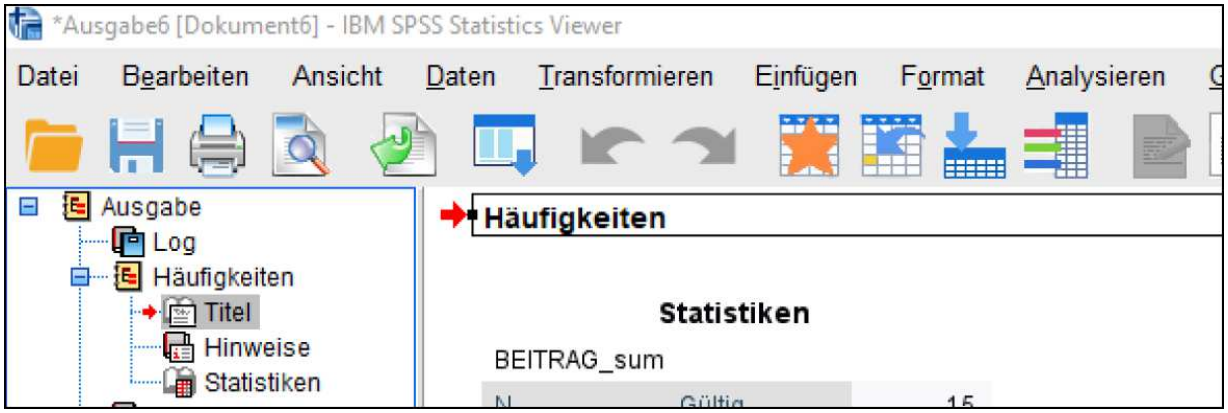


**Abbildung 5.73:** Rubrik »Titel« im Navigationsbereich des Viewers (in Bezug auf *AGGR.SAV*)

Wenn Sie im Navigationsbereich nunmehr »Hinweise« (beim ersten »Häufigkeiten«-Ausgabenblock) anklicken, öffnet sich im Inhaltsbereich ein Rahmen, in dem verschiedene Hintergrundinformationen zum hier relevanten Ausgabeblock zu finden sind. Diese Informationen beziehen sich etwa auf die Art der verwendeten Datendatei (in unserem Beispiel finden Sie hier die Information »Aggregated File«), auf eventuell gesetzte Filter, auf etwaige Datengewichtungen und so weiter (siehe Abbildung 5.74).

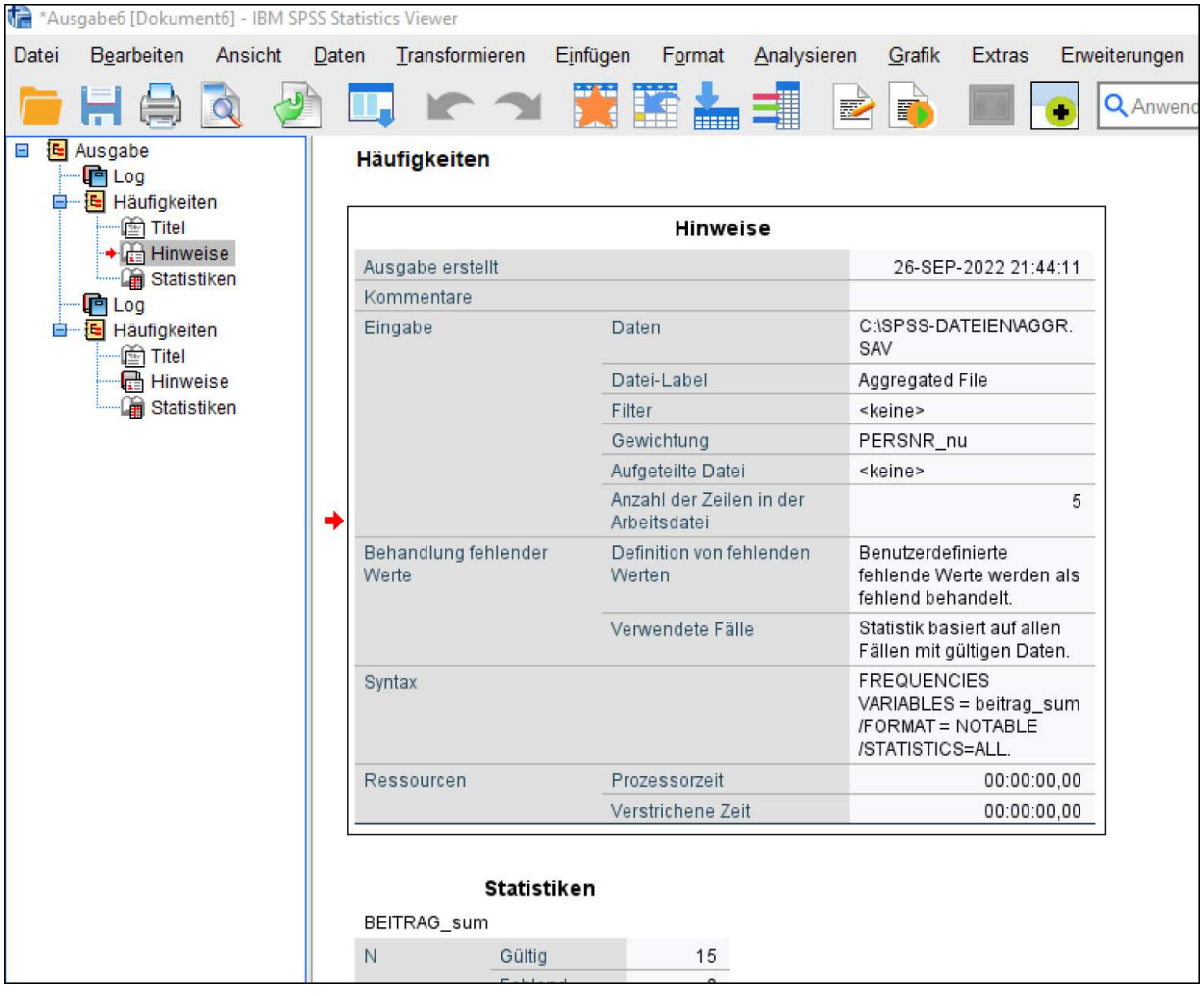



**Abbildung 5.74:** Rubrik »Hinweise« im Navigationsbereich des Viewers (in Bezug auf *AGGR.SAV*)

Schließlich klicken Sie bitte im Navigationsbereich beim ersten Ausgabeblock »Häufigkeiten« auch noch auf die dritte Rubrik »Statistiken«. Offenkundig (siehe Abbildung 5.75) bezieht sich diese Rubrik auf die Ergebnisse der entsprechenden (FREQUENCIES-)Statistikprozedur. Dies geht aus der betreffenden Umrahmung im Inhaltsbereich hervor.

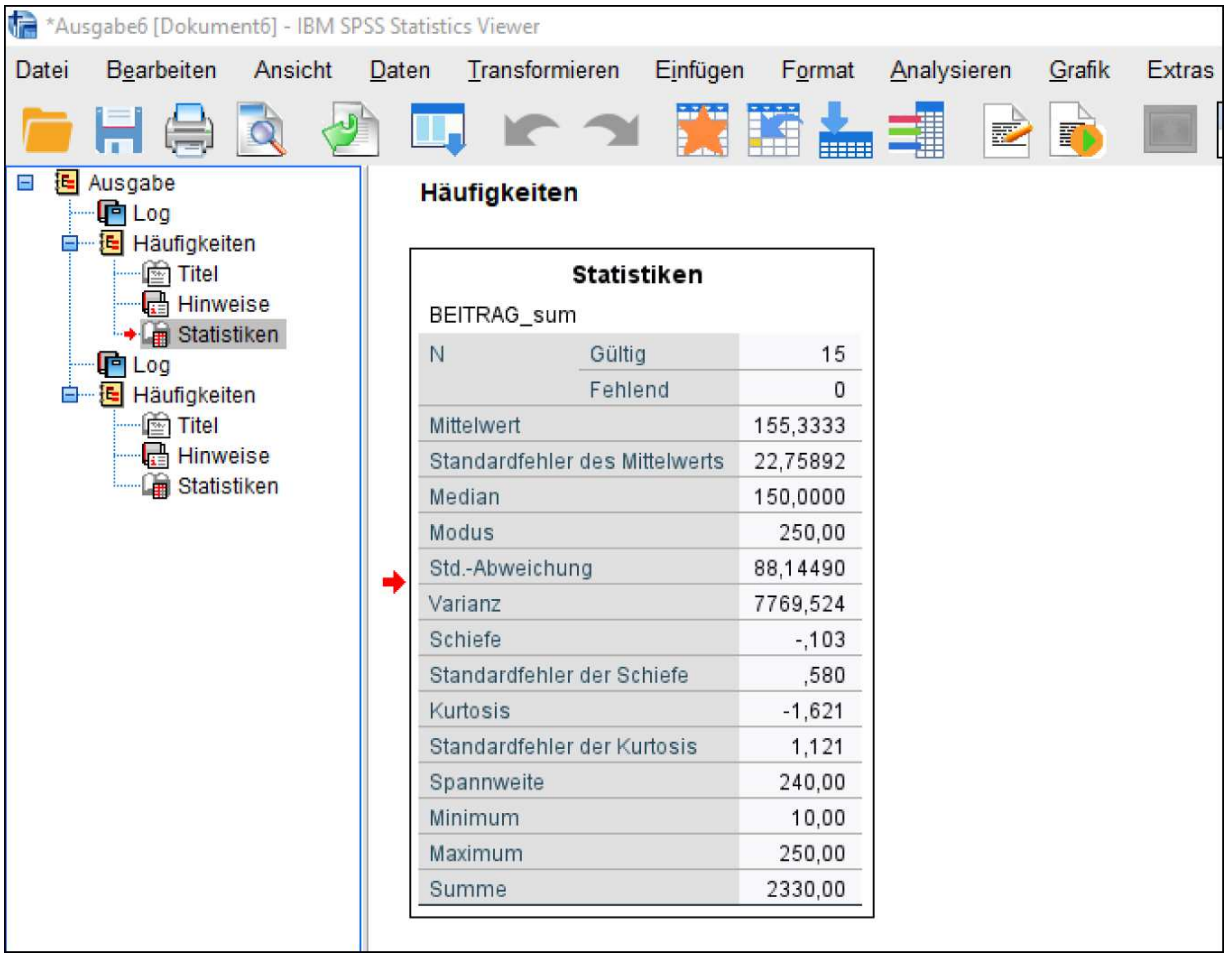



**Abbildung 5.75:** Eigentliche Auswertungsrubrik »Statistiken« im Navigationsbereich des Viewers (in Bezug auf *AGGR.SAV*)

Sie können innerhalb des Navigationsbereichs die einzelnen Elemente in der Ihnen (sicherlich) vertrauten (Windows®-)Weise

- ✔ markieren,
- ✔ ausschneiden,
- ✔ verschieben oder

- ✔ kopieren.

Dies ist in Abbildung 5.76 angedeutet, in der der Ausgabebereich »Statistiken« markiert worden ist. Anschließend ist mit der rechten Maustaste für diesen Bereich ein Fenster geöffnet worden, aus dem die vorstehend genannten Optionen des Ausschneidens, Kopierens dieses Bereichs und so weiter hervorgehen.

Im Wesentlichen dient der (im Übrigen auch Gliederungsbereich genannte) Navigationsbereich, wie der Name schon aussagt, in erster Linie dem schnellen Navigieren bezüglich der Elemente des Inhaltsbereichs. Dabei wird durch einen kleinen roten Pfeil innerhalb des Navigationsbereichs auf die zugeordnete Ausgabe verwiesen. Es ist leicht möglich, mittels dieses kleinen Pfeils den Fokus auf einen anderen Ausgabebereich zu lenken.

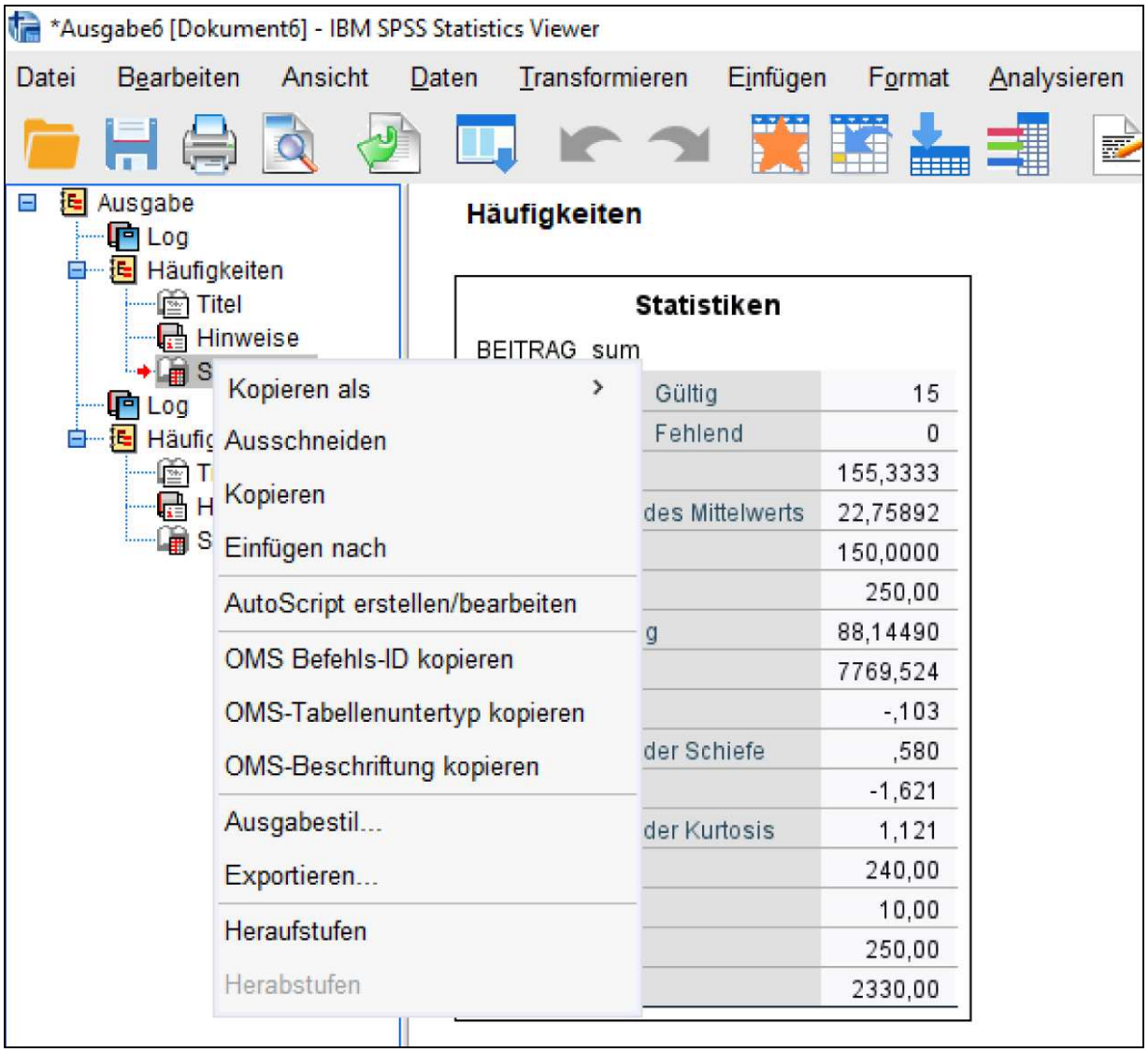



**Abbildung 5.76:** Optionen des Ausschneidens, Verschiebens oder Kopierens im Navigationsbereich

# *Inhaltsbereich*

Im **Inhaltsbereich** auf der rechten Viewer-Seite werden Ihre Berechnungsergebnisse angezeigt. Dies geschieht in Tabellenform oder in Form von Grafiken.

Darüber hinaus werden im Inhaltsbereich üblicherweise – bei entsprechender Einstellung – auch noch abgebildet:

- ✔ SPSS-Kommandos (inklusive Kommentare),

- ✔ Titel,
- ✔ Warnungen und
- ✔ Fehlermeldungen.

## *Warnungen*

Auf die beiden ersten beiden Punkte ist bereits im vorhergehenden Unterabschnitt eingegangen worden. **Warnungen** werden dann ausgegeben, wenn Syntaxvorschriften nicht eingehalten werden, dadurch aber das weitere Ausführen eines Syntaxprogramms nicht weiter beeinträchtigt wird.

Wenn Sie zum Beispiel das Syntaxfile *SYNTAX19A.SPS* mit der Zeile

```
TITLE 'Häufigkeitstabelle für die Beitragszahlungen – nach Geschlecht getrennt,
Faik-Beispieldatensatz, erstellt am 26.09.2022 von Jürgen Faik, der gerne lange
Titel kreiert und Spaß an Statistik hat, gerade an Häufigkeitstabellen, und auch
an ökonomischen Zusammenhängen interessiert ist'.
```

laden, werden Sie im Inhaltsbereich eine Warnung erhalten, dass der vorgesehene Titel mehr als die zulässigen 256 Zeichen umfasst. Der im Viewer ausgegebene Titel wird auf 256 Zeichen »gestutzt«. Mehr passiert aber nicht (siehe Abbildung 5.77).

**Abbildung 5.77:** Warnung im Viewer-Inhaltsbereich (in Bezug auf *PERSONEN4A.SAV*)

## *Fehlermeldungen*

Viel problematischer als Warnungen sind **Fehlermeldungen** im Viewer-Inhaltsbereich, weil hierdurch zumindest die betroffene Prozedur beeinträchtigt wird. Möglicherweise wird sogar das gesamte Syntaxprogramm gestoppt.

Betrachten Sie hierzu bitte das nachfolgende Syntaxprogramm (*SYNTAX22.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\CORRELATIONS.SAV'.
COMPUTE z1 = x1 + y1.
KOMPUTE z2 = z1 + x2.
FREQUENCIES VARIABLES = z1 z2
/FORMAT = NOTABLE
/STATISTICS = MEDIAN MEAN.
```

Ganz offenkundig ist der zweite COMPUTE-Befehl falsch geschrieben worden. Lassen Sie jetzt *SYNTAX22.SPS* »laufen«, ist zu erwarten, dass SPSS Ihnen deswegen eine Fehlermeldung (unter der Überschrift »Warnungen«) zukommen lässt. Was noch passiert, ist in Abbildung 5.78 dokumentiert.

Natürlich sind wir alle nur Menschen, und Irren ist bekanntlich menschlich. SPSS ist aber an dieser Stelle wenig nachsichtig mit uns, wie aus Abbildung 5.78 hervorgeht. SPSS stellt die falsche Schreibweise des zweiten COMPUTE-Befehls als Fehler fest und bestraft Sie für diesen Fehler insofern gnadenlos, als dass die statistischen Kennziffern Median und Arithmetisches Mittel weder für die (mit einem korrekten COMPUTE-Befehl definierte) Variable *z1* noch für die (mit einem falschen COMPUTE-Befehl definierte) Variable *z2* berechnet werden. Das ist hart – oder? Wenn Sie also Fehler vermeiden können, dann tun Sie das!

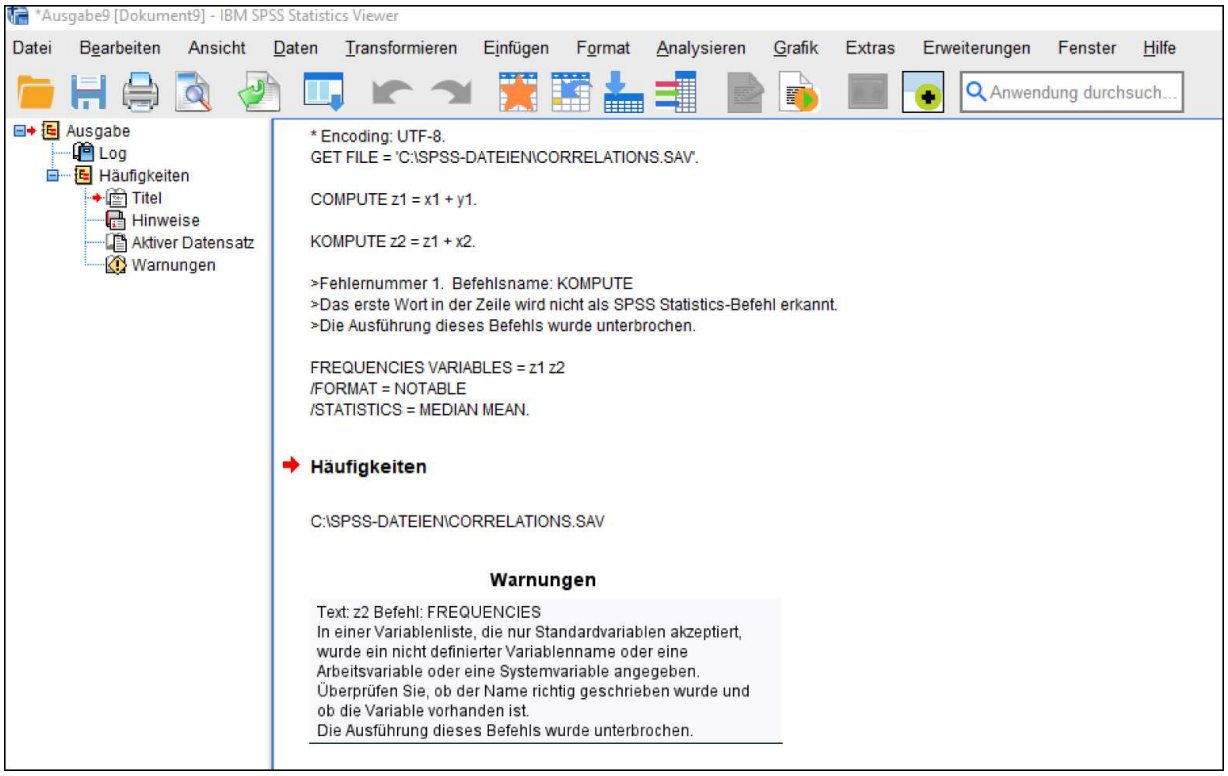



**Abbildung 5.78:** Fehlermeldung im Viewer-Inhaltsbereich (in Bezug auf *CORRELATIONS.SAV*)

## *Pivottabellen*

Ihre Ergebnisse werden in der Regel in Tabellen- oder in Diagrammform dargestellt. Die betreffenden Ergebnistabellen werden auch als **Pivottabellen** bezeichnet.

Lassen Sie bitte in diesem Zusammenhang *SYNTAX23.SPS* (jetzt mit zwei korrekten COMPUTE-Anweisungen – man ist ja lernfähig …) »laufen«, also folgendes Syntaxprogramm:

```
GET FILE = 'C:\SPSS-DATEIEN\CORRELATIONS.SAV'.
COMPUTE z1 = x1 + y1.
COMPUTE z2 = z1 + x2.
FREQUENCIES VARIABLES = z1 z2
/FORMAT = NOTABLE
/STATISTICS = MEDIAN MEAN.
```

Sie erhalten in diesem Fall als Ergebnis eine Pivottabelle, bei der in den Zeilen die verwendeten statistischen Kennziffern, in den Spalten die ausgewählten Variablen und in den Zellen die Ergebnisse stehen (siehe Abbildung 5.79).

Es kann nun sein, dass Ihnen diese Pivottabelle etwa von ihrem Design her nicht besonders gefällt. Dann können Sie sie durch einen Doppelklick auf die Tabelle verändern (siehe hierzu den in Abbildung 5.80 angezeigten Hinweis).

Wenn Sie in Ihrem nächsten Schritt auf die Pivottabelle doppelt klicken, können Sie diese Tabelle gewissermaßen nach eigenem Gutdünken umgestalten. Ich habe zum Beispiel die erste Variable umbenannt (von »*z1*« in »Erste Variable«) und zudem die Spaltenbreite für diese erste Variable durch Ziehen mit der Maus (massiv) erweitert (siehe Abbildung 5.81).

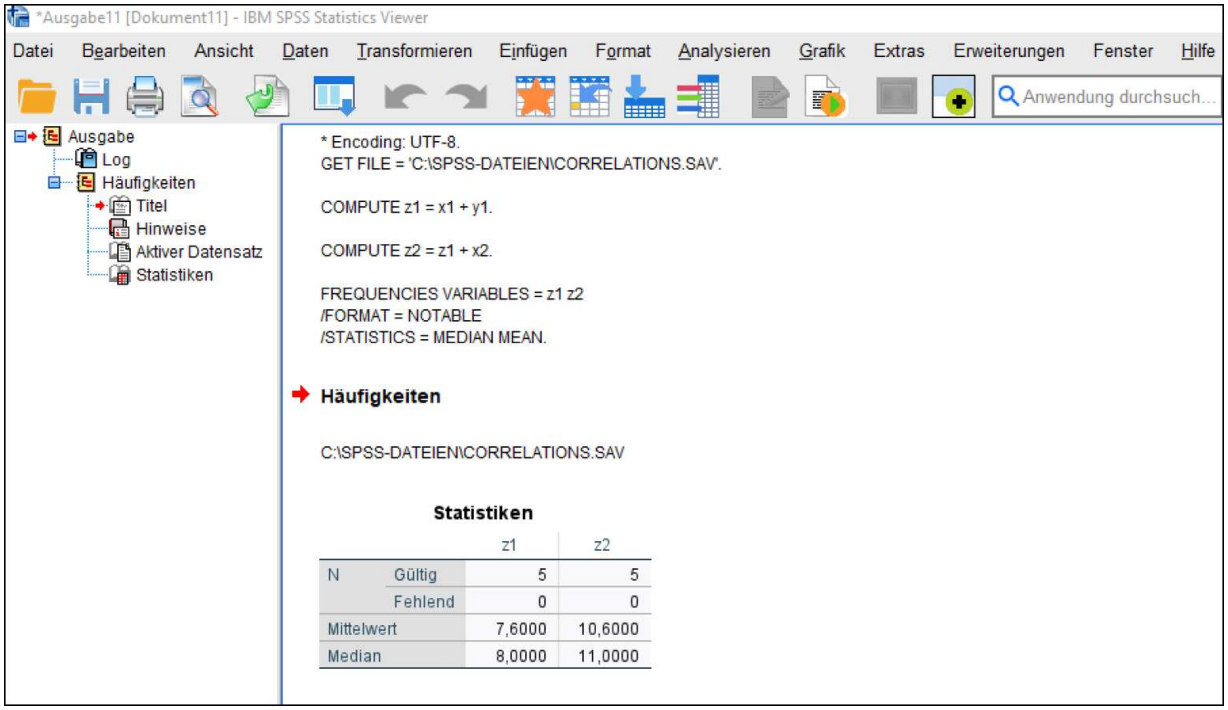



**Abbildung 5.79:** Pivottabelle im Viewer-Inhaltsbereich (in Bezug auf *CORRELATIONS.SAV*)

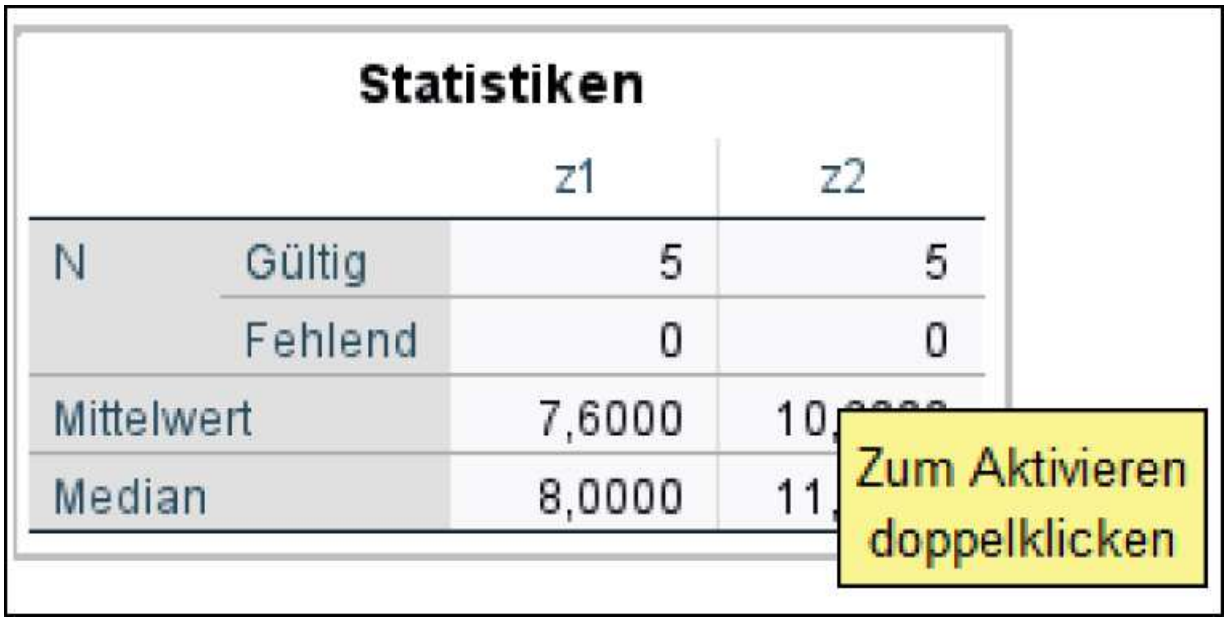



**Abbildung 5.80:** Doppelklick zum Verändern einer Pivottabelle im Viewer-Inhaltsbereich (hier in Bezug auf *CORRELATIONS.SAV*)

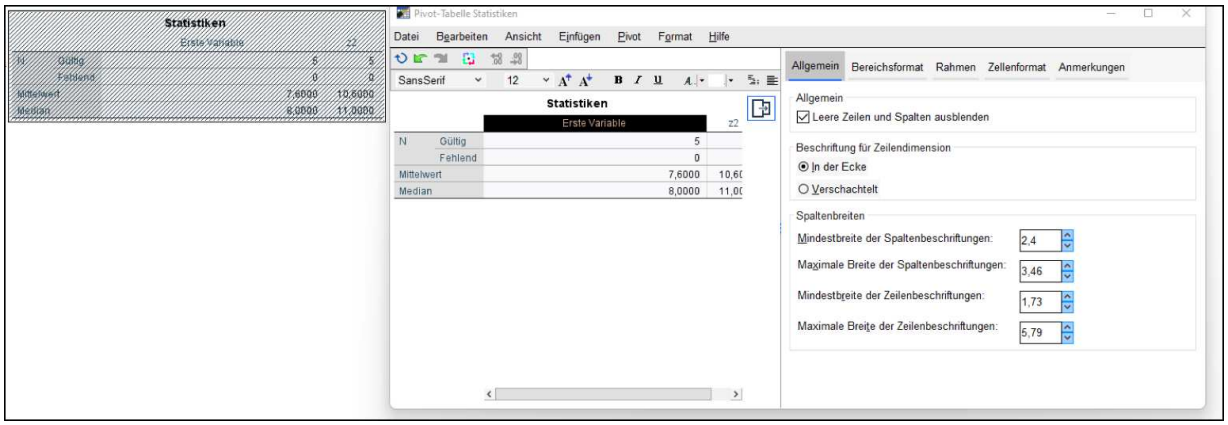



**Abbildung 5.81:** Veränderung einer Pivottabelle im Viewer-Inhaltsbereich (hier in Bezug auf *CORRELATIONS.SAV*)

Wenn Sie eine Pivottabelle einmal mit der rechten Maustaste anklicken, öffnet sich das in Abbildung 5.82 dargestellte Fenster. Sie können eine Pivottabelle entsprechend zum Beispiel ausschneiden beziehungsweise kopieren und dann anderweitig nutzen. Ich kopiere gerne SPSS-Pivottabellen nach EXCEL und verarbeite sie dort weiter (was ich in diesem Buch schon einmal erwähnt hatte).

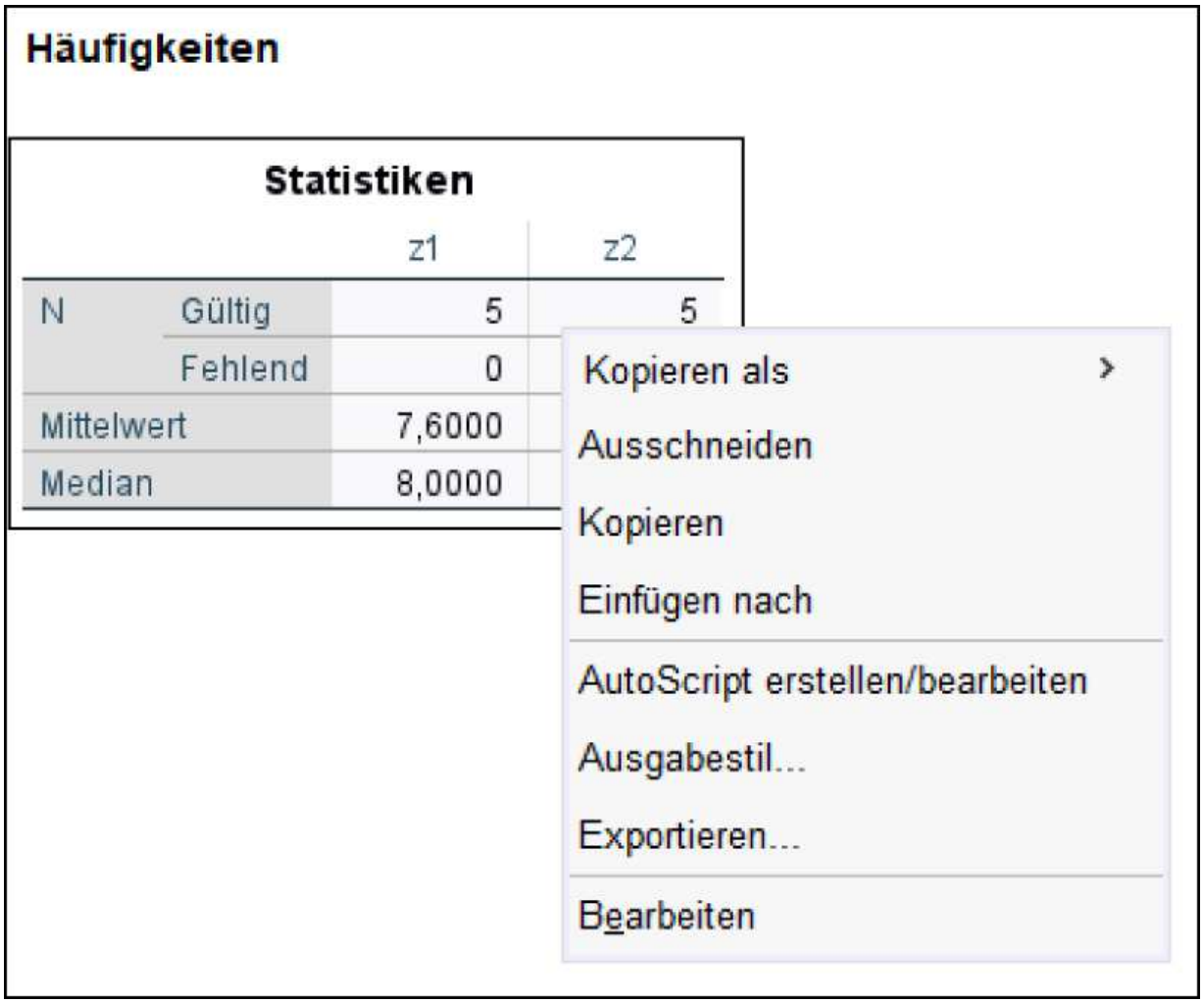



**Abbildung 5.82:** Erweiterte Verarbeitung-smöglichkeiten einer Pivottabelle im Viewer-Inhaltsbereich (hier in Bezug auf *CORRELATIONS.SAV*)

## *Diagramme*

Alternativ oder additiv zu einer Darstellung Ihrer Ergebnisse in Tabellenform können Sie sich **Diagramme** »auswerfen« lassen.

Angenommen, Sie möchten sich für die Variable *x1* aus dem Datensatz *CORRELATIONS.SAV* ein Balkendiagramm von SPSS erstellen lassen, dann könnten Sie Folgendes programmieren (siehe *SYNTAX24.SPS*):

```
GET FILE = 'C:\SPSS-DATEIEN\CORRELATIONS.SAV'.
FREQUENCIES VARIABLES = x1
/FORMAT = NOTABLE
/BARCHART.
```

Sie erhalten dann das in Abbildung 5.83 abgebildete Diagramm.

Durch einen Doppelklick auf ein Diagramm im Viewer-Inhaltsbereich können Sie das betreffende Diagramm umformatieren. Es öffnet sich der Diagrammeditor, in dem Sie entsprechende Veränderungen vornehmen können (siehe Abbildung 5.84, in der mit dem Shortcut Strg+T ein allerdings etwas eingeschränktes Eigenschaftsfenster geöffnet wurde). Ein umfangreicheres Eigenschaftsfenster als in Abbildung 5.84 erhalten Sie in unserem Beispiel durch erneuten Doppelklick auf die Balken (siehe hierzu Abbildung 5.85).

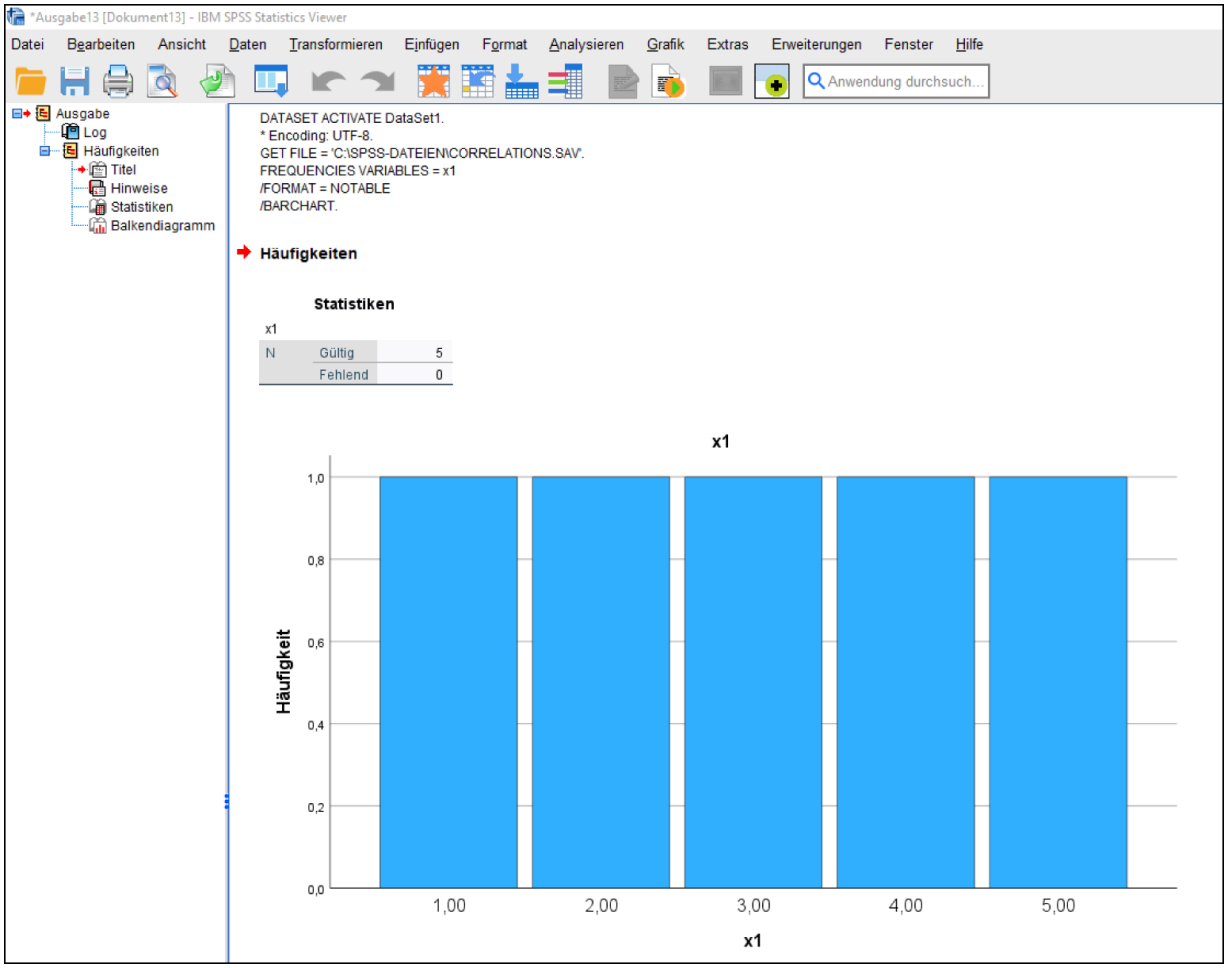



**Abbildung 5.83:** Diagramm im Viewer-Inhaltsbereich (hier in Bezug auf *CORRELATIONS.SAV*)

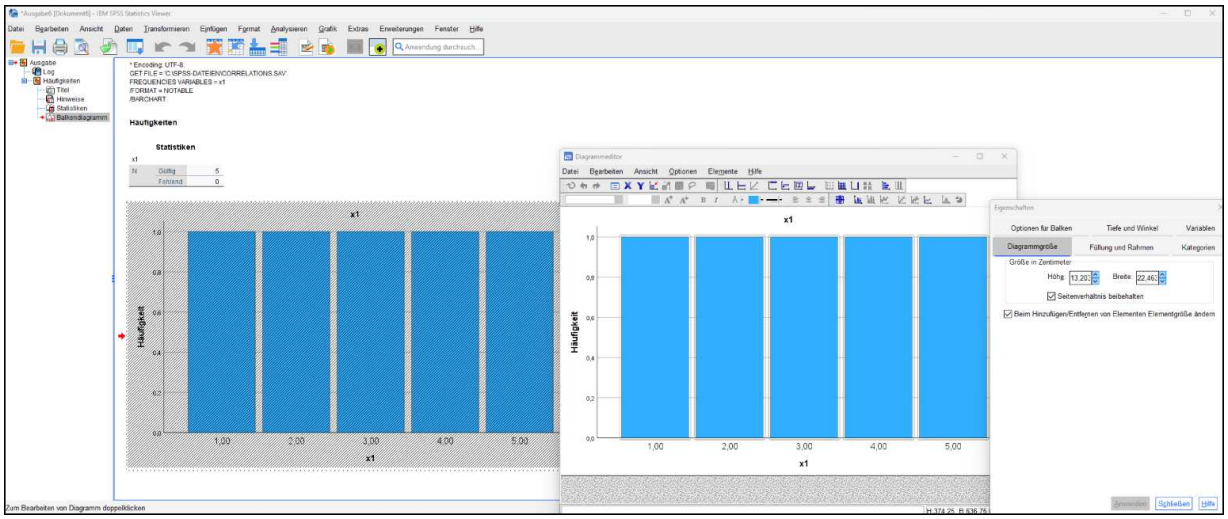


**Abbildung 5.84:** Öffnen des Diagrammeditors aus dem Viewer-Inhaltsbereich heraus (hier in Bezug auf *CORRELATIONS.SAV*)

Es steht Ihnen im Diagrammeditor ein reichhaltiges Menü zur Veränderung Ihrer Grafik zur Verfügung. Am besten spielen Sie einfach mal ein bisschen mit den verschiedenen Veränderungsvarianten! Ich habe in Abbildung 5.85 auf die Schnelle in der Grafik die Farbe der Balken verändert, von blau zu gelb.

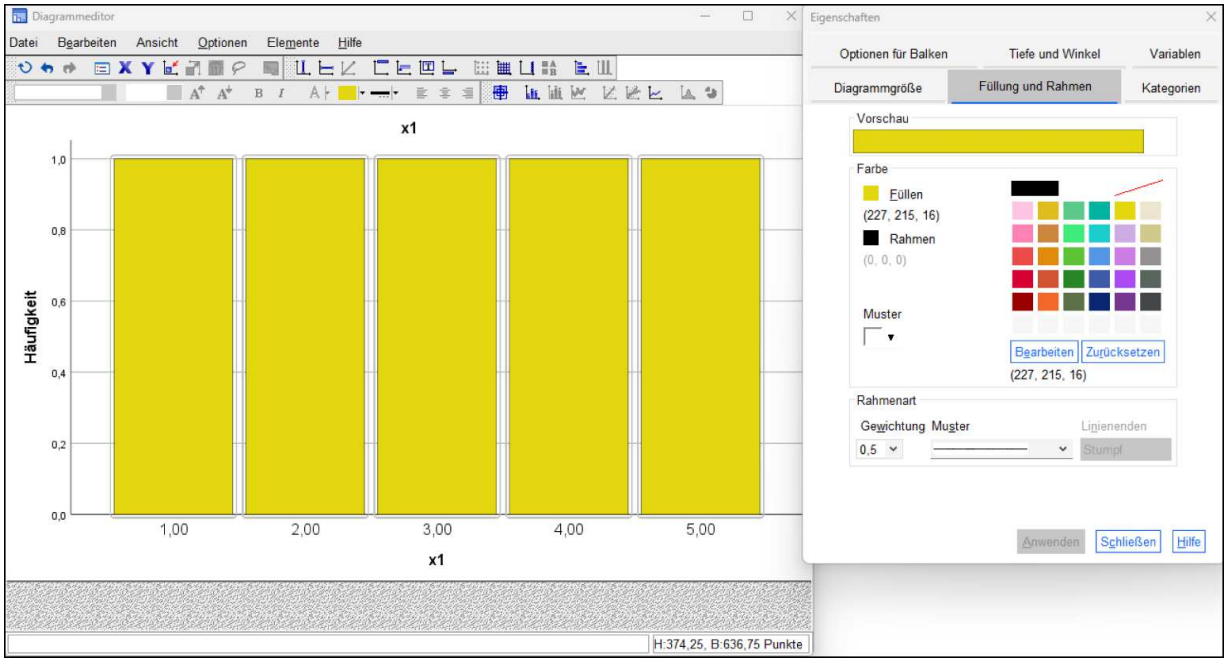



**Abbildung 5.85:** Verändern eines Diagramms im Diagrammeditor

Analog zu Pivottabellen können Sie auch Diagramme in anderen Programmen als SPSS weiterverwenden beziehungsweise weiterbearbeiten. Dazu müssen Sie – nach einem Klick mit der rechten Maustaste auf das Diagramm – lediglich die betreffenden Inhalte über die Zwischenablage und über BEARBEITEN|KOPIEREN beziehungsweise mittels der Tastenkombination Strg+C sowie anschließend mit BEARBEITEN|EINFÜGEN (oder Strg+V) an die gewünschte Stelle in dem anderen Programm kopieren.

## *Speichern des Viewer-Inhaltsbereichs*

Ihren Output können Sie als *SPV*-Datei in SPSS abspeichern. Dann können Sie jederzeit auf den betreffenden Output zurückgreifen, sich also Ihre alten Ergebnisse noch einmal anschauen. Ihren Output speichern Sie ganz einfach via DATEI|SPEICHERN (UNTER) als *SPV*-Datei. Die betreffende Datei können Sie dann jederzeit wieder in SPSS öffnen mit: DATEI|ÖFFNEN|AUSGABE oder – wenn es noch nicht allzu lange her ist, dass die Datei geöffnet wurde – mit DATEI|ZULETZT GEÖFFNETE DATEIEN.

Eine von mir in Kapitel 1 bereits ausgesprochene massive Warnung muss ich allerdings an dieser Stelle noch einmal unbedingt loswerden (bitte gestatten Sie mir, dass ich mich hier ausnahmsweise mal selbst zitiere, weil ich das in diesem Buch schon mal geschrieben hatte, aber das folgende Anliegen ist mir sehr wichtig und: »Doppelt hält bekanntlich besser.«):

Ab SPSS-Version 16 sind frühere SPSS-Outputfiles in den neueren SPSS-Versionen nicht mehr unmittelbar in SPSS einlesbar, das heißt letztlich nicht mehr (unkompliziert) nutzbar: Die entsprechende Extension wurde seinerzeit von *.SPO* auf *.SPV* verändert. Nur mittels der Installation einer zusätzlichen Applikation – der sogenannten »**Legacy-Viewer«-Applikation** – ist an dieser Stelle eine Abwärtskompatibilität von neueren SPSS-Versionen als Version 16 gegenüber eben dieser Version 16 (beziehungsweise gegenüber Versionen vor SPSS 16) herstellbar.

> Auf der glücklichen Seite des statistischen Lebens ist derjenige, der früher die SPSS-Outputfiles, was möglich ist, als EXCEL-Datei (oder als WORD-Datei) abgespeichert hatte und solcherart sich seine alten Ergebnisse in EXCEL anschauen (beziehungsweise eventuell weiterverarbeiten) kann.

Sie müssen Ihren Output allerdings nicht als SPSS-Outputfile abspeichern, sondern Sie können ihn auch komplett in ein anderes Programm transferieren, und zwar über DATEI|EXPORTIEREN als:

- ✔ HTML-Datei,
- ✔ PDF-Datei,
- ✔ MS WORD-/RTF-Datei,
- ✔ MS-PowerPoint-Datei oder
- ✔ EXCEL-Datei.

Dieser Transfer wird im Folgenden für EXCEL und WORD demonstriert. Zu diesem Zweck lassen Sie bitte noch einmal das Syntaxprogramm *SYNTAX23.SPS* »laufen«:

```
GET FILE = 'C:\SPSS-DATEIEN\CORRELATIONS.SAV'.
COMPUTE z1 = x1 + y1.
COMPUTE z2 = z1 + x2.
FREQUENCIES VARIABLES = z1 z2
/FORMAT = NOTABLE
/STATISTICS = MEDIAN MEAN.
```

Den erhaltenen Output können Sie nun in ein anderes Programm als SPSS exportieren. In Abbildung 5.86 ist als Beispiel der Export nach EXCEL gezeigt. Nachdem Sie über DATEIEN|EXPORTIEREN das Fenster AUSGABE EXPORTIEREN geöffnet haben, geben Sie in diesem Zusammenhang bei TYP ein EXCEL-Format ein. Des Weiteren fügen Sie bitte in der Mitte der Dialogbox den von Ihnen gewünschten Verzeichnisort und Dateinamen ein (siehe ebenfalls Abbildung 5.86).

Ihr Output wird nach Klicken auf der OK-Taste am gewünschten Ort unter der gewünschten Dateibezeichnung abgespeichert. Der Inhalt des Outputs wird im vorliegenden Fall wie in Abbildung 5.87 in EXCEL

dargestellt und steht zur Weiterverarbeitung durch das Tabellenverarbeitungsprogramm EXCEL zu Ihrer Verfügung.

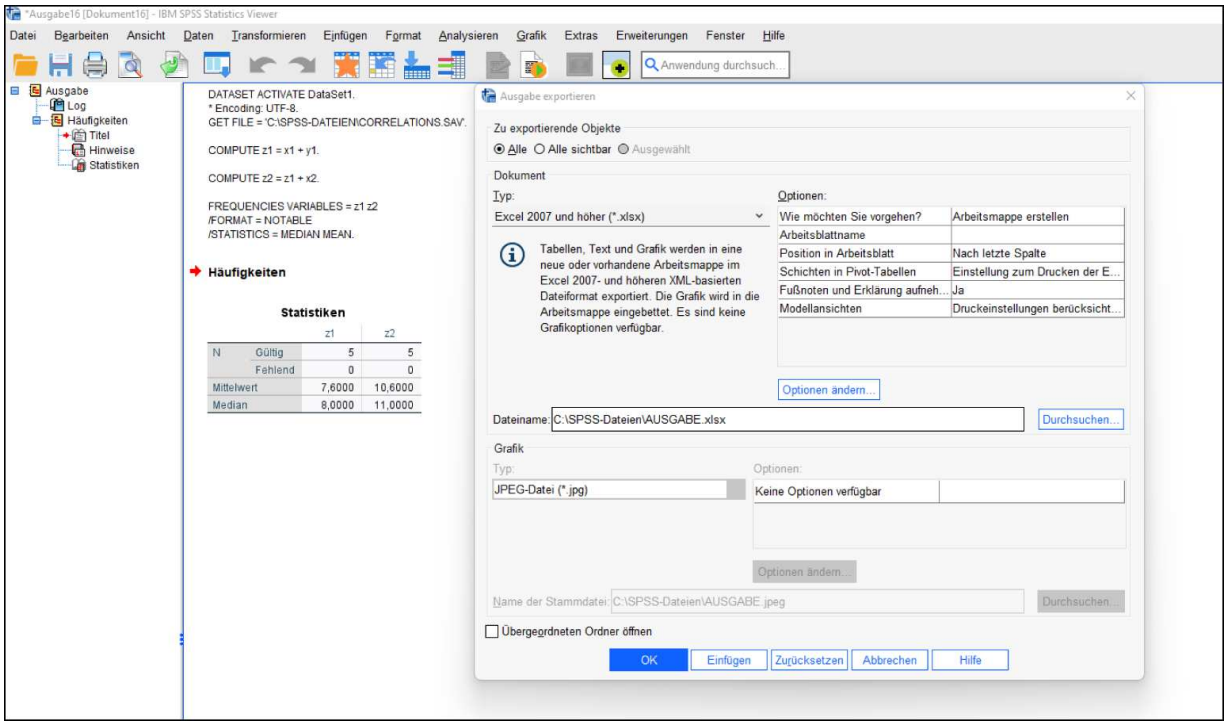


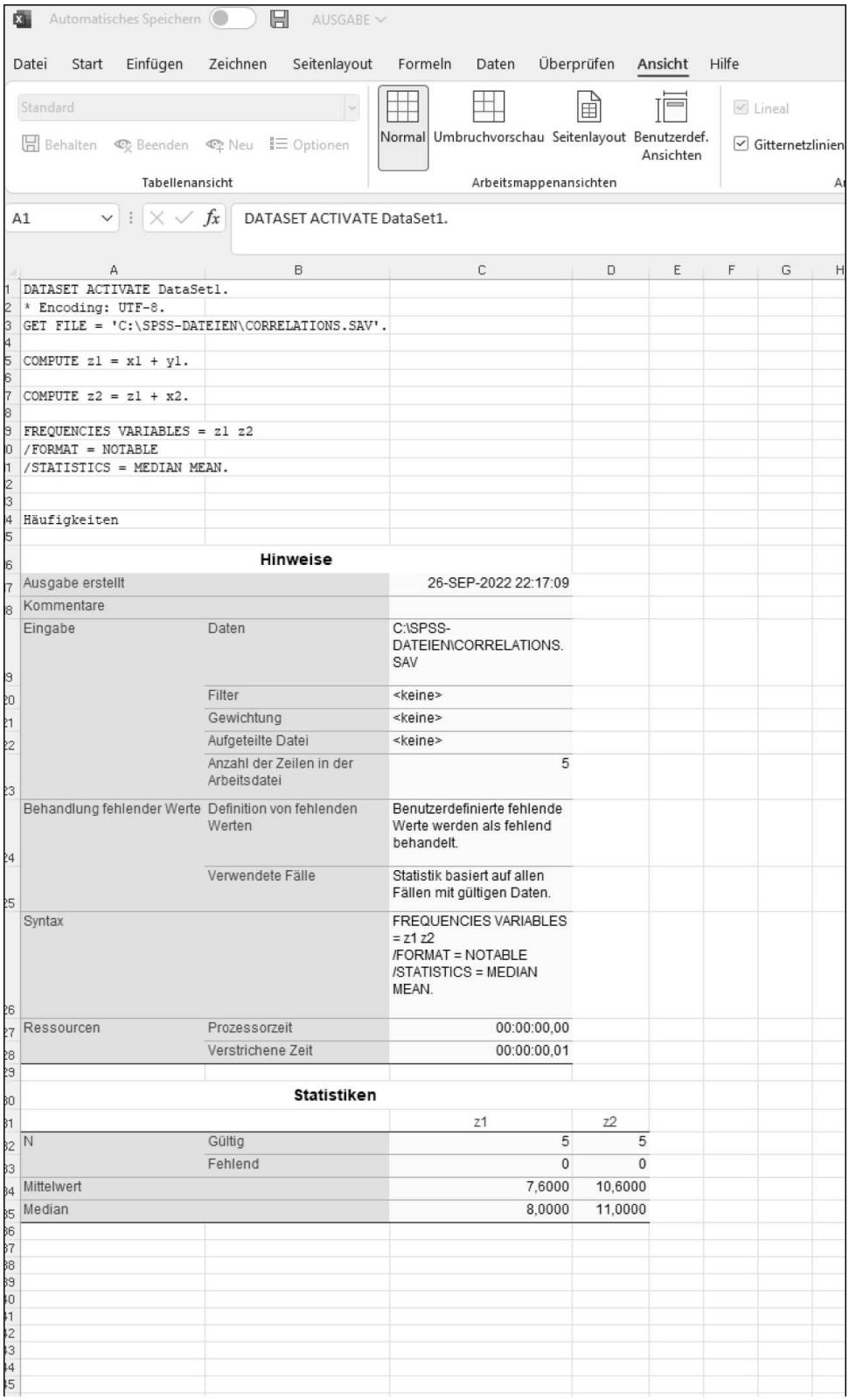

**Abbildung 5.86:** Abspeichern von SPSS-Output in einer EXCEL-Datei (hier in Bezug auf *CORRELATIONS.SAV*)

Eine andere Möglichkeit des Abspeicherns von SPSS-Output in einem anderen Programm ist etwa das Abspeichern als WORD-Datei (siehe Abbildung 5.88). Nachdem Sie auch hier über DATEIEN|EXPORTIEREN das Fenster AUSGABE EXPORTIEREN geöffnet haben, geben Sie bei TYP ein WORD-Format ein und notieren zudem in der Mitte der Dialogbox den gewünschten Verzeichnisort und den Datennamen.

Nach Klicken auf der OK-Taste wird Ihr Output auch in diesem Fall am gewünschten Ort unter der gewünschten Dateibezeichnung abgespeichert. Wie das Ganze in WORD aussieht, zeigt Abbildung 5.89. Mit dem in WORD dargestellten SPSS-Output können Sie im Rahmen des Textverarbeitungsprogramms WORD weiterarbeiten.

Wenn Sie sich im Übrigen von der Philosophie des »papierlosen Büros« verabschiedet haben [;-)], dann können Sie die Viewer-Inhalte über

DATEI|DRUCKEN entweder komplett oder ausschnittsweise in Form markierter Abschnitte ausdrucken.




```
DATASET ACTIVATE DataSet1.
* Encoding: UTF-8.
GET FILE = 'C:\SPSS-DATEIEN\CORRELATIONS.SAV'.

COMPUTE z1 = x1 + y1.

COMPUTE z2 = z1 + x2.

FREQUENCIES VARIABLES = z1 z2
/FORMAT = NOTABLE
/STATISTICS = MEDIAN MEAN.


Häufigkeiten
```

**Hinweise**

| | | |
|---|---|---|
| Ausgabe erstellt | | 26-SEP-2022 22:17:09 |
| Kommentare | | |
| Eingabe | Daten | C:\SPSS-DATEIEN\CORRELATIONS.SAV |
| | Filter | <keine> |
| | Gewichtung | <keine> |
| | Aufgeteilte Datei | <keine> |
| | Anzahl der Zeilen in der Arbeitsdatei | 5 |
| Behandlung fehlender Werte | Definition von fehlenden Werten | Benutzerdefinierte fehlende Werte werden als fehlend behandelt. |
| | Verwendete Fälle | Statistik basiert auf allen Fällen mit gültigen Daten. |
| Syntax | | FREQUENCIES VARIABLES = z1 z2 /FORMAT = NOTABLE /STATISTICS = MEDIAN MEAN. |
| Ressourcen | Prozessorzeit | 00:00:00,00 |
| | Verstrichene Zeit | 00:00:00,01 |

**Statistiken**

| | | z1 | z2 |
|---|---|---|---|
| N | Gültig | 5 | 5 |
| | Fehlend | 0 | 0 |
| Mittelwert | | 7,6000 | 10,6000 |
| Median | | 8,0000 | 11,0000 |

**Abbildung 5.87:** SPSS-Output in einer EXCEL-Datei (hier in Bezug auf *CORRELATIONS.SAV*)

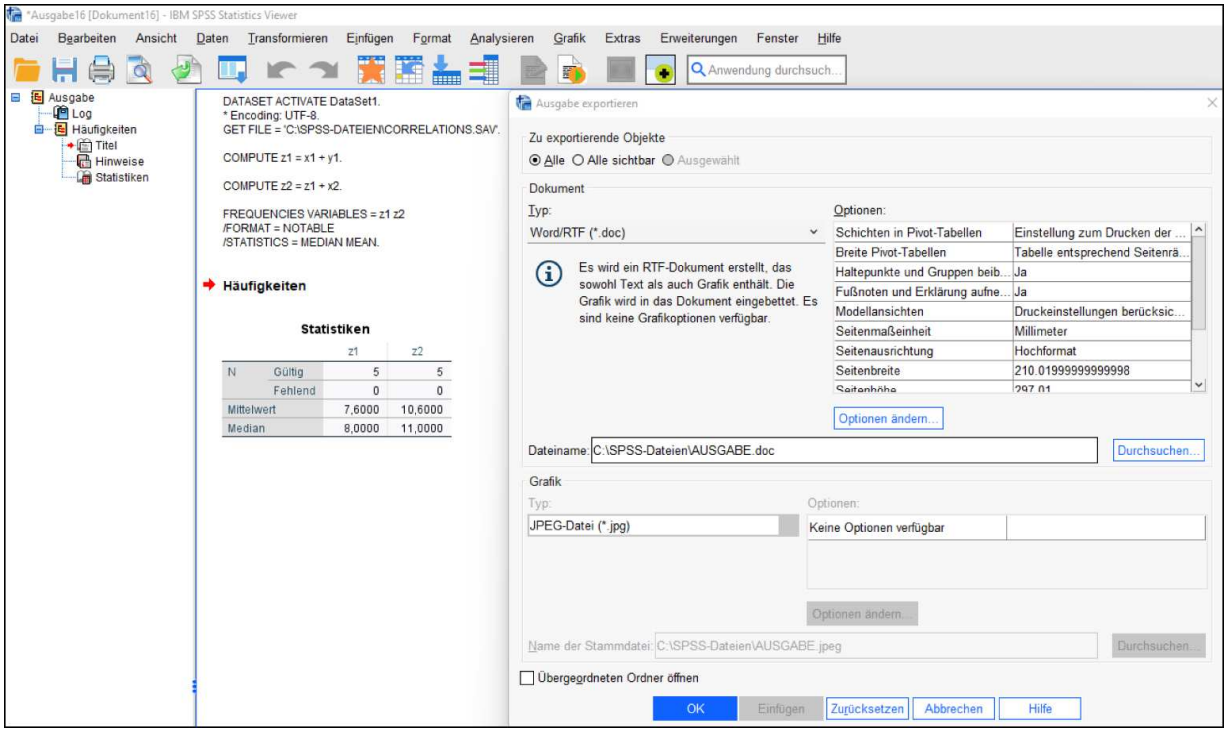
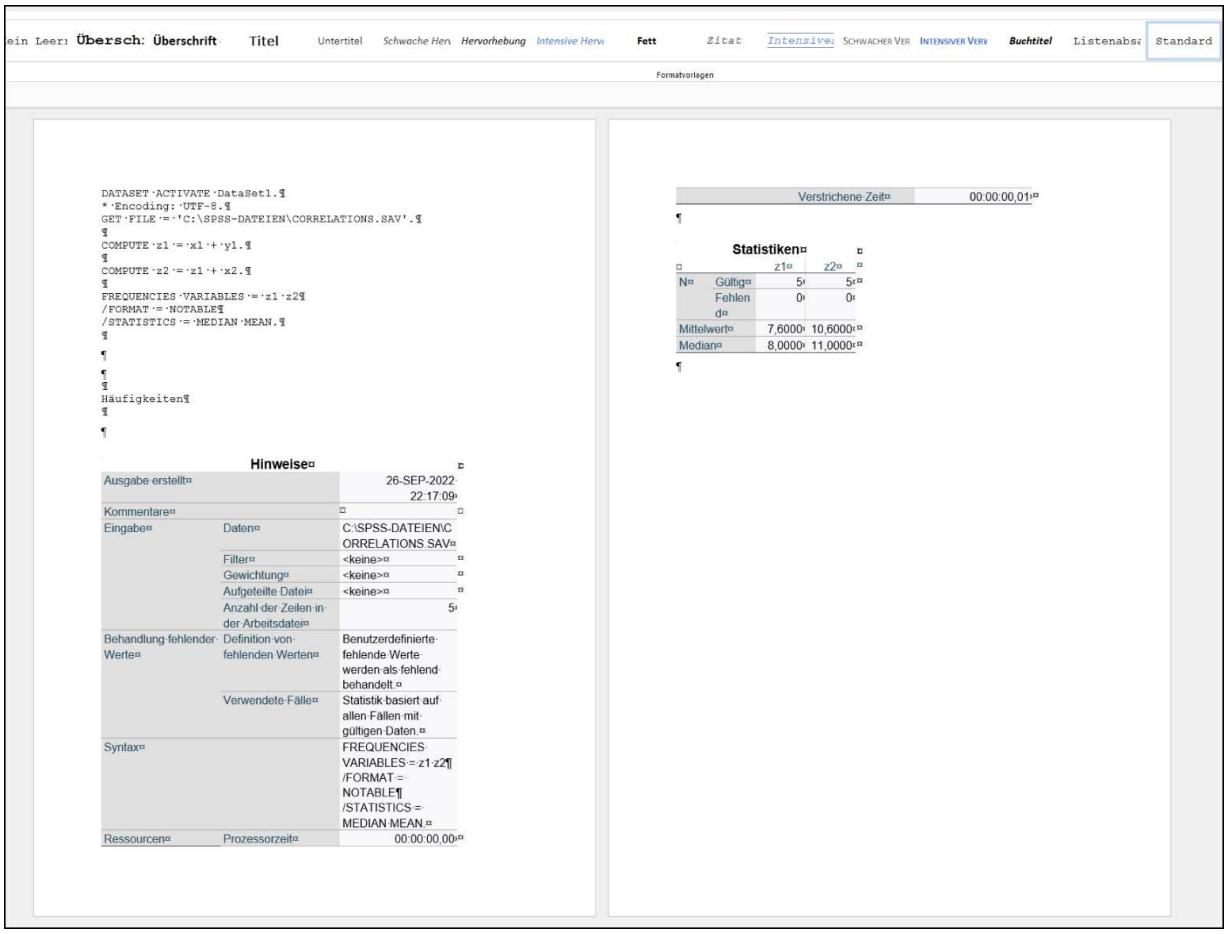



**Abbildung 5.88:** Abspeichern von SPSS-Output in einer WORD-Datei (in Bezug auf *CORRELATIONS.SAV*)



```
DATASET ACTIVATE DataSet1.
* Encoding: UTF-8.
GET FILE = 'C:\SPSS-DATEIEN\CORRELATIONS.SAV'.

COMPUTE z1 = x1 + y1.

COMPUTE z2 = z1 + x2.

FREQUENCIES VARIABLES = z1 z2
/FORMAT = NOTABLE
/STATISTICS = MEDIAN MEAN.
```

Häufigkeiten

**Hinweise**

| | | |
|---|---|---|
| Ausgabe erstellt | | 26-SEP-2022 22:17:09 |
| Kommentare | | |
| Eingabe | Daten | C:\SPSS-DATEIEN\CORRELATIONS.SAV |
| | Filter | <keine> |
| | Gewichtung | <keine> |
| | Aufgeteilte Datei | <keine> |
| | Anzahl der Zeilen in der Arbeitsdatei | 5 |
| Behandlung fehlender Werte | Definition von fehlenden Werten | Benutzerdefinierte fehlende Werte werden als fehlend behandelt. |
| | Verwendete Fälle | Statistik basiert auf allen Fällen mit gültigen Daten. |
| Syntax | | FREQUENCIES VARIABLES = z1 z2 /FORMAT = NOTABLE /STATISTICS = MEDIAN MEAN |
| Ressourcen | Prozessorzeit | 00:00:00,00 |
| | Verstrichene Zeit | 00:00:00,01 |

**Statistiken**

| | | z1 | z2 |
|---|---|---|---|
| N | Gültig | 5 | 5 |
| | Fehlend | 0 | 0 |
| Mittelwert | | 7,6000 | 10,6000 |
| Median | | 8,0000 | 11,0000 |

**Abbildung 5.89:** SPSS-Output in einer WORD-Datei (hier in Bezug auf *CORRELATIONS.SAV*)

# Teil II

# Deskriptive SPSS-Statistiken

# IN DIESEM TEIL …

- ✔ Eindimensionale Häufigkeitsverteilungen
- ✔ Eindimensionale Kennziffern

# Kapitel 6

# Von der Schönheit des Anhäufens: Eindimensionale Häufigkeitsverteilungen

## IN DIESEM KAPITEL

Zum Häufigkeitsbegriff

Zur Fehler- und Ausreißeranalyse

Häufigkeitsdiagramme

Kapitel 6 befasst sich mit eindimensionalen Häufigkeitsverteilungen, das heißt mit den Häufigkeiten (im Sinne von Fallzahlen) jeweils nur einer Variablen. Darüber hinaus präsentiert Kapitel 6 – auf Basis der entsprechenden Häufigkeitsauszählungen – Möglichkeiten, Eingabefehler bei den Daten sowie auch Ausreißerwerte zu erkennen. Schließlich wird in Kapitel 6 auf diverse Diagrammoptionen – hier im Kontext von Häufigkeitsauszählungen – eingegangen.

## *Zum Häufigkeitsbegriff*

Um bestimmte Sachverhalte statistisch übersichtlich darzustellen, können Sie die Merkmalsausprägungen auszählen und auf diese Art und Weise einzelne Merkmale beschreiben. Daher gehört dieses Auszählen (ebenso wie die weiteren Kapitel von Teil II) grundsätzlich zur **deskriptiven Statistik** (auf Deutsch: zur beschreibenden Statistik).

In den meisten Lehrbüchern wird die deskriptive Statistik – in Abgrenzung zur ausdrücklich auf Stichproben bezogenen induktiven Statistik (auf Deutsch: zur schließenden Statistik) – auf statistische (Grund-)Gesamtheiten bezogen. Daher werden in diesen Lehrbüchern bei der deskriptiven Statistik üblicherweise auch keine wahrscheinlichkeitstheoretischen Aussagen im Sinne des statistischen Gehalts (mit einem Fachbegriff: der Signifikanz) beziehungsweise der Repräsentativität entsprechender Befunde getroffen. Dies ist bei SPSS anders: Hier wird auch im Rahmen typischer deskriptiver Kennziffern von einer Stichprobenziehung ausgegangen. Entsprechend werden von SPSS an den betreffenden Stellen Signifikanzaussagen getroffen; außerdem werden bestimmte Kennziffern wie zum Beispiel die Varianz (siehe hierzu Kapitel 7) leicht anders definiert und daher berechnet, als Sie das möglicherweise aus Lehrbüchern zur deskriptiven Statistik kennen.

Die eingangs erwähnten Auszählungen dienen auch dem Erkennen von Fehlern und Ausreißern, aber auch der Diskussion fehlender Werte im Datensatz, wie in diesem Kapitel noch näher umrissen wird. Allgemein spricht man bei den genannten Auszählungen davon, dass man **Häufigkeiten** für das Vorkommen der einzelnen Merkmalsausprägungen bestimmt. Häufigkeiten können Sie bereits für Merkmale mit Nominalskalenniveau angeben.

Sie können absolute und relative Häufigkeiten bilden und diese zudem schrittweise addieren. Im letztgenannten Fall spricht man vom Kumulieren beziehungsweise von kumulierten (absoluten oder relativen) Häufigkeiten.

Bei einer **absoluten Häufigkeit** geben Sie für das Vorkommen einer Merkmalsausprägung eine ganze Zahl an, zum Beispiel: In der vergangenen Statistik-Klausur haben 2 von 20 Teilnehmern die Note 1 erreicht, 5 der 20 Teilnehmer eine 2, 10 der 20 Teilnehmer eine 3 und so weiter. Eine Häufigkeit können Sie insofern auch relativ formulieren, als Sie für das Vorkommen einer

Merkmalsausprägung einen Prozentwert (im Sinne eines Anteilswerts) angeben; im letztgenannten Beispiel also: 10 Prozent der Klausurteilnehmer haben die Note 1 erzielt, 25 Prozent eine 2, 50 Prozent eine 3 und so weiter.

Eine **kumulierte Häufigkeit** addiert von der niedrigsten bis zu einer vorgegebenen Ausprägung die Fallzahlen (bei einer kumulierten absoluten Häufigkeit) beziehungsweise die Prozentwerte (bei einer kumulierten relativen Häufigkeit). Wegen dieser unterschiedlichen Wertigkeit der einzelnen Merkmalsausprägungen sollten Sie kumulierte Häufigkeiten erst ab Ordinalskalenniveau verwenden. Im vorgenannten Beispiel mit der Statistik-Klausur erhalten Sie die Antwort auf die Fragestellung, wie viele Personen höchstens die Note 3 geschrieben haben, dadurch, dass Sie die Fallzahlen beziehungsweise die Prozentwerte für die Notenstufen 1, 2 und 3 zusammenzählen. Entsprechend haben im Beispielfall 17 Teilnehmer (kumulierte absolute Häufigkeit) beziehungsweise 17 von 20 Teilnehmern, also 85 Prozent der Teilnehmer (kumulierte relative Häufigkeit) als schlechteste Note eine 3 erreicht (endlich mal eine Statistik-Klausur, die super ausgefallen ist – aber es handelt sich hier leider nur um ein fiktives Beispiel).

## *Häufigkeitstabelle*

Über ANALYSIEREN|DESKRIPTIVE STATISTIKEN|HÄUFIGKEITEN… kommen Sie in eine SPSS-Dialogbox, mit deren Hilfe Sie eindimensionale (= univariate) **Häufigkeitsverteilungen** anfordern können. Eine Häufigkeitsverteilung spiegelt die den einzelnen Merkmalsausprägungen zugeordneten Häufigkeiten wider.

Derartige Häufigkeitsverteilungen werden in diesem Kapitel jeweils separat für ein einzelnes Merkmal betrachtet – daher auch der Begriff der eindimensionalen beziehungsweise univariaten Häufigkeitsverteilung. Auch in Kapitel 7 wird sich nur auf die statistische Analyse eines einzelnen Merkmals bezogen. Ab Kapitel 8

werden dann aber auch die Zusammenhänge zwischen zwei und mehr Merkmalen betrachtet.

Was sich in Abbildung 6.1 – im Übrigen für die Datendatei *KORBTREFFER.SAV* – geöffnet hat, ist ein **Dialogfeld**. Dieses konkrete Dialogfeld (alternative Bezeichnung: Dialogbox) ist typisch für statistische Operationen. Bei solchen »statistischen« Dialogboxen finden Sie üblicherweise die Anordnung der verfügbaren Variablen links in der Dialogbox. Falls Ihnen die Variablenanzeige nicht breit genug ist, können Sie die Dialogbox mittels »Ziehen« mit der Maus einfach vergrößern. Die von Ihnen für die entsprechende Analyse ausgewählten Variablen müssen Sie lediglich mit der Maus in die Rubrik VARIABLE(N) ziehen. Sie können sich in derartigen Dialogfeldern entweder den Variablennamen oder das Variablenlabel anzeigen lassen.

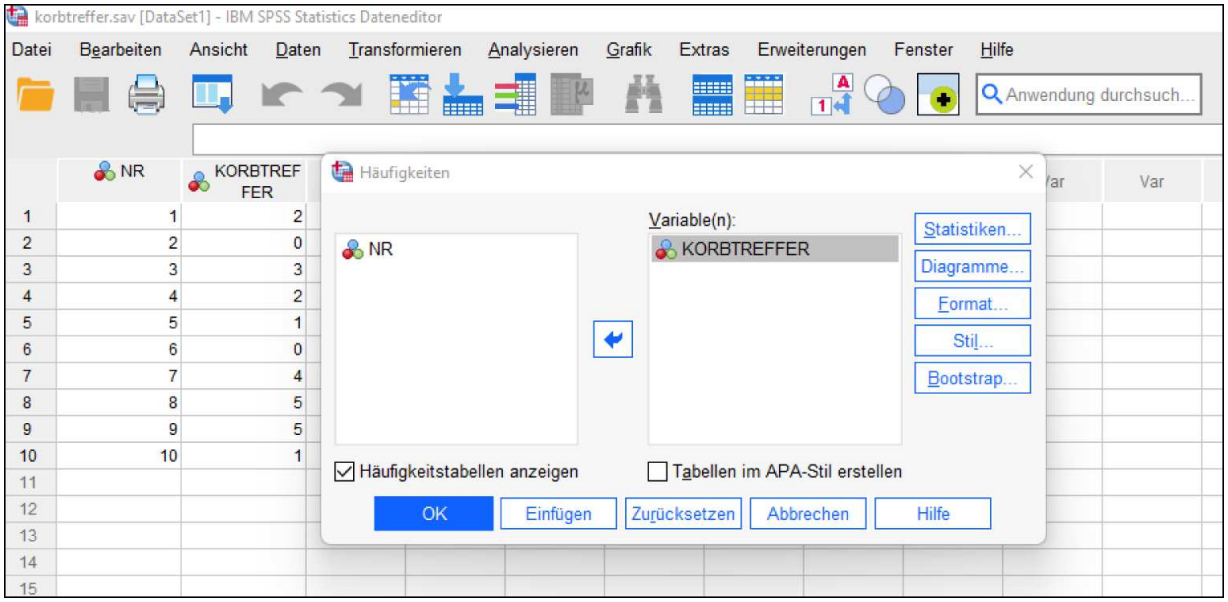



**Abbildung 6.1:** Fenster für HÄUFIGKEITEN (hier für *KORBTREFFER.SAV*)

Das vorliegende Dialogfeld weist zwei Hauptkomponenten auf:

- ✔ die Liste der Quellvariablen,
- ✔ die Zielvariablenliste(n).

Dialogfelder haben in der Regel fünf Steuerungselemente:

- ✔ OK,
- ✔ EINFÜGEN,
- ✔ ZURÜCKSETZEN,
- ✔ ABBRECHEN,
- ✔ HILFE.

Damit jetzt bei dieser »Häufigkeits-Dialogbox« die Häufigkeitsverteilung in Form einer **Häufigkeitstabelle** auch angezeigt wird, lassen Sie einfach das Häkchen bei HÄUFIGKEITSTABELLE ANZEIGEN stehen. Dies empfiehlt sich naheliegenderweise nur bei Variablen mit einer überschaubaren Anzahl an Ausprägungen (wie im »Basketball-Beispiel« im nächsten Abschnitt dieses Kapitels). Sinnvollerweise sollten Sie nämlich eine **Häufigkeitstabelle** nicht für außerordentlich viele Werte erstellen, da das Ganze dann leicht sehr schnell unübersichtlich wird.

## *Das »Basketball-Beispiel« ohne fehlende Werte*

Lassen Sie uns gemeinsam die Funktionsweise von HÄUFIGKEITEN… an einem Beispiel erläutern. Hierzu habe ich mir einfach mal ein fiktives Beispiel aus dem Basketballsport überlegt. Stellen Sie sich vor, Sie lassen 10 Personen jeweils fünfmal auf einen Basketballkorb werfen und zählen anschließend die Anzahl der Treffer für jede Person. Dann könnten Sie etwa das in Tabelle 6.1 angegebene Ergebnis erhalten.

| *Person Nr.* | *Anzahl der Korbtreffer* |
|---|---|
| 1 | 2 |
| 2 | 0 |
| 3 | 3 |
| 4 | 2 |
| 5 | 1 |
| 6 | 0 |
| 7 | 4 |
| 8 | 5 |

| *Person Nr.* | *Anzahl der Korbtreffer* |
|---|---|
| 9 | 5 |
| 10 | 1 |

**Tabelle 6.1:** »Basketball-Beispiel« zu Häufigkeiten

Wenn Sie jetzt manuell die Anzahl der Treffer auszählen, werden Sie folgendes Ergebnis erhalten:

- ✔ 0 Treffer: 2-mal,
- ✔ 1 Treffer: 2-mal,
- ✔ 2 Treffer: 2-mal,
- ✔ 3 Treffer: 1-mal,
- ✔ 4 Treffer: 1-mal und
- ✔ 5 Treffer: 2-mal.

Den vorstehenden Datensatz finden Sie unter *KORBTREFFER.SAV*. Analog zu Abbildung 6.1 wird nun für diese Datendatei im SPSS-Menü auf ANALYSIEREN|DESKRIPTIVE STATISTIKEN|HÄUFIGKEITEN geklickt. Anschließend geben Sie bitte – wie ebenfalls bereits in Abbildung 6.1 dargestellt – als Zielvariable *korbtreffer* ein und belassen das Häkchen bei HÄUFIGKEITSTABELLE ANZEIGEN. Aus STATISTIKEN wählen Sie an dieser Stelle bitte noch nichts aus. Mit diesen Setzungen erhalten Sie als Output für das »Basketball-Beispiel« die in Abbildung 6.2 angegebene Häufigkeitstabelle.

**Häufigkeiten**

[DataSet1] C:\SPSS-Dateien\korbtreffer.sav

| Statistiken | | |
|---|---|---|
| KORBTREFFER | | |
| N | Gültig | 10 |
| | Fehlend | 0 |

| KORBTREFFER | | | | | |
|---|---|---|---|---|---|
| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
| Gültig | 0 | 2 | 20,0 | 20,0 | 20,0 |
| | 1 | 2 | 20,0 | 20,0 | 40,0 |
| | 2 | 2 | 20,0 | 20,0 | 60,0 |
| | 3 | 1 | 10,0 | 10,0 | 70,0 |
| | 4 | 1 | 10,0 | 10,0 | 80,0 |
| | 5 | 2 | 20,0 | 20,0 | 100,0 |
| | Gesamt | 10 | 100,0 | 100,0 | |

**Abbildung 6.2:** HÄUFIGKEITEN-Output für »Basketball-Beispiel« (*KORBTREFFER.SAV*)

Dieser Output dürfte Ihnen vertraut vorkommen. SPSS weist die Häufigkeiten dabei sowohl absolut als auch relativ (in Prozent) aus. Beim relativen Ausweis wird unterschieden in »Prozent« und in »Gültige Prozente«. In unserem Fall ergeben sich offenkundig diesbezüglich keine Unterschiede. Dies liegt daran, dass im »Basketball-Beispiel« alle Werte gültig sind; das heißt, dass bei keinem Fall ein Wert fehlt. Ich werde gleich ein Beispiel mit fehlenden Werten mit Ihnen diskutieren.

Außerdem weist SPSS in der letzten Spalte noch »Kumulierte Prozente« aus im Sinne der kumulierten relativen Häufigkeiten. So erkennen Sie

anhand von Abbildung 6.2 etwa, dass 40 Prozent der Teilnehmer am Korbwurf-Wettbewerb höchstens einen Treffer erzielt haben (das heißt: gar nicht oder nur einmal getroffen haben).

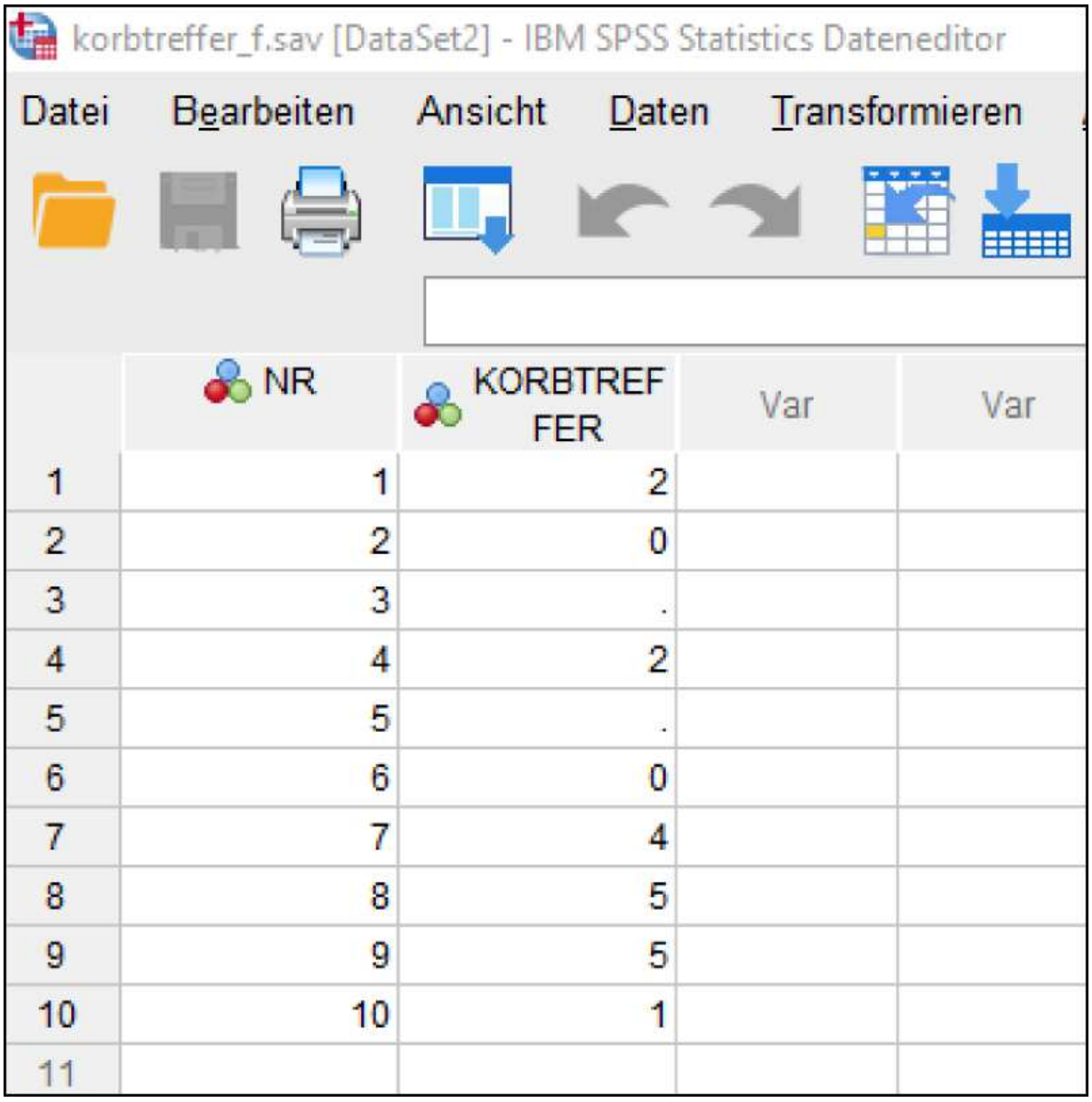
Den in Abbildung 6.2 dargestellten Output hätten Sie auch über ein kleines Syntaxprogramm realisieren können:

```
GET FILE='C:\SPSS-DATEIEN\KORBTREFFER.SAV'.
FREQUENCIES VARIABLES=korbtreffer.
```

Dieser FREQUENCIES-Befehl kommt vielleicht dem einen oder anderen aus Kapitel 5 bekannt vor. Dort hatten Sie ihn bereits ohne nähere Erläuterung an einzelnen Stellen verwendet. Dies geschah dort alleinig unter dem Gesichtspunkt, dass ich Ihnen ein bisschen »Butter bei die Fische« geben wollte, soll heißen: dass ich Ihnen dort schon ein paar statistische Auswertungen aus Veranschaulichkeitsgründen präsentieren wollte.

# *Das »Basketball-Beispiel« mit fehlenden Werten*

So, nun kommen Sie zu einem abgewandelten Datensatz für das »Basketball-Beispiel«, wie er sich in *KORBTREFFER_F.SAV* wiederfindet. In diesem Datensatz sind für die dritte und für die fünfte Person keine Angaben bezüglich ihrer Trefferzahl vorhanden (siehe Abbildung 6.3).




| | NR | KORBTREFFER | Var | Var |
|---|---|---|---|---|
| 1 | 1 | 2 | | |
| 2 | 2 | 0 | | |
| 3 | 3 | . | | |
| 4 | 4 | 2 | | |
| 5 | 5 | . | | |
| 6 | 6 | 0 | | |
| 7 | 7 | 4 | | |
| 8 | 8 | 5 | | |
| 9 | 9 | 5 | | |
| 10 | 10 | 1 | | |
| 11 | | | | |

**Abbildung 6.3:** Datensatz *KORBTREFFER_F.SAV* mit zwei fehlenden Werten

Wenn Sie sich nun über ANALYSIEREN|DESKRIPTIVE STATISTIKEN|HÄUFIGKEITEN und mittels des bei HÄUFIGKEITSTABELLE ANZEIGEN gesetzten Häkchens die Häufigkeitstabelle für *korbtreffer* erstellen lassen, wird vorab zunächst einmal korrekt im Viewer angegeben, dass die Variable *korbtreffer* zwei fehlende Werte hat.

Darüber hinaus erkennen Sie in Abbildung 6.4, dass es nunmehr bezüglich der relativen Häufigkeitsverteilung in den Spalten »Prozent«

und »Gültige Prozente« Abweichungen gibt. So findet sich beispielsweise bei der Ausprägung »1 Treffer« in der »Prozent«-Spalte der relative Häufigkeitswert 10,0 Prozent, während in der »Gültige Prozente«-Spalte bei dieser Merkmalsausprägung 12,5 Prozent steht. Dieser Unterschied ergibt sich daraus, dass die »Prozent«-Häufigkeiten sich auf alle 10 Fälle (also inklusive der beiden fehlenden Werte) beziehen, während die »Gültige Prozente«-Häufigkeiten als Bezugsgröße nur die gültigen acht Werte (das heißt: alle Werte abzüglich der beiden fehlenden Werte) haben.

**Häufigkeiten**

[DataSet2] C:\SPSS-Dateien\korbtreffer_f.sav

| Statistiken | | |
|---|---|---|
| KORBTREFFER | | |
| N | Gültig | 8 |
| | Fehlend | 2 |

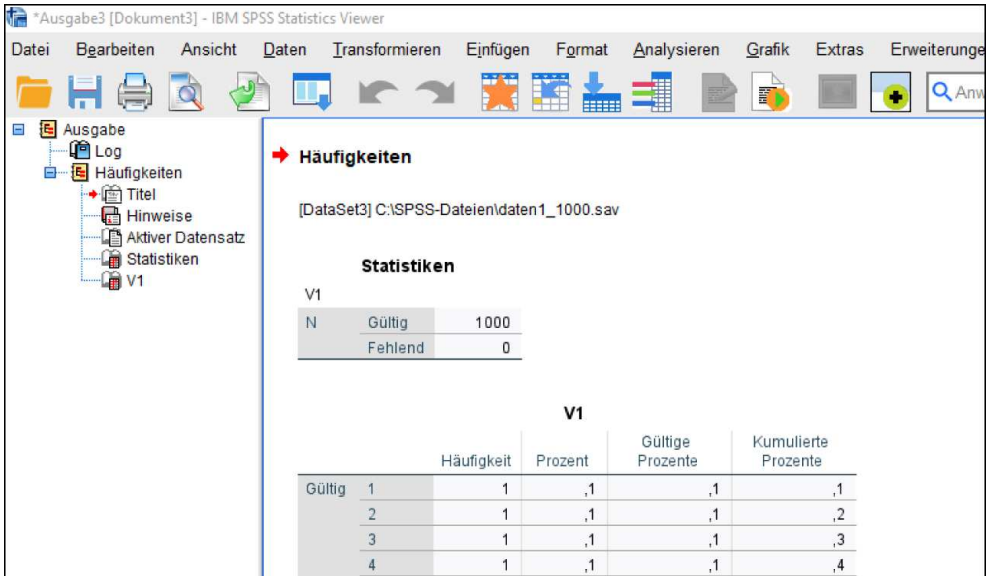
| KORBTREFFER | | | | | |
|---|---|---|---|---|---|
| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
| Gültig | 0 | 2 | 20,0 | 25,0 | 25,0 |
| | 1 | 1 | 10,0 | 12,5 | 37,5 |
| | 2 | 2 | 20,0 | 25,0 | 62,5 |
| | 4 | 1 | 10,0 | 12,5 | 75,0 |
| | 5 | 2 | 20,0 | 25,0 | 100,0 |
| | Gesamt | 8 | 80,0 | 100,0 | |
| Fehlend | System | 2 | 20,0 | | |
| Gesamt | | 10 | 100,0 | | |

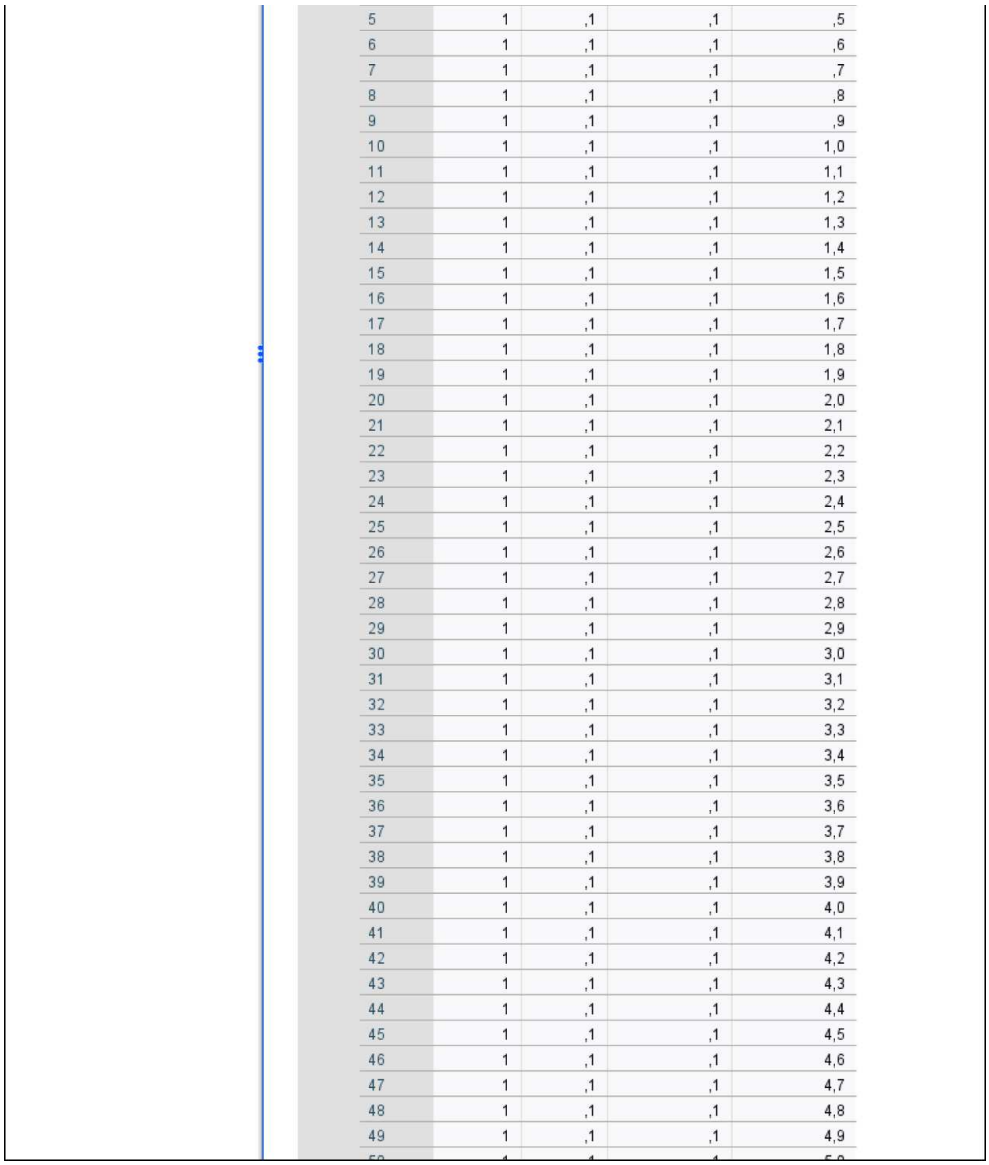
**Abbildung 6.4:** Häufigkeitstabelle für eine Variable mit (zwei) fehlenden Werten (*KORBTREFFER_F.SAV*)

Wie Sie der letzten Spalte der Häufigkeitstabelle in Abbildung 6.4 entnehmen können, werden die kumulierten relativen Häufigkeiten auf der Grundlage der relativen Häufigkeiten ohne die beiden fehlenden Werte berechnet, also auf der Grundlage der »Gültige Prozente«-Spalte.

## *Viele Einzelwerte als Problem einer Häufigkeitstabelle*

Um Ihnen zu illustrieren, dass eine Häufigkeitstabelle bei sehr vielen Werten leicht unübersichtlich wird, habe ich einfach mal die Datendatei *DATEN1_1000.SAV* erstellt. Sie besteht aus einer einzigen Variablen *v1* und – was hier wesentlicher ist – aus insgesamt 1.000 unterschiedlichen Werten (ich habe die natürlichen Zahlen von 1 bis 1.000 genommen). Wenn Sie sich nunmehr über ANALYSIEREN|DESKRIPTIVE STATISTIKEN|HÄUFIGKEITEN und mittels des bei HÄUFIGKEITSTABELLE ANZEIGEN gesetzten Häkchens die zugehörige Häufigkeitstabelle »auswerfen« lassen, erhalten Sie die in Abbildung 6.5 auszugsweise angegebene Häufigkeitstabelle, in der jeder Merkmalsausprägung eine absolute Häufigkeit von 1 zugewiesen ist. Sie müssen im Viewer über mehrere Seiten scrollen, um sich alle Werte anschauen zu können. Das ist gewiss nicht sehr übersichtlich – ganz abgesehen von der geringen Information, die man aus so einer Tabelle zieht. Sie sollten sich vor der Erstellung einer Häufigkeitstabelle also gut überlegen, ob Sie wirklich so viele Merkmalsausprägungen benötigen.





| | | | | |
|---|---|---|---|---|
| 5 | 1 | ,1 | ,1 | ,5 |
| 6 | 1 | ,1 | ,1 | ,6 |
| 7 | 1 | ,1 | ,1 | ,7 |
| 8 | 1 | ,1 | ,1 | ,8 |
| 9 | 1 | ,1 | ,1 | ,9 |
| 10 | 1 | ,1 | ,1 | 1,0 |
| 11 | 1 | ,1 | ,1 | 1,1 |
| 12 | 1 | ,1 | ,1 | 1,2 |
| 13 | 1 | ,1 | ,1 | 1,3 |
| 14 | 1 | ,1 | ,1 | 1,4 |
| 15 | 1 | ,1 | ,1 | 1,5 |
| 16 | 1 | ,1 | ,1 | 1,6 |
| 17 | 1 | ,1 | ,1 | 1,7 |
| 18 | 1 | ,1 | ,1 | 1,8 |
| 19 | 1 | ,1 | ,1 | 1,9 |
| 20 | 1 | ,1 | ,1 | 2,0 |
| 21 | 1 | ,1 | ,1 | 2,1 |
| 22 | 1 | ,1 | ,1 | 2,2 |
| 23 | 1 | ,1 | ,1 | 2,3 |
| 24 | 1 | ,1 | ,1 | 2,4 |
| 25 | 1 | ,1 | ,1 | 2,5 |
| 26 | 1 | ,1 | ,1 | 2,6 |
| 27 | 1 | ,1 | ,1 | 2,7 |
| 28 | 1 | ,1 | ,1 | 2,8 |
| 29 | 1 | ,1 | ,1 | 2,9 |
| 30 | 1 | ,1 | ,1 | 3,0 |
| 31 | 1 | ,1 | ,1 | 3,1 |
| 32 | 1 | ,1 | ,1 | 3,2 |
| 33 | 1 | ,1 | ,1 | 3,3 |
| 34 | 1 | ,1 | ,1 | 3,4 |
| 35 | 1 | ,1 | ,1 | 3,5 |
| 36 | 1 | ,1 | ,1 | 3,6 |
| 37 | 1 | ,1 | ,1 | 3,7 |
| 38 | 1 | ,1 | ,1 | 3,8 |
| 39 | 1 | ,1 | ,1 | 3,9 |
| 40 | 1 | ,1 | ,1 | 4,0 |
| 41 | 1 | ,1 | ,1 | 4,1 |
| 42 | 1 | ,1 | ,1 | 4,2 |
| 43 | 1 | ,1 | ,1 | 4,3 |
| 44 | 1 | ,1 | ,1 | 4,4 |
| 45 | 1 | ,1 | ,1 | 4,5 |
| 46 | 1 | ,1 | ,1 | 4,6 |
| 47 | 1 | ,1 | ,1 | 4,7 |
| 48 | 1 | ,1 | ,1 | 4,8 |
| 49 | 1 | ,1 | ,1 | 4,9 |

**Abbildung 6.5:** Häufigkeitstabelle für sehr viele (1.000) Merkmalsausprägungen einer Variablen *v1* in *DATEN1_1000.SAV* (Output-Auszug)

# *Zur Fehler- und Ausreißeranalyse*

Eine Häufigkeitstabelle beziehungsweise Häufigkeiten ganz allgemein können in Bezug auf Ihr Datenmaterial sowohl für eine Fehler- als auch für eine Ausreißeranalyse genutzt werden, wie im Folgenden kurz gezeigt wird.

## *Fehleranalyse*

Häufigkeitsauszählungen können also verwendet werden, um **Fehleingaben** zu überprüfen.

Fügen Sie zu diesem Zweck einfach mal in den Datensatz *KORBTREFFER_F.SAV* an den beiden Stellen mit den fehlenden Werten jeweils den Wert 6 ein (ich habe das Ganze für Sie als *KORBTREFFER_F2.SAV* abgespeichert). Wenn Sie sich jetzt über das SPSS-Menü in bewährter Weise eine Häufigkeitstabelle ausgeben lassen, sieht das auf den ersten Blick recht schlüssig aus (siehe Abbildung 6.6).

**Häufigkeiten**

[DataSet4] C:\SPSS-Dateien\korbtreffer_f2.sav

| Statistiken | | |
|---|---|---|
| KORBTREFFER | | |
| N | Gültig | 10 |
| | Fehlend | 0 |

| KORBTREFFER | | | | | |
|---|---|---|---|---|---|
| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
| Gültig | 0 | 2 | 20,0 | 20,0 | 20,0 |
| | 1 | 1 | 10,0 | 10,0 | 30,0 |
| | 2 | 2 | 20,0 | 20,0 | 50,0 |
| | 4 | 1 | 10,0 | 10,0 | 60,0 |
| | 5 | 2 | 20,0 | 20,0 | 80,0 |
| | 6 | 2 | 20,0 | 20,0 | 100,0 |
| | Gesamt | 10 | 100,0 | 100,0 | |

**Abbildung 6.6:** Fehleranalyse anhand einer Häufigkeitstabelle (*KORBTREFFER_F2.SAV*)

Aber hoppla! Kann es die Ausprägung »6 Treffer« überhaupt geben? Jede Person wirft doch im Beispiel nur fünfmal auf den Korb, und selbst

Dirk Nowitzki dürfte Schwierigkeiten haben, bei fünf Würfen sechsmal zu treffen. Hier liegt offensichtlich eine nicht mögliche Merkmalsausprägung und damit ein Eingabefehler vor. Dies ist ein Beispiel dafür, einen Fehler in Ihrem Datensatz anhand einer Häufigkeitstabelle zu erkennen.

Möchten Sie – aufgrund einer vorangegangenen Fehlerbetrachtung – fehlerhafte Fälle korrigieren, so können Sie das recht bequem bewerkstelligen. Hierzu müssen Sie nur in der Datenansicht die betreffenden Variablen markieren. Bei sehr, sehr vielen Variablen können Sie über BEARBEITEN|GEHE ZU VARIABLE… die gewünschte Variable relativ leicht finden. In Abbildung 6.7 wird die Variable *korbtreffer* im Datensatz *KORBTREFFER_F2.SAV* gesucht.

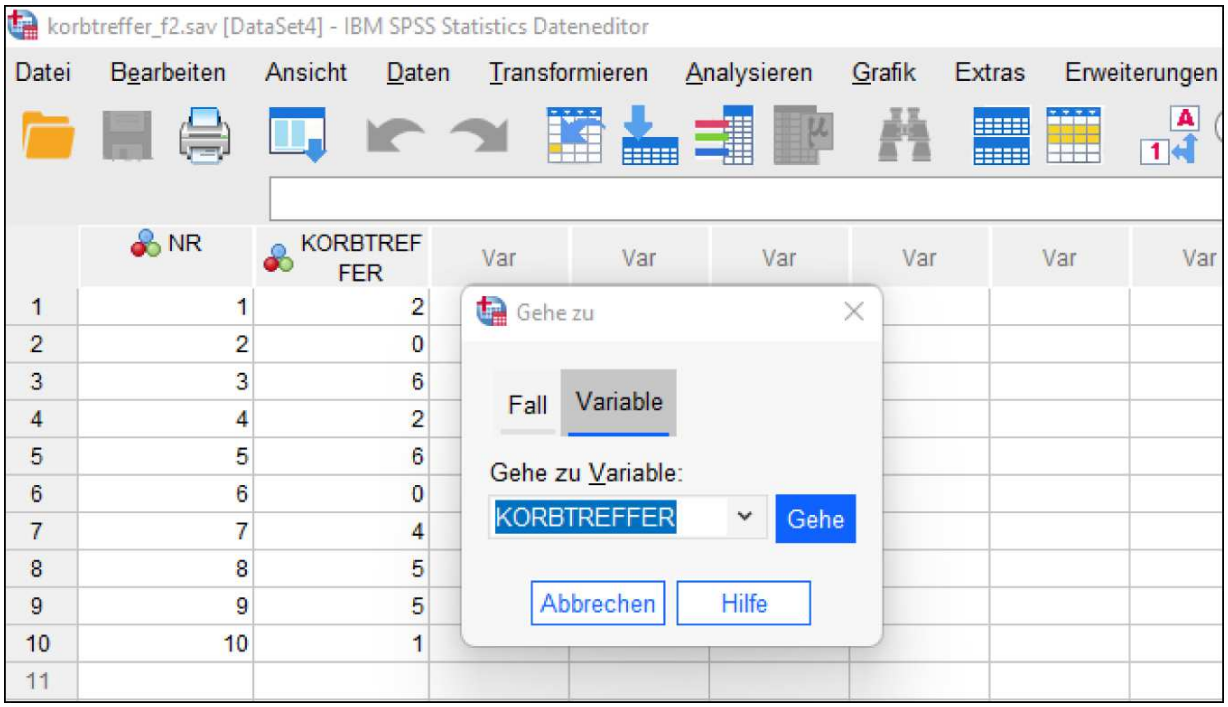



**Abbildung 6.7:** Suche nach einer bestimmten Variablen in einem Datensatz (*KORBTREFFER_F2.SAV*)

Anschließend können Sie für diese Variable über BEARBEITEN|SUCHEN… (oder über den Shortcut Strg+F) einen bestimmten Wert suchen. Suchen Sie im Übrigen nach systemfehlenden (SYSMIS-)Werten, müssen Sie hierbei einfach einen Punkt eingeben (weil SPSS systemfehlende Werte durch einen Punkt symbolisiert).

Allerdings setzt das Ersetzen der alten (beziehungsweise der fehlenden) Werte durch einen neuen Wert selbstverständlich voraus, dass Ihr neuer Wert vernünftig zustande gekommen ist, also nicht nur auf Zuruf oder per Würfeln.

In Abbildung 6.8 wird nach dem Wert »6« gesucht, da dieser – wie oben dargelegt – ja fehlerhaft ist. Dieser Wert kann jetzt durch einen neuen, realistischen Wert ersetzt werden. Wenn Sie keinen solchen realistischen Wert zur Hand haben, sollten Sie in die entsprechenden Tabellenzellen nichts einsetzen, also die betreffenden Tabellenzellen mit »fehlend« belegen.

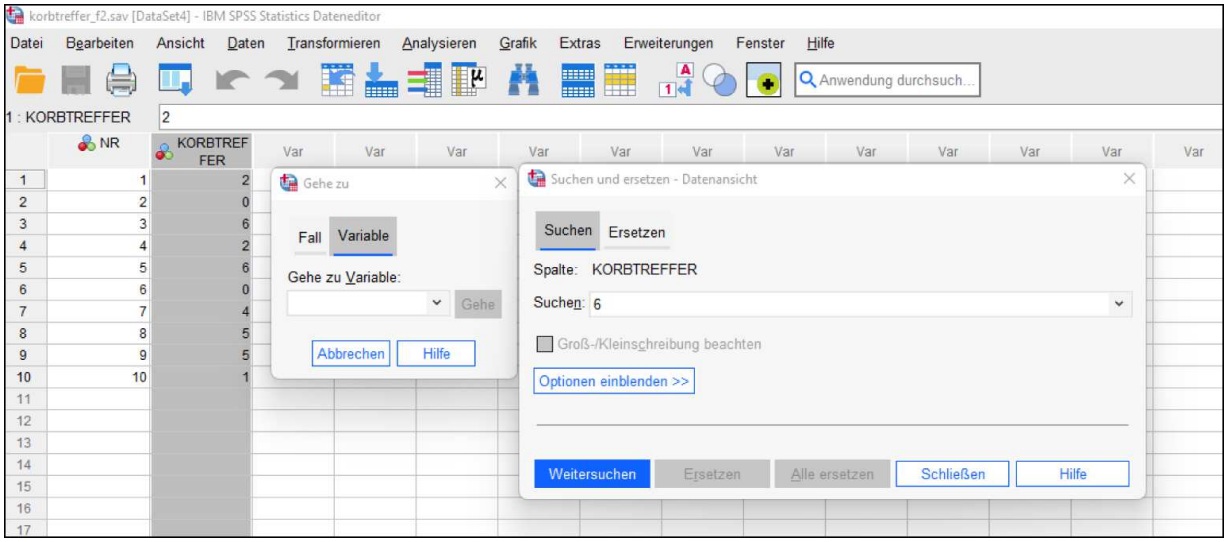



**Abbildung 6.8:** Suche nach einem bestimmten Wert einer bestimmten Variablen in einem Datensatz (*KORBTREFFER_F2.SAV*)

## *Ausreißeranalyse*

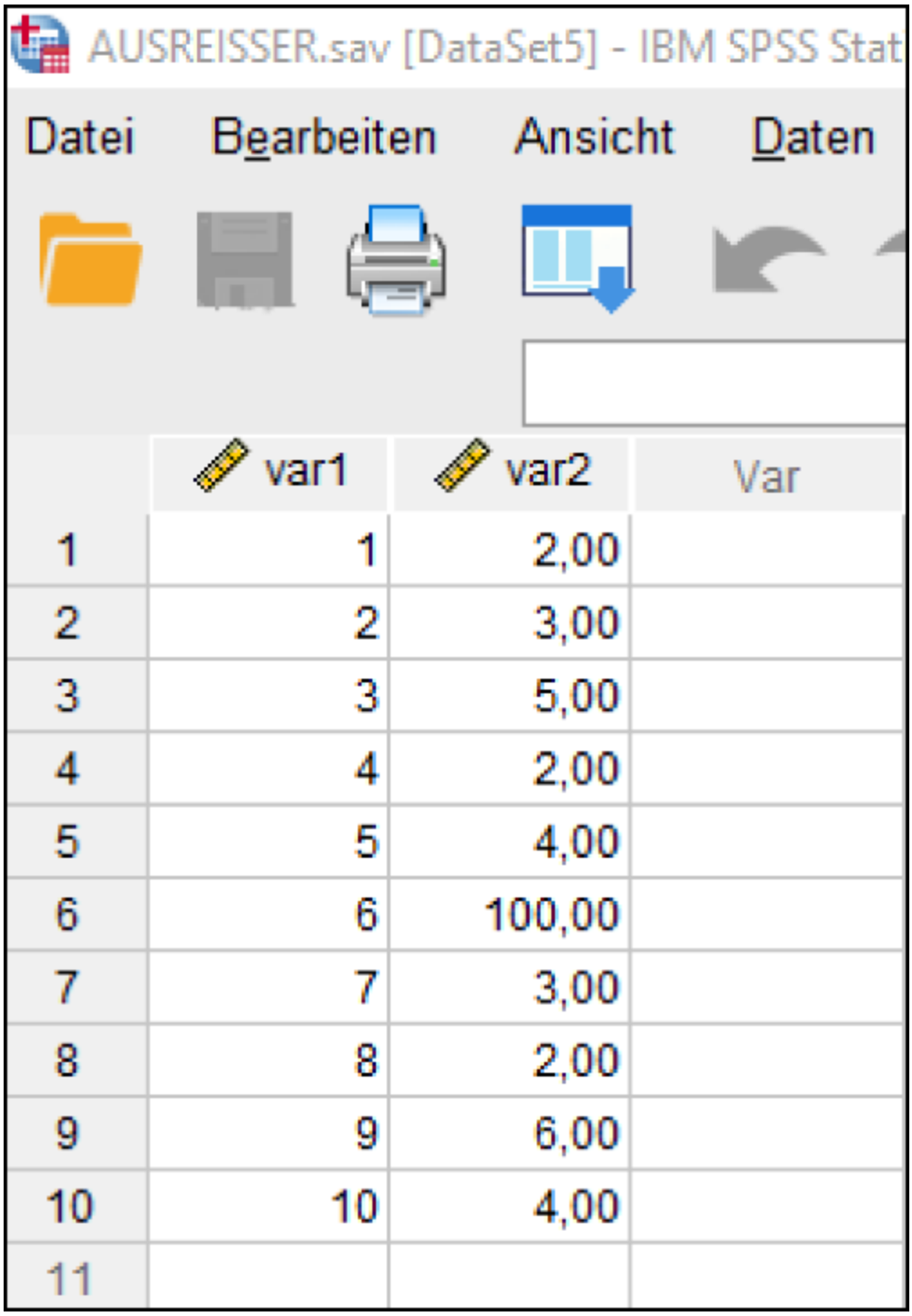
Zur **Ausreißer-Überprüfung** eignet sich zwar meines Erachtens die explorative Datenanalyse via ANALYSIEREN|DESKRIPTIVE STATISTIKEN|EXPLORATIVE DATENANALYSE am besten; gleichwohl können Sie bereits anhand einer Häufigkeitstabelle Hinweise auf extreme Merkmalsausprägungen erhalten.

**Ausreißer** können als Ursache haben:

- ✔ Erhebungs- beziehungsweise Erfassungsfehler,

- ✔ Besonderheiten der Merkmalsträger (beispielsweise eine Befragung eines Teils der Stichprobenteilnehmer nach deren Einschätzung der Terrorgefahr am Tage eines Terroranschlags).

Um Ihnen die Grundproblematik von Ausreißerwerten näherzubringen, habe ich die Datendatei *AUSREISSER.SAV* erstellt (siehe Abbildung 6.9). Bei der Variablen *var2* tanzt der sechste Fall aus der Reihe. Während sich die Ausprägungen von *var2* in allen anderen Fällen im Wertebereich von 2,00 bis 6,00 bewegen, ist beim sechsten Fall als Ausprägung von *var2* 100,00 angegeben.




| | var1 | var2 | Var |
|---|---|---|---|
| 1 | 1 | 2,00 | |
| 2 | 2 | 3,00 | |
| 3 | 3 | 5,00 | |
| 4 | 4 | 2,00 | |
| 5 | 5 | 4,00 | |
| 6 | 6 | 100,00 | |
| 7 | 7 | 3,00 | |
| 8 | 8 | 2,00 | |
| 9 | 9 | 6,00 | |
| 10 | 10 | 4,00 | |
| 11 | | | |

**Abbildung 6.9:** »Ausreißer-Datensatz« *AUSREISSER.SAV*

In diesem Fall mit wenigen Ausprägungen können Sie sich wieder eine Häufigkeitstabelle anzeigen lassen und werden vermutlich schon auf den

ersten Blick den Ausreißer erkennen. Aus Abbildung 6.10 ist der Ausreißerwert »100« sofort ersichtlich.

Liegen hingegen weitaus mehr Variablenausprägungen vor, empfiehlt sich das eingangs zu diesem Unterabschnitt bereits erwähnte Vorgehen über ANALYSIEREN|DESKRIPTIVE STATISTIKEN|EXPLORATIVE DATENANALYSE. Hierbei müssen Sie die zu untersuchenden Variablen der Liste der abhängigen Variablen hinzufügen. Unter STATISTIK steht Ihnen in diesem Zusammenhang außerdem AUSREISSER zur Verfügung (siehe Abbildung 6.11).

**Häufigkeiten**

[DataSet5] C:\SPSS-Dateien\AUSREISSER.sav

**Statistiken**

var2

| | | |
|---|---|---|
| N | Gültig | 10 |
| | Fehlend | 0 |

**var2**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | 2,00 | 3 | 30,0 | 30,0 | 30,0 |
| | 3,00 | 2 | 20,0 | 20,0 | 50,0 |
| | 4,00 | 2 | 20,0 | 20,0 | 70,0 |
| | 5,00 | 1 | 10,0 | 10,0 | 80,0 |
| | 6,00 | 1 | 10,0 | 10,0 | 90,0 |
| | 100,00 | 1 | 10,0 | 10,0 | 100,0 |
| | Gesamt | 10 | 100,0 | 100,0 | |

**Abbildung 6.10:** Häufigkeitsauszählung in SPSS mit Ausreißerwert »100« (*AUSREISSER.SAV*)

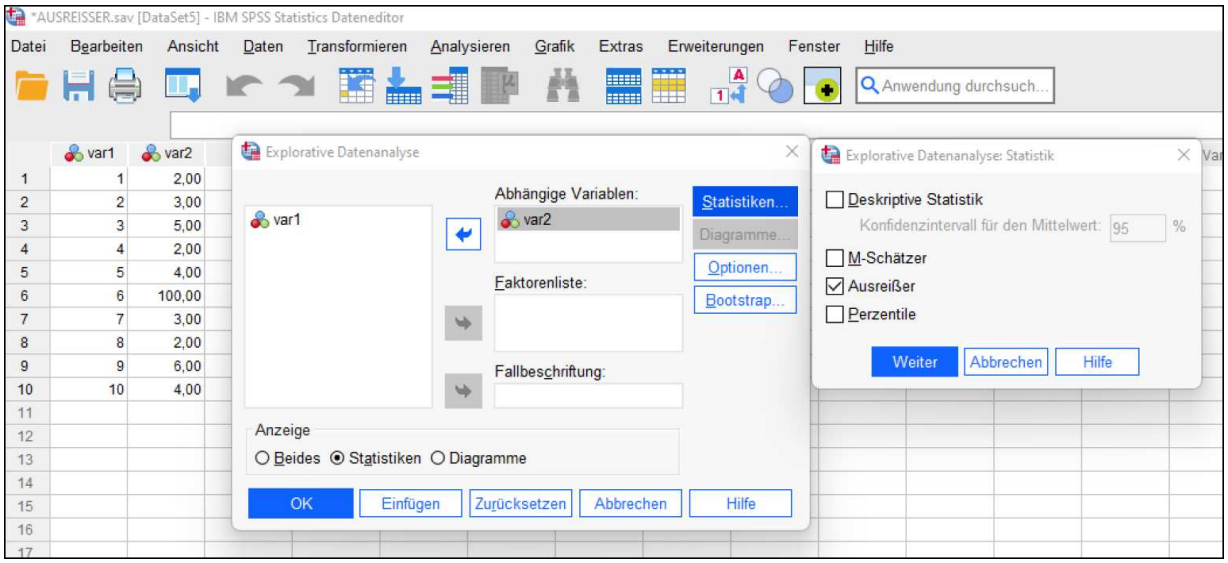



**Abbildung 6.11:** Starten einer Ausreißer-Analyse mithilfe von EXPLORATIVE DATENANALYSE|STATISTIK (für *AUSREISSER.SAV* )

Haben Sie die vorstehenden Klicks gemäß Abbildung 6.11 gesetzt (und zudem bei ANZEIGE in der Dialogbox in der Mitte auf STATISTIKEN geklickt und in der rechten Dialogbox AUSREISSER angeklickt), erhalten Sie die in Abbildung 6.12 angegebene Auswertung. Es werden grundsätzlich bei einer solchen SPSS-Analyse die fünf größten und die fünf kleinsten Werte mit entsprechenden Fallzuordnungen ausgewiesen. (Da unser Datensatz zufällig genau zehn Fälle aufweist, werden also in unserem Beispiel alle Fälle gelistet.) Auch hierbei sticht der Wert 100 ins Auge und kann somit als Ausreißer identifiziert werden.

**Explorative Datenanalyse**

**Verarbeitete Fälle**

| | Fälle | | | | | |
|---|---|---|---|---|---|---|
| | Gültig | | Fehlend | | Gesamt | |
| | N | Prozent | N | Prozent | N | Prozent |
| var2 | 10 | 100,0% | 0 | 0,0% | 10 | 100,0% |

**Extremwerte**

| | | | Fallnummer | Wert |
|---|---|---|---|---|
| var2 | Größte Werte | 1 | 6 | 100,00 |
| | | 2 | 9 | 6,00 |
| | | 3 | 3 | 5,00 |
| | | 4 | 5 | 4,00 |
| | | 5 | 10 | 4,00 |
| | Kleinste Werte | 1 | 8 | 2,00 |
| | | 2 | 4 | 2,00 |
| | | 3 | 1 | 2,00 |
| | | 4 | 7 | 3,00 |
| | | 5 | 2 | 3,00 |

**Abbildung 6.12:** Ergebnis einer tabellarischen Ausreißeranalyse mithilfe von EXPLORATIVE DATENANALYSE|STATISTIKEN (für *AUSREISSER.SAV*)

Alternativ hätten Sie in Abbildung 6.11 in der oberen Dialogbox anstelle von STATISTIK auf DIAGRAMME oder auf BEIDES klicken können. In diesem Fall weist Ihnen SPSS sowohl ein **Stem-Leaf-Diagramm** (auf Deutsch: Stamm-Blatt-Diagramm) als auch einen **Boxplot** aus (entweder separat oder zusammen mit STATISTIK-Angaben).

Ein *Stem-Leaf-Diagramm* ist im Grunde genommen eine etwas unkonventionell dargestellte **Häufigkeitstabelle** mit zunächst den

Häufigkeiten, dann den Werten und schließlich einer grafikähnlichen Darstellung der Häufigkeiten.

Bei einem *Boxplot* handelt es sich ebenfalls um die grafische Veranschaulichung einer Häufigkeitsverteilung. Ein Boxplot ist eine Rechteckdarstellung mit im Wesentlichen einer Untergrenze (sogenannter unterer Quartilswert; das heißt: der Wert an der Stelle von 25 Prozent bei der kumulierten relativen Häufigkeitsverteilung) und einer Obergrenze (sogenannter oberer Quartilswert; das heißt: der Wert an der Stelle von 75 Prozent bei der kumulierten relativen Häufigkeitsverteilung). Des Weiteren wird in das Rechteck der sogenannte Medianwert als waagerechte Linie eingezeichnet. Der Medianwert findet sich an der Stelle von 50 Prozent bei der kumulierten relativen Häufigkeitsverteilung. Schließlich weist ein Boxplot auch noch den größten und den kleinsten Wert der Merkmalsausprägungen aus – und damit auch etwaige Ausreißerwerte, sofern sie (deutlich) von dem Boxplot-Rechteck entfernt liegen.

Im vorstehenden Sinne finden Sie in Abbildung 6.13 sowohl ein Stem-Leaf-Diagramm als auch einen Boxplot für den Beispieldatensatz *AUSREISSER.SAV*. Bei dem entsprechenden Stem-Leaf-Diagramm werden zum einen die Häufigkeiten »quasi-grafisch« durch Nullen in der Spalte »Blatt« ganz rechts repräsentiert. Zum anderen – und das ist hier wichtiger – wird ausdrücklich darauf hingewiesen, dass in einem Fall ein Ausreißerwert vorliegt, was ja auch bei Fall Nummer 6 (mit der Ausprägung 100,00) so ist.

Besser ist dieser Ausreißerwert im Boxplot-Teil der Abbildung 6.13 zu erkennen. In der entsprechenden Grafik liegt ein Wert (eben der Wert 100,00) deutlich oberhalb des Rechtecks. Dabei gibt SPSS auch die zugehörige Fallnummer (Nummer 6) in dem betreffenden Diagramm an. Nachteilig an dem großen Abstand des Ausreißers von den anderen Werten ist in Abbildung 6.13 offensichtlich, dass das Boxplot-Rechteck ziemlich »gestaucht« ist und seine Kennziffern von ihrem Wert her kaum zu erkennen sind.

**var2**

var2 Stengel-Blatt-Diagramm

| Häufigkeit | Stem & | Blatt |
|---|---|---|
| 3,00 | 2 . | 000 |
| 2,00 | 3 . | 00 |
| 2,00 | 4 . | 00 |
| 1,00 | 5 . | 0 |
| 1,00 | 6 . | 0 |
| 1,00 Extremwerte (>=100,0) | | |

Stammbreite: 1,00
Jedes Blatt: 1 Fälle

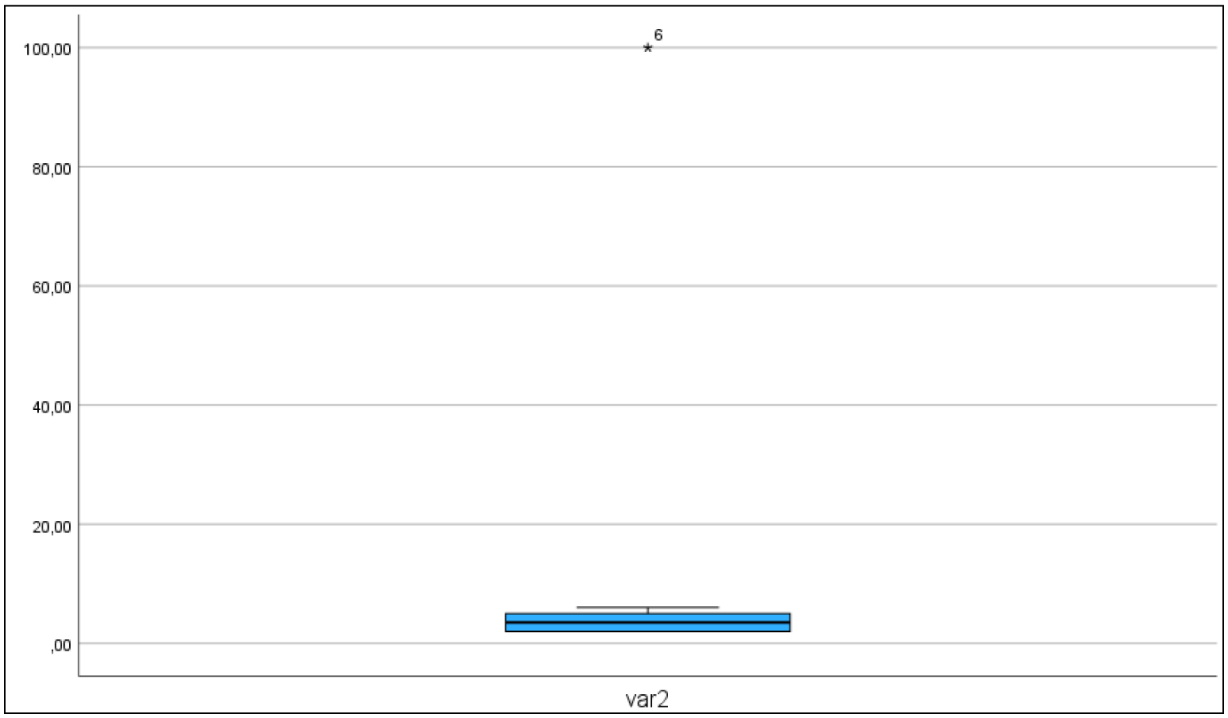



**Abbildung 6.13:** Ergebnis einer grafischen Ausreißer-Analyse mithilfe von EXPLORATIVE DATENANALYSE|DIAGRAMME (*AUSREISSER.SAV*)

Beim Entfernen von Ausreißerdaten müssen Sie vorsichtig sein: Sie sollten sich schon sehr sicher sein, dass diesbezüglich eine Irregularität vorliegt. Andernfalls könnten Sie sich dem Vorwurf ausgesetzt sehen, Ihre Daten zu Ihren Gunsten zu manipulieren, um die Ihnen »passenden« Ergebnisse zu erhalten. Wenn Sie sich für das Löschen eines Ausreißers entschieden haben, können Sie diesen als fehlenden Wert deklarieren (entweder als systemfehlend: SYSMIS oder als benutzerdefiniert) oder – sozusagen im Sinne der »großen Keule« – gleich den gesamten Fall aus Ihrem Datensatz löschen.

# *Häufigkeitsdiagramme*

Zur Beurteilung einer eindimensionalen Häufigkeitsverteilung ist auch eine Abbildung nicht schlecht. Manchmal erschließen sich dem Betrachter durch Grafiken Informationen besser als durch Tabellen, gerade bei vielen Merkmalsausprägungen.

In der Subdialogbox (von HÄUFIGKEITEN:) DIAGRAMME stehen Ihnen verschiedene Möglichkeiten für entsprechende **Diagramme** zur Verfügung:

- ✔ Balken-, Kreisdiagramme, Histogramme,
- ✔ Normalverteilungskurve und
- ✔ Streudiagramme.

Während ich in diesem Abschnitt im Folgenden auf die beiden ersten Diagrammarten näher eingehe, gilt dies nicht für die Streudiagramme. Der Grund dafür ist, dass diese Diagrammart insbesondere in Kapitel 9 bei der Regressionsanalyse eine wichtige Rolle spielt und daher erst dort behandelt wird. Ich bitte diesbezüglich daher noch um etwas Geduld – danke!

Die Schönheit eines Bilds liegt natürlich im Auge des Betrachters! Für mich sind die nachfolgend behandelten SPSS-Diagramme zwar nicht schlecht; sie sind aber meines Erachtens recht aufwendig zu bearbeiten. Daher präferiere ich es, meinen SPSS-Output zum Beispiel nach EXCEL zu überführen und dort die entsprechenden Diagramme zu erstellen und zu verändern. Vielleicht ist dies bei Ihnen aber anders! Dann sollten Sie die Ausführungen der nächsten Abschnitte ausführlich und gewissenhaft lesen.

## *Balken-, Kreisdiagramme, Histogramme*

Klicken Sie doch bitte im Zusammenhang mit der Datendatei *KORBTREFFER.SAV* mal »Balkendiagramme« an (siehe Abbildung 6.14), und Sie erhalten ein schönes Bildchen für unsere Korbwürfe in Form eines **Balkendiagramms** (siehe Abbildung 6.15). Um nicht schon wieder die zugehörige Häufigkeitstabelle dargestellt zu bekommen, schalten Sie bitte HÄUFIGKEITSTABELLE ANZEIGEN aus, was bei der SPSS-Syntax dem Unterbefehl /FORMAT=NOTABLE entspricht.

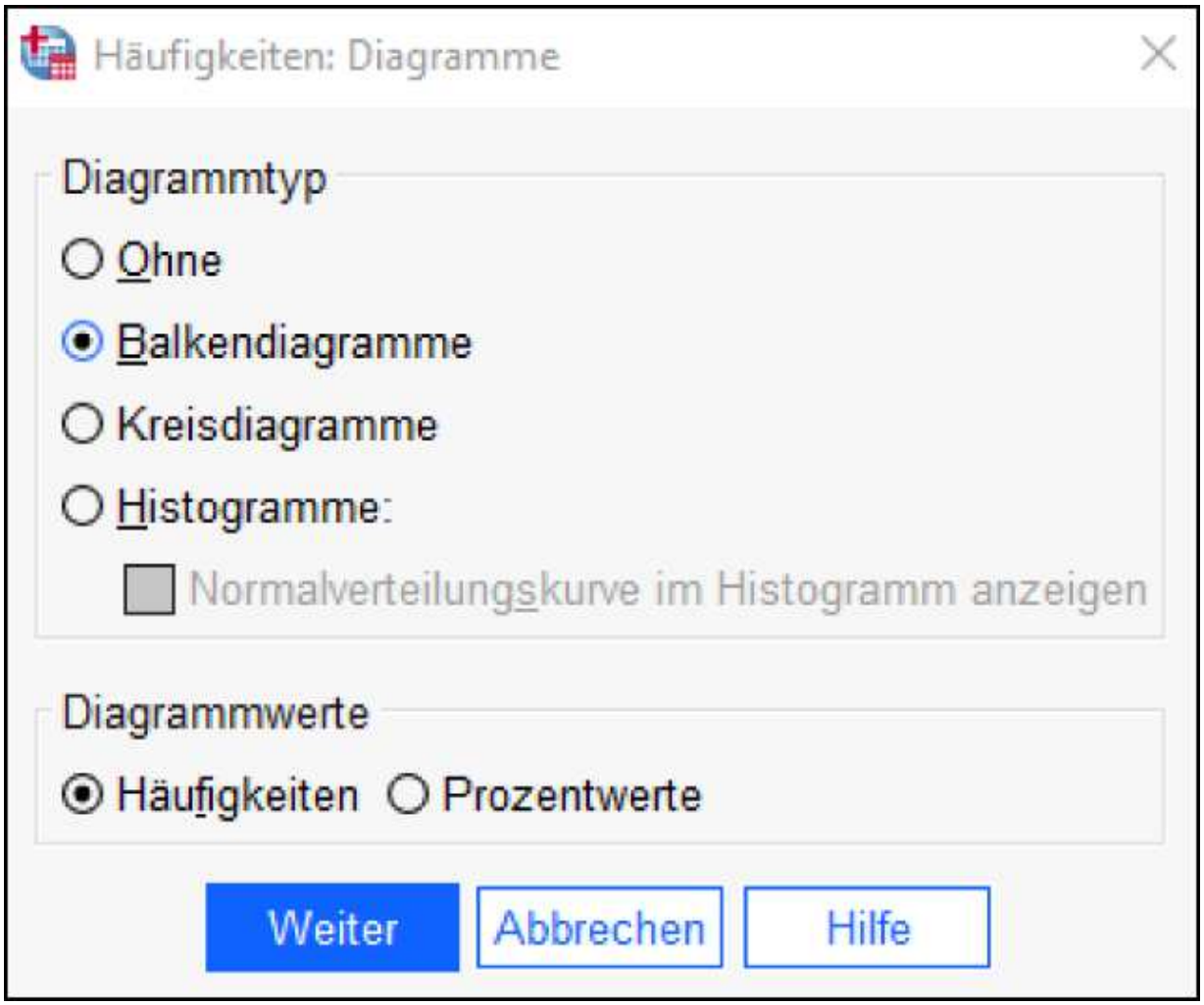



**Abbildung 6.14:** DIAGRAMME bei HÄUFIGKEITEN (in Bezug auf *KORBTREFFER.SAV*)

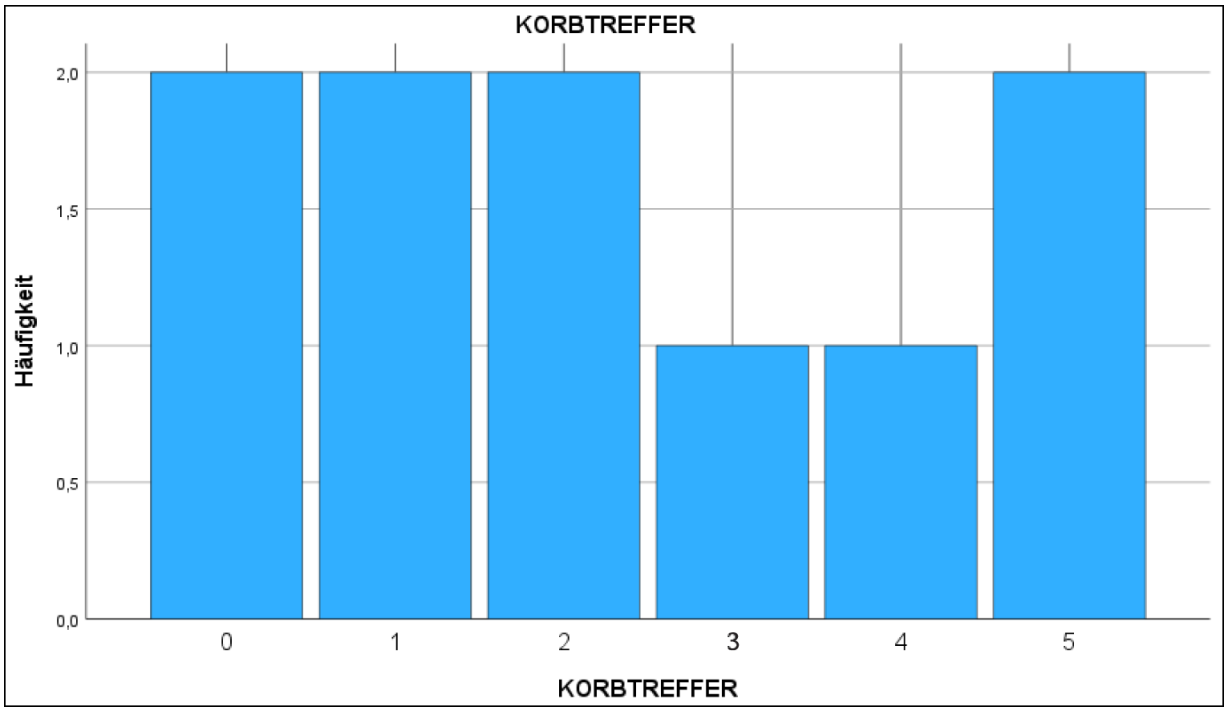



**Abbildung 6.15:** Balkendiagramm als Output (*KORBTREFFER.SAV*)

Das betreffende (Balken-)Bildchen erhalten Sie nämlich auch mittels SPSS-Syntax:

```
FREQUENCIES VARIABLES=korbtreffer
     /FORMAT=NOTABLE
     /BARCHART.
```

Wenn Sie kein großer Fan von Balkendiagrammen sind, sondern lieber ein **Kreisdiagramm** erstellen möchten, ist dies für das vorliegende Beispiel ganz leicht. Hierzu müssen Sie in der aus Abbildung 6.14 bekannten Subdialogbox nicht auf BALKENDIAGRAMME, sondern auf KREISDIAGRAMME klicken, und schwuppdiwupp präsentiert Ihnen SPSS im Viewer ein Kreisdiagramm (siehe Abbildung 6.16).

**Häufigkeiten**

| Statistiken | | |
|---|---|---|
| KORBTREFFER | | |
| N | Gültig | 10 |
| | Fehlend | 0 |

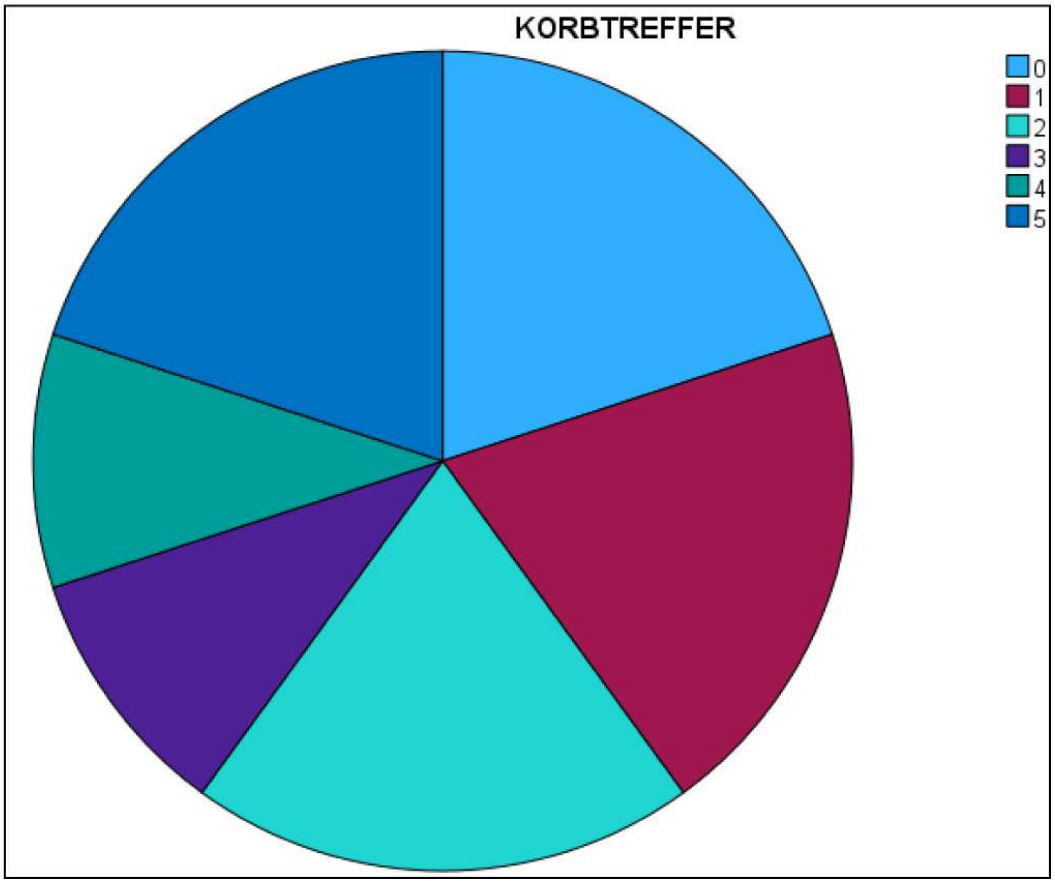



**Abbildung 6.16:** Kreisdiagramm als Output (*KORBTREFFER.SAV*)

Wenn Sie sich das vorstehende Kreisdiagramm lieber über ein kleines Syntaxprogramm erzeugen lassen möchten, müssen Sie in den Syntaxeditor Folgendes eingeben:

```
FREQUENCIES VARIABLES=korbtreffer
    /FORMAT=NOTABLE
    /PIECHART.
```

Gegenüber dem obigen, auf ein Balkendiagramm bezogenen Syntaxprogramm ist in der letzten Programmzeile lediglich »BARCHART« durch »PIECHART« ersetzt worden.

Liegen Ihre Daten klassifiziert vor, das heißt in Form verschiedener Klassen jeweils mit einer Klassenunter- und einer Klassenobergrenze, können Sie ein sogenanntes **Histogramm** erstellen. Die Balkenbreite ist dabei durch die Differenz aus jeweiliger Klassenobergrenze und Klassenuntergrenze gegeben. Interpretieren Sie doch einfach einmal – ungeachtet der (fehlenden) inhaltlichen Sinnhaftigkeit – für unser »Basketball-Beispiel« mit *KORBTREFFER.SAV* als Datengrundlage die einzelnen Ausprägungen (0 bis 5) als Bezeichnungen einzelner Klassen. Auf dieser (inhaltlich natürlich etwas fragwürdigen) Grundlage können Sie nun in der aus Abbildung 6.14 bekannten Dialogbox anstelle von auf BALKENDIAGRAMME nunmehr auf HISTOGRAMME klicken. In der Folge »wirft« Ihnen SPSS als Output in der Tat unter der Überschrift »Histogramm« eine entsprechende Grafik in der neuen Abbildung 6.17 »aus« (die in unserem speziellen Fall der Balkendiagramm-Darstellung gleicht, was an unserer inhaltlich fragwürdigen Gleichsetzung der Korbtreffer mit Klassenbezeichnungen liegt – aber seien Sie nicht allzu irritiert: Es geht an dieser Stelle nur ums Prinzip; außerdem gibt es im Buch noch Histogramm-Darstellungen auf der Basis »echter« Klassen, zum Beispiel Abbildung 14.5 in Kapitel 14 mit dann ebenfalls keinen Abständen auf der X-Achse zwischen den Klassengrenzen).

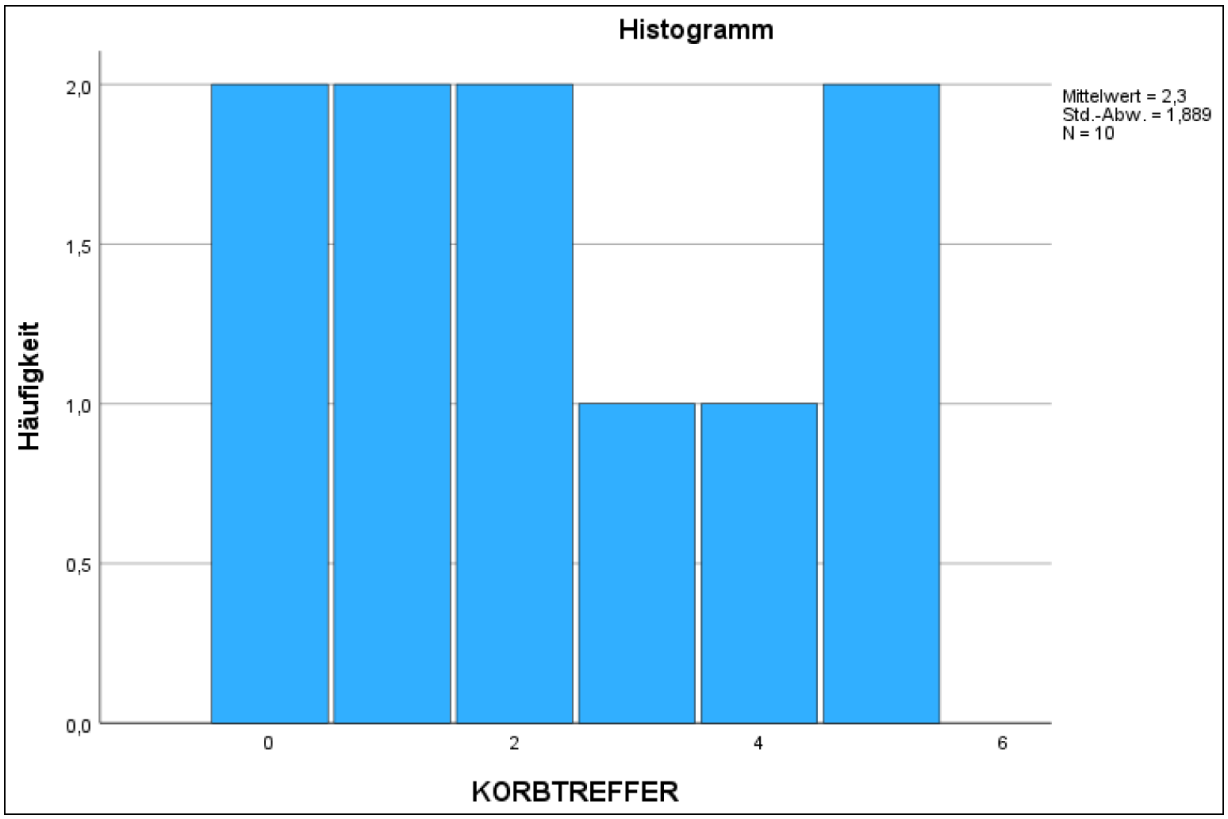



**Abbildung 6.17:** Histogramm als Output (*KORBTREFFER.SAV*)

Die Grafik in Abbildung 6.17 hätten Sie auch über einen SPSS-Syntaxbefehl erhalten, und zwar über:

```
FREQUENCIES VARIABLES=korbtreffer
     /FORMAT=NOTABLE
     /HISTOGRAM.
```

Liegen Klassen mit unterschiedlicher Breite vor, so sollten Sie diese Unterschiede bei den relativen Häufigkeiten durch Normierungen ausgleichen: Man dividiert dann üblicherweise die relativen Häufigkeiten durch die Breite der jeweiligen Klasse. Dabei ist die Breite einer Klasse als Differenz aus der Klassenobergrenze und der Klassenuntergrenze definiert.

Alternativ können Sie eine bestimmte Klassenbreite als Normierungsgrundlage verwenden und die (normierten) relativen Häufigkeiten von Klassen mit anderer Klassenbreite an diese »Basis-Klassenbreite« anpassen. Haben Sie zum Beispiel zwei

Klassen »0 bis unter 200 Euro« und »200 bis unter 600 Euro« mit einer relativen Häufigkeit von jeweils 50 Prozent vorliegen, können Sie die Klassenbreite der unteren Klasse (= 200 Euro) zur Normierung verwenden. Sie können nunmehr die normierte relative Häufigkeit für diese untere Klasse bei 50 Prozent belassen und diejenige der oberen Klasse wegen der doppelten Klassenbreite auf 25 Prozent halbieren. Alternativ können Sie nach dem ersten genannten Normierungsverfahren die relativen Häufigkeiten der beiden Klassen einfach durch die jeweilige Klassenbreite dividieren und erhalten hierdurch für die untere Klasse als normierte relative Häufigkeit $\frac{0{,}50}{200} = 0{,}0025$ und für die obere Klasse $\frac{0{,}50}{400} = 0{,}00125.$ Aus Darstellungsgründen können Sie diese Werte noch mit einem konstanten Faktor – zum Beispiel mit 1.000 – multiplizieren.

SPSS weist derartige normierte relative Häufigkeiten leider nicht aus, sodass Sie hier auf eigene Berechnungen dieser normierten relativen Häufigkeiten angewiesen sind. Diese Problematik können Sie zum Beispiel derart bewältigen, dass Sie die betreffende SPSS-Häufigkeitstabelle nach EXCEL kopieren und dort entsprechend bearbeiten.

## *Normalverteilungskurve*

Im Zusammenhang mit der Erstellung eines Histogramms (siehe hierzu die Ausführungen im vorstehenden Unterabschnitt) können Sie sich zusätzlich zu dem Histogramm von SPSS auch noch eine Normalverteilungskurve in das betreffende Diagramm einzeichnen lassen. Solcherart können Sie per Augenschein beurteilen, ob Ihre empirische Merkmalsverteilung zumindest näherungsweise einer **Normalverteilung** – umgangssprachlich: einer »Glockenkurve« – folgt. Diese Beurteilungen sind gerade dann wichtig, wenn Sie für Ihre statistischen Operationen (wie zum Beispiel beim statistischen Testen, das weiter hinten im Buch in Teil IV behandelt wird) auf die Normalverteilungsannahme angewiesen sind.

Sie können sich die Normalverteilungskurve in Ihr Diagramm aber nur dann einzeichnen lassen, wenn Sie anstelle eines Balken- oder Kreisdiagramms ein Histogramm als Diagrammform ausgewählt haben.

Um sich nun auch noch die Normalverteilungskurve in Ihr Diagramm, das heißt in Ihr Histogramm, einzeichnen zu lassen, klicken Sie in der Subdialogbox DIAGRAMME zusätzlich zu HISTOGRAMME auf NORMALVERTEILUNGSKURVE IM HISTOGRAMM ANZEIGEN, wie in Abbildung 6.18 dargelegt.

In unserem Fall mit *KORBTREFFER.SAV* als Datenbasis erhalten Sie dann den in Abbildung 6.19 angegebenen Output. Wie ersichtlich, verläuft das Histogramm nicht sonderlich eng an der Normalverteilungskurve, weist also nicht gerade einen glockenförmigen Verlauf auf. Na ja, man kann nicht alles haben!

Die zugehörige SPSS-Syntax sieht so aus (auch hier ist die Anzeige der Häufigkeitstabelle ausgeschaltet):

```
FREQUENCIES VARIABLES=korbtreffer
     /FORMAT=NOTABLE
     /HISTOGRAM NORMAL.
```

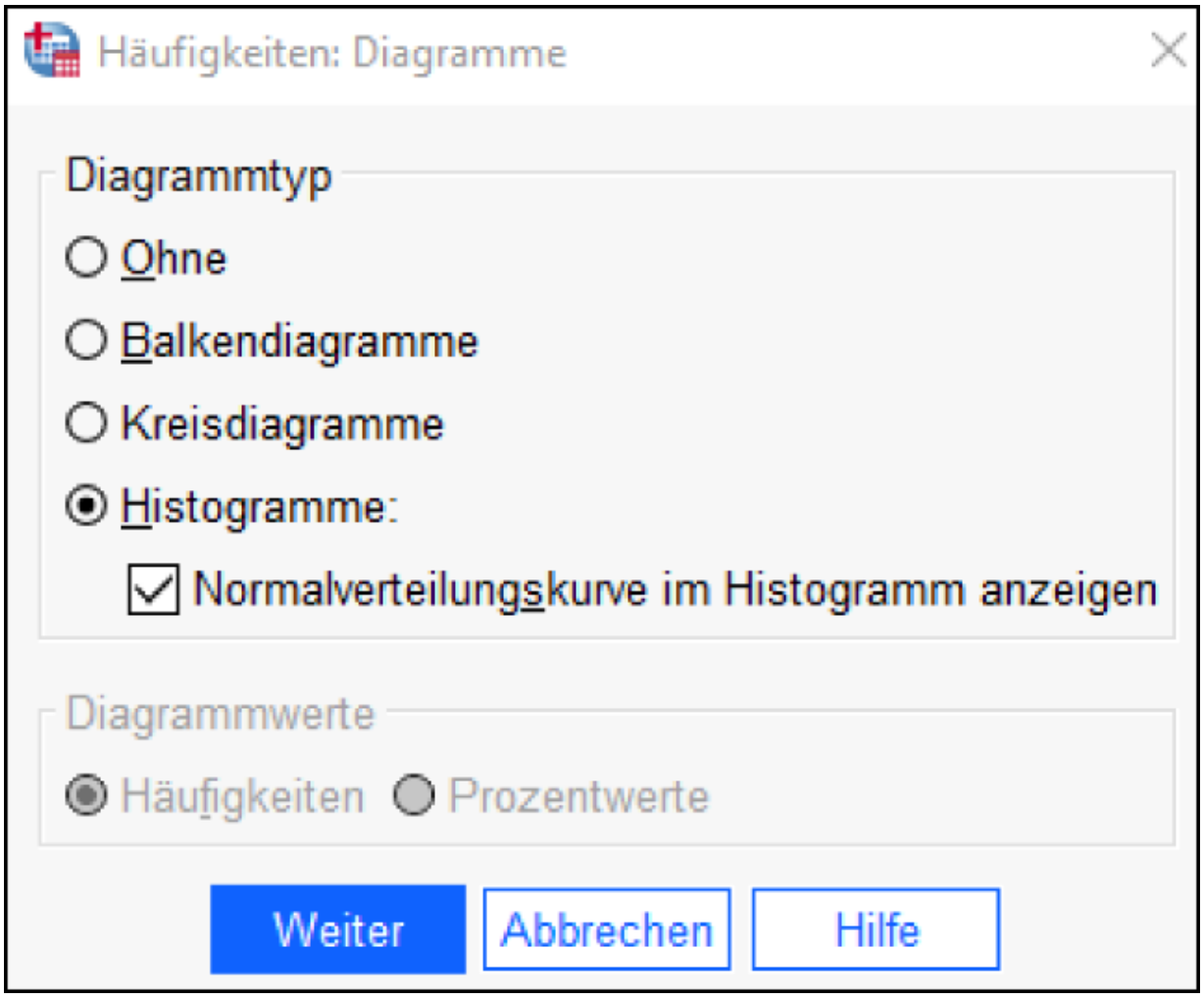



**Abbildung 6.18:** Normalverteilungskurve anzeigen lassen (hier in Bezug auf *KORBTREFFER.SAV*)

**Statistiken**

KORBTREFFER

| | | |
|---|---|---|
| N | Gültig | 10 |
| | Fehlend | 0 |

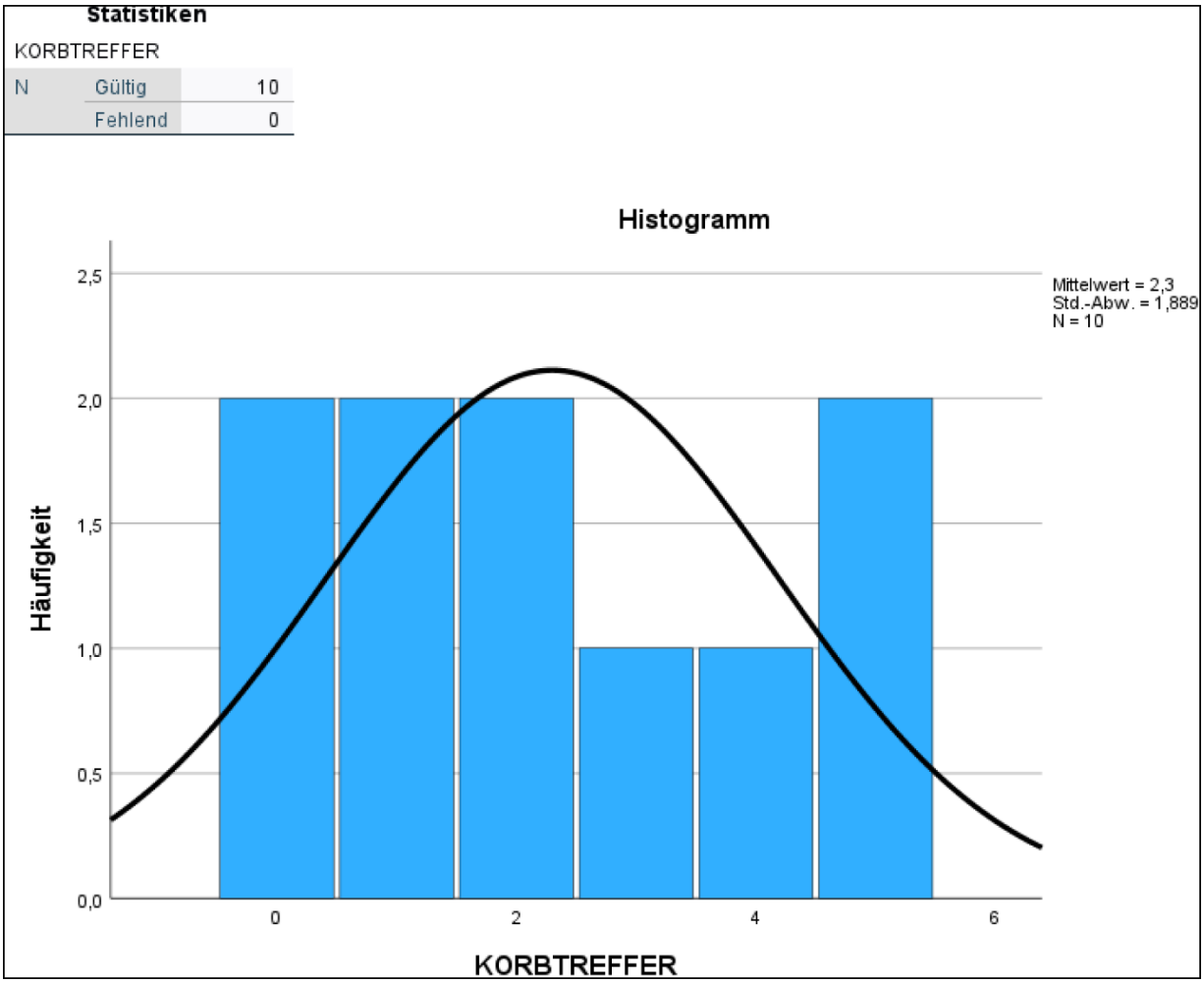



**Abbildung 6.19:** Histogramm und Normalverteilungskurve als Output (*KORBTREFFER.SAV*)

# Kapitel 7

# Viele Wege führen nach Rom: Eindimensionale deskriptive Statistiken

## IN DIESEM KAPITEL

Unterschiede zwischen statistischen Prozeduren und statistischen Funktionen

Art der statistischen Kennziffern bei den einzelnen Statistikprozeduren

Häufigkeits-Prozedur (FREQUENCIES)

DESCRIPTIVES-Prozedur

Explorative Datenanalyse (EXAMINE)

Mittelwertvergleichs-Prozedur (MEANS)

Fallzusammenfassungs-Prozedur (SUMMARIZE)

Zur Beurteilung der Datenqualität sind vielfach ein paar »knackige« statistische Kennziffern wie Mittelwerte oder Streuungsangaben hilfreich. In SPSS gibt es hierbei an vielen Stellen die Möglichkeit, entsprechende Angaben zu erhalten, wie in diesem Kapitel gezeigt wird. Diese Möglichkeiten ergeben sich über verschiedene **statistische Prozeduren**. Von diesen statistischen Prozeduren sind die statistischen Funktionen in SPSS scharf abzugrenzen, wie nachfolgend gezeigt wird.

## *Unterschiede zwischen statistischen Prozeduren und statistischen Funktionen*

Zwischen den **statistischen Funktionen** in SPSS und den SPSS-Statistikprozeduren besteht ein grundsätzlicher Unterschied. Mit einer Statistikprozedur wie zum Beispiel DESCRIPTIVES (siehe unten in diesem Kapitel) wird *über alle Fälle hinweg* für eine Variable etwa das arithmetische Mittel berechnet. Demgegenüber berechnen Sie mit einer statistischen Funktion wie MEAN *für einen Fall* über mehrere Variablen hinweg den entsprechenden Mittelwert aus den Ausprägungen der betreffenden Variablen. Während sich die Statistikprozeduren also sozusagen senkrecht durch die Datenmatrix durcharbeiten, tun dies die statistischen Funktionen demnach sozusagen waagerecht.

Um die Wirkweise statistischer Funktionen zu zeigen, wird von dem nachfolgenden »Basketball-Beispiel« ausgegangen, das Sie grundsätzlich vielleicht schon aus Kapitel 6 kennen (siehe hier jetzt die Datendatei *KORBTREFFER4.SAV* ). In diesem Beispiel sind die

Basketball-Wurfergebnisse von zehn Leuten in den früheren Jahren 2014, 2015 und 2016 angegeben, und zwar bei jeweils fünf Würfen in jedem Jahr (siehe Tabelle 7.1).

| *Person Nr.* | *Korbtreffer 2014* | *Korbtreffer 2015* | *Korbtreffer 2016* |
|---|---|---|---|
| 1 | 2 | 2 | 3 |
| 2 | 0 | 1 | 1 |
| 3 | 3 | 2 | 3 |
| 4 | 2 | 4 | 5 |
| 5 | 1 | 1 | 1 |
| 6 | 0 | 1 | 2 |
| 7 | 4 | 0 | 0 |
| 8 | 5 | 3 | 2 |
| 9 | 5 | 4 | 3 |
| 10 | 1 | 2 | 3 |

**Tabelle 7.1**: Statistische Funktionen: erweitertes »Basketball-Beispiel«

Lassen Sie uns jetzt für dieses Beispiel und für die neue Variable *korbtreffer2014_2016* für jeden Fall den arithmetischen Mittelwert berechnen. Hierzu müssen Sie zunächst im Menü TRANSFORMIEREN|VARIABLE BERECHNEN auf der linken Seite bei NEUE VARIABLE Ihre neue Variable *korbtreffer2014_2016* angeben. Dann müssen Sie in die Funktionsgruppe STATISTISCH gehen. Dort wählen Sie bitte von den angegebenen Funktionen MEAN aus, wie in Abbildung 7.1 angedeutet.

Der arithmetische Mittelwert ist als Durchschnitt definiert, und zwar als Merkmalssumme dividiert durch die Fallzahl. (Zum Beispiel: Für die fünf Ausprägungen einer bestimmten Variablen 1, 2, 4, 7 und 9 beträgt das arithmetische Mittel $\frac{1+2+4+7+9}{5} = 4,6.$)

Im nächsten Schritt geben Sie bitte unter NUMERISCHER AUSDRUCK mittels einer Formel ein, dass Sie das arithmetische Mittel aus den Variablen *korbtreffer2014, korbtreffer2015* und *korbtreffer2016* berechnen möchten, also

$$korbtreffer2014_2016 = MEAN(korbtreffer2014, korbtreffer2015, korbtreffer2016).$$

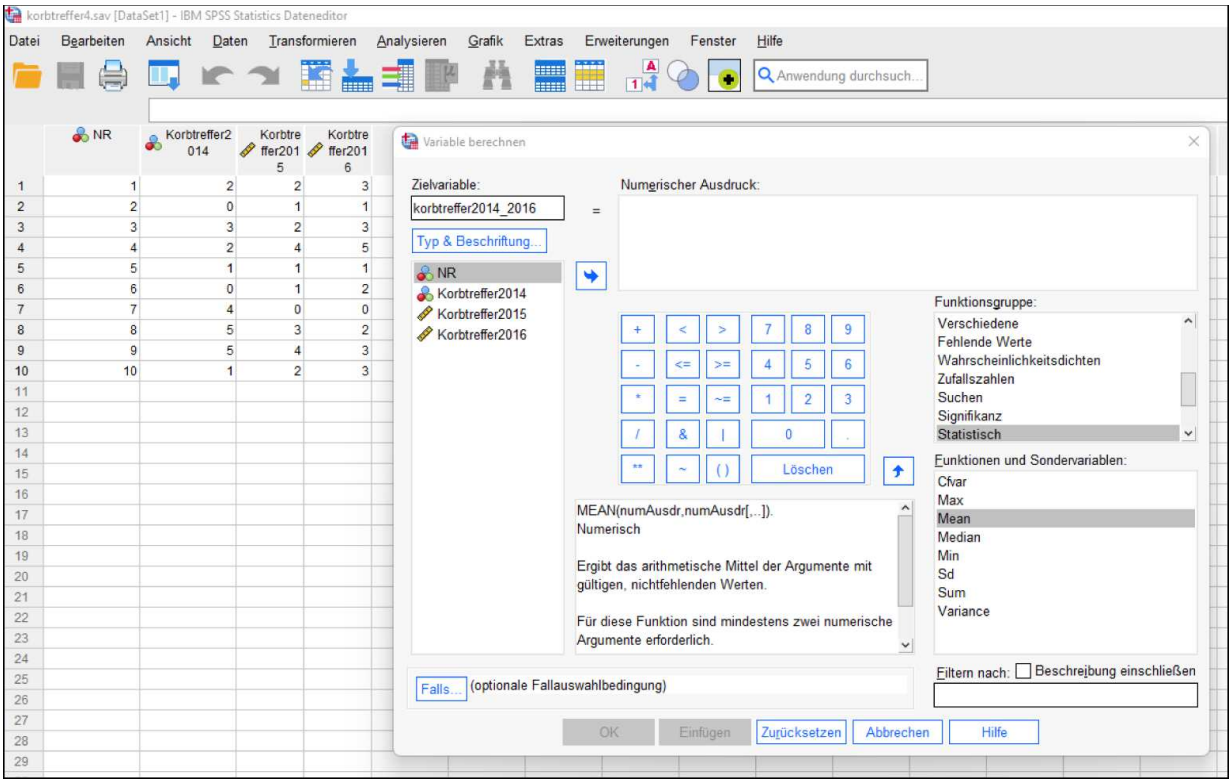



**Abbildung 7.1:** Statistische Funktion MEAN (für *KORBTREFFER4.SAV*)

Dies ist in Abbildung 7.2 dargestellt.

Wenn Sie nun erwartet haben, dass Ihre Mittelwerte in einem Outputfile dargestellt werden, sehen Sie sich getäuscht. Es wurde ja eine neue Variable definiert, deren Werte jeweils die Mittelwerte aus drei bestehenden Variablen sind. Daher werden Ihre Ergebnisse an dieser Stelle im Dateneditor in der Variablenspalte für *korbtreffer2014_2016* abgebildet, wie in Abbildung 7.3 angegeben. Für Fall 1 beispielsweise erhalten Sie also einen Durchschnittswert in Höhe von 2,33 Treffern.

Als statistische Funktionen stehen in SPSS in der entsprechenden Statistik-Funktionsgruppe des Weiteren zur Verfügung:

- CFVAR (Variationskoeffizient im Sinne des Verhältnisses aus Standardabweichung und arithmetischem Mittelwert),
- MAX (maximaler Wert),
- MEDIAN (genau in der Mitte der Verteilung liegender Wert),
- MIN (minimaler Wert),
- SD (Standardabweichung als Quadratwurzel aus der Varianz; zur Varianz-Erläuterung siehe die übernächste Funktion),
- SUM (Summe) und
- VARIANCE (Varianz als mittlere quadratische Abweichung der einzelnen Werte vom arithmetischen Mittelwert).

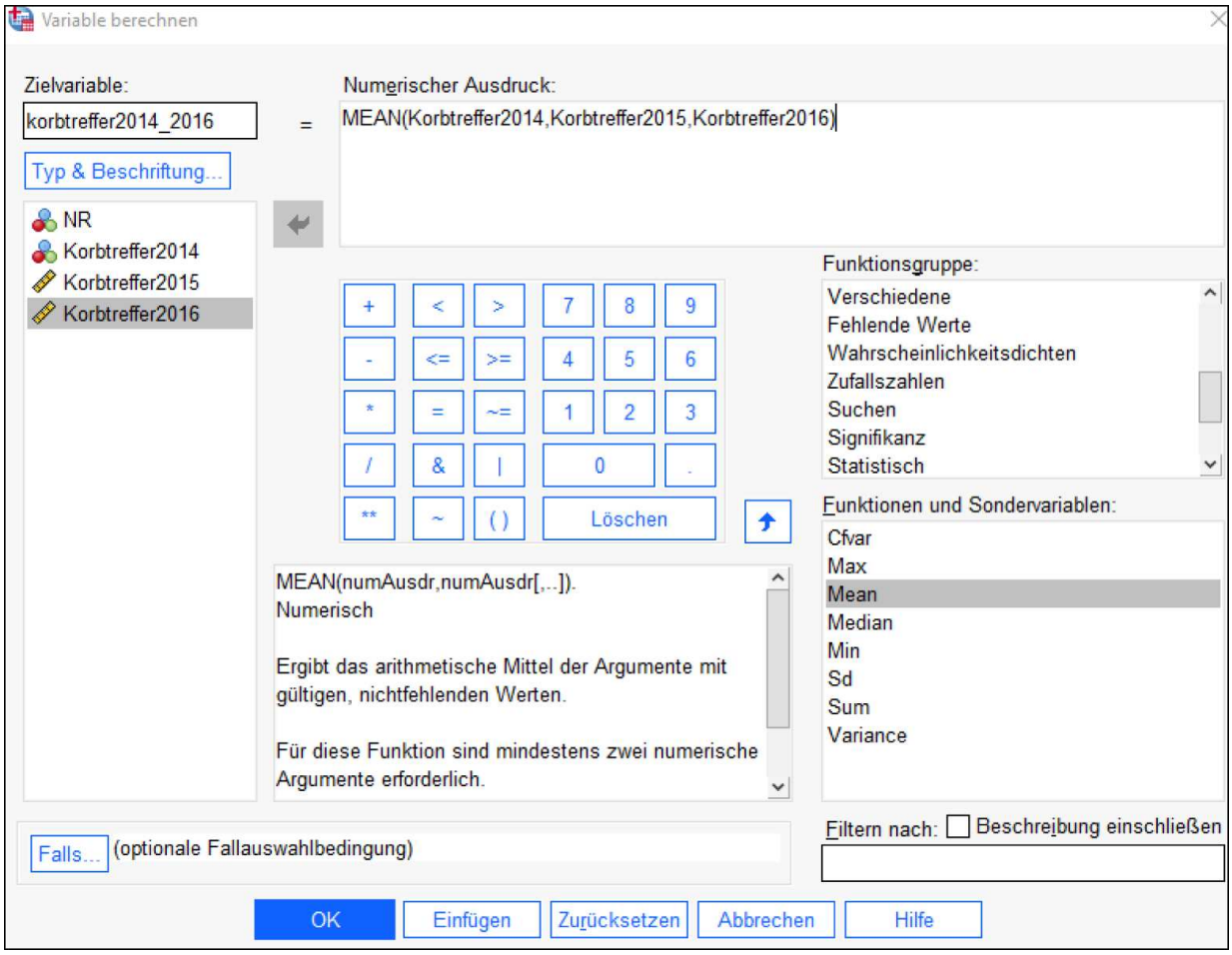
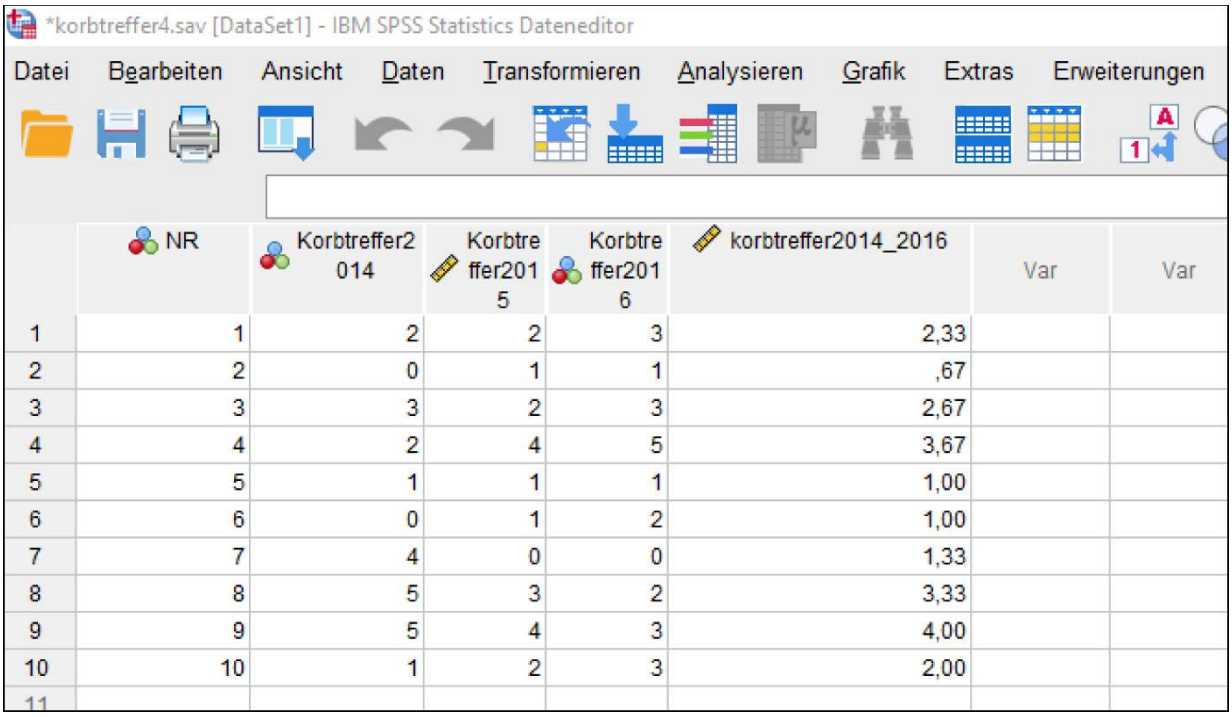



**Abbildung 7.2:** Anwendung der MEAN-Funktion auf die Variable *korbtreffer2014_2016* (*KORBTREFFER4.SAV*)




| | NR | Korbtreffer2014 | Korbtreffer2015 | Korbtreffer2016 | korbtreffer2014_2016 | Var | Var |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 2 | 3 | 2,33 | | |
| 2 | 2 | 0 | 1 | 1 | ,67 | | |
| 3 | 3 | 3 | 2 | 3 | 2,67 | | |
| 4 | 4 | 2 | 4 | 5 | 3,67 | | |
| 5 | 5 | 1 | 1 | 1 | 1,00 | | |
| 6 | 6 | 0 | 1 | 2 | 1,00 | | |
| 7 | 7 | 4 | 0 | 0 | 1,33 | | |
| 8 | 8 | 5 | 3 | 2 | 3,33 | | |
| 9 | 9 | 5 | 4 | 3 | 4,00 | | |
| 10 | 10 | 1 | 2 | 3 | 2,00 | | |
| 11 | | | | | | | |

**Abbildung 7.3:** Neue Variable *korbtreffer2014_2016* auf Basis der MEAN-Funktion (*KORBTREFFER4.SAV*)

Interessieren Sie sich im »Basketball-Beispiel« anstelle von MEAN für die fallweise Summe, müssen Sie anstelle von MEAN einfach SUM eingeben, wie in Abbildung 7.4 und Abbildung 7.5 gezeigt. Ihre genutzte Formel lautet in diesem Fall:

$$korbtreffer2014_2016 = \mathrm{SUM}\left(korbtreffer2014,\ korbtreffer2015,\ korbtreffer2016\right).$$

Gleichwertig hierzu hätten Sie natürlich auch definieren können:

$$korbtreffer2014_2016 = korbtreffer2014 + korbtreffer2015 + korbtreffer2016.$$

Sie erkennen sicherlich leicht, dass die Summations-Funktion in der SPSS-Syntax mit dem SPSS-Befehl COMPUTE korrespondiert:

$$\mathrm{COMPUTE}\, korbtreffer2014_2016 = korbtreffer2014 + korbtreffer2015 + korbtreffer2016.$$

Für Fall 1 beispielsweise erhalten Sie demnach eine Summe von 7 Treffern (in den drei Jahren 2014 bis 2016; siehe Abbildung 7.5).

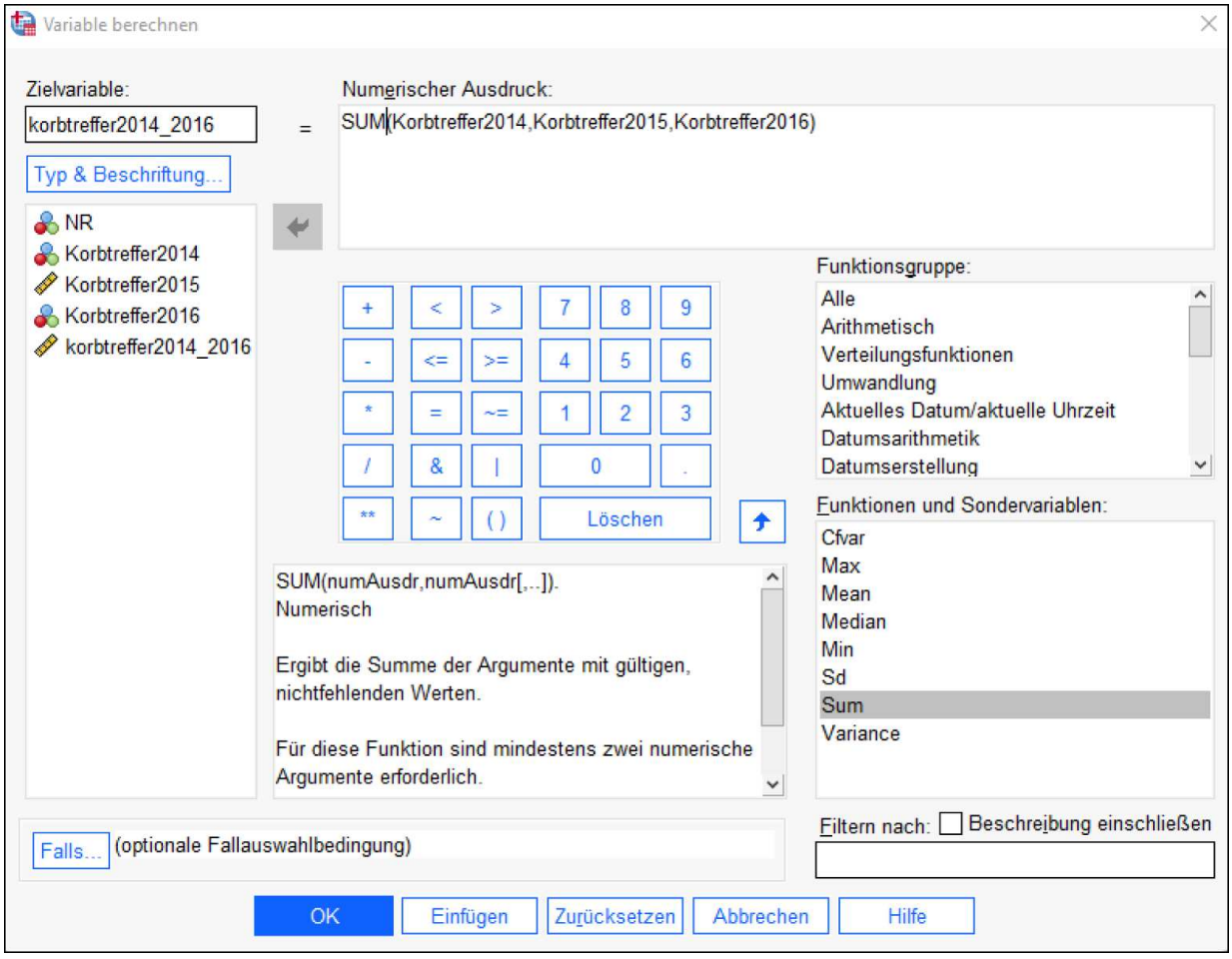
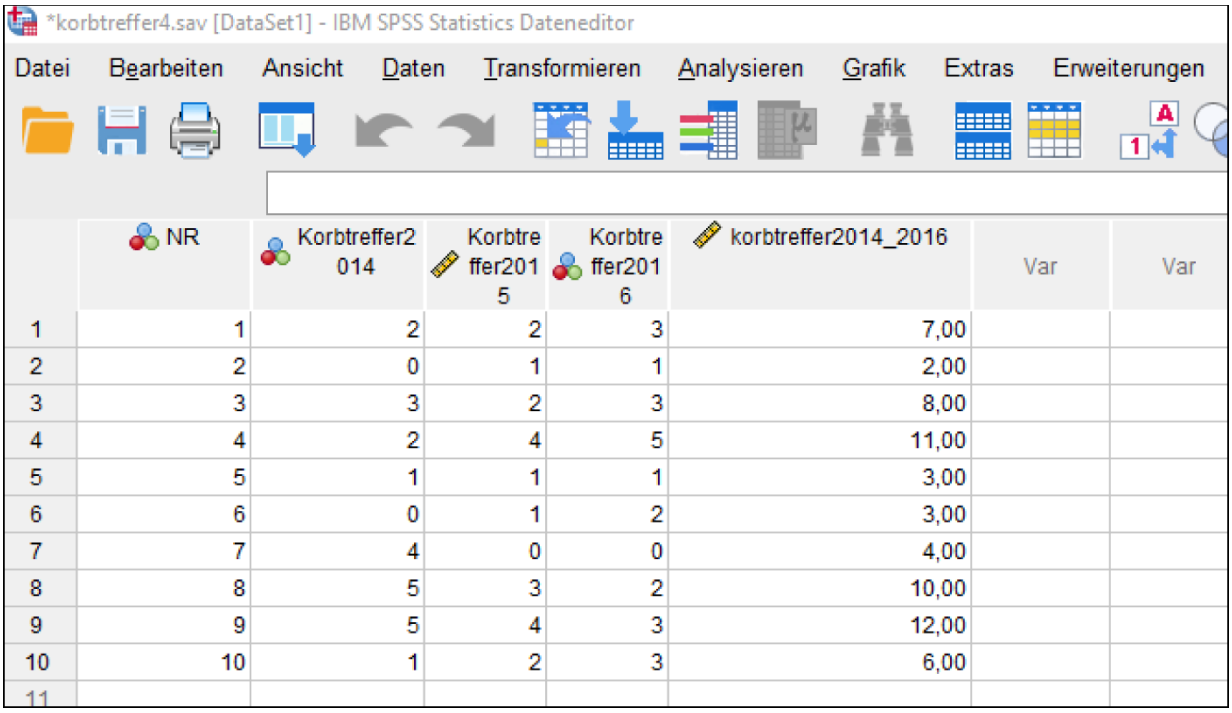



**Abbildung 7.4:** Anwendung der SUM-Funktion auf die Variable *korbtreffer2014_2016* (*KORBTREFFER4.SAV*)




| | NR | Korbtreffer2014 | Korbtreffer2015 | Korbtreffer2016 | korbtreffer2014_2016 | Var | Var |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 2 | 3 | 7,00 | | |
| 2 | 2 | 0 | 1 | 1 | 2,00 | | |
| 3 | 3 | 3 | 2 | 3 | 8,00 | | |
| 4 | 4 | 2 | 4 | 5 | 11,00 | | |
| 5 | 5 | 1 | 1 | 1 | 3,00 | | |
| 6 | 6 | 0 | 1 | 2 | 3,00 | | |
| 7 | 7 | 4 | 0 | 0 | 4,00 | | |
| 8 | 8 | 5 | 3 | 2 | 10,00 | | |
| 9 | 9 | 5 | 4 | 3 | 12,00 | | |
| 10 | 10 | 1 | 2 | 3 | 6,00 | | |
| 11 | | | | | | | |

**Abbildung 7.5:** Neue Variable *korbtreffer2014_2016* für die SUM-Funktion (*KORBTREFFER4.SAV*)

In den Funktionsgruppen VERTEILUNGSFUNKTIONEN, QUANTILFUNKTIONEN, WAHRSCHEINLICHKEITSDICHTEN, ZUFALLSZAHLEN und SIGNIFIKANZ finden Sie weitere Funktionen ausdrücklich für statistische Zwecke. Ohne hierauf besonders ausführlich einzugehen (da die betreffenden speziellen SPSS-Funktionen – insbesondere die sogenannten Pseudozufallsgeneratoren – im Grunde genommen erst in der »Königsklasse« der Inferenzstatistik – das heißt hier: erst in Teil IV dieses Buchs – eine Rolle spielen), liefert dabei zum Beispiel Rv.NORMAL im Sinne eines Pseudozufallsgenerators normalverteilte Zufallszahlen. Rv.UNIFORM beispielsweise führt demgegenüber zu gleichverteilten Zufallszahlen.

# *Art der statistischen Kennziffern bei den einzelnen Statistikprozeduren*

Die **Statistikauswertungen** in SPSS für einzelne Variablen über alle Fälle hinweg erfolgen nach dem allgemeinen Schema PROZEDUR|VARIABLE(N)|OPTIONEN|AUSFÜHREN. Bei der betreffenden Variablenauswahl ist unbedingt auf die Skalierung beziehungsweise auf den Variablentyp zu achten und damit darauf, welche statistischen Kennziffern überhaupt analytisch angemessen sind. Wie Tabelle 7.2 verdeutlicht, stehen die einzelnen statistischen Kennziffern zum Teil in unterschiedlicher Zusammenstellung in den verschiedenen statistischen Prozeduren zur Verfügung. In der nachfolgenden Tabelle sind in den Spalten die statistischen Prozeduren mit ihrer Syntaxbezeichnung angegeben.

Die in Tabelle 7.2 angegebenen statistischen Kennziffern sind wie folgt definiert (siehe auch die Zusammenstellung statistischer Kennziffern bereits in Kapitel 2):

- *Summe*:Hierunter versteht man das Ergebnis der Addition von Merkmalsausprägungen.

| *Kennziffer* | *Deskriptive Statistiken (DESCRIPTIVES)* | *Häufigkeiten (FREQUENCIES)* | *Explorative Statistiken (EXAMINE)* | *Fälle zusammenfassen (SUMMARIZE)/Mittelwerte vergleichen (MEANS)* |
|---|---|---|---|---|
| Summe | X | X | | X |
| Arithmetisches Mittel | X | X | X | X |
| (Gruppierter) Median | | X | X | X |
| Modus | | X | | |
| Harmonisches Mittel | | | | X |
| Geometrisches Mittel | | | | X |
| Quantile | | X | | |
| Perzentile | | X | X | |
| Interquartilbereich | | | X | |
| Minimum | X | X | X | X |
| Maximum | X | X | X | X |
| Spannweite (»Bereich«) | X | X | X | X |
| Varianz | X | X | X | X |
| Standardabweichung | X | X | X | X |
| Schiefe | X | X | X | X |
| Wölbung (Kurtosis) | X | X | X | X |
| Standardfehler des arithmetischen Mittelwerts | X | X | X | X |
| Standardfehler der Schiefe | | | | X |
| Standardfehler der Wölbung | | | | X |

**Tabelle 7.2:** Statistische SPSS-Prozeduren und verschiedene statistische Kennziffern

- *Arithmetischer Mittelwert*: Er wird auch als Durchschnitt bezeichnet. In diesem Sinne spiegelt er die Merkmalssumme dividiert durch die Fallzahl wider (zum Beispiel: Für die fünf Ausprägungen 1, 2, 4, 7 und 9 beträgt das arithmetische Mittel $\frac{1+2+4+7+9}{5} = 4{,}6$).

- *Median*:Hiermit ist die genau in der Mitte der geordneten Liste der Merkmalsausprägungen liegende Merkmalsausprägung gemeint (zum Beispiel: Bei fünf Ausprägungen 1, 2, 4, 7 und 9 ist der Wert 4 der in der Mitte liegende Wert, also der Median).
- *Modus*: Dies ist die am häufigsten vorkommende Merkmalsausprägung.
- *Harmonischer Mittelwert*:Dieser Mittelwert ist für Verhältnisse aus zwei Größen definiert, wobei – im Unterschied zum arithmetischen Mittel – die Zählergröße einen festen Wert hat und die Nennergröße variabel ist; das harmonische Mittel ist daher – etwas verschwurbelt ausgedrückt – der Kehrwert des arithmetischen Mittelwerts der Kehrwerte der Merkmalsausprägungen (zum Beispiel: Berechnung der Fahrzeit bei gegebener Wegstrecke, also etwa 2 Stunden Fahrzeit für die ersten 100 Kilometer und dann 3 Stunden Fahrzeit für die weiteren 100 Kilometer Wegstrecke; das harmonische Mittel beträgt in diesem Fall: $\frac{100\ \text{km}+100\ \text{km}}{2\ \text{h}+3\ \text{h}}=40\ \frac{\text{km}}{\text{h}}$).
- *Geometrischer Mittelwert*:Für prozentuale Veränderungsraten ist der arithmetische Mittelwert nicht geeignet. An dieser Stelle kommt der geometrische Mittelwert zum Einsatz, und zwar als N-te Wurzel aus dem Produkt der Veränderungsfaktoren (N: Anzahl der Veränderungen insgesamt; zum Beispiel: Für die drei Veränderungsraten von 5 Prozent, 4 Prozent und 2 Prozent beträgt das geometrische Mittel für die Veränderungsfaktoren $\sqrt[3]{1{,}05\cdot 1{,}04\cdot 1{,}02}=1{,}0366,$ und damit ist die mittlere Veränderungsrate + 3,66 Prozent).
- *Quantile*: Hierunter wird die Aufteilung der gesamten Häufigkeitsverteilung in gleich große prozentuale Abschnitte verstanden: *Perzentile*: in 1-Prozent-Bereiche (also in 100 Bereiche beziehungsweise in Prozentbereiche nach eigener Vorgabe), Dezile: in 10-Prozent-Bereiche (also in 10 Bereiche), Quintile: in 20-Prozent-Bereiche (also in 5 Bereiche) und Quartile: in 25-Prozent-Bereiche (also in 4 Bereiche).
- *Interquartilbereich*:Dies kennzeichnet die Differenz aus dem Wert an der Stelle der Obergrenze des dritten Quartils und an der Stelle der Untergrenze des ersten Quartils.
- *Minimum*:Dies ist schlichtweg der niedrigste Wert.
- *Maximum*: Das ist der größte Wert.
- *Spannweite*:Sie ist als Differenz aus der höchsten und der niedrigsten Merkmalsausprägung definiert.
- *Varianz, Standardabweichung*: Varianz und Standardabweichung geben Auskunft über die Streuung der Merkmalsausprägungen. Die Varianz ist dabei die durchschnittliche quadratische Abweichung der jeweiligen Merkmalsausprägungen vom arithmetischen Mittelwert. Die Standardabweichung ist die Quadratwurzel aus der Varianz.
- *Schiefe, Wölbung*:Anhand der Schiefe können Sie erkennen, ob eine Verteilung der Merkmalsausprägungen symmetrisch oder asymmetrisch ist. Die Wölbung (auch »Kurtosis« genannt) gibt gegenüber einer Referenzverteilung an (üblicherweise wird

hier die »Glockenkurve« der Normalverteilung unterstellt), ob die Verteilung der Merkmalsausprägungen von ihrer Form her gewölbter (»ausgebauchter«) ist oder nicht.

- ✔ *Standardfehler (zum Beispiel) des arithmetischen Mittelwertes*: Hierunter ist, vereinfacht ausgedrückt, zu verstehen, wie sehr der arithmetische Mittelwert einer Stichprobe vom arithmetischen Mittelwert in der Grundgesamtheit abweicht. (Es handelt sich hierbei um eine Kennziffer aus der induktiven Statistik, die erst in Teil IV dieses Buchs von größerer Bedeutung ist.)

# Häufigkeits-Prozedur (FREQUENCIES)

Gemäß Tabelle 7.2 lassen sich für die **»Häufigkeits-Prozedur«** ziemlich viele statistische Kennziffern berechnen, die zum Teil insofern interessant sind, als sie wie der Median lediglich mindestens Ordinalskalenniveau voraussetzen. Bezüglich des ebenfalls von dieser Prozedur ausgewiesenen Modus können Sie sogar mit lediglich nominalskalierten Variablen arbeiten. Zugang zur »Häufigkeits-Prozedur« erhalten Sie über ANALYSIEREN|DESKRIPTIVE STATISTIKEN|HÄUFIGKEITEN|STATISTIKEN. Der zugehörige SPSS-Syntaxbefehl lautet FREQUENCIES. (Bereits in den Kapiteln 5 und 6 wurde zumindest zum Teil auf FREQUENCIES eingegangen, vielfach dort aber nur zur Darstellung von Häufigkeitsverteilungen in einer Häufigkeitstabelle.)

## FREQUENCIES-Statistiken am »Basketball-Beispiel«

Zur Illustration soll wieder das »Basketball-Beispiel« dienen mit zehn Korbwerfern, deren Treffer bei jeweils fünf Würfen sich wie folgt darstellen:

Die Häufigkeitsverteilung setzt sich aus jeweils zweimal 0, 1, 2 und 5 Treffern sowie aus jeweils einmal 3 und 4 Treffern zusammen (siehe auch Tabelle 7.3). Den höchsten Häufigkeitswert im Sinne des **Modus** (= 2-mal) erhalten Sie demnach für 0 Treffer, für 1 Treffer, für 2 Treffer und für 5 Treffer. Der genau in der Mitte der Verteilung liegende Wert, der **Median**, befindet sich bei insgesamt 10 Fällen zwischen dem fünften und dem sechsten Häufigkeitswert, in diesem Fall also beim Trefferwert 2 Treffer. Als **arithmetischen Mittelwert** $\bar{X}$ im Sinne des Durchschnitts erhalten Sie manuell:

| *Person Nr.* | *Anzahl Korbtreffer* |
|---|---|
| 1 | 2 |
| 2 | 0 |
| 3 | 3 |
| 4 | 2 |
| 5 | 1 |
| 6 | 0 |
| 7 | 4 |
| 8 | 5 |
| 9 | 5 |

| *Person Nr.* | *Anzahl Korbtreffer* |
|---|---|
| 10 | 1 |

Tabelle 7.3: »Basketball-Beispiel«

$$\bar{X} = \frac{0\cdot 2 + 1\cdot 2 + 2\cdot 2 + 3\cdot 1 + 4\cdot 1 + 5\cdot 2}{10} = \frac{23}{10} = 2{,}3.$$

Als **Varianz** $S^2$ beziehungsweise als **Standardabweichung** S lassen sich hieraus berechnen:

$$S^2 = \frac{1}{10}\cdot\Big[(0-2{,}3)^2\cdot 2 + (1-2{,}3)^2\cdot 2 + (2-2{,}3)^2\cdot 2 + (3-2{,}3)^2\cdot 1 + (4-2{,}3)^2\cdot 1 + (5-2{,}3)^2\cdot 2\Big] = \frac{32{,}1}{10} = 3{,}21$$

$$S = \sqrt[2]{3{,}21} \approx 1{,}7916.$$

Die **Schiefe** SCH und die **Wölbung** (Kurtosis) WB lassen sich manuell wie folgt berechnen, wobei hier bei der Wölbung der Wert 3 abgezogen wird, der die Wölbung der hier als Referenz genutzten Normalverteilung ist:

$$SCH = \frac{\frac{1}{N}\cdot\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^3}{S^3};$$

$$\text{das heißt: } SCH = \frac{1}{10\cdot 1{,}7916^3}\cdot\Big[(0-2{,}3)^3\cdot 2 + (1-2{,}3)^3\cdot 2 + (2-2{,}3)^3\cdot 2 + (3-2{,}3)^3\cdot 1 + (4-2{,}3)^3\cdot 1 + (5-2{,}3)^3\cdot 2\Big] = \frac{15{,}84}{10\cdot 1{,}7916^3} = +0{,}2754$$

$$WB = \frac{\frac{1}{N}\cdot\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^4}{S^4} - 3;$$

$$\text{das heißt: } WB = \frac{1}{10\cdot 1{,}7916^4}\cdot\Big[(0-2{,}3)^4\cdot 2 + (1-2{,}3)^4\cdot 2 + (2-2{,}3)^4\cdot 2 + (3-2{,}3)^4\cdot 1 + (4-2{,}3)^4\cdot 1 + (5-2{,}3)^4\cdot 2\Big] - 3 = \frac{176{,}577}{10\cdot 1{,}7916^4} - 3 = -1{,}2862.$$

Der Vollständigkeit halber: Manuell ergibt sich als **Minimalwert** 0 (für 0 Treffer) und als **Maximalwert** 5 (für 5 Treffer), sodass die **Spannweite** aus Maximal- und Minimalwert 5 (= 5 − 0) beträgt.

So, jetzt wollen wir mal sehen, was SPSS an dieser Stelle aus dem Datenmaterial macht. Hierzu rufen Sie bitte die Datei *KORBTREFFER.SAV* auf. Anschließend bringen Sie im

HÄUFIGKEITEN-Dialogfenster *korbtreffer* in die VARIABLE(N)-Rubrik und klicken unter STATISTIKEN so gut wie alles an, wie in Abbildung 7.6 dargelegt.

Für ANALYSIEREN|DESKRIPTIVE STATISTIKEN|HÄUFIGKEITEN… stehen standardmäßig folgende Statistiken zur Verfügung (wie auch bereits in Tabelle 7.2 dargestellt):

- ✔ (Arithmetischer) Mittelwert
- ✔ Median
- ✔ Modalwert
- ✔ Standardabweichung

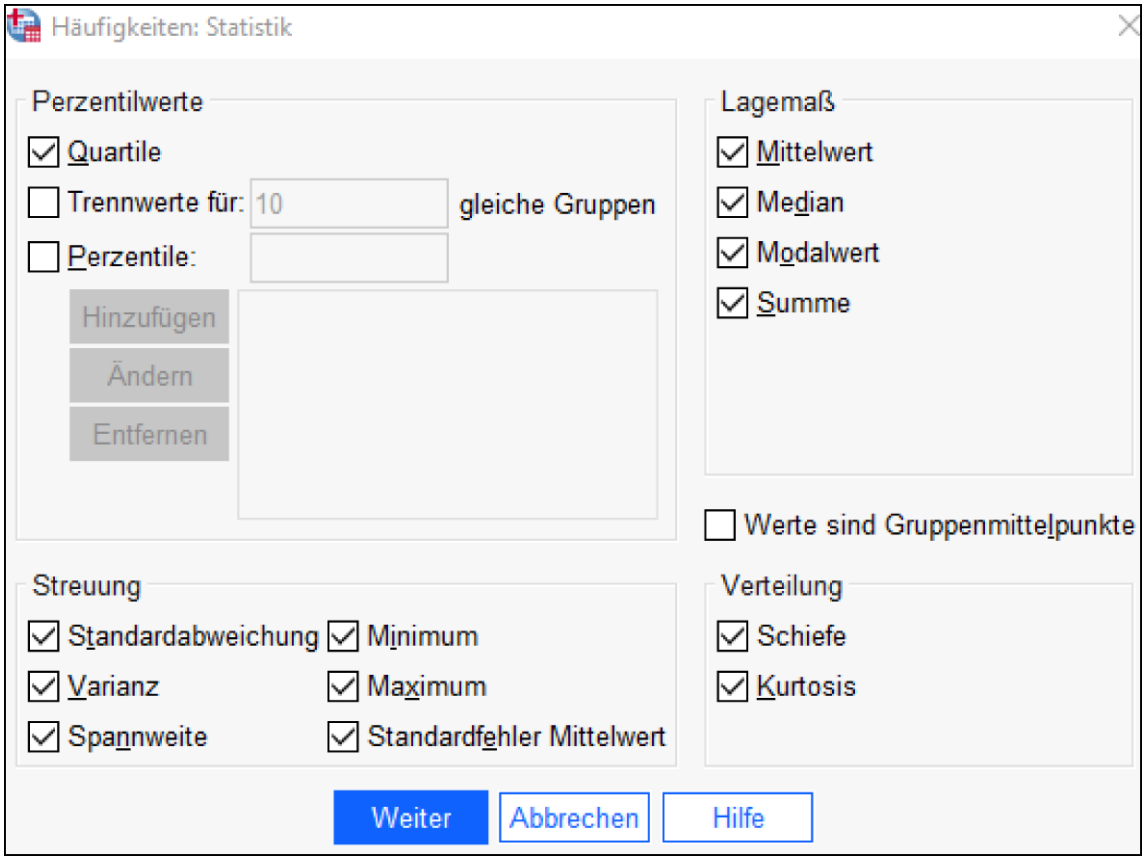



**Abbildung 7.6:** HÄUFIGKEITEN|STATISTIKEN für das »Basketball-Beispiel« (*KORBTREFFER.SAV*)

- ✔ Varianz
- ✔ Minimum
- ✔ Maximum
- ✔ Schiefe
- ✔ Kurtosis

Daneben können Sie noch zwischen verschiedenen Quantileinstellungen auswählen:

- ✔ Perzentile (1-Prozent-Bereiche, also 100 Bereiche),
- ✔ Dezile (10-Prozent-Bereiche, also 10 Bereiche),
- ✔ Quintile (20-Prozent-Bereiche, also 5 Bereiche) und
- ✔ Quartile (25-Prozent-Bereiche, also 4 Bereiche).

Außerdem nehmen Sie bitte – wie in Abbildung 7.7 gezeigt – einfach das Häkchen bei HÄUFIGKEITSTABELLE ANZEIGEN weg. Anschließend klicken Sie auf die OK-Taste, und in Sekundenschnelle wirft Ihnen SPSS dann verschiedene statistische Kennziffern aus.

Das Ganze hätten Sie im Übrigen auch nicht über das Menü, sondern als Syntaxprogramm bewerkstelligen können. Da Sie ja keine Häufigkeitstabelle »ausgeworfen« bekommen möchten, geben Sie einfach nach der FREQUENCIES-Hauptanweisung /FORMAT=NOTABLES ein. Mit der weiteren Spezifikation »/NTILES=4« fordern Sie hier die Werte für die Verteilungsquartile an. Mit /NTILES=5 hätten Sie Verteilungsquintile, mit /NTILES=10 Verteilungsdezile und mit /NTILES=100 Verteilungsperzentile angefordert.

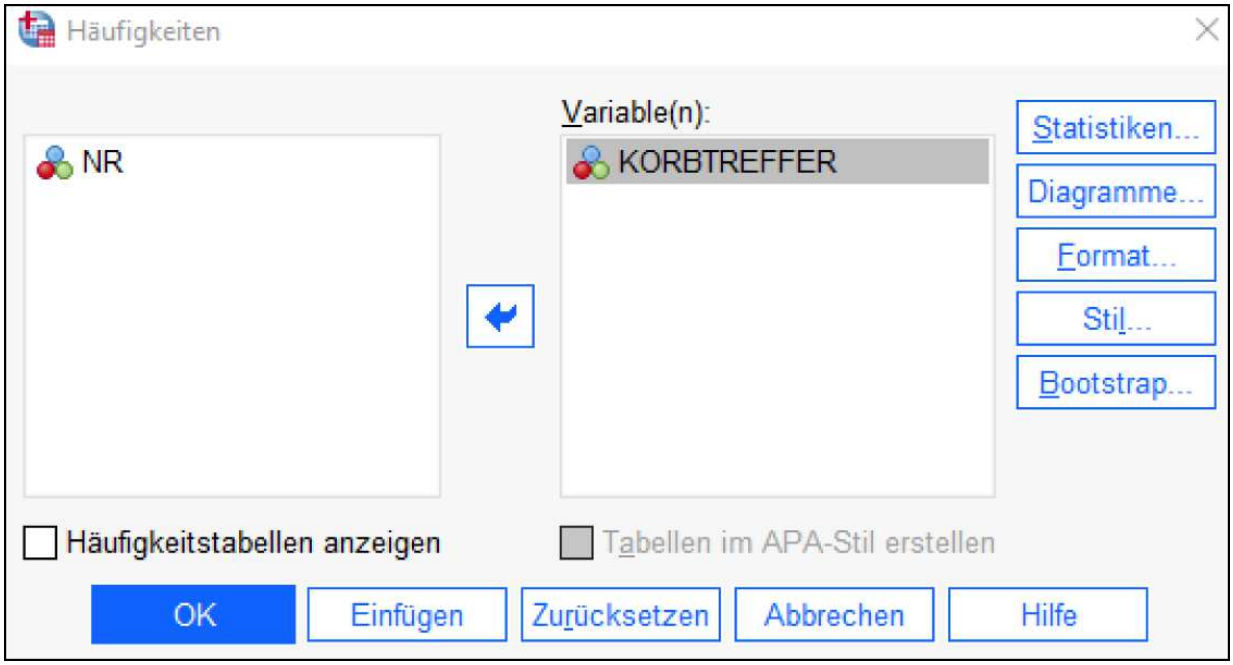



**Abbildung 7.7:** Unterdrückung der Häufigkeitstabelle bei HÄUFIGKEITEN|STATISTIKEN (*KORBTREFFER.SAV*)

Konkret müssen Sie in Bezug auf die Datendatei *KORBTREFFER.SAV* folgende Syntax eingeben:

```
FREQUENCIES VARIABLES=korbtreffer
        /FORMAT=NOTABLES
        /NTILES=4
        /STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM SEMEAN
MEAN MEDIAN MODE SUM SKEWNESS SESKEW KURTOSIS SEKURT.
```

Abbildung 7.8 gibt einen Überblick über die berechneten statistischen Kennziffern. Die durchschnittliche Trefferanzahl beispielsweise beträgt 2,30 Treffer.

Bezogen auf die von SPSS in der betreffenden Abbildung »ausgeworfenen« statistischen Kennziffern, wird beim Modus nur ein Wert angegeben – dies aber verbunden mit dem Hinweis, dass mehrere Modalwerte vorliegen und hier nur der niedrigste Wert angegeben ist. Des Weiteren stimmen die Werte für den (arithmetischen) Mittelwert, den Median, die Spannweite, das Minimum, das Maximum und die Summe mit Ihren manuellen Berechnungen überein.

Abweichungen gibt es hingegen bei der Varianz und bei der Standardabweichung. Die Unterschiede ergeben sich daraus, dass Sie oben durch alle 10 Fälle dividiert haben, SPSS aber nur durch (10 − 1 =) 9 Fälle geteilt hat. Dies liegt daran, dass bei Ihrer manuellen Varianz- beziehungsweise Standardabweichungsberechnung Abweichungen berechnet worden sind, die auf eine Grundgesamtheit bezogen sind, während SPSS jeweils eine Schätzfunktion für die betreffenden Abweichungen verwendet.

**Statistiken**

KORBTREFFER

| | | |
|---|---|---|
| N | Gültig | 10 |
| | Fehlend | 0 |
| Mittelwert | | 2,30 |
| Standardfehler des Mittelwerts | | ,597 |
| Median | | 2,00 |
| Modus | | 0[a] |
| Std.-Abweichung | | 1,889 |
| Varianz | | 3,567 |
| Schiefe | | ,327 |
| Standardfehler der Schiefe | | ,687 |
| Kurtosis | | -1,310 |
| Standardfehler der Kurtosis | | 1,334 |
| Spannweite | | 5 |
| Minimum | | 0 |
| Maximum | | 5 |
| Summe | | 23 |
| Perzentile | 25 | ,75 |
| | 50 | 2,00 |
| | 75 | 4,25 |

a. Mehrere Modi vorhanden. Der kleinste Wert wird angezeigt.

**Abbildung 7.8:** Statistische Kennziffern (FREQUENCIES) für das »Basketball-Beispiel« (*KORBTREFFER.SAV*)

Dass SPSS für die Standardabweichung beziehungsweise die Varianz stets die betreffenden Schätzfunktionen nutzt, heißt: Bei diesen Kennziffern dividiert SPSS nicht durch die Fallzahl, sondern durch die Fallzahl minus 1. Der Grund hierfür ist, dass diese empirische Streuung für die induktive Statistik gerade im Rahmen der dortigen Testverfahren von sehr großer Bedeutung ist und dadurch Signifikanzaussagen ermöglicht werden. Bei sehr großen Fallzahlen unterscheiden sich die verschiedenen Varianzformulierungen kaum, bei kleinen Fallzahlen hingegen schon merklich. Liegen etwa – wie im »Basketball-Beispiel« – nur zehn Fälle vor und beträgt die Summe der quadrierten Abweichungen vom arithmetischen Mittelwert 32,1 Einheiten, erhalten Sie bei der Division durch 10 eine Varianz in Höhe von 3,21 Einheiten, bei der Division durch 9 (= 10 − 1) hingegen eine solche in Höhe von 3,567 Einheiten. Don't worry: Sie können händisch den von SPSS angegebenen Varianzwert einfach dadurch in den anderen Varianzwert überführen, dass Sie den SPSS-Wert mit 9 (= 10 − 1), also mit (N-1), multiplizieren und dann durch 10, also durch N, dividieren:

$$3{,}567 \cdot \frac{9}{10} = 3{,}21.$$

Grundsätzlich gibt SPSS immer die empirischen (Stichproben-)Kennziffern mit entsprechenden Signifikanzangaben aus. Das wird gleich auch bezüglich der Schiefe und der Wölbung deutlich.

Es ergeben sich nämlich auch hinsichtlich der von SPSS dargelegten Schiefe- und Wölbungswerte Abweichungen zu Ihren manuellen Berechnungen.

Für die Schiefe SCH und die Wölbung WB verwendet SPSS allgemein folgende komplexe Formeln (mit: N = Fallzahl; $X_i$ = i-te Ausprägung der Variablen X, $\bar{X}$ = Arithmetischer Mittelwert):

$$SCH = \frac{N}{(N-1)\cdot(N-2)} \cdot \frac{\cdot\sum_{i=1}^{N}(X_i - \bar{X})^3}{S^3};$$

$$WB = \frac{N\cdot(N+1)}{(N-1)\cdot(N-2)\cdot(N-3)} \cdot \frac{\frac{1}{N}\cdot\sum_{i=1}^{N}(X_i - \bar{X})^4}{S^4} - \frac{3\cdot(N-1)^2}{(N-2)\cdot(N-3)}.$$

Für das »Basketball-Beispiel« lauten demnach der SPSS-Schiefe- beziehungsweise der SPSS-Wölbungswert:

$$SCH = \frac{10}{9 \cdot 8} \cdot \frac{1}{1{,}8886^3} \cdot \Big[ (0-2{,}3)^3 \cdot 2 + (1-2{,}3)^3 \cdot 2 + (2-2{,}3)^3 \cdot 2 + (3-2{,}3)^3 \cdot 1 + (4-2{,}3)^3 \cdot 1 + (5-2{,}3)^3 \cdot 2 \Big] = +0{,}327;$$

$$WB = \frac{10 \cdot 11}{9 \cdot 8 \cdot 7 \cdot 1{,}8886^4} \cdot \Big[ (0-2{,}3)^4 \cdot 2 + (1-2{,}3)^4 \cdot 2 + (2-2{,}3)^4 \cdot 2 + (3-2{,}3)^4 \cdot 1 + (4-2{,}3)^4 \cdot 1 + (5-2{,}3)^4 \cdot 2 \Big] - \frac{3 \cdot 9^2}{8 \cdot 7} = -1{,}310.$$

Wenn Sie (auch) mit der Statistiksoftware STATA arbeiten, muss ich Sie leider noch auf eine andere Berechnungsart der Schiefe und der Wölbung als in SPSS hinweisen. STATA unterscheidet sich nämlich bei den entsprechenden Berechnungen von SPSS dadurch, dass es sich grundsätzlich an den weiter oben angegebenen Formeln für die Schiefe und die Wölbung orientiert und zusätzlich bei der Wölbung den Wert 3 nicht abzieht:

$$SCH = \frac{\frac{1}{N} \cdot \sum_{i=1}^{N} (X_i - \bar{X})^3}{S^3}; \quad WB = \frac{\frac{1}{N} \cdot \sum_{i=1}^{N} (X_i - \bar{X})^4}{S^4}.$$

Wie Sie Abbildung 7.8 ferner entnehmen können, weist – gemäß unserer Anforderung in Abbildung 7.6 – SPSS auch noch die Quartilgrenzen aus. Konkret werden im »Basketball-Beispiel« in Bezug auf die Verteilung von *korbtreffer* für die Obergrenze des untersten, das heißt des 25-Prozentquartils 0,75 Treffer, für die Obergrenze des mittleren, das heißt des 50-Prozentquartils (das im Übrigen mit dem Median gleichzusetzen ist) 2,00 Treffer und für die Obergrenze des zweitobersten, das heißt des 75-Prozentquartils (= der Untergrenze des obersten Quartils) 4,25 Treffer angegeben.

Allgemein gelten für die Grenzen $X_Q$ eines Quantils Q folgende Formeln (die eckige Klammer repräsentiert hierbei die sogenannte Gaußklammer und bedeutet hier ein Abrunden des in den Klammern stehenden Produkts auf die nächstkleinere ganze Zahl):

$$X_Q = X_{[N \cdot Q]+1} \text{ für } (N \cdot Q) \text{ nicht ganzzahlig}$$

beziehungsweise

$$X_Q = \frac{1}{2} \cdot (X_{(N \cdot Q)} + X_{(N \cdot Q+1)}) \qquad \text{für } (N \cdot Q) \text{ ganzzahlig.}$$

Im »Basketball-Beispiel« ist N – das heißt: die Fallzahl – gleich 10. Für Q setzen Sie bitte alternativ 0,25 (25-Prozentquartil), 0,50 (50-Prozentquartil) und 0,75 (75-Prozentquartil) ein, sodass Sie in diesem Beispiel für (N · Q) im 25-Prozentquartil 2,5 erhalten, was nicht ganzzahlig ist. Für die Grenze $X_{Q0,,25}$ nutzen Sie also bitte die erste der beiden eben

angegebenen Quantilformeln, wobei gemäß Gaußklammer hier von 2,5 auf 2 abgerundet wird und die Grenze $X_{Q0,25}$ folglich bei $X_3$, also beim dritten Wert der Verteilung liegt. Um diesen Wert aus unserer Verteilung der Korbwürfe ablesen zu können, schauen Sie sich doch einfach noch einmal die zugehörige absolute Häufigkeitsverteilung in unserem Beispiel an (siehe hierzu Tabelle 7.4):

Dem dritten Wert in dieser geordneten Liste ist demnach 1 Treffer zugeordnet – und nicht 0,75 wie bei SPSS (siehe Abbildung 7.8).

Für das 50-Prozentquartil beträgt (N · Q) 5 (also: 10 · 0,5), was ganzzahlig ist, sodass die untere der beiden obigen Quantilformeln zur Anwendung kommt. $X_{Q0,5}$ ist daher gleich: $\frac{X_5 + X_6}{2}$. Laut Tabelle 7.4 ist der fünfte Wert gleich 2, und der sechste Wert ist ebenfalls gleich 2, sodass sich für $X_{Q0,,5}$ 2 ergibt. Dies stimmt an dieser Stelle mit dem obigen SPSS-Ergebnis überein.

| *Treffer* | *Absolute Häufigkeit* |
|---|---|
| 0 | 2 |
| 1 | 2 |
| 2 | 2 |
| 3 | 1 |
| 4 | 1 |
| 5 | 2 |

**Tabelle 7.4**: Absolute Häufigkeitsverteilung im »Basketball-Beispiel«

Sie erhalten schließlich für das 75-Prozentquartil in Bezug auf (N · Q) das nicht-ganzzahlige Ergebnis 7,5, sodass Sie dieses Mal bitte wieder die obere Quantilformel verwenden. Die Verwendung dieser Formel ergibt $X_8$. Sie müssen also nach dem achten X-Wert in der Häufigkeitstabelle schauen. Dieser liegt bei 4 Treffern. SPSS hingegen hat für $X_{Q0,75}$ mit 4,25 Treffern einen anderen Wert berechnet (siehe Abbildung 7.8).

Wie kommen nun die unterschiedlichen Quartilwerte zwischen Ihren und den SPSS-Berechnungen zustande? Dies liegt daran, dass SPSS ein **lineares Interpolationsverfahren** verwendet, bei dem sich die Gewichtungsfaktoren von Quartil zu Quartil schrittweise, sozusagen »gleitend« verändern. Im vorliegenden »Basketball-Beispiel« liegen folgende gleitende Veränderungen vor, woraus die oben bereits dokumentierten SPSS-Quartilwerte resultieren:

$$X_{Q0,25} = 0 \cdot 0,25 + 1 \cdot 0,75 = 0,75;$$

$$X_{Q0,50} = 2 \cdot 0,50 + 2 \cdot 0,50 = 2,00;$$

$$X_{Q0,75} = 4 \cdot 0,75 + 5 \cdot 0,25 = 4,25.$$

## *FREQUENCIES-Bootstrapping*

In SPSS gibt es – über ein Erweiterungsmodul – noch eine weitere ganz interessante Berechnungsvariante, nämlich das sogenannte **Bootstrapping**. Allgemein erstellen Sie mit Bootstrapping ein Konfidenzintervall (also, vereinfacht ausgedrückt, einen mit einer bestimmten Wahrscheinlichkeit für möglich gehaltenen Wertebereich) für Ihre jeweilige statistische Maßzahl. Vorgegeben ist von SPSS ein 95-Prozent-Intervall, für das von SPSS jeweils der untere und der obere geschätzte Wert ausgegeben werden. Um dieses Intervall zu konstruieren, zieht SPSS aus Ihrer Datenbasis n Stichproben (standardmäßig: 1.000 Stichproben) und erstellt auf dieser Basis die jeweiligen Konfidenzintervalle.

Bootstrapping können Sie grundsätzlich in SPSS an vielen Stellen einsetzen:

- ✔ Bei den deskriptiven Statistiken (FREQUENCIES, DESCRIPTIVES, EXAMINE, MEANS, nicht aber bei SUMMARIZE!),
- ✔ bei Kreuztabellierungen (siehe Kapitel 8),
- ✔ bei Korrelationen (siehe Kapitel 8),
- ✔ bei Regressionen (siehe Kapitel 9 und 10) und
- ✔ bei verschiedenen Tests (t-Tests, ANOVA; siehe Teil IV).

Im Folgenden führen Sie bitte ein Bootstrapping für das »Basketball-Beispiel« durch. Hierzu müssen Sie bei HÄUFIGKEITEN in die Subdialogbox BOOTSTRAP gehen (siehe Abbildung 7.9). Dort müssen Sie lediglich BOOTSTRAPPING DURCHFÜHREN anklicken, und los geht's! Allerdings ist bei lediglich 10 Fällen wie im »Basketball-Beispiel« das Bootstrapping meines Erachtens nicht ganz so aussagekräftig …

Sei es drum – es geht hier nur ums Prinzip. In Form von SPSS-Syntax müssen Sie für das Einschalten des Bootstrappings zunächst eingeben:

```
BOOTSTRAP
        /SAMPLING METHOD=SIMPLE
        /VARIABLES INPUT=korbtreffer
        /CRITERIA CILEVEL=95 CITYPE=PERCENTILE NSAMPLES=1000
        /MISSING USERMISSING=EXCLUDE.
```

Es resultiert – vom Grundsatz her (die Werte hängen natürlich von den konkreten Ziehungsergebnissen beim Bootstrapping ab) – der in Abbildung 7.10 dargestellte Output. Grundsätzlich werden für die einzelnen Kennziffern (Konfidenz-, also Vertrauens-)Intervalle ausgewiesen. Für das arithmetische Mittel beispielsweise ergibt sich, etwas schlicht formuliert, dass es mit 95-prozentiger Wahrscheinlichkeit im Wertebereich von 1,20 Treffern und 3,50 Treffern liegt.

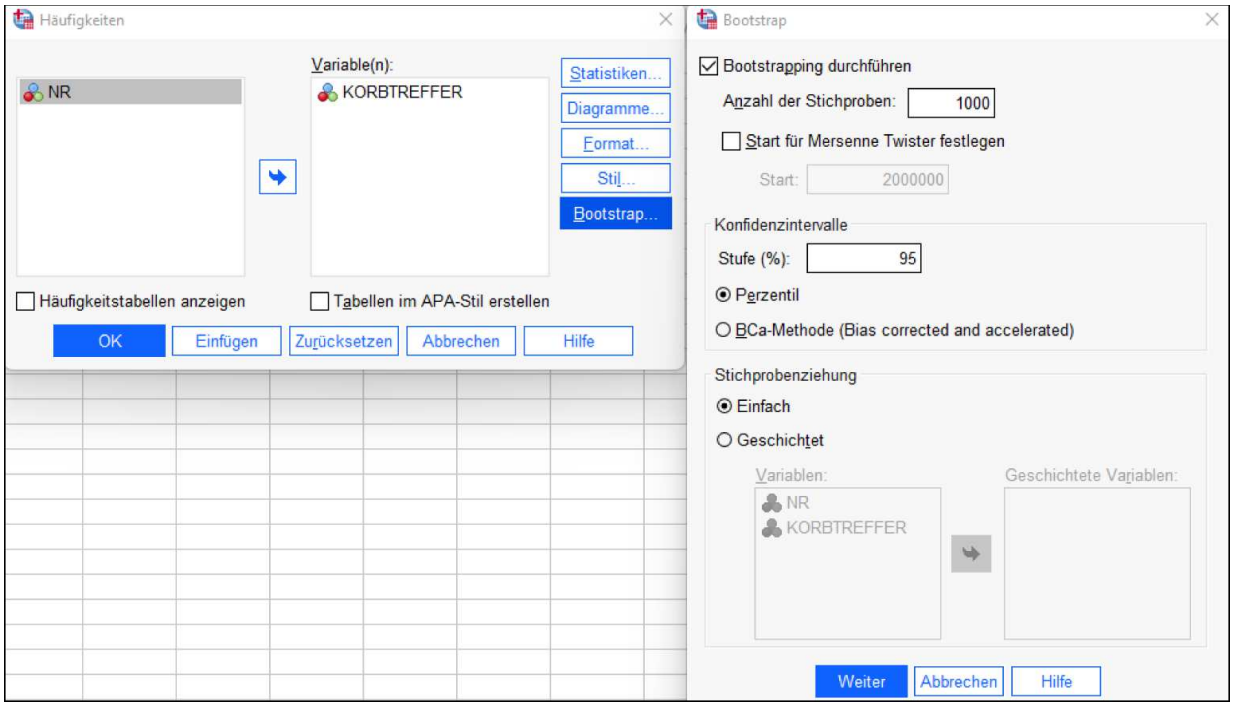



**Abbildung 7.9:** BOOTSTRAPPING DURCHFÜHREN bei HÄUFIGKEITEN (*KORBTREFFER.SAV*)

**Häufigkeiten**

**Statistiken**

KORBTREFFER

| | | | Bootstrap[b] | | | |
|---|---|---|---|---|---|---|
| | | | | | 95% Konfidenzintervall | |
| | | Statistik | Verzerrung | Std.-Fehler | Unterer | Oberer |
| N | Gültig | 10 | 0 | 0 | 10 | 10 |
| | Fehlend | 0 | 0 | 0 | 0 | 0 |
| Mittelwert | | 2,30 | ,01 | ,57 | 1,20 | 3,50 |
| Standardfehler des Mittelwerts | | ,597 | | | | |
| Median | | 2,00 | ,11 | ,90 | 1,00 | 4,00 |
| Modus | | 0[a] | | | | |
| Std.-Abweichung | | 1,889 | -,108 | ,288 | 1,075 | 2,234 |
| Varianz | | 3,567 | -,313 | ,978 | 1,156 | 4,989 |
| Schiefe | | ,327 | -,010 | ,590 | -,910 | 1,536 |
| Standardfehler der Schiefe | | ,687 | | | | |
| Kurtosis | | -1,310 | ,494 | 1,212 | -2,123 | 2,628 |
| Standardfehler der Kurtosis | | 1,334 | | | | |
| Spannweite | | 5 | | | | |
| Minimum | | 0 | | | | |
| Maximum | | 5 | | | | |
| Summe | | 23 | | | | |
| Perzentile | 25 | ,75 | ,00 | ,64 | ,00 | 2,00 |
| | 50 | 2,00 | ,11 | ,90 | 1,00 | 4,00 |
| | 75 | 4,25 | -,29 | 1,00 | 2,00 | 5,00 |

a. Mehrere Modi vorhanden. Der kleinste Wert wird angezeigt.

b. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

**Abbildung 7.10:** Bootstrapping-Output bei HÄUFIGKEITEN (*KORBTREFFER.SAV*)

# *DESCRIPTIVES-Prozedur*

Auch über ANALYSIEREN|DESKRIPTITVE STATISTIKEN|DESKRIPTIVE STATISTIK können Sie sich – Nomen est omen! – ein paar beschreibende (= deskriptive) statistische Kennziffern wie Mittelwerte oder Streuungsangaben besorgen. Der zugehörige SPSS-Befehl lautet **DESCRIPTIVES**. Vom Grundsatz her werden Ihnen bei DESKRIPTIVE STATISTIK ähnliche statistische Kennziffern »ausgeworfen« wie bei HÄUFIGKEITEN.

Über DESCRIPTIVES stehen Ihnen folgende Kennziffern zur Verfügung (siehe auch obige Tabelle 7.2):

- ✔ Summe,
- ✔ Arithmetisches Mittel,
- ✔ Minimum, Maximum, Spannweite,
- ✔ Varianz und Standardabweichung,
- ✔ Standardfehler für den arithmetischen Mittelwert sowie
- ✔ Schiefe und Wölbung.

Verglichen mit FREQUENCIES, erfordern die statistischen Auswertungen mittels DESCRIPTIVES metrisches Skalenniveau, sodass hier Kennziffern wie Modus, Median oder Quartile, die auch bei Nominal- beziehungsweise bei Ordinalskalenniveau berechnet werden können, nicht verfügbar sind.

## *DESCRIPTIVES-Statistiken am »Basketball-Beispiel«*

Prinzipiell sieht die Menü-Maske für ANALYSIEREN|DESKRIPTITVE STATISTIKEN|DESKRIPTIVE STATISTIK wie in Abbildung 7.11 aus.

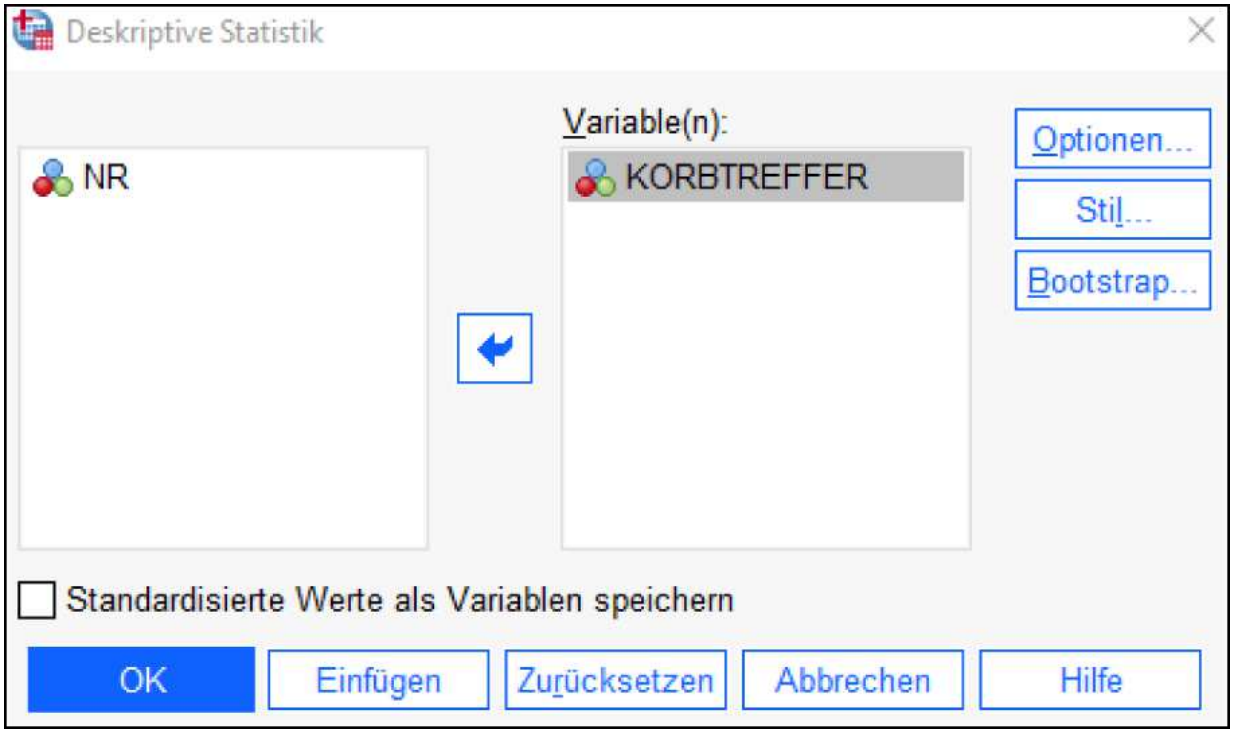



**Abbildung 7.11**: Menü-Maske für DESKRIPTIVE STATISTIK (in Bezug auf *KORBTREFFER.SAV*)

Bezogen auf das »Basketball-Beispiel«, klicken Sie jetzt bitte für die Datei *KORBTREFFER.SAV* unter OPTIONEN alle Kennziffern an, wie in Abbildung 7.12

dargestellt (wobei sich die Kennziffern auch hier wieder auf die Variable *korbtreffer* beziehen).

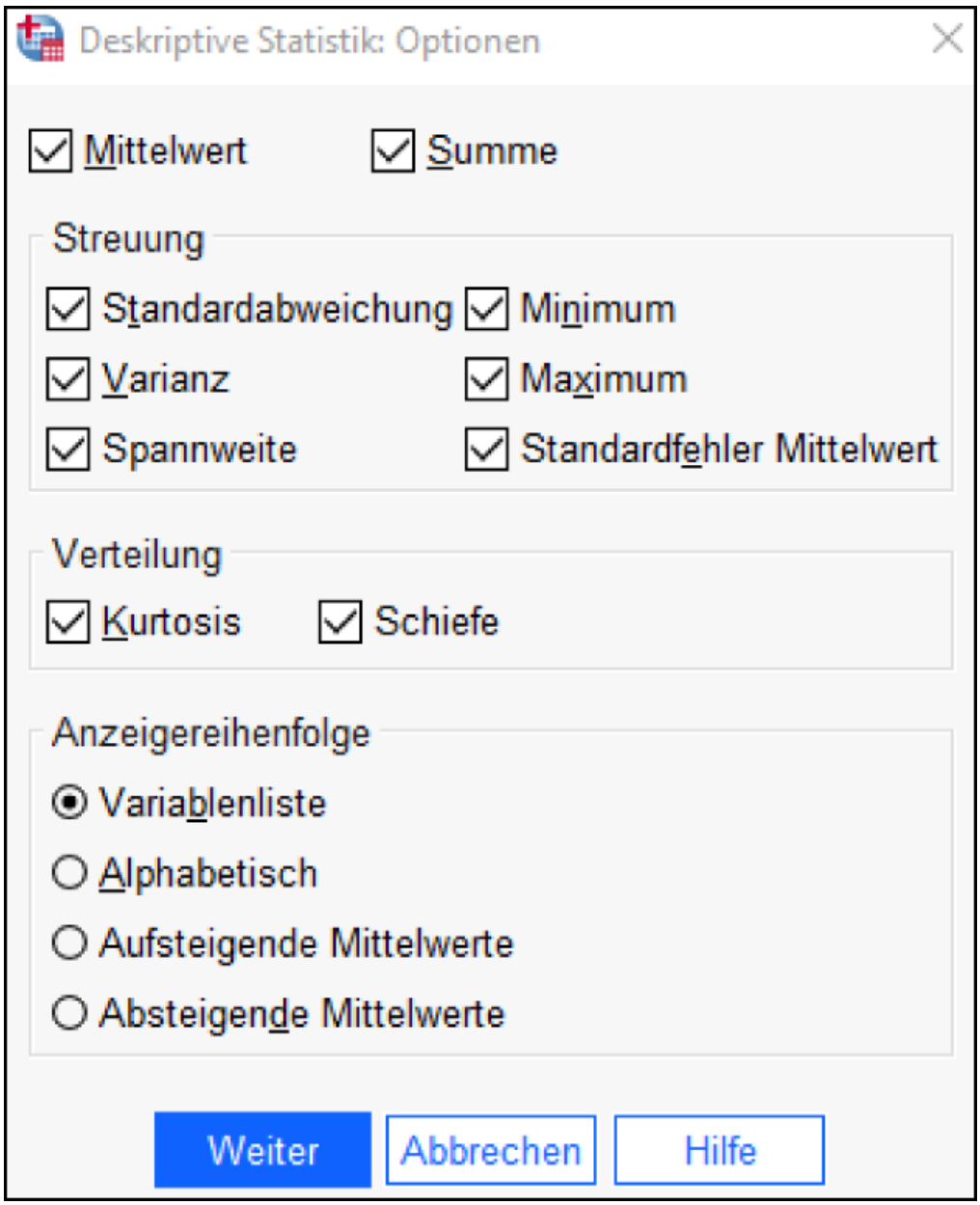



**Abbildung 7.12:** OPTIONEN bei DESKRIPTIVE STATISTIK (hier in Bezug auf *KORBTREFFER.SAV*)

Die sich aus dem vorstehenden Ablauf ergebenden statistischen Kennziffern sind in Abbildung 7.13 zusammengestellt. Standardmäßig stellt SPSS die Kennziffern in den Spalten der zugehörigen Pivottabelle dar, was bei vielen Kennziffern wie hier unübersichtlich werden kann. Wenn Sie die betreffende Pivottabelle anklicken, können Sie die Zeilen und die Spalten vertauschen, sodass das Ganze übersichtlicher wird.

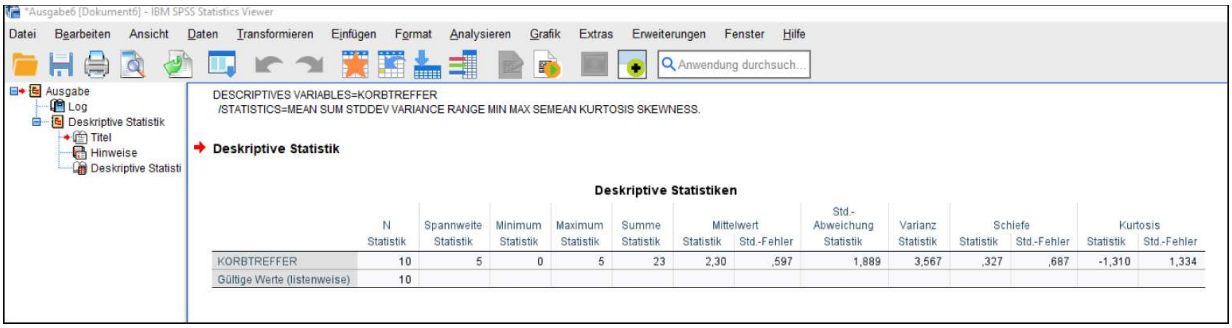



**Abbildung 7.13:** Output bei DESKRIPTIVE STATISTIK (*KORBTREFFER.SAV*)

Das vorstehende Ergebnis hätten Sie auch über die SPSS-Syntax erhalten:

```
DESCRIPTIVES VARIABLES=korbtreffer
        /STATISTICS=MEAN SUM STDDEV VARIANCE RANGE MIN MAX SEMEAN KURTOSIS
SKEWNESS.
```

## *DESCRIPTIVES-Bootstrapping*

Auch bei DESKRIPTIVE STATISTIK gibt es die Rubrik BOOTSTRAP als Subdialogbox. Über diese können Sie auch hier wieder das **Bootstrapping** starten (siehe Abbildung 7.14).

Wichtig ist an dieser Stelle der Hinweis, dass sich SPSS Ihre BOOTSTRAP-Einstellungen innerhalb einer statistischen Prozedur und einer SPSS-Session merkt. Das heißt: Wenn Sie kurz zuvor bei DESKRIPTIVE STATISTIK ein Bootstrapping durchgeführt und das entsprechende Häkchen in der Subdialogbox BOOTSTRAP nicht entfernt haben, führt SPSS auch beim nächstfolgenden Statistikvorgang bei DESKRIPTIVE STATISTIK automatisch ein Bootstrapping durch.

Sie erhalten dann für die Datendatei *KORBTREFFER.SAV* den in Abbildung 7.15 angegebenen (oder einen ähnlichen) BOOTSTRAP-Output. Es sind statistische Kennziffern mit Werten innerhalb eines 95-Konfidenzintervalls ausgewiesen, beispielsweise die Schiefe im Wertebereich von −0,871 bis 1,576.

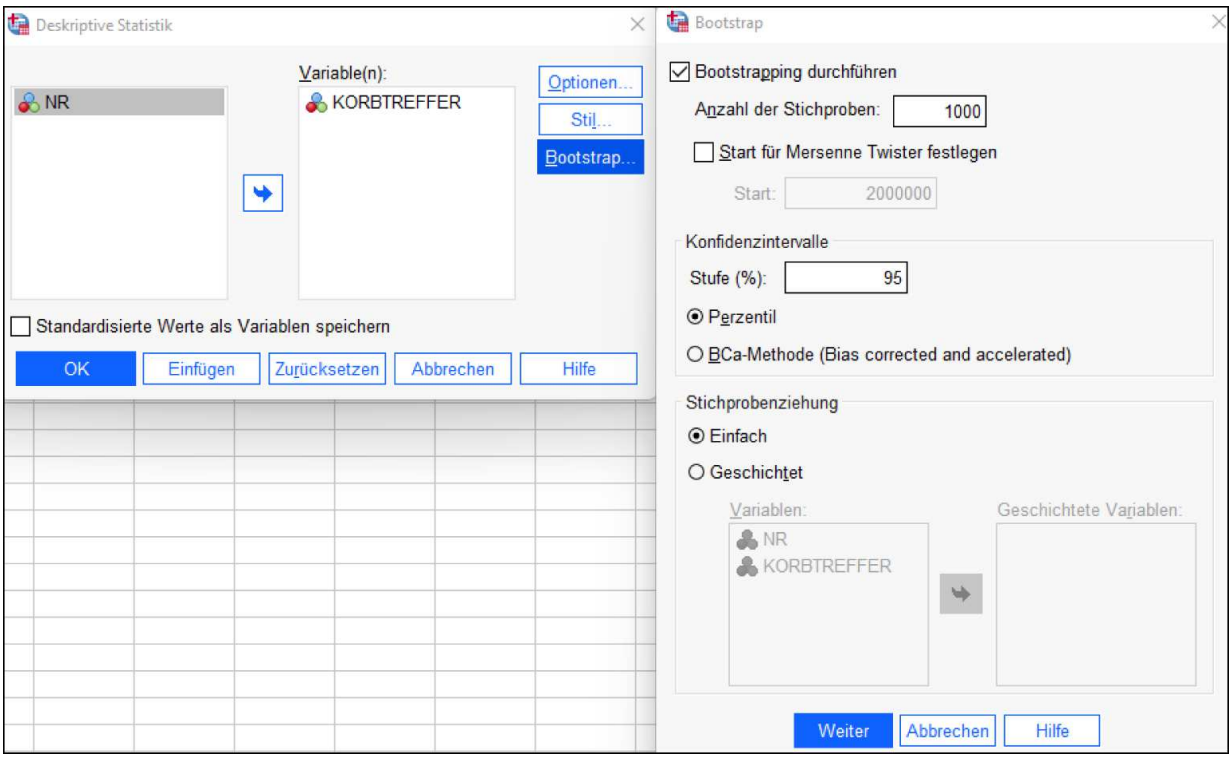



**Abbildung 7.14:** BOOTSTRAP bei DESKRIPTIVE STATISTIKEN (in Bezug auf *KORBTREFFER.SAV*)

**Deskriptive Statistik**

**Deskriptive Statistiken**

| | | | | Bootstrap[a] | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | 95% Konfidenzintervall | |
| | | Statistik | Std.-Fehler | Verzerrung | Std.-Fehler | Unterer | Oberer |
| KORBTREFFER | N | 10 | | 0 | 0 | 10 | 10 |
| | Spannweite | 5 | | | | | |
| | Minimum | 0 | | | | | |
| | Maximum | 5 | | | | | |
| | Summe | 23 | | | | | |
| | Mittelwert | 2,30 | ,597 | -,03 | ,58 | 1,20 | 3,40 |
| | Std.-Abweichung | 1,889 | | -,139 | ,289 | 1,059 | 2,234 |
| | Varianz | 3,567 | | -,422 | ,969 | 1,122 | 4,989 |
| | Schiefe | ,327 | ,687 | ,025 | ,598 | -,871 | 1,576 |
| | Kurtosis | -1,310 | 1,334 | ,603 | 1,283 | -2,153 | 3,294 |
| Gültige Werte (listenweise) | N | 10 | | 0 | 0 | 10 | 10 |

a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

**Abbildung 7.15:** BOOTSTRAP-Output bei DESKRIPTIVE STATISTIKEN (*KORBTREFFER.SAV*)

# *Explorative Datenanalyse (EXAMINE)*

Deskriptive Statistiken erhalten Sie auch mithilfe der **explorativen Datenanalyse** über ANALYSIEREN|DESKRIPTIVE STATISTIKEN|EXPLORATIVE

DATENANALYSE|STATISTIK. Der zugehörige SPSS-Syntaxbefehl lautet EXAMINE.

## *EXAMINE-Statistiken am »Basketball-Beispiel«*

Im Einzelnen stehen Ihnen bei EXPLORATIVE DATENANALYSE vor allem folgende deskriptive Statistiken zur Verfügung (siehe auch Tabelle 7.2):

- ✔ Arithmetischer Mittelwert (inklusive Konfidenzintervall (»95 %«) und sogenanntem 5-Prozent-getrimmten arithmetischen Mittelwert, bei dem die fünf Prozent kleinsten und die fünf Prozent größten Werte nicht in die Berechnung eingehen),
- ✔ Median,
- ✔ Varianz und Standardabweichung,
- ✔ Minimum, Maximum, Spannweite und Interquartilabstand,
- ✔ Perzentile sowie
- ✔ Schiefe und Kurtosis.

Die vorstehenden Statistiken fordern Sie, wie gerade geschrieben, über das Menü ANALYSIEREN|DESKRIPTIVE STATISTIKEN|EXPLORATIVE DATENANALYSE|STATISTIK an. Zuvor müssen Sie in der entsprechenden Maske *korbtreffer* als zu analysierende Variable auswählen (unter ABHÄNGIGE VARIABLEN). Außerdem müssen Sie noch ankreuzen, ob Sie STATISTIKEN, DIAGRAMME oder BEIDES als Output erhalten möchten. In Abbildung 7.16 ist zunächst einmal nur STATISTIKEN angekreuzt worden.

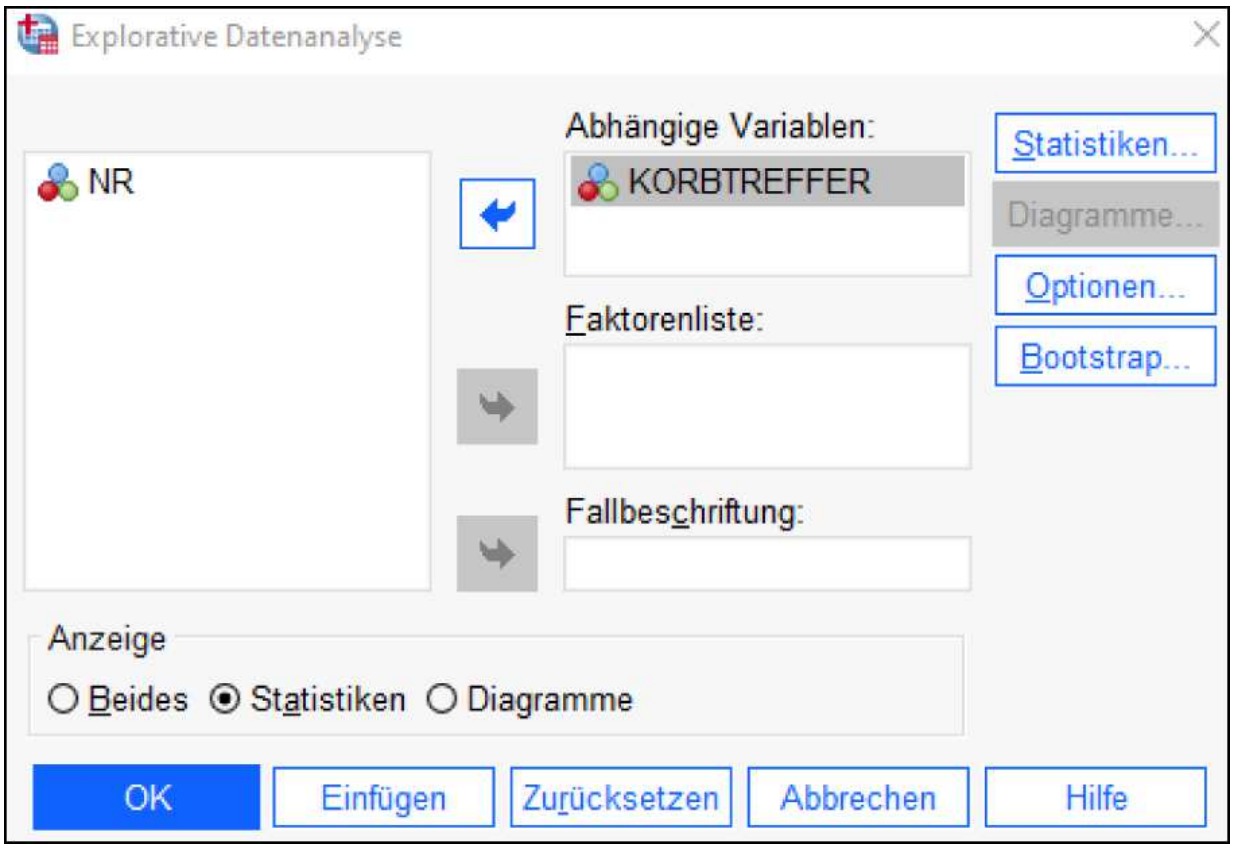



**Abbildung 7.16:** Start-Dialog bei EXPLORATIVE DATENANALYSE (*KORBTREFFER.SAV*)

Im nächsten Schritt wählen Sie unter STATISTIKEN die gewünschten Statistiken aus, bitte genauso wie in Abbildung 7.17. (Dort ist im Übrigen auch der sogenannte *MSchätzer* ausgewählt worden. Dieser – auf der *Maximum-Likelihood-Methode* aufbauende Schätzansatz – begrenzt grundsätzlich den Einfluss von problematischen Fällen wie Ausreißerwerten. Er bezieht sich hier auf die geschätzte (mittlere) Anzahl an Korbtreffern. Er wird in Abbildung 7.18 in mehreren Varianten ausgewiesen, spielt aber hier bei den weiteren Erläuterungen keine große Rolle mehr.)

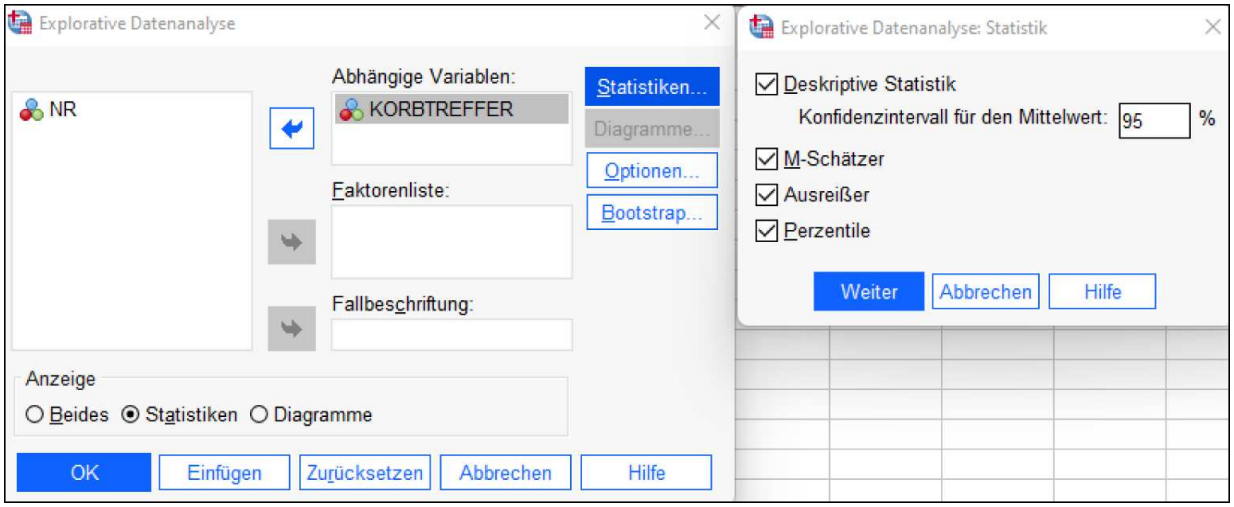
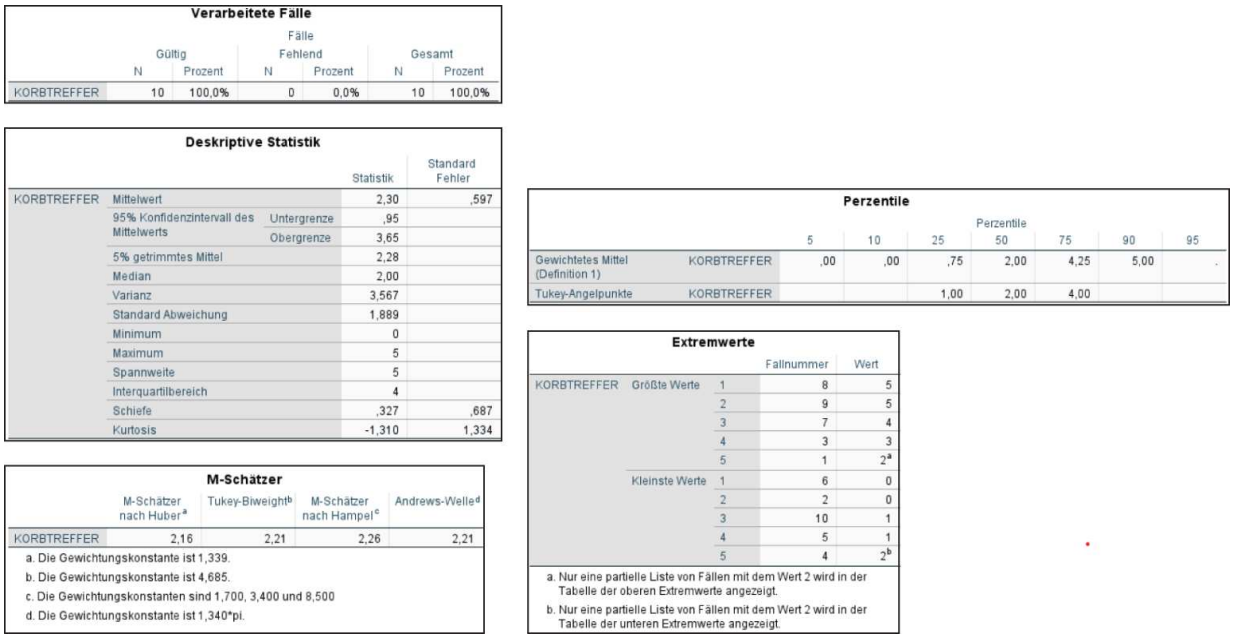


**Abbildung 7.17:** STATISTIKEN bei EXPLORATIVE DATENANALYSE (hier in Bezug auf *KORBTREFFER.SAV*)

Mit den vorstehenden Angaben erhalten Sie den Output gemäß Abbildung 7.18.



**Verarbeitete Fälle**

| | Fälle | | | | | |
|---|---|---|---|---|---|---|
| | Gültig | | Fehlend | | Gesamt | |
| | N | Prozent | N | Prozent | N | Prozent |
| KORBTREFFER | 10 | 100,0% | 0 | 0,0% | 10 | 100,0% |

**Deskriptive Statistik**

| | | | Statistik | Standard Fehler |
|---|---|---|---|---|
| KORBTREFFER | Mittelwert | | 2,30 | ,597 |
| | 95% Konfidenzintervall des Mittelwerts | Untergrenze | ,95 | |
| | | Obergrenze | 3,65 | |
| | 5% getrimmtes Mittel | | 2,28 | |
| | Median | | 2,00 | |
| | Varianz | | 3,567 | |
| | Standard Abweichung | | 1,889 | |
| | Minimum | | 0 | |
| | Maximum | | 5 | |
| | Spannweite | | 5 | |
| | Interquartilbereich | | 4 | |
| | Schiefe | | ,327 | ,687 |
| | Kurtosis | | -1,310 | 1,334 |

**M-Schätzer**

| | M-Schätzer nach Huber[a] | Tukey-Biweight[b] | M-Schätzer nach Hampel[c] | Andrews-Welle[d] |
|---|---|---|---|---|
| KORBTREFFER | 2,16 | 2,21 | 2,26 | 2,21 |

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.

**Perzentile**

| | | Perzentile | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 5 | 10 | 25 | 50 | 75 | 90 | 95 |
| Gewichtetes Mittel (Definition 1) | KORBTREFFER | ,00 | ,00 | ,75 | 2,00 | 4,25 | 5,00 | . |
| Tukey-Angelpunkte | KORBTREFFER | | | 1,00 | 2,00 | 4,00 | | |

**Extremwerte**

| | | | Fallnummer | Wert |
|---|---|---|---|---|
| KORBTREFFER | Größte Werte | 1 | 8 | 5 |
| | | 2 | 9 | 5 |
| | | 3 | 7 | 4 |
| | | 4 | 3 | 3 |
| | | 5 | 1 | 2[a] |
| | Kleinste Werte | 1 | 6 | 0 |
| | | 2 | 2 | 0 |
| | | 3 | 10 | 1 |
| | | 4 | 5 | 1 |
| | | 5 | 4 | 2[b] |

a. Nur eine partielle Liste von Fällen mit dem Wert 2 wird in der Tabelle der oberen Extremwerte angezeigt.
b. Nur eine partielle Liste von Fällen mit dem Wert 2 wird in der Tabelle der unteren Extremwerte angezeigt.

**Abbildung 7.18:** Statistik-Output bei EXPLORATIVE DATENANALYSE (*KORBTREFFER.SAV*)

Die zugehörige Syntax für die vorstehende explorative Auswertung lautet:

```
EXAMINE VARIABLES=korbtreffer
     /PLOT NONE
     /MESTIMATORS HUBER (1.339) ANDREW(1.34) HAMPEL(1.7,3.4,8.5) TUKEY(4.685)
     /PERCENTILES (5,10,25,50,75,90,95) HAVERAGE
     /STATISTICS DESCRIPTIVES EXTREME
     /CINTERVAL 95
     /MISSING LISTWISE
     /NOTOTAL.
```

## *EXAMINE-Bootstrapping*

Führen Sie anhand des vorstehenden »Basketball-Beispiels« auch im Rahmen der explorativen Datenanalyse ein **Bootstrapping** über den Button BOOTSTRAP durch, erhalten Sie das nachstehende (oder von den Wertebereichen her: ein ähnliches) Bild (siehe Abbildung 7.19 und Abbildung 7.20). Dabei wird in einer Spalte beim Bootstrapping übrigens eine »*Verzerrung*« angegeben. Damit ist die Differenz zwischen dem Durchschnittswert der in der entsprechenden Zeile angegebenen statistischen Kennziffer in allen Bootstrap-Stichproben auf der einen Seite und dem Wert in der Spalte »Statistik« auf der anderen Seite gemeint.

**Verarbeitete Fälle**

| | Fälle | | | | | |
|---|---|---|---|---|---|---|
| | Gültig | | Fehlend | | Gesamt | |
| | N | Prozent | N | Prozent | N | Prozent |
| KORBTREFFER | 10 | 100,0% | 0 | 0,0% | 10 | 100,0% |

**Deskriptive Statistik**

| | | | Statistik | Standard Fehler | Bootstrap[a] Verzerrung | Bootstrap[a] Standard Fehler | Bootstrap[a] 95% Konfidenzintervall Unterer | Bootstrap[a] 95% Konfidenzintervall Oberer |
|---|---|---|---|---|---|---|---|---|
| KORBTREFFER | Mittelwert | | 2,30 | ,597 | -,03 | ,56 | 1,10 | 3,40 |
| | 95% Konfidenzintervall des Mittelwerts | Untergrenze | ,95 | | | | | |
| | | Obergrenze | 3,65 | | | | | |
| | 5% getrimmtes Mittel | | 2,28 | | -,03 | ,61 | 1,06 | 3,50 |
| | Median | | 2,00 | | ,04 | ,86 | ,50 | 4,00 |
| | Varianz | | 3,567 | | -,356 | 1,024 | ,991 | 5,167 |
| | Standard Abweichung | | 1,889 | | -,123 | ,304 | ,995 | 2,273 |
| | Minimum | | 0 | | | | | |
| | Maximum | | 5 | | | | | |
| | Spannweite | | 5 | | | | | |
| | Interquartilbereich | | 4 | | 0 | 1 | 1 | 5 |
| | Schiefe | | ,327 | ,687 | ,020 | ,582 | -,870 | 1,546 |
| | Kurtosis | | -1,310 | 1,334 | ,540 | 1,258 | -2,150 | 2,628 |

a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

**Abbildung 7.19:** BOOTSTRAP-Output bei EXPLORATIVE DATENANALYSE (*KORBTREFFER.SAV*) (1)

**M-Schätzer[e]**

| | | Statistik | Bootstrap[f] Verzerrung | Bootstrap[f] Standard Fehler | Bootstrap[f] 95% Konfidenzintervall Unterer | Bootstrap[f] 95% Konfidenzintervall Oberer |
|---|---|---|---|---|---|---|
| KORBTREFFER | M-Schätzer nach Huber[a] | 2,16 | -,01[g] | ,75[g] | ,67[g] | 3,86[g] |
| | Tukey-Biweight[b] | 2,21 | -,09[g] | ,85[g] | ,53[g] | 4,04[g] |
| | M-Schätzer nach Hampel[c] | 2,26 | -,08[g] | ,74[g] | ,68[g] | 3,72[g] |
| | Andrews-Welle[d] | 2,21 | -,10[g] | ,86[g] | ,51[g] | 4,05[g] |

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.
e. Einige M-Schätzer in einer oder mehreren aufgeteilten Dateien können aufgrund der extrem zentralisierten Verteilung um den Median nicht berechnet werden.
f. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben
g. Basierend auf 971 Stichproben

**Perzentile**

| | | Perzentile | Perzentil | Bootstrap[a] Verzerrung | Bootstrap[a] Standard Fehler | Bootstrap[a] 95% Konfidenzintervall Unterer | Bootstrap[a] 95% Konfidenzintervall Oberer |
|---|---|---|---|---|---|---|---|
| Gewichtetes Mittel (Definition 1) | KORBTREFFER | 5 | ,00 | ,10 | ,31 | ,00 | 1,00 |
| | | 10 | ,00 | ,13 | ,31 | ,00 | 1,00 |
| | | 25 | ,75 | -,03 | ,63 | ,00 | 2,00 |
| | | 50 | 2,00 | ,04 | ,86 | ,50 | 4,00 |
| | | 75 | 4,25 | -,35 | 1,02 | 2,00 | 5,00 |
| | | 90 | 5,00 | -,21 | ,50 | 3,00 | 5,00 |
| | | 95 | . | [b] | [b] | [b,c] | [b,c] |
| Tukey-Angelpunkte | KORBTREFFER | 25 | 1,00 | -,18 | ,70 | ,00 | 2,00 |
| | | 50 | 2,00 | ,04 | ,86 | ,50 | 4,00 |
| | | 75 | 4,00 | -,27 | 1,14 | 2,00 | 5,00 |

a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben
b. Basierend auf 0 Stichproben
c. Ein 95%-Konfidenzintervall benötigt mindestens 39 Bootstrap-Stichproben.

**Extremwerte[c]**

| | | | Fallnummer | Wert |
|---|---|---|---|---|
| KORBTREFFER | Größte Werte | 1 | 8 | 5 |
| | | 2 | 9 | 5 |
| | | 3 | 7 | 4 |
| | | 4 | 3 | 3 |
| | | 5 | 1 | 2[a] |
| | Kleinste Werte | 1 | 6 | 0 |
| | | 2 | 2 | 0 |
| | | 3 | 10 | 1 |
| | | 4 | 5 | 1 |
| | | 5 | 4 | 2[b] |

a. Nur eine partielle Liste von Fällen mit dem Wert 2 wird in der Tabelle der oberen Extremwerte angezeigt.
b. Nur eine partielle Liste von Fällen mit dem Wert 2 wird in der Tabelle der unteren Extremwerte angezeigt.
c. Für eine oder mehrere aufgeteilte Dateien überschreitet die angeforderte Anzahl der Extremwerte die Anzahl der Datenpunkte. Eine kleinere Anzahl von Extremwerten wird angezeigt.

**Abbildung 7.20:** BOOTSTRAP-Output bei EXPLORATIVE DATENANALYSE (*KORBTREFFER.SAV*) (2)

# *Mittelwertvergleichs-Prozedur (MEANS)*

Eine weitere Möglichkeit, statistische Kennziffern zu berechnen, bietet SPSS mit dem **Mittel-wertvergleich** über ANALYSIEREN|MITTELWERTE UND PROPORTIONEN VERGLEICHEN|MITTELWERTE an. Das zugehörige SPSS-Syntaxbefehlswort lautet

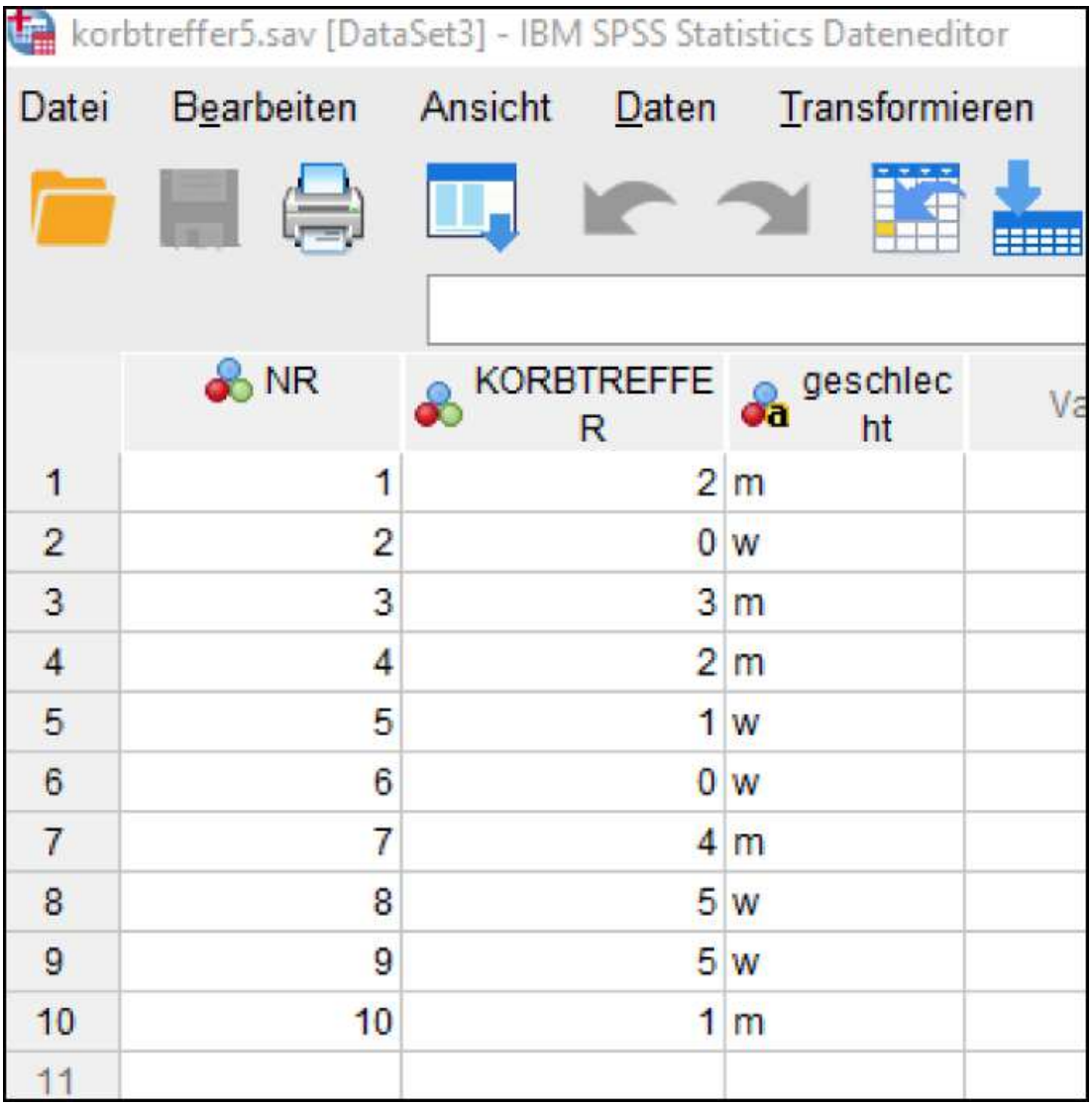
MEANS. Entgegen der Prozedur- beziehungsweise Befehlsbezeichnung gibt SPSS über Mittelwerte hinaus weitere deskriptive Kennziffern aus. (Immerhin ist in SPSS 29 die Prozedur-Bezeichnung von MITTELWERTE VERGLEICHEN in MITTELWERTE UND PROPORTIONEN VERGLEICHEN umbenannt worden.)

Wie aus Tabelle 7.2 zu Beginn dieses Kapitels hervorgeht, können Sie in SPSS über MEANS folgende statistische Kennziffern berechnen:

- ✔ Summe,
- ✔ Arithmetisches Mittel,
- ✔ Median,
- ✔ Gruppierter Median (das heißt: Median bei gruppierten Daten),
- ✔ Harmonisches Mittel,
- ✔ Geometrisches Mittel,
- ✔ Minimum, Maximum und Spannweite,
- ✔ Varianz und Standardabweichung,
- ✔ Schiefe und Wölbung sowie
- ✔ Standardfehler für arithmetisches Mittel, Schiefe und Wölbung.

## *MEANS-Statistiken am »Basketball-Beispiel«*

Anders als bei den vorstehend behandelten statistischen Prozeduren werden bei MEANS deskriptive Statistiken grundsätzlich für eine Variable in *Abhängigkeit* von einer anderen Variablen dargestellt. Hierzu wird die Korbtreffer-Datendatei um eine (String-)Variable *geschlecht* erweitert (*KORBTREFFER5.SAV*), wie in Abbildung 7.21 angegeben. Die Abkürzung »m« steht dabei für »männlich« und »w« für »weiblich«.




| | NR | KORBTREFFER | geschlecht | Va |
|---|---|---|---|---|
| 1 | 1 | 2 | m | |
| 2 | 2 | 0 | w | |
| 3 | 3 | 3 | m | |
| 4 | 4 | 2 | m | |
| 5 | 5 | 1 | w | |
| 6 | 6 | 0 | w | |
| 7 | 7 | 4 | m | |
| 8 | 8 | 5 | w | |
| 9 | 9 | 5 | w | |
| 10 | 10 | 1 | m | |
| 11 | | | | |

**Abbildung 7.21:** Um die Variable *geschlecht* erweiterte Korbtreffer-Datendatei *KORBTREFFER5.SAV*

Im Folgenden stellen Sie bitte die deskriptiven Statistiken für *korbtreffer* in Abhängigkeit von *geschlecht* dar. Dies können Sie – wie in Abbildung 7.22 angegeben – in der betreffenden Subdialogbox MITTELWERTE (von ANALYSIEREN|MITTELWERTE UND PROPORTIONEN VERGLEICHEN) in die Wege leiten.

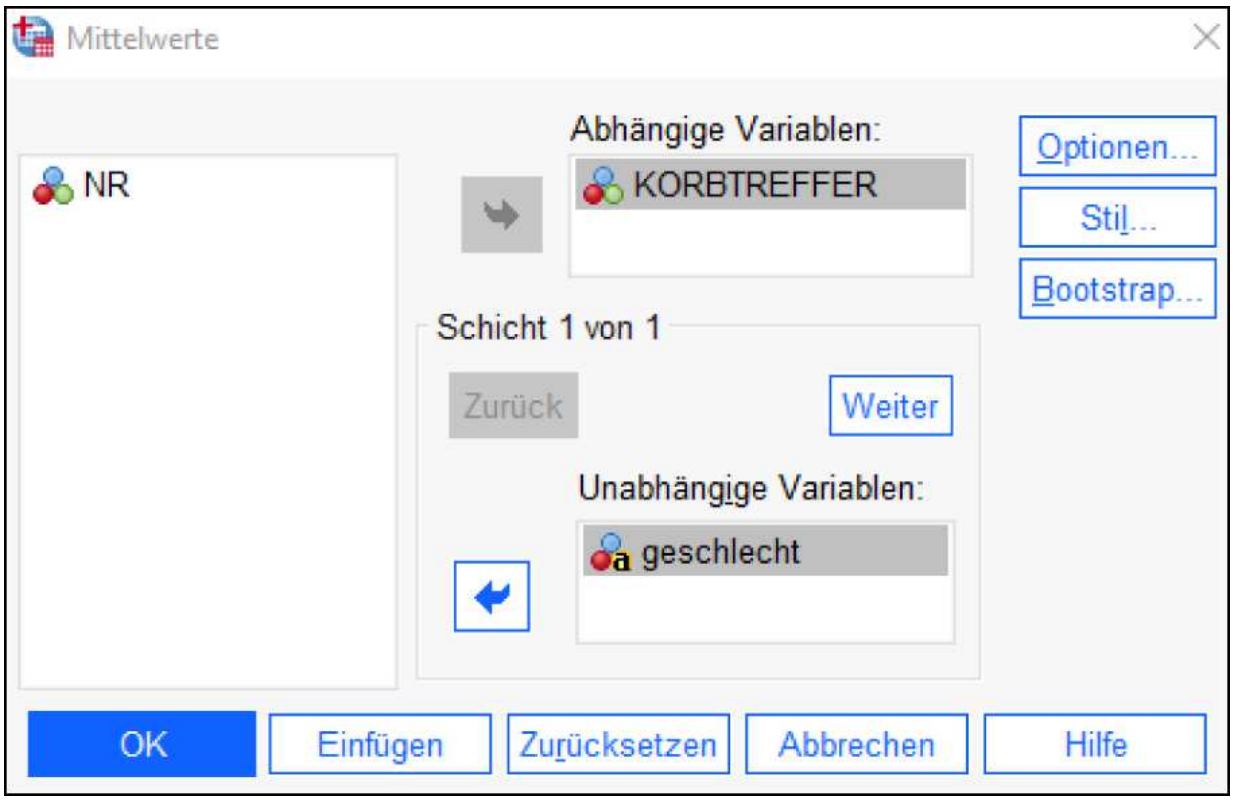



**Abbildung 7.22:** MEANS-Berechnungen für *korbtreffer* in Abhängigkeit von *geschlecht* (*KORBTREFFER5.SAV*)

Als Nächstes wählen Sie unter OPTIONEN die gewünschten Statistiken aus (siehe hierzu Abbildung 7.23). Das geht hier nicht durch Anklicken, sondern durch Ziehen der betreffenden Statistiken von links nach rechts. Nicht ausgewählt werden hier das harmonische und das geometrische Mittel. Das liegt darin begründet, dass sich diese beiden Mittelwerte auf andere Datensachverhalte als hier beziehen.

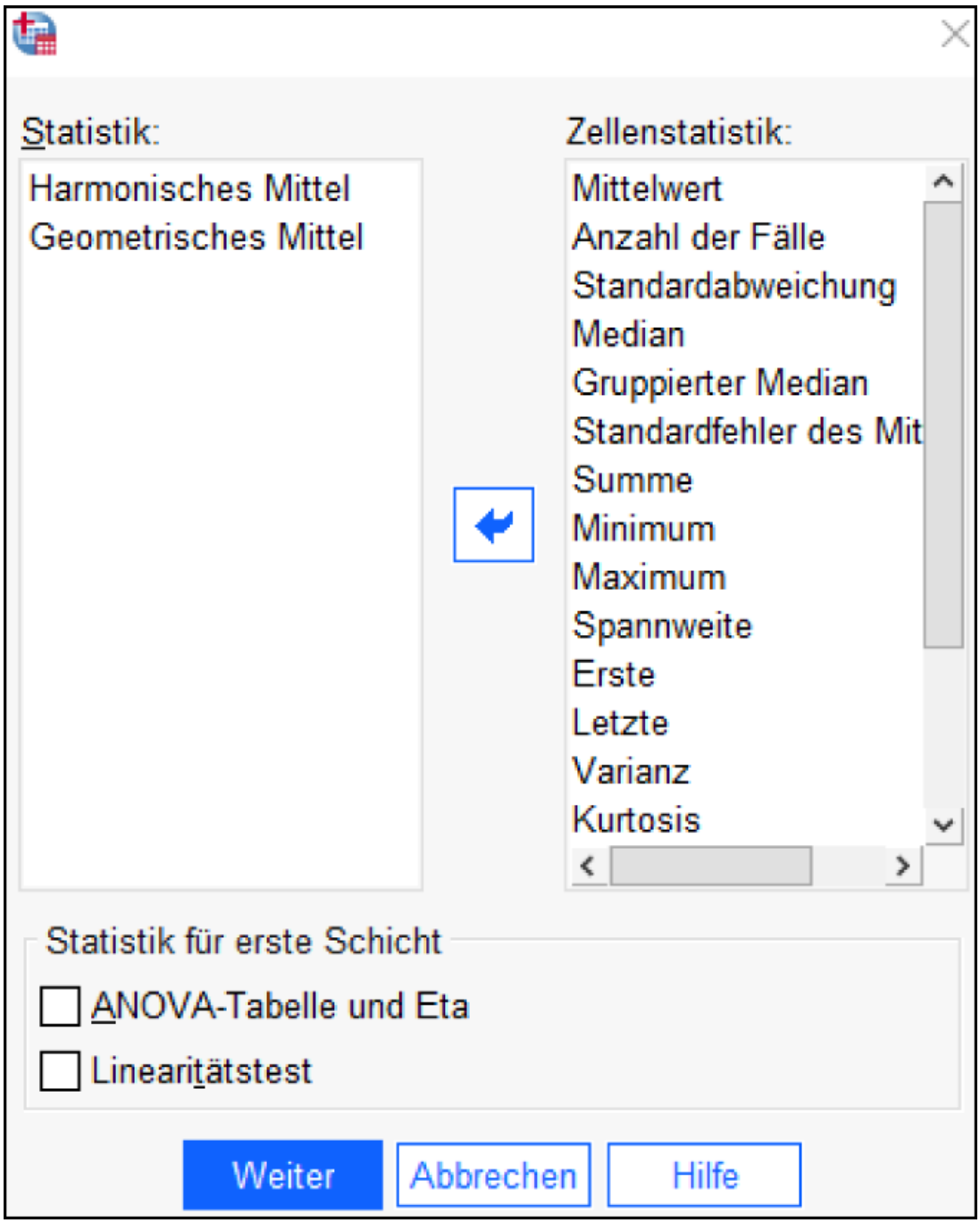



**Abbildung 7.23:** MEANS-Statistiken für *korbtreffer* in Abhängigkeit von *geschlecht* (*KORBTREFFER5.SAV*)

Die zugehörige Syntax ist:

```
MEANS TABLES=korbtreffer BY geschlecht
    /CELLS=MEAN COUNT STDDEV MEDIAN SUM MIN MAX RANGE GMEDIAN
    SEMEAN FIRST LAST VAR KURT SEKURT SKEW SESKEW SPCT NPCT.
```

Mit den vorstehenden Angaben ergibt sich der in Abbildung 7.24 dargestellte Output. Bei den statistischen Kennziffern wird in den beiden ersten Zeilen nach »männlich« (beziehungsweise nach »m«) und »weiblich« (beziehungsweise »w«) differenziert, ehe in

der dritten Zeile die Kennziffern für alle Fälle (also für Frauen und Männer zusammengenommen) angegeben sind.

**Verarbeitete Fälle**

| | Fälle | | | | | |
|---|---|---|---|---|---|---|
| | Eingeschlossen | | Ausgeschlossen | | Insgesamt | |
| | N | Prozent | N | Prozent | N | Prozent |
| KORBTREFFER * geschlecht | 10 | 100,0% | 0 | 0,0% | 10 | 100,0% |

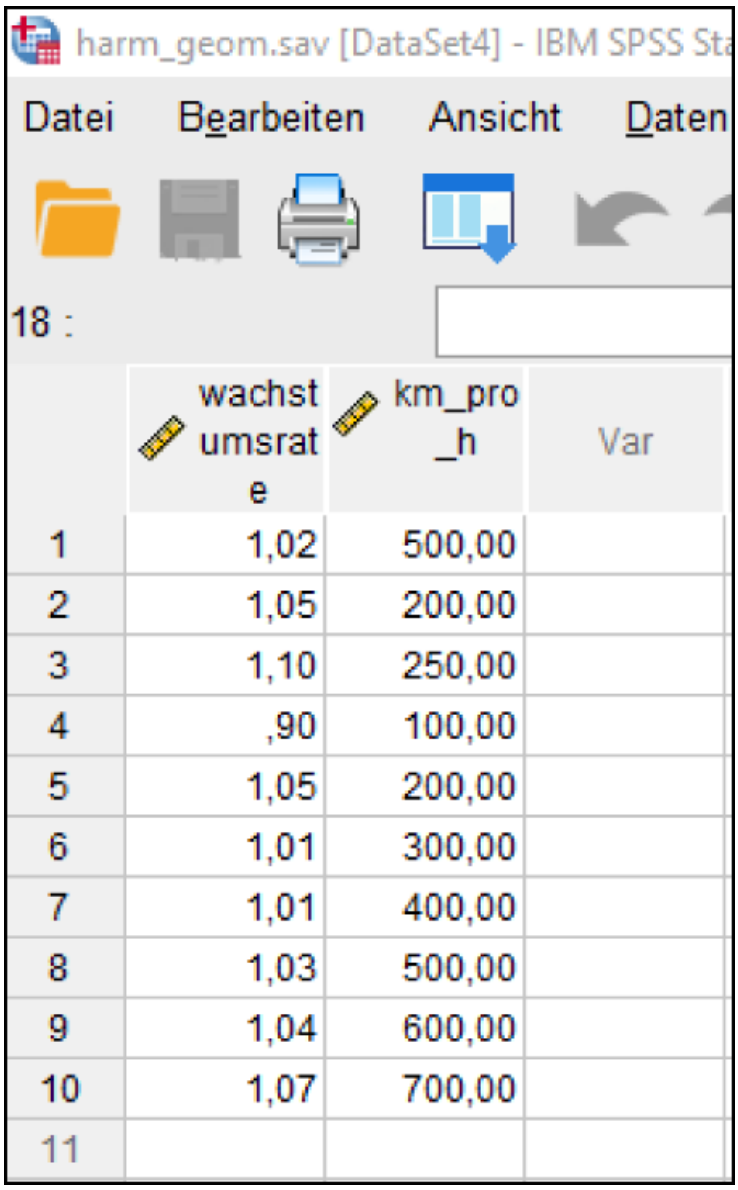
**Bericht**

KORBTREFFER

| geschlecht | Mittelwert | N | Std.-Abweichung | Median | Gruppierter Median | Standardfehler des Mittelwerts | Summe | Minimum | Maximum | Spannweite | Erste | Letzte | Varianz | Kurtosis | Standardfehler der Kurtosis | Schiefe | % der Gesamtsumme | % der Gesamtanzahl | Standardfehler der Schiefe |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| m | 2,40 | 5 | 1,140 | 2,00 | 2,33 | ,510 | 12 | 1 | 4 | 3 | 2 | 1 | 1,300 | -,178 | 2,000 | ,405 | 52,2% | 50,0% | ,913 |
| w | 2,20 | 5 | 2,588 | 1,00 | 1,00 | 1,158 | 11 | 0 | 5 | 5 | 0 | 5 | 6,700 | -3,215 | 2,000 | ,502 | 47,8% | 50,0% | ,913 |
| Insgesamt | 2,30 | 10 | 1,889 | 2,00 | 2,00 | ,597 | 23 | 0 | 5 | 5 | 2 | 1 | 3,567 | -1,310 | 1,334 | ,327 | 100,0% | 100,0% | ,687 |

**Abbildung 7.24:** MEANS-Statistikoutput für *korbtreffer* in Abhängigkeit von *geschlecht* (*KORBTREFFER5.SAV*)

Um auch noch die MEANS-Wirkweise für das harmonische und das geometrische Mittel darzulegen, wird im Folgenden auf den Datensatz *HARM_GEOM.SAV* Bezug genommen (siehe Abbildung 7.25).




| | wachstumsrate | km_pro_h | Var |
|---|---|---|---|
| 1 | 1,02 | 500,00 | |
| 2 | 1,05 | 200,00 | |
| 3 | 1,10 | 250,00 | |
| 4 | ,90 | 100,00 | |
| 5 | 1,05 | 200,00 | |
| 6 | 1,01 | 300,00 | |
| 7 | 1,01 | 400,00 | |
| 8 | 1,03 | 500,00 | |
| 9 | 1,04 | 600,00 | |
| 10 | 1,07 | 700,00 | |
| 11 | | | |

**Abbildung 7.25:** Die Datendatei *HARM_GEOM.SAV*

Da das harmonische beziehungsweise das geometrische Mittel über alle Fälle berechnet werden soll, wird an dieser Stelle ein kleiner Trick eingeführt: Sie können einfach in der Rubrik der unabhängigen Variablen nichts eingeben, und Sie erhalten ein Gesamtergebnis. Alternativ hätten Sie aber auch als unabhängige, das heißt als Gruppierungsvariable die im Datensatz enthaltene Variable *alle* verwenden können. Auch dann wird über alle Fälle hinweg das harmonische beziehungsweise das geometrische Mittel berechnet.

Mit diesen Setzungen resultiert als geometrisches Mittel für *wachstumsrate* der Output gemäß Abbildung 7.26.

Das harmonische Mittel für *km_pro_h* ist mit der entsprechenden Vorgehensweise auf MEANS-Basis wie in Abbildung 7.27 dargestellt berechnet worden.

**Abbildung 7.26:** Geometrisches Mittel für *wachstumsrate* (MEANS) (*HARM_GEOM.SAV*)

**Abbildung 7.27:** Harmonisches Mittel für *km_pro_h* (MEANS) (*HARM_GEOM.SAV*)

## *MEANS-Bootstrapping*

Wenn Sie anhand des vorstehenden, um *geschlecht* erweiterten »Basketball-Beispiels« (*KORBTREFFER5.SAV*) ein MEANS-**Bootstrapping** über den Button BOOTSTRAP durchführen, sieht das Ganze von den Werten her zumindest so ähnlich wie in Abbildung 7.28 aus.

**Bericht**

KORBTREFFER

| geschlecht | | Statistik | Bootstrap[a] Verzerrung | Std.-Fehler | 95% Konfidenzintervall Unterer | Oberer |
|---|---|---|---|---|---|---|
| m | Mittelwert | 2,40 | ,01 | ,49 | 1,50 | 3,40 |
| | N | 5 | 0 | 2 | 2 | 8 |
| | Std.-Abweichung | 1,140 | -,169[c] | ,340[c] | ,303[c] | 1,528[c] |
| | Median | 2,00 | ,35 | ,65 | 1,00 | 4,00 |
| | Gruppierter Median | 2,33 | ,06 | ,54 | 1,50 | 3,50 |
| | Standardfehler des Mittelwerts | ,510 | | | | |
| | Summe | 12 | | | | |
| | Minimum | 1 | | | | |
| | Maximum | 4 | | | | |
| | Spannweite | 3 | | | | |
| | Erste | 2 | | | | |
| | Letzte | 1 | | | | |
| | Varianz | 1,300 | -,241[c] | ,664[c] | ,115[c] | 2,333[c] |
| | Kurtosis | -,178 | ,018[g] | 2,489[g] | -6,000[g] | 5,000[g] |
| | Standardfehler der Kurtosis | 2,000 | | | | |
| | Schiefe | ,405 | -,226[e] | 1,008[e] | -2,000[e] | 2,000[e] |
| | % der Gesamtsumme | 52,2% | | | | |
| | % der Gesamtanzahl | 50,0% | | | | |
| | Standardfehler der Schiefe | ,913 | | | | |
| w | Mittelwert | 2,20 | ,00[b] | 1,09[b] | ,00[b] | 4,20[b] |
| | N | 5 | 0[b] | 2[b] | 2[b] | 8[b] |
| | Std.-Abweichung | 2,588 | -,386[d] | ,731[d] | ,000[d] | 2,887[d] |
| | Median | 1,00 | 1,00[b] | 1,97[b] | ,00[b] | 5,00[b] |
| | Gruppierter Median | 1,00 | ,86[b] | 1,24[b] | ,00[b] | 4,20[b] |
| | Standardfehler des Mittelwerts | 1,158 | | | | |
| | Summe | 11 | | | | |
| | Minimum | 0 | | | | |
| | Maximum | 5 | | | | |
| | Spannweite | 5 | | | | |
| | Erste | 0 | | | | |
| | Letzte | 5 | | | | |
| | Varianz | 6,700 | -1,316[d] | 2,338[d] | ,000[d] | 8,333[d] |
| | Kurtosis | -3,215 | 2,671[h] | 3,701[h] | -6,000[h] | 6,000[h] |
| | Standardfehler der Kurtosis | 2,000 | | | | |
| | Schiefe | ,502 | -,119[f] | 1,298[f] | -2,236[f] | 2,250[f] |
| | % der Gesamtsumme | 47,8% | | | | |
| | % der Gesamtanzahl | 50,0% | | | | |
| | Standardfehler der Schiefe | ,913 | | | | |
| Insgesamt | Mittelwert | 2,30 | ,02 | ,56 | 1,30 | 3,40 |
| | N | 10 | 0 | 0 | 10 | 10 |
| | Std.-Abweichung | 1,889 | -,112 | ,293 | 1,059 | 2,263 |
| | Median | 2,00 | ,16 | ,89 | 1,00 | 4,00 |
| | Gruppierter Median | 2,00 | ,18 | ,78 | ,86 | 4,00 |
| | Standardfehler des Mittelwerts | ,597 | | | | |
| | Summe | 23 | | | | |
| | Minimum | 0 | | | | |
| | Maximum | 5 | | | | |
| | Spannweite | 5 | | | | |
| | Erste | 2 | | | | |
| | Letzte | 1 | | | | |
| | Varianz | 3,567 | -,324 | ,992 | 1,122 | 5,122 |
| | Kurtosis | -1,310 | ,442 | 1,113 | -2,123 | 2,224 |
| | Standardfehler der Kurtosis | 1,334 | | | | |
| | Schiefe | ,327 | -,039 | ,563 | -,753 | 1,360 |
| | % der Gesamtsumme | 100,0% | | | | |
| | % der Gesamtanzahl | 100,0% | | | | |
| | Standardfehler der Schiefe | ,687 | | | | |

a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

b. Basierend auf 998 Stichproben

c. Basierend auf 991 Stichproben

d. Basierend auf 986 Stichproben

e. Basierend auf 934 Stichproben

f. Basierend auf 929 Stichproben

g. Basierend auf 817 Stichproben

h. Basierend auf 832 Stichproben

**Abbildung 7.28:** BOOTSTRAP-Output bei MITTELWERTE VERGLEICHEN (*KORBTREFFER5.SAV*)

# *Fallzusammenfassungs-Prozedur (SUMMARIZE)*

Über die **Fallzusammenfassungs-Prozedur** ANALYSIEREN|BERICHTE|FALLZUSAMMENFASSUNGEN stehen Ihnen ebenfalls statistische Messziffern zur Verfügung. Allerdings benötigen Sie an dieser Stelle – analog zu MITTELWERTE UND PROPORTIONEN VERGLEICHEN – grundsätzlich eine Gruppierungsvariable. Das SPSS-Syntaxbefehlswort heißt SUMMARIZE.

Wie aus Tabelle 7.2 hervorgeht, können Sie in SPSS über SUMMARIZE die gleichen statistischen Kennziffern wie bei MEANS berechnen:

- ✔ Summe,
- ✔ Arithmetisches Mittel,
- ✔ Median,
- ✔ Gruppierter Median,
- ✔ Harmonisches Mittel,

- ✔ Geometrisches Mittel,
- ✔ Minimum, Maximum und Spannweite,
- ✔ Varianz und Standardabweichung,
- ✔ Schiefe und Wölbung sowie
- ✔ Standardfehler für arithmetisches Mittel, Schiefe und Wölbung.

Mit Ausnahme des harmonischen und geometrischen Mittels sind via SUMMARIZE-Prozedur für das »Basketball-Beispiel« (Datendatei *KORBTREFFER5.SAV*) die vorstehenden Statistiken berechnet worden. Wichtig ist dabei, dass Sie in der Fallzusammenfassungs-Dialogbox das Häkchen bei FÄLLE ANZEIGEN entfernen, um über alle Fälle zu rechnen (siehe Abbildung 7.29).

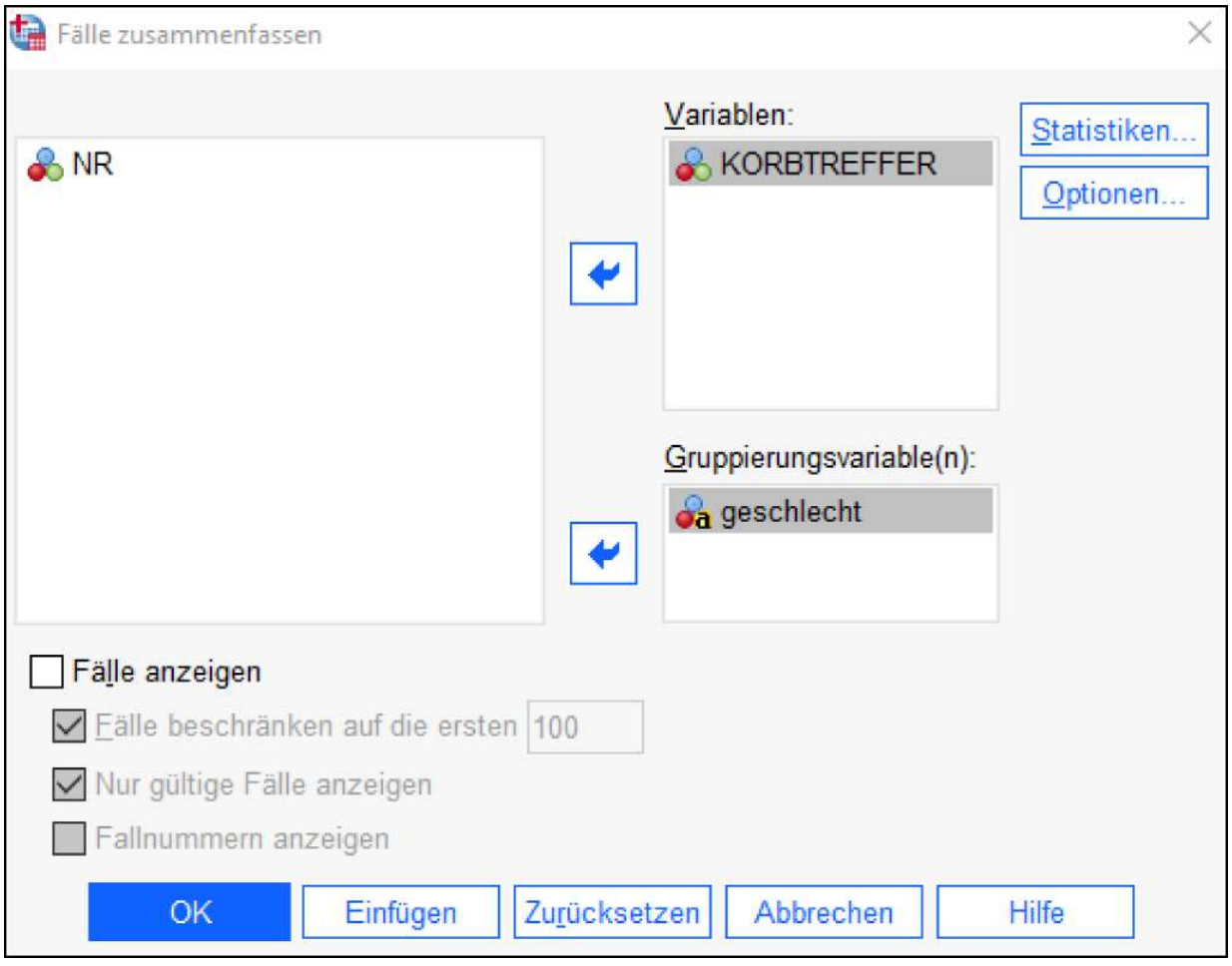



**Abbildung 7.29:** Alle Fälle berücksichtigen bei SUMMARIZE (hier in Bezug auf *KORBTREFFER5.SAV*)

Die Ergebnisse der SUMMARIZE-Berechnungen sind in Abbildung 7.30 zusammengestellt.

Die Syntax für den betreffenden Output sieht so aus:

```
SUMMARIZE
    /TABLES=korbtreffer BY geschlecht
    /FORMAT=NOLIST TOTAL
    /TITLE='Zusammenfassung von Fällen'
    /MISSING=VARIABLE
    /CELLS=COUNT MEAN MEDIAN GMEDIAN SEMEAN SUM MIN MAX RANGE
    FIRST LAST STDDEV VAR KURT SEKURT SKEW SESKEW SPCT NPCT.
```

**Zusammenfassung von Fällen**

KORBTREFFER

| geschlecht | N | Mittelwert | Median | Gruppierter Median | Standardfehler des Mittelwerts | Summe | Minimum | Maximum | Spannweite | Erste | Letzte | Std.-Abweichung | Varianz | Kurtosis | Standardfehler der Kurtosis | Schiefe | Standardfehler der Schiefe | % der Gesamtsumme | % der Gesamtanzahl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| m | 5 | 2,40 | 2,00 | 2,33 | ,510 | 12 | 1 | 4 | 3 | 2 | 1 | 1,140 | 1,300 | -,178 | 2,000 | ,405 | ,913 | 52,2% | 50,0% |
| w | 5 | 2,20 | 1,00 | 1,00 | 1,158 | 11 | 0 | 5 | 5 | 0 | 5 | 2,588 | 6,700 | -3,215 | 2,000 | ,502 | ,913 | 47,8% | 50,0% |
| Insgesamt | 10 | 2,30 | 2,00 | 2,00 | ,597 | 23 | 0 | 5 | 5 | 2 | 1 | 1,889 | 3,567 | -1,310 | 1,334 | ,327 | ,687 | 100,0% | 100,0% |

**Abbildung 7.30:** SUMMARIZE-Statistikoutput für *korbtreffer* in Abhängigkeit von *geschlecht* (*KORBTREFFER5.SAV*)

Für SUMMARIZE steht das **Bootstrapping** nicht zur Verfügung, was meiner Meinung nach plausibel ist, da Sie bei SUMMARIZE bestimmte Fälle auswählen können, was dem Bootstrapping-Grundgedanken widerspricht, da sich das Bootstrapping auf eine Zufallsauswahl aus *allen* Fällen bezieht.

In Teil III geht es anschließend um Methoden, die man entweder der deskriptiven Statistik (nämlich bei Vorliegen von Daten aus Totalerhebungen) oder der induktiven Statistik (nämlich bei Verwendung von Stichprobendaten) begründet zuordnen kann. Im Einzelnen werden Ihnen bi- beziehungsweise multivariate Zusammenhänge nähergebracht, und zwar vornehmlich im Rahmen der Korrelations- und der Regressionsanalyse, aber auch etwa in Form der Diskriminanz-, der Cluster- und der Faktoranalyse.

# Teil III

# Deskriptive und induktive SPSS-Statistiken gemischt

# IN DIESEM TEIL …

- ✔ Mehrdimensionale Häufigkeitsverteilungen
- ✔ Korrelationen, Regressionen
- ✔ Cluster-, Diskriminanz-, Faktorenanalyse
- ✔ Überlebensanalysen

# Kapitel 8

# Was alles mehr oder weniger miteinander zusammenhängt: Mehrdimensionale Häufigkeitsverteilungen

## IN DIESEM KAPITEL

Kreuztabellen

Assoziationen

Korrelationen

Das Thema dieses Kapitels sind mehrdimensionale Häufigkeitsverteilungen. Im Wesentlichen geht es hierbei um die statistischen Zusammenhänge zwischen zwei (und mehr) Variablen, die man grundsätzlich in einer Tabelle, einer sogenannten Kreuztabelle, darstellen kann. In Bezug auf die statistischen Zusammenhänge zwischen zwei nominalskalierten Merkmalen spricht man von Assoziationen, bezüglich der statistischen Zusammenhänge zwischen zwei mindestens ordinalskalierten Merkmalen von Korrelationen.

Dieses Kapitel ist ebenso wie die nachfolgenden Kapitel 9 bis 12 Teil III: »Deskriptive und induktive SPSS-Statistiken gemischt« zugeordnet worden. Der Grund hierfür ist, dass die in den betreffenden Kapiteln vorgestellten Methoden (begründet) bei Bezugnahme auf Grundgesamtheiten als Datenbasis deskriptiv (also: beschreibend) und bei Bezugnahme auf Stichproben als Datengrundlage induktiv (im Sinne des statistischen Schließens von den Stichprobenergebnissen auf die Grundgesamtheitswerte) konzipiert sind.

# *Vorbemerkungen zu den Zusammenhängen zwischen mehreren Variablen*

Bislang wurde in den vorstehenden Kapiteln jeweils nur die Verteilung einer einzelnen Variablen diskutiert. Sie können aber statistische Zusammenhänge auch zwischen mehreren Variablen betrachten. Eine typische Einteilung für mehrdimensionale Häufigkeitsverteilungen mit den entsprechenden Analysemöglichkeiten ist folgende:

- ✔ Zusammenhänge zwischen zwei Variablen (bivariate Zusammenhänge):
  - Assoziation (erforderliches Skalenniveau: nominalskalierte Variablen),
  - Rangkorrelation (erforderliches Skalenniveau: ordinalskalierte Variablen),
  - Korrelation (erforderliches Skalenniveau in der Regel: metrisch skalierte Variablen, ohne dass eine Ursache-Wirkungs-Beziehung eine Rolle spielt),
  - Regression (erforderliches Skalenniveau typischerweise: metrisch skalierte Variablen; im Sinne einer Ursache-Wirkungs-Beziehung).
- ✔ Zusammenhänge zwischen mehr als zwei Variablen (multivariate Zusammenhänge):
  - Multiple Regression (erforderliches Skalenniveau typischerweise: metrisch skalierte Variablen, aber auch Dummies als unabhängige Variablen möglich),
  - Logistische Regression und Ähnliches (erforderliches Skalenniveau: dichotome Variablen als abhängige Variablen) sowie
  - Diskriminanz-, Cluster- und Faktorenanalyse.

- Von den vorstehend genannten Verfahren wird auf die Regressionsverfahren erst im nächsten und übernächsten Kapitel eingegangen, auf die Diskriminanz-, Cluster- und Faktorenanalyse in Kapitel 11. In diesem Kapitel geht es demgegenüber um die Zusammenhänge zwischen zwei Variablen im Sinne von Assoziationen und Korrelationen.

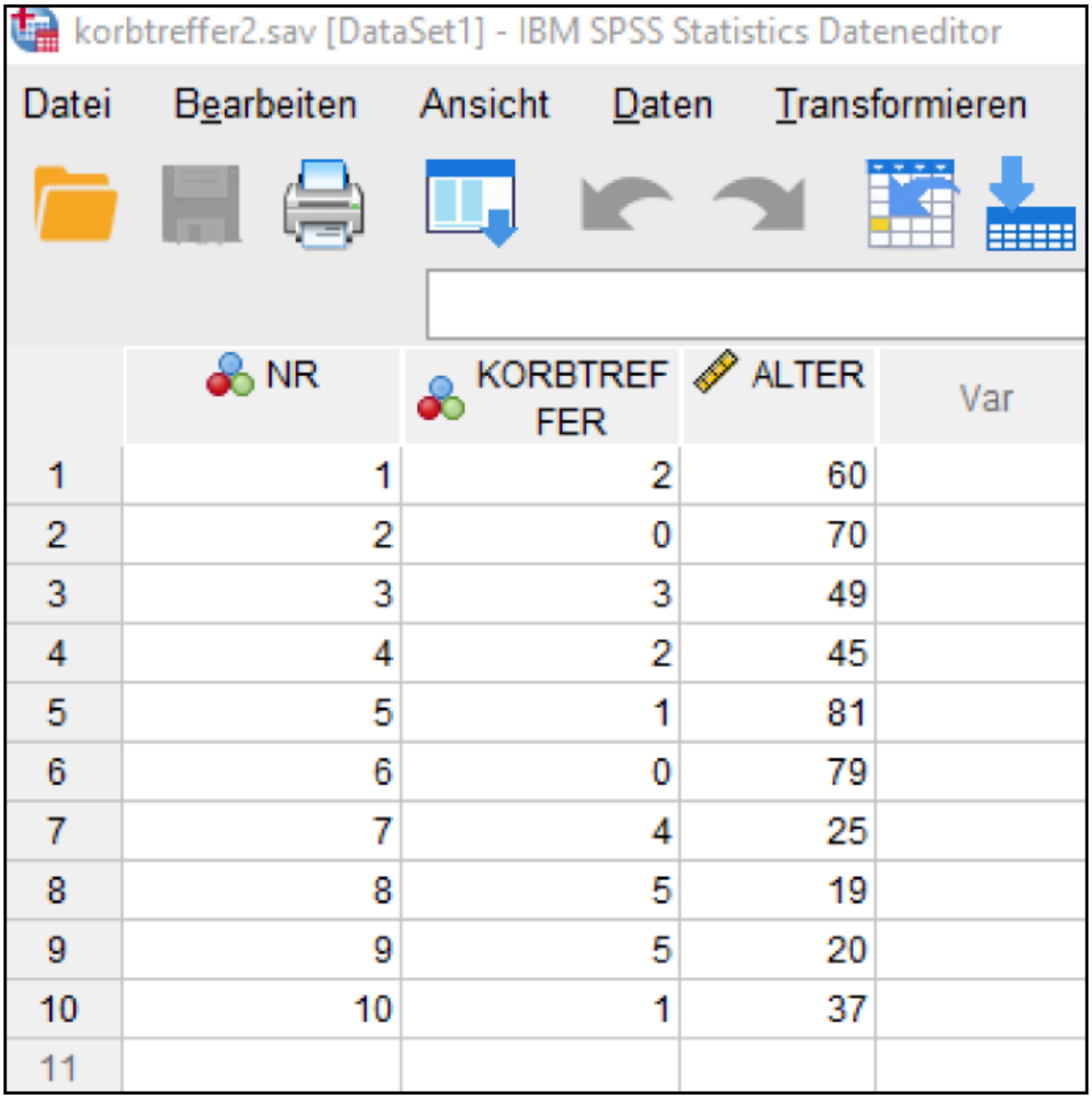
# *Kreuztabelle*

Ausgangspunkt für Assoziations- und Korrelationsbetrachtungen ist im bivariaten Zusammenhang vielfach eine **Kreuztabelle**. Die Ausprägungen der einen Variablen stehen dabei in den Zeilen, die der anderen Variablen in den Spalten. In den Zellen befinden sich dann die absoluten oder die relativen Häufigkeiten.

## *Aufbau einer Kreuztabelle in SPSS*

In SPSS erhalten Sie eine bivariate Häufigkeitsverteilung im SPSS-Menü über ANALYSIEREN|DESKRIPTIVE STATISTIKEN|KREUZTABELLEN. Diese Kreuztabelle können Sie über die Subdialogbox ZELLEN gestalten, wie im Folgenden noch anhand von Beispielen gezeigt wird. Der korrespondierende SPSS-Befehl zum genannten SPSS-Menü ist CROSSTABS.

Um Ihnen die Wirkweise des KREUZTABELLEN-Menüs zu illustrieren, habe ich das bereits mehrfach in den vorhergehenden Kapiteln erwähnte »Basketball-Beispiel« um eine Variable *alter* erweitert (siehe Abbildung 8.1; die zugehörige SPSS-Datendatei heißt *KORBTREFFER2.SAV* ).




| | NR | KORBTREFFER | ALTER | Var |
|---|---|---|---|---|
| 1 | 1 | 2 | 60 | |
| 2 | 2 | 0 | 70 | |
| 3 | 3 | 3 | 49 | |
| 4 | 4 | 2 | 45 | |
| 5 | 5 | 1 | 81 | |
| 6 | 6 | 0 | 79 | |
| 7 | 7 | 4 | 25 | |
| 8 | 8 | 5 | 19 | |
| 9 | 9 | 5 | 20 | |
| 10 | 10 | 1 | 37 | |
| 11 | | | | |

**Abbildung 8.1:** Erweiterte »Basketball-Beispiel«-Datendatei (*KORBTREFFER2.SAV*)

Kreuztabellierungen sind nur dann sinnvoll, wenn Sie sowohl für die X- als auch für die Y-Variable relativ wenige Ausprägungen vorliegen haben. Ansonsten wird die Darstellung (viel) zu unübersichtlich. Kreuztabellen sind daher vor allem für kategoriale, also für gruppierte Variablen geeignet, letztlich also für nominale, ordinale oder diskrete Variablen mit vergleichsweise wenigen Ausprägungen, aber eben auch für klassifizierte Variablen (also Variablen, die in verschiedene Klassen eingeteilt sind).

Würden Sie nämlich zum Beispiel eine Altersvariable mit Werten von 0 Jahren (Neugeborene) bis 100 Jahre mit der für jedes Alter zu erwartenden durchschnittlichen Restlebenserwartung kreuztabellieren, erhielten Sie eine Tabelle, die aus 101 Zeilen und ebenfalls 101 Spalten besteht. Das ist natürlich alles andere als übersichtlich!

Für *korbtreffer* – das heißt für die Anzahl der Korbtreffer einer Person bei jeweils fünf Versuchen – haben Sie ja nur sechs Ausprägungen – sozusagen sechs Kategorien – von 0 bis 5. Damit die Altersvariable eine kategoriale Größe ist, wird sie im Folgenden klassifiziert zur Variablen *alter_klasse* mit den Altersklassen »Bis 24 Jahre«, »25 bis 44 Jahre«, »45 bis 64 Jahre« und »65 Jahre und älter«. Dies ist über die nachfolgende SPSS-Syntax realisiert worden (und in *KORBTREFFER3.SAV* abgespeichert worden):

```
RECODE alter (0 thru 24=1) (25 thru 44=2) (45 thru 64=3) (ELSE=4) INTO
alter_klasse.
EXECUTE.
```

Im nächsten Schritt können jetzt bereits – in *KORBTREFFER3.SAV* – die Variablen *korbtreffer* und *alter_klasse* miteinander kreuztabelliert werden. Das tun Sie einfach, indem Sie im Rahmen des bereits erwähnten SPSS-Menüs ANALYSIEREN|DESKRIPTIVE STATISTIKEN|KREUZTABELLEN die beiden Variablennamen eingeben (siehe Abbildung 8.2).

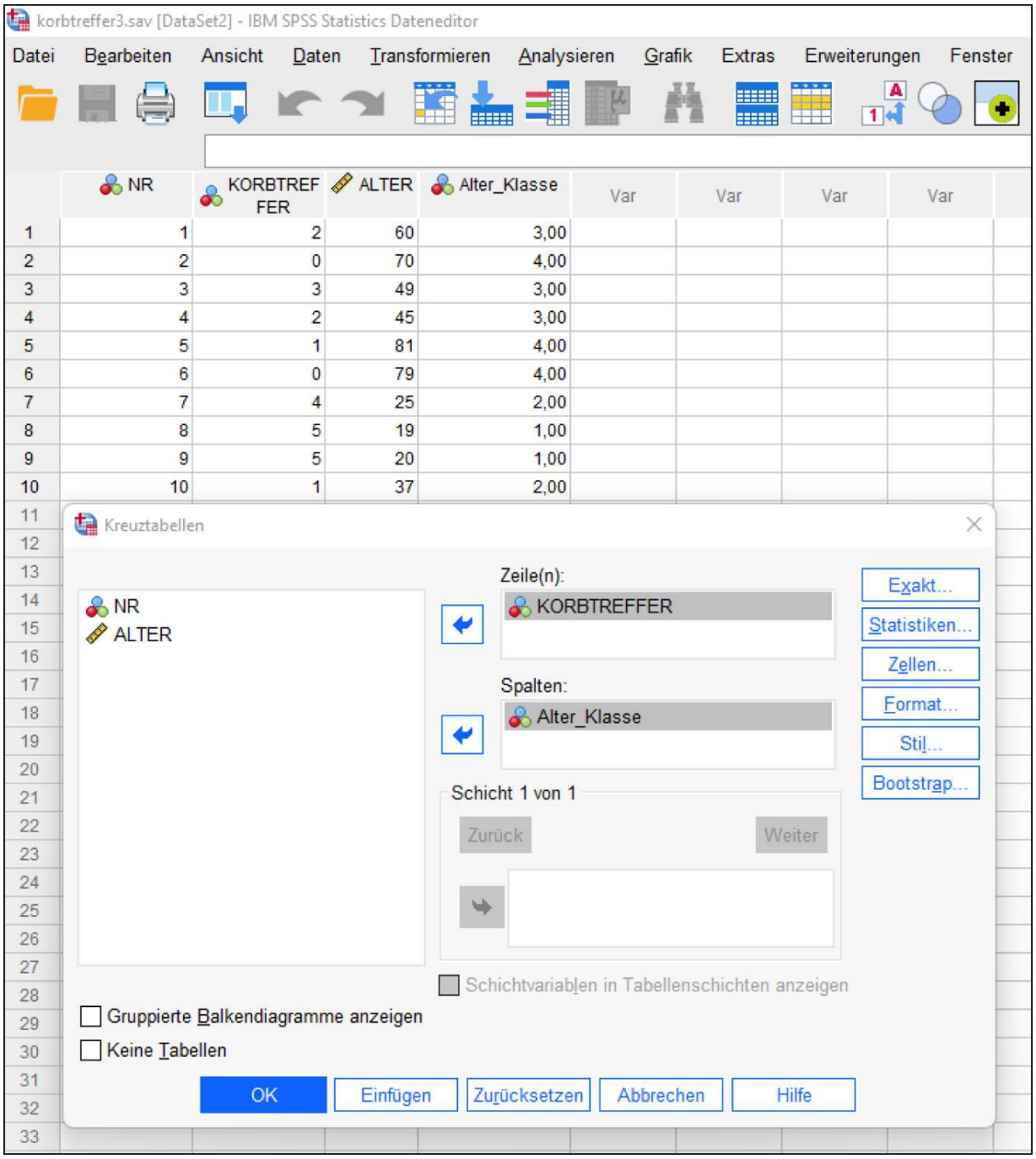



**Abbildung 8.2:** Kreuztabellierungsanweisung für *korbtreffer* und *alter_klasse* (*KORBTREFFER3.SAV* )

## *Absolute Häufigkeiten*

Als Nächstes können Sie in der Subdialogbox ZELLEN ANZEIGEN verschiedene Zelleninhalte auswählen, wobei Sie grundsätzlich zwischen absoluten und relativen Häufigkeiten unterscheiden können.

Wählen Sie zum Beispiel, wie in Abbildung 8.3 angegeben, nur die **beobachteten Häufigkeiten** aus, werden Ihnen für den Treffer-Alters-Zusammenhang die entsprechenden absoluten Häufigkeiten ausgegeben.

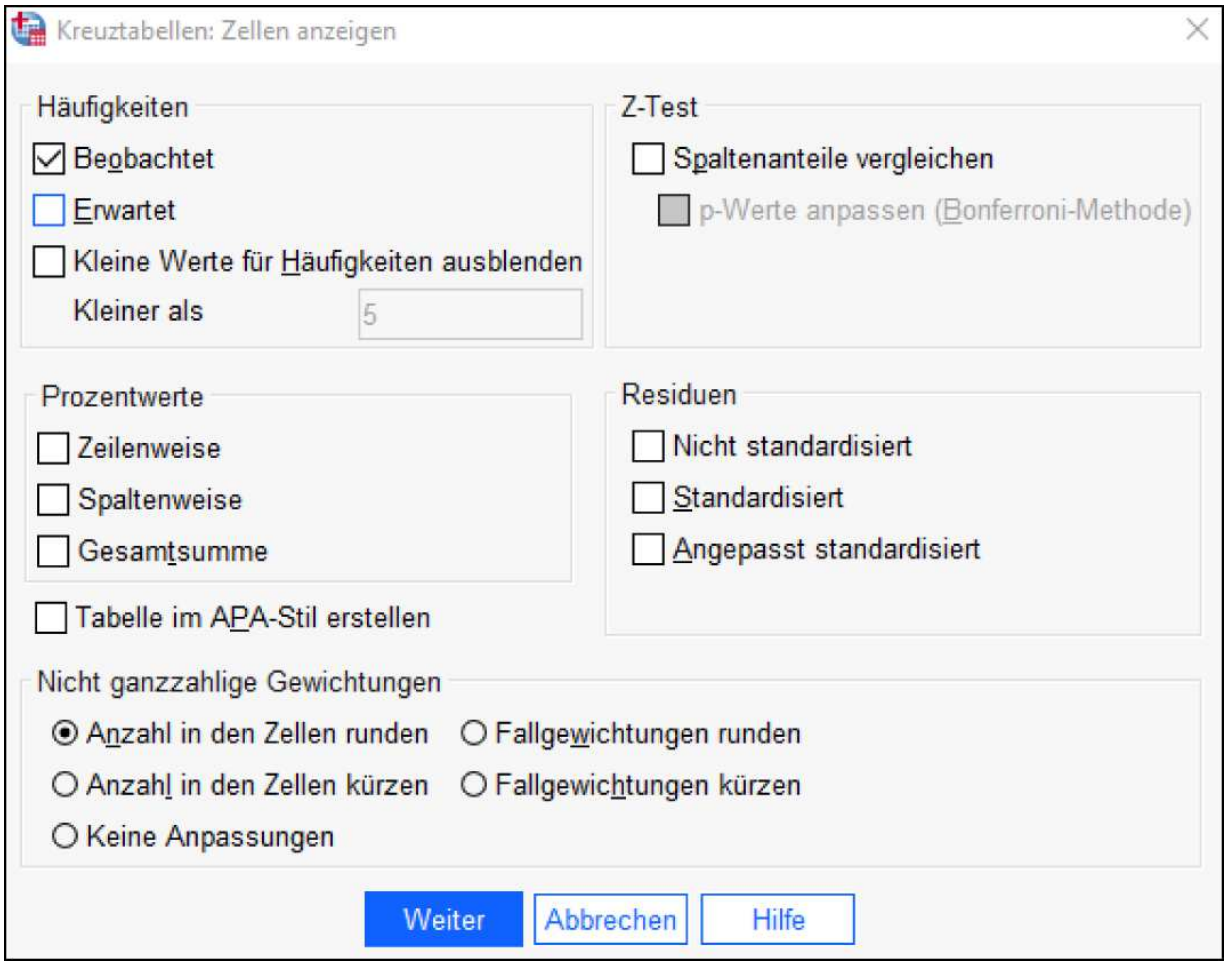



**Abbildung 8.3:** Auswahl der beobachteten Fallzahlen in einer Kreuztabelle (*KORBTREFFER3.SAV*)

Als entsprechende SPSS-Syntax hätten Sie schreiben müssen:

```
CROSSTABS
    /TABLES=korbtreffer BY alter_klasse
    /CELLS=COUNT.
```

Die sich ergebende Kreuztabelle zwischen Treffern und Altersklassen ist in Abbildung 8.4 dokumentiert. Dieser Abbildung ist zum Beispiel zu entnehmen, dass genau ein Korbtreffer (bei fünf Versuchen) sowohl in der zweiten (»25 bis 44 Jahre«) als auch in der vierten Altersklasse (»65 Jahre und älter«) aufgetreten ist.

**KORBTREFFER * Alter_Klasse Kreuztabelle**

Anzahl

| | | Alter_Klasse | | | | |
|---|---|---|---|---|---|---|
| | | 1,00 | 2,00 | 3,00 | 4,00 | Gesamt |
| KORBTREFFER | 0 | 0 | 0 | 0 | 2 | 2 |
| | 1 | 0 | 1 | 0 | 1 | 2 |
| | 2 | 0 | 0 | 2 | 0 | 2 |
| | 3 | 0 | 0 | 1 | 0 | 1 |
| | 4 | 0 | 1 | 0 | 0 | 1 |
| | 5 | 2 | 0 | 0 | 0 | 2 |
| Gesamt | | 2 | 2 | 3 | 3 | 10 |

**Abbildung 8.4:** Kreuztabellenoutput für die beobachteten Fallzahlen (*KORBTREFFER3.SAV*)

Die Zeilen- und die Spaltensummen geben die **Randverteilungen** der beiden betrachteten Variablen an. Dabei stellen die Zeilensummen in unserem Beispiel die Fallzahlensummen für die Variable *korbtreffer* und die Spaltensummen die Fallzahlensummen für die Variable *alter_klasse* dar. Aus der entsprechenden Randverteilung für *korbtreffer* können Sie also ablesen, dass 0, 1, 2 und 5 Treffer jeweils zweimal sowie 3 und 4 Treffer jeweils einmal aufgetreten sind. Die Betrachtung der Randverteilung für *alter_klasse* ergibt, dass zwei Leute in die erste Altersklasse (»Bis 24 Jahre«), ebenfalls zwei Personen in die zweite Altersklasse (»25 bis 44 Jahre«) und je drei Personen in die dritte (»45 bis 64 Jahre«) und die vierte Altersklasse (»65 Jahre und älter«) fallen. Ganz unten rechts findet sich in einer Kreuztabelle die gesamte Fallzahl, in unserem Beispiel ja bekanntlich 10 Fälle.

Grundsätzlich können Sie eine Tabelle wie die in Abbildung 8.4 angegebene auch mit dem CTABLES-Befehl in SPSS erzeugen. Um die vorstehende Tabelle (ohne die Randsummen) zu erstellen, hätten Sie entsprechend auch folgende SPSS-Syntax eingeben können:

```
CTABLES
/TABLE korbtreffer BY alter_klasse.
```

Sie können in einer SPSS-Kreuztabelle auch die sogenannten **erwarteten Häufigkeiten** angeben. Hierzu klicken Sie, wie in Abbildung 8.5 gezeigt, in der Kreuztabellen-Subdialogbox ZELLEN ANZEIGEN bitte zusätzlich noch auf ERWARTET.

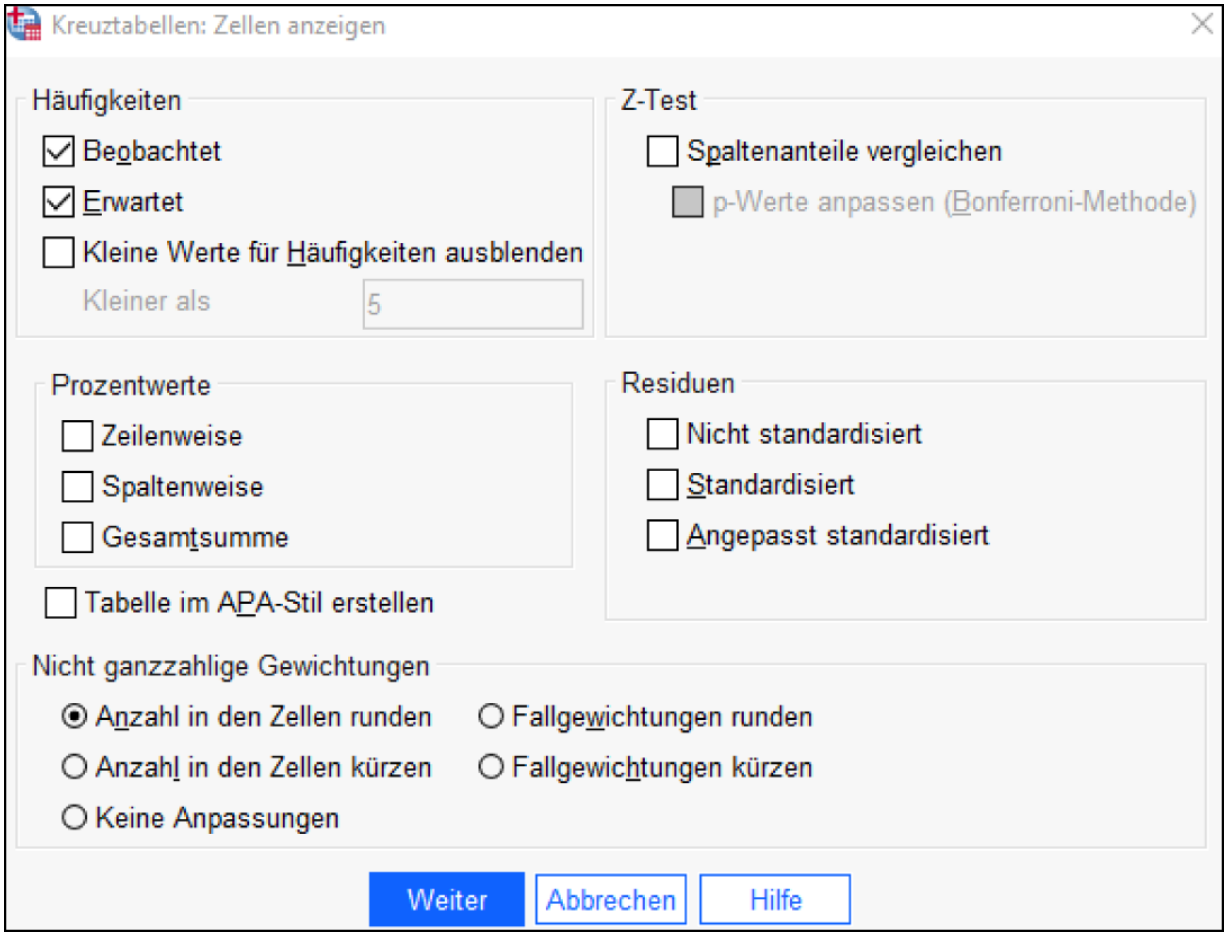



**Abbildung 8.5:** Auswahl der beobachteten und der erwarteten Fallzahlen in einer Kreuztabelle (hier in Bezug auf *KORBTREFFER3.SAV*)

Was genau versteht man unter erwarteten Häufigkeiten? Diesem Begriff liegt die Vorstellung der **statistischen Unabhängigkeit** zwischen den beiden miteinander tabellierten Variablen zugrunde. Statistische Unabhängigkeit meint, dass die beiden betrachteten Variablen sich nicht gegenseitig beeinflussen. In unserem Beispiel heißt dies etwa, dass die Korbtreffer unabhängig vom Alter der Werfer sind (beziehungsweise umgekehrt, dass das Alter der Werfer unabhängig von der Anzahl der Korbtreffer ist). Entsprechend ergeben sich die erwarteten Häufigkeiten

als Produkte der jeweiligen Randverteilungswerte, die noch durch die Gesamtfallzahl dividiert werden. Zur Berechnung der erwarteten absoluten Häufigkeit für die Kombination »0 Treffer und Altersklasse 1« beispielsweise berechnet sich diese erwartete absolute Häufigkeit als $\frac{2 \cdot 2}{10} = 0,4$. Diese Zahl mit einer Nachkommastelle ist hier also die erwartete Anzahl für diese Kombination. Die weiteren berechneten erwarteten (absoluten) Häufigkeiten sind in Abbildung 8.6 ausgewiesen.

Wenn bei Ihren Berechnungen der erwarteten absoluten Häufigkeiten – wie hier – keine ganzen Zahlen herauskommen, runden Sie bitte keineswegs auf ganze Zahlen. Würden Sie dies nämlich tun, würden Ihre Berechnungen der nachfolgend diskutierten Zusammenhangsmaße (viel) zu ungenau!

Aus dem Vergleich der beobachteten mit den (bei statistischer Unabhängigkeit) erwarteten Fallzahlen können Rückschlüsse auf den Grad der statistischen Abhängigkeit zwischen den beiden betrachteten Variablen gezogen worden. Hierzu sind Kennziffern entwickelt worden, auf die im nächsten Abschnitt dieses Kapitels noch näher eingegangen wird.

Die Syntax für den Output der Abbildung 8.6 lautet im Übrigen:

```
CROSSTABS
    /TABLES=korbtreffer BY alter_klasse
    /CELLS=COUNT EXPECTED.
```

**KORBTREFFER * Alter_Klasse Kreuztabelle**

| | | | Alter_Klasse | | | | |
|---|---|---|---|---|---|---|---|
| | | | 1,00 | 2,00 | 3,00 | 4,00 | Gesamt |
| KORBTREFFER | 0 | Anzahl | 0 | 0 | 0 | 2 | 2 |
| | | Erwartete Anzahl | ,4 | ,4 | ,6 | ,6 | 2,0 |
| | 1 | Anzahl | 0 | 1 | 0 | 1 | 2 |
| | | Erwartete Anzahl | ,4 | ,4 | ,6 | ,6 | 2,0 |
| | 2 | Anzahl | 0 | 0 | 2 | 0 | 2 |
| | | Erwartete Anzahl | ,4 | ,4 | ,6 | ,6 | 2,0 |
| | 3 | Anzahl | 0 | 0 | 1 | 0 | 1 |
| | | Erwartete Anzahl | ,2 | ,2 | ,3 | ,3 | 1,0 |
| | 4 | Anzahl | 0 | 1 | 0 | 0 | 1 |
| | | Erwartete Anzahl | ,2 | ,2 | ,3 | ,3 | 1,0 |
| | 5 | Anzahl | 2 | 0 | 0 | 0 | 2 |
| | | Erwartete Anzahl | ,4 | ,4 | ,6 | ,6 | 2,0 |
| Gesamt | | Anzahl | 2 | 2 | 3 | 3 | 10 |
| | | Erwartete Anzahl | 2,0 | 2,0 | 3,0 | 3,0 | 10,0 |

**Abbildung 8.6:** Kreuztabellenoutput für die beobachteten und die erwarteten Fallzahlen (*KORBTREFFER3.SAV*)

## *Relative Häufigkeiten*

Alternativ zu den vorstehend behandelten absoluten Häufigkeiten können Sie sich in der Kreuztabellen-Subdialogbox ZELLEN ANZEIGEN unter PROZENTWERTE von SPSS auch relative Häufigkeiten »auswerfen« lassen. Dies wird nachfolgend wieder anhand der Datendatei *KORBTREFFER3.SAV* für die beiden Variablen *korbtreffer* und *alter_klasse* demonstriert. Zunächst klicken Sie bitte ZEILENWEISE unter PROZENTWERTE an (siehe Abbildung 8.7).

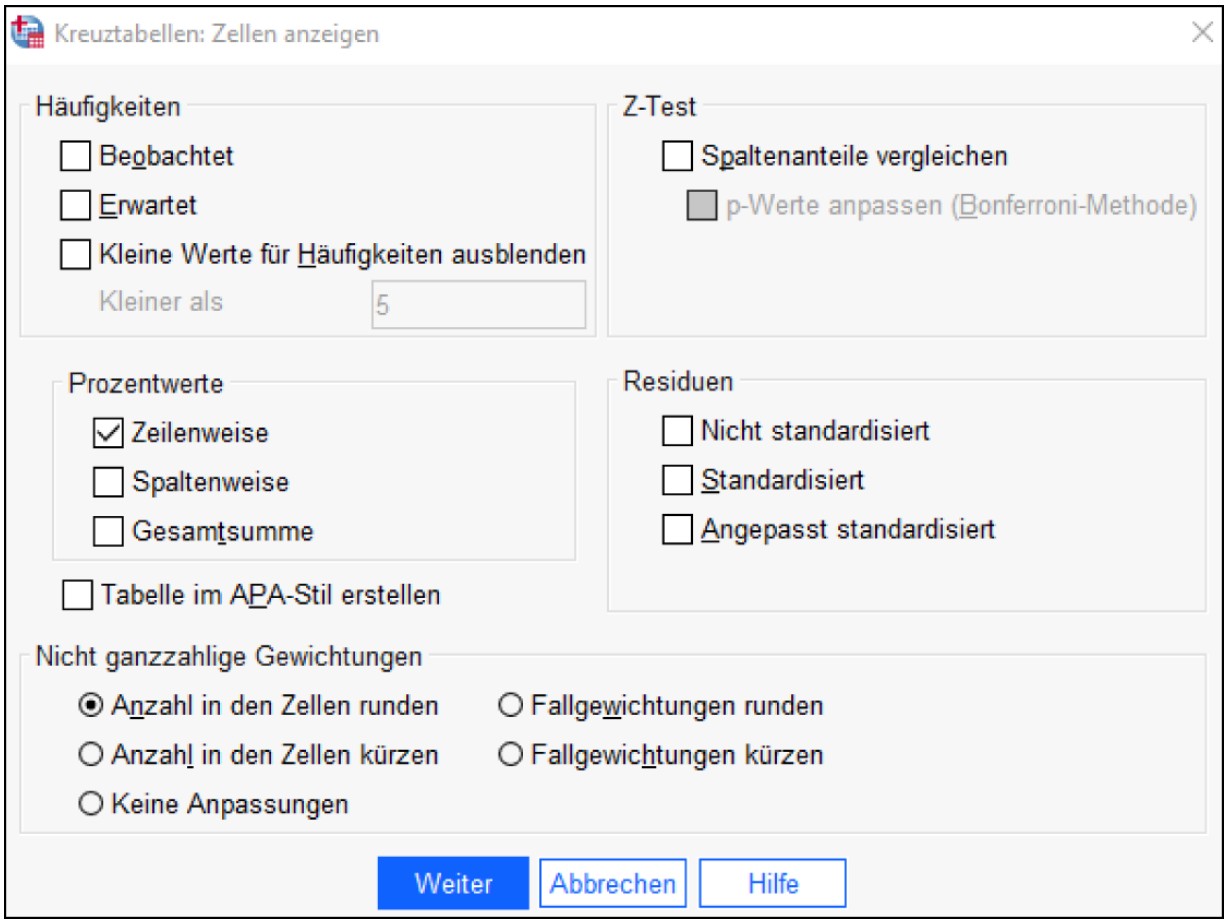



**Abbildung 8.7:** Auswahl der zeilenweise berechneten relativen Häufigkeiten in einer Kreuztabelle (hier in Bezug auf *KORBTREFFER3.SAV*)

In Syntaxform ergibt sich hier:

```
CROSSTABS
    /TABLES=korbtreffer BY alter_klasse
    /CELLS= ROW.
```

Mit der vorstehenden Entscheidung lassen Sie sich von SPSS in den Zellen der betreffenden Kreuztabelle relative Häufigkeiten in Bezug auf die Fallzahlen-Randverteilung der in den Zeilen dargestellten Variablen (hier: *korbtreffer*) berechnen. Entsprechend ergibt die zeilenweise Summe dieser relativen Häufigkeiten jeweils 100 Prozent (siehe Abbildung 8.8). Wie lässt sich nun der nachfolgende Output bezüglich der Zellen interpretieren? Lassen Sie mich hierzu bitte einfach ein Beispiel geben: 50 Prozent der Fälle mit einem Treffer entfallen auf die Altersklasse 2 und weitere 50 Prozent der Fälle mit einem Treffer auf die Altersklasse 4.

**KORBTREFFER * Alter_Klasse Kreuztabelle**

% von KORBTREFFER

| | | Alter_Klasse | | | | |
|---|---|---|---|---|---|---|
| | | 1,00 | 2,00 | 3,00 | 4,00 | Gesamt |
| KORBTREFFER | 0 | | | | 100,0% | 100,0% |
| | 1 | | 50,0% | | 50,0% | 100,0% |
| | 2 | | | 100,0% | | 100,0% |
| | 3 | | | 100,0% | | 100,0% |
| | 4 | | 100,0% | | | 100,0% |
| | 5 | 100,0% | | | | 100,0% |
| Gesamt | | 20,0% | 20,0% | 30,0% | 30,0% | 100,0% |

**Abbildung 8.8:** Kreuztabellenoutput für die zeilenweise berechneten relativen Häufigkeiten (*KORBTREFFER3.SAV*)

Analog zu Ihrer Berechnung von relativen Häufigkeiten auf Basis der Zeilenvariablen *korbtreffer* können Sie auch relative Häufigkeiten in einer Kreuztabelle auf Basis der Spaltenvariablen *alter_klasse* berechnen. Hierzu müssen Sie in der Kreuztabellen-Subdialogbox ZELLEN ANZEIGEN lediglich auf SPALTENWEISE klicken (siehe Abbildung 8.9).

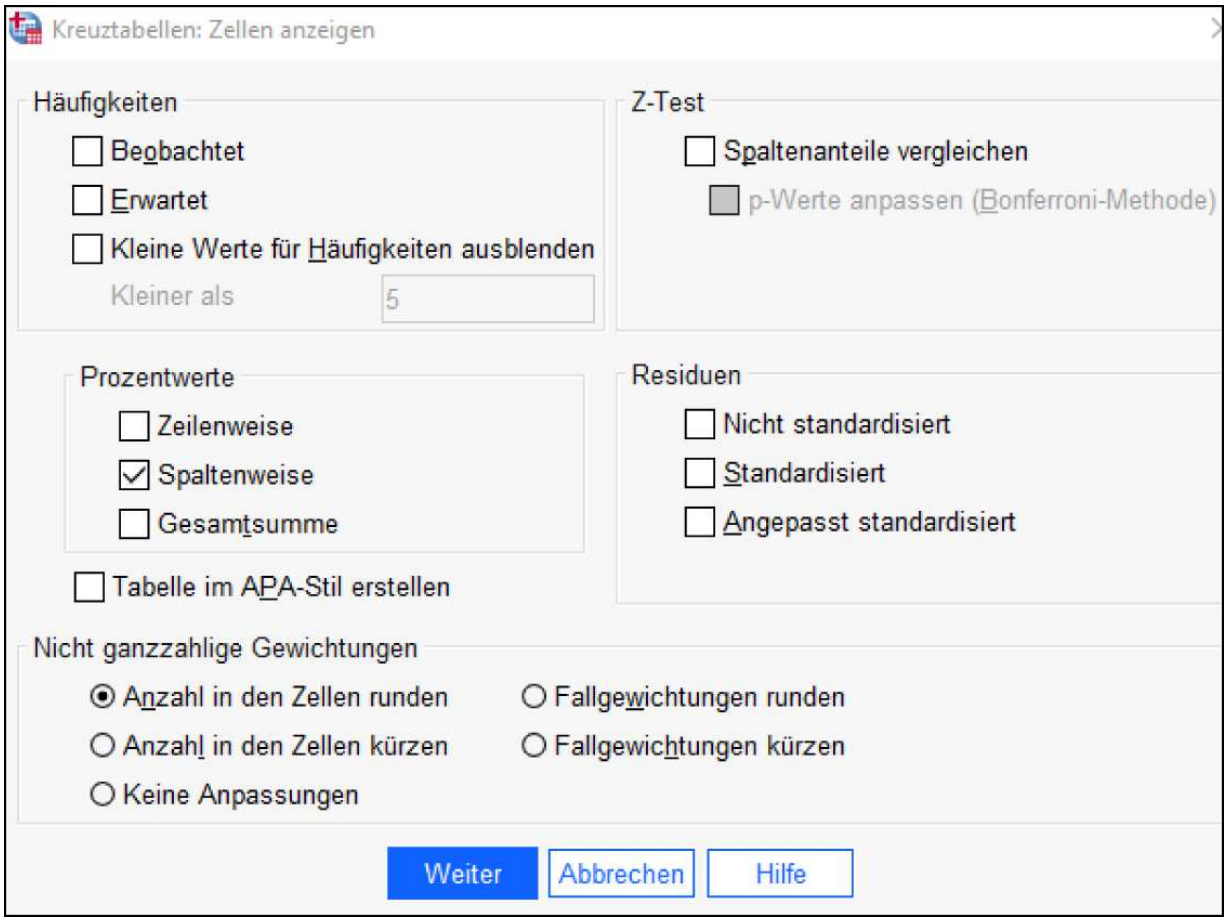



**Abbildung 8.9:** Auswahl der spaltenweise berechneten relativen Häufigkeiten in einer Kreuztabelle (hier in Bezug auf *KORBTREFFER3.SAV* )

Die zugehörige Syntax lautet:

```
CROSSTABS
    /TABLES=korbtreffer BY alter_klasse
    /CELLS= COLUMN.
```

Durch diese Festlegung werden nunmehr von SPSS für die Zellen der Kreuztabelle relative Häufigkeiten bezüglich der Fallzahlen-Randverteilung der Spaltenvariablen (hier: *alter_klasse*) berechnet, sodass hier die spaltenweise Summe dieser relativen Häufigkeiten jeweils 100 Prozent ergibt (siehe Abbildung 8.10). Bezüglich der Interpretation des nachfolgenden Zellen-Outputs gebe ich auch hier ein Beispiel: 66,7 Prozent der Personen in Altersklasse 4 haben keinen Treffer erzielt und die restlichen 33,3 Prozent dieser Altersklasse (lediglich) einen Treffer.

**KORBTREFFER * Alter_Klasse Kreuztabelle**

% von Alter_Klasse

| | | Alter_Klasse | | | | |
|---|---|---|---|---|---|---|
| | | 1,00 | 2,00 | 3,00 | 4,00 | Gesamt |
| KORBTREFFER | 0 | | | | 66,7% | 20,0% |
| | 1 | | 50,0% | | 33,3% | 20,0% |
| | 2 | | | 66,7% | | 20,0% |
| | 3 | | | 33,3% | | 10,0% |
| | 4 | | 50,0% | | | 10,0% |
| | 5 | 100,0% | | | | 20,0% |
| Gesamt | | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% |

**Abbildung 8.10:** Kreuztabellenoutput für die spaltenweise berechneten relativen Häufigkeiten (*KORBTREFFER3.SAV*)

Wenn sich Ihre relativen Häufigkeiten in einer Kreuztabelle weder auf die Fallzahlen der Zeilenvariablen noch auf die Fallzahlen der Spaltenvariablen beziehen sollen, sondern auf die Gesamtfallzahl (hier: 10 Personen), dann müssen Sie in der Kreuztabellen-Subdialogbox ZELLEN ANZEIGEN auf GESAMTSUMME klicken (siehe Abbildung 8.11).

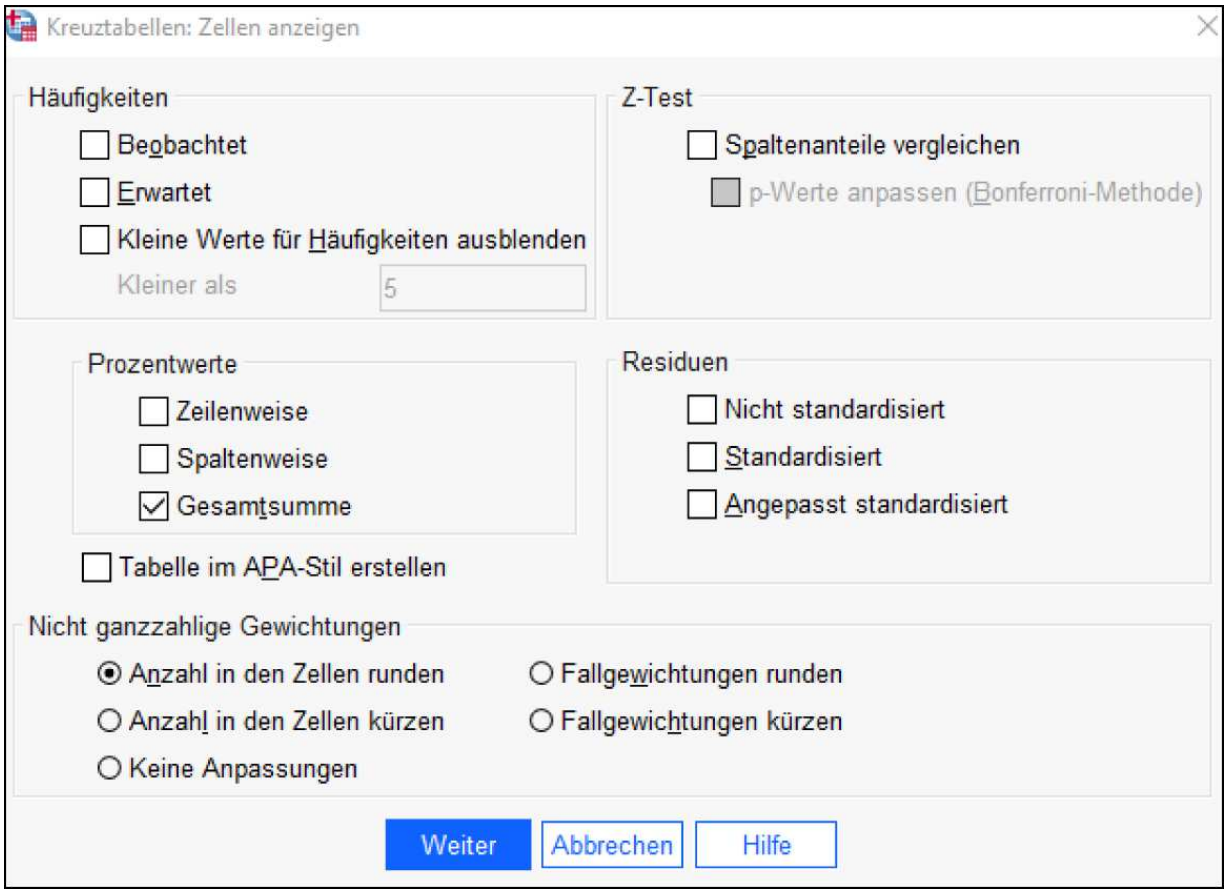



**Abbildung 8.11:** Auswahl der auf die Gesamtfallzahl bezogenen relativen Häufigkeiten in einer Kreuztabelle (hier in Bezug auf *KORBTREFFER3.SAV*)

Anstelle der Nutzung des SPSS-Menüs hätten Sie auch folgendes kleines Syntaxprogramm an dieser Stelle schreiben können:

```
CROSSTABS
    /TABLES=korbtreffer BY alter_klasse
    /CELLS= TOTAL.
```

Mit der vorstehenden Setzung erhalten Sie den in Abbildung 8.12 dargestellten Output. Die Prozentwerte in den Zellen der Kreuztabelle sind jetzt auf die Gesamtfallzahl von 10 Personen bezogen. Beispielsweise befinden sich 20 Prozent aller Fälle in der Altersklasse 4 und haben zugleich keinen Korbtreffer verbucht. Oder: 10 Prozent aller Fälle sind in der Altersklasse 3 zu finden und haben zugleich 3 Treffer erzielt.

## *Absolute und relative Häufigkeiten zusammen*

Wenn Sie sozusagen das Maximum an Informationen aus einer Kreuztabellierung herausholen möchten, klicken Sie in der Kreuztabellen-Subdialogbox ZELLEN ANZEIGEN doch einfach mal – wie in Abbildung 8.13 festgehalten – auf (fast) »alles«, das heißt auf:

- ✔ beobachtet,
- ✔ erwartet,

**KORBTREFFER * Alter_Klasse Kreuztabelle**

% der Gesamtzahl

| | | Alter_Klasse | | | | |
|---|---|---|---|---|---|---|
| | | 1,00 | 2,00 | 3,00 | 4,00 | Gesamt |
| KORBTREFFER | 0 | | | | 20,0% | 20,0% |
| | 1 | | 10,0% | | 10,0% | 20,0% |
| | 2 | | | 20,0% | | 20,0% |
| | 3 | | | 10,0% | | 10,0% |
| | 4 | | 10,0% | | | 10,0% |
| | 5 | 20,0% | | | | 20,0% |
| Gesamt | | 20,0% | 20,0% | 30,0% | 30,0% | 100,0% |

**Abbildung 8.12:** Kreuztabellenoutput für auf die Gesamtfallzahl bezogene relative Häufigkeiten (*KORBTREFFER3.SAV*)

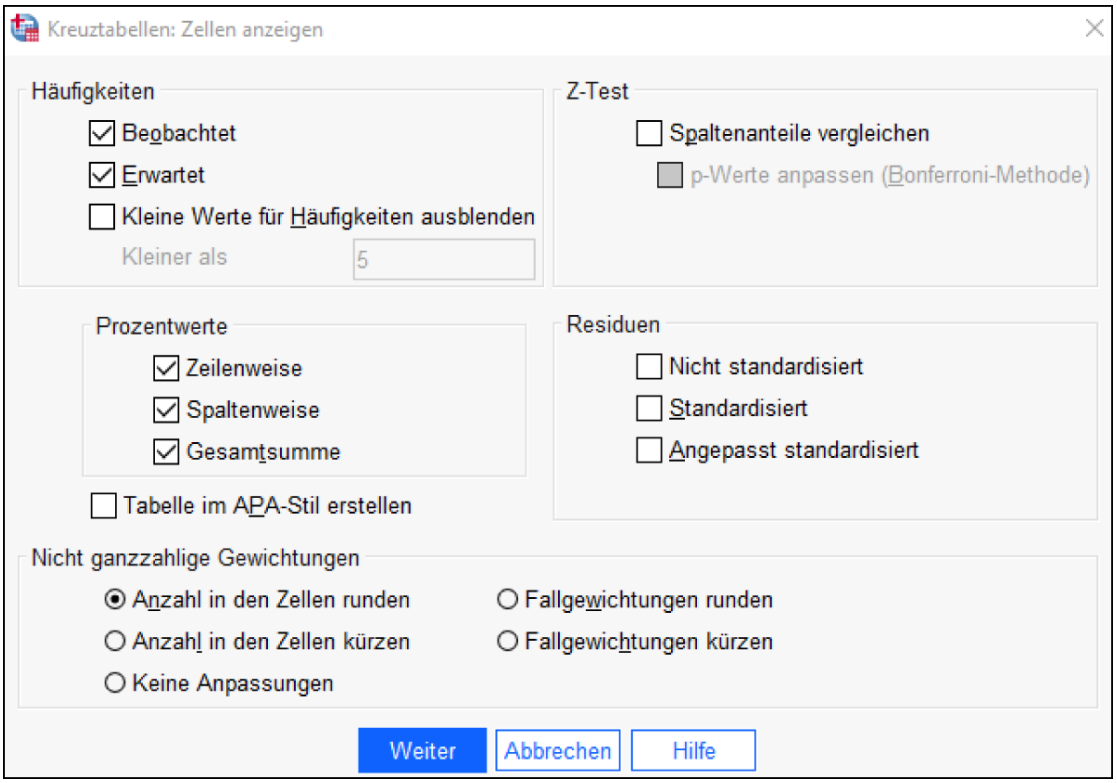



**Abbildung 8.13:** Auswahl von absoluten und relativen Häufigkeiten in einer Kreuztabelle (hier in Bezug auf *KORBTREFFER3.SAV*)

- ✔ zeilenweise,
- ✔ spaltenweise und
- ✔ Gesamtsumme.

Die korrespondierende Syntax lautet dann:

```
CROSSTABS
      /TABLES=korbtreffer BY alter_klasse
      /CELLS=COUNT EXPECTED ROW COLUMN TOTAL.
```

Als Zellenwerte erhalten Sie in diesem Fall in einer Kreuztabelle sämtliche in den beiden vorhergehenden Unterabschnitten dieses Kapitels einzeln dargestellten absoluten und relativen Häufigkeiten (siehe Abbildung 8.14). Ich finde das Ganze nicht sehr übersichtlich und rate Ihnen, sich auf weniger Zellenwerte zu konzentrieren. Wenn Sie aber mit einem solchen »Informations-Overflow« klarkommen, soll es mir natürlich auch recht sein.

**KORBTREFFER * Alter_Klasse Kreuztabelle**

| | | | Alter_Klasse | | | | |
|---|---|---|---|---|---|---|---|
| | | | 1,00 | 2,00 | 3,00 | 4,00 | Gesamt |
| KORBTREFFER | 0 | Anzahl | 0 | 0 | 0 | 2 | 2 |
| | | Erwartete Anzahl | ,4 | ,4 | ,6 | ,6 | 2,0 |
| | | % von KORBTREFFER | 0,0% | 0,0% | 0,0% | 100,0% | 100,0% |
| | | % von Alter_Klasse | 0,0% | 0,0% | 0,0% | 66,7% | 20,0% |
| | | % der Gesamtzahl | 0,0% | 0,0% | 0,0% | 20,0% | 20,0% |
| | 1 | Anzahl | 0 | 1 | 0 | 1 | 2 |
| | | Erwartete Anzahl | ,4 | ,4 | ,6 | ,6 | 2,0 |
| | | % von KORBTREFFER | 0,0% | 50,0% | 0,0% | 50,0% | 100,0% |
| | | % von Alter_Klasse | 0,0% | 50,0% | 0,0% | 33,3% | 20,0% |
| | | % der Gesamtzahl | 0,0% | 10,0% | 0,0% | 10,0% | 20,0% |
| | 2 | Anzahl | 0 | 0 | 2 | 0 | 2 |
| | | Erwartete Anzahl | ,4 | ,4 | ,6 | ,6 | 2,0 |
| | | % von KORBTREFFER | 0,0% | 0,0% | 100,0% | 0,0% | 100,0% |
| | | % von Alter_Klasse | 0,0% | 0,0% | 66,7% | 0,0% | 20,0% |
| | | % der Gesamtzahl | 0,0% | 0,0% | 20,0% | 0,0% | 20,0% |
| | 3 | Anzahl | 0 | 0 | 1 | 0 | 1 |
| | | Erwartete Anzahl | ,2 | ,2 | ,3 | ,3 | 1,0 |
| | | % von KORBTREFFER | 0,0% | 0,0% | 100,0% | 0,0% | 100,0% |
| | | % von Alter_Klasse | 0,0% | 0,0% | 33,3% | 0,0% | 10,0% |
| | | % der Gesamtzahl | 0,0% | 0,0% | 10,0% | 0,0% | 10,0% |
| | 4 | Anzahl | 0 | 1 | 0 | 0 | 1 |
| | | Erwartete Anzahl | ,2 | ,2 | ,3 | ,3 | 1,0 |
| | | % von KORBTREFFER | 0,0% | 100,0% | 0,0% | 0,0% | 100,0% |
| | | % von Alter_Klasse | 0,0% | 50,0% | 0,0% | 0,0% | 10,0% |
| | | % der Gesamtzahl | 0,0% | 10,0% | 0,0% | 0,0% | 10,0% |
| | 5 | Anzahl | 2 | 0 | 0 | 0 | 2 |
| | | Erwartete Anzahl | ,4 | ,4 | ,6 | ,6 | 2,0 |
| | | % von KORBTREFFER | 100,0% | 0,0% | 0,0% | 0,0% | 100,0% |
| | | % von Alter_Klasse | 100,0% | 0,0% | 0,0% | 0,0% | 20,0% |
| | | % der Gesamtzahl | 20,0% | 0,0% | 0,0% | 0,0% | 20,0% |
| Gesamt | | Anzahl | 2 | 2 | 3 | 3 | 10 |
| | | Erwartete Anzahl | 2,0 | 2,0 | 3,0 | 3,0 | 10,0 |
| | | % von KORBTREFFER | 20,0% | 20,0% | 30,0% | 30,0% | 100,0% |
| | | % von Alter_Klasse | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% |
| | | % der Gesamtzahl | 20,0% | 20,0% | 30,0% | 30,0% | 100,0% |

**Abbildung 8.14:** Kreuztabellenoutput für absolute und relative Häufigkeiten (*KORBTREFFER3.SAV*)

## *Erweiterte Kreuztabellen*

SPSS bietet Ihnen die Möglichkeit, Kreuztabellen noch differenzierter als oben beschrieben darzustellen. So können Sie die Zeilenvariable noch nach weiteren Variablen in verschiedene Subgruppen unterscheiden.

Um dies zu veranschaulichen, öffnen Sie bitte zunächst den in Abbildung 8.15 dargestellten Datensatz *GESCHL_FARBE2.SAV*, der neben der Personennummer *persnr* vier Stringvariablen enthält: *geschlecht* (Geschlecht einer Person mit den Ausprägungen »männlich« und »weiblich«), *augenfarbe* (Augenfarbe einer Person mit den Ausprägungen »grün«, »braun«, »blau«, »grau« und »andere« – diese letztgenannte Ausprägung sozusagen als »Rest-Kategorie«, also etwa Menschen mit zwei unterschiedlichen Augenfarben), *qualifikation* (berufliche Qualifikation einer Person mit den Ausprägungen »gering«, »mittel« und »hoch«) sowie *nationalität* (Nationalität einer Person mit den Ausprägungen »deutsch« und »nichtdeutsch«). (Die nachfolgende Häufigkeitsverteilung ist übrigens mit der empirischen, weltweit festgestellten Augenfarbenverteilung insofern nicht ganz kompatibel, als in der Realität nur etwa zwei Prozent der Weltbevölkerung die Augenfarbe »grün« haben. Bitte sehen Sie mir diese kleine Ungenauigkeit nach!)

| | persnr | geschlecht | augenfarbe | qualifikation | nationalität | Var |
|---|---|---|---|---|---|---|
| 1 | 1 | männlich | grün | hoch | deutsch | |
| 2 | 2 | weiblich | blau | mittel | nichtdeutsch | |
| 3 | 3 | männlich | braun | gering | nichtdeutsch | |
| 4 | 4 | männlich | andere | mittel | deutsch | |
| 5 | 5 | männlich | grau | hoch | deutsch | |
| 6 | 6 | weiblich | blau | hoch | deutsch | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 7 | 7 | männlich | grau | hoch | deutsch | |
| 8 | 8 | weiblich | grün | hoch | deutsch | |
| 9 | 9 | männlich | grün | mittel | deutsch | |
| 10 | 10 | weiblich | braun | gering | deutsch | |
| 11 | 11 | weiblich | blau | gering | deutsch | |
| 12 | 12 | männlich | braun | hoch | deutsch | |
| 13 | 13 | weiblich | braun | gering | deutsch | |
| 14 | 14 | weiblich | andere | gering | deutsch | |
| 15 | 15 | weiblich | braun | gering | deutsch | |
| 16 | 16 | männlich | grün | hoch | deutsch | |
| 17 | 17 | weiblich | braun | gering | deutsch | |
| 18 | 18 | weiblich | grün | gering | deutsch | |
| 19 | 19 | weiblich | braun | gering | deutsch | |
| 20 | 20 | weiblich | blau | gering | deutsch | |
| 21 | 21 | männlich | braun | hoch | nichtdeutsch | |
| 22 | 22 | männlich | braun | gering | nichtdeutsch | |
| 23 | 23 | weiblich | braun | mittel | nichtdeutsch | |
| 24 | 24 | weiblich | grün | mittel | nichtdeutsch | |
| 25 | 25 | weiblich | blau | hoch | nichtdeutsch | |
| 26 | | | | | | |

**Abbildung 8.15:** Datendatei *GESCHL_FARBE2.SAV* mit den Variablen *geschlecht*, *augenfarbe*, *qualifikation* und *nationalität*

Ohne näher auf die inhaltliche Sinnhaftigkeit des diskutierten Zusammenhangs einzugehen, soll im Folgenden grundsätzlich die Variable *geschlecht* als Zeilenvariable und die Variable *augenfarbe* als Spaltenvariable in der betreffenden Kreuztabelle genutzt werden. Allerdings sollen die beiden »Zeilengruppen« der Männer und Frauen noch nach den beiden anderen Stringvariablen *qualifikation* und *nationalität* differenziert werden. Das heißt: Es sollen Gruppen wie »mittelqualifizierte deutsche Männer« oder »hochqualifizierte nichtdeutsche Frauen« gebildet werden, und diese Gruppen werden dann mit der Variablen *augenfarbe* kreuztabelliert.

In einem ersten Schritt wird deshalb in der KREUZTABELLEN-Dialogbox unter SCHICHT 1 VON 1 *qualifikation* eingetragen (siehe Abbildung 8.16).

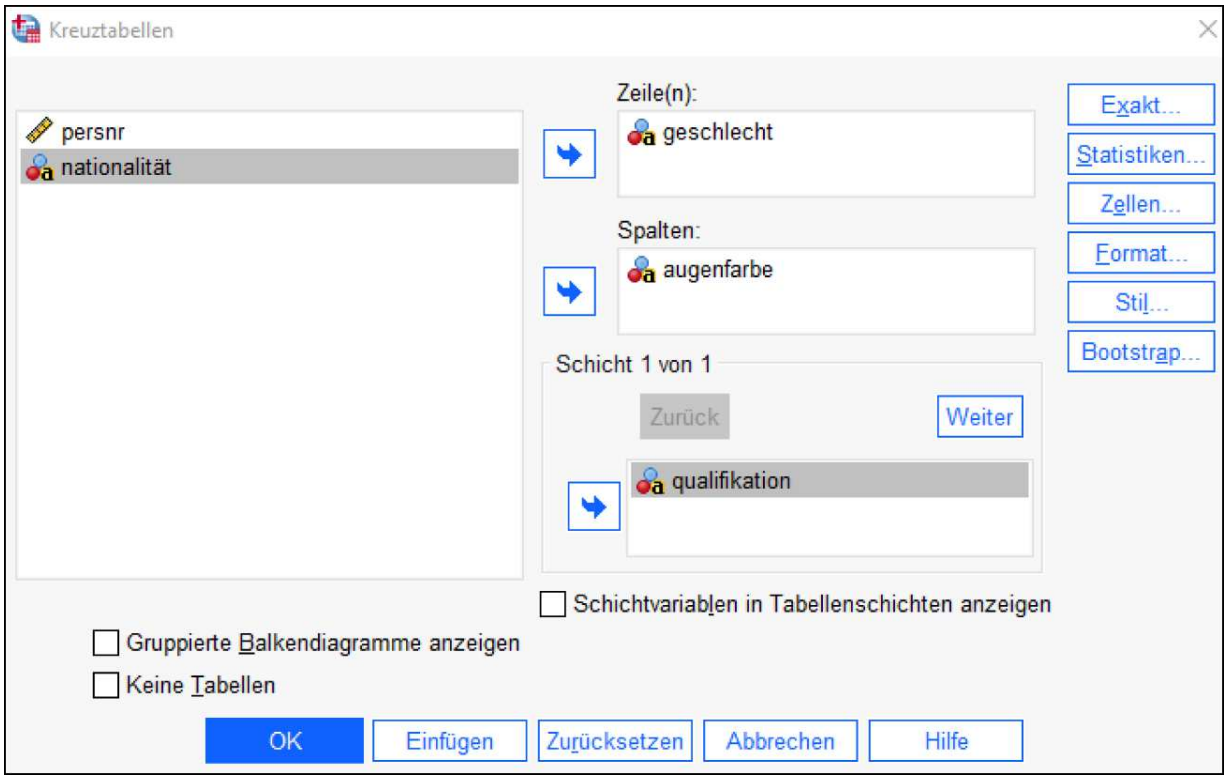



**Abbildung 8.16:** Anweisung für eine erweiterte Kreuztabelle: erste Unterteilung der Zeilenvariablen *geschlecht* (*GESCHL_FARBE2.SAV*)

Um jetzt auch noch die Nationalität ins Spiel zu bringen, müssen Sie in Abbildung 8.16 auf WEITER klicken, und Sie erhalten postwendend die Rubrik SCHICHT 2 VON 2, in die Sie bitte *nationalität* eintragen (siehe Abbildung 8.17). SPSS bietet Ihnen im Übrigen die Möglichkeit noch weiterer Zeilendifferenzierungen, doch, wie Sie aus der Outputdarstellung in Abbildung 8.18 sicherlich erkennen, ist bereits in unserem kleinen Beispiel die Übersichtlichkeit der entsprechenden (erweiterten) Kreuztabelle doch stark eingeschränkt – oder was meinen Sie?

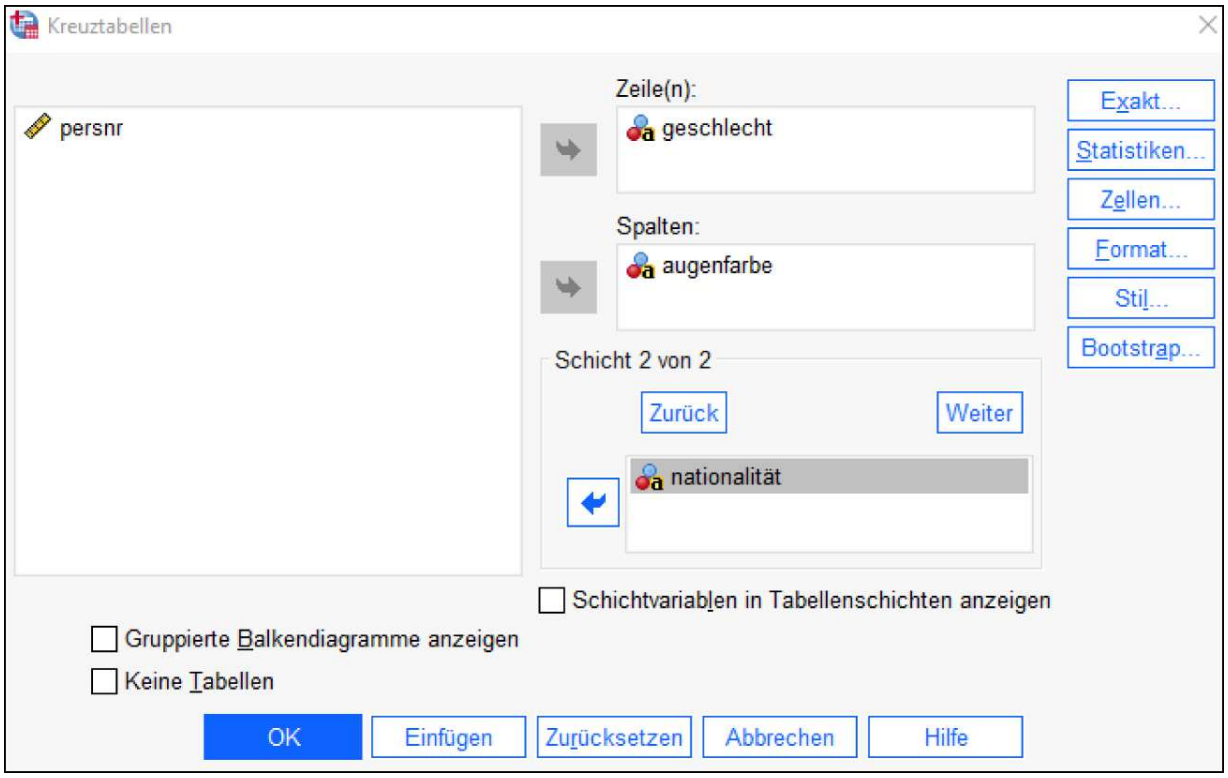



**Abbildung 8.17:** Anweisung für eine erweiterte Kreuztabelle: zweite Unterteilung der Zeilenvariablen *geschlecht* (*GESCHL_FARBE2.SAV*)

**geschlecht * augenfarbe * qualifikation * nationalität Kreuztabelle**

Anzahl

| nationalität | qualifikation | | | augenfarbe andere | blau | braun | grau | grün | Gesamt |
|---|---|---|---|---|---|---|---|---|---|
| deutsch | gering | geschlecht | weiblich | 1 | 2 | 5 | | 1 | 9 |
| | | Gesamt | | 1 | 2 | 5 | | 1 | 9 |
| | hoch | geschlecht | männlich | | 0 | 1 | 2 | 2 | 5 |
| | | | weiblich | | 1 | 0 | 0 | 1 | 2 |
| | | Gesamt | | | 1 | 1 | 2 | 3 | 7 |
| | mittel | geschlecht | männlich | 1 | | | | 1 | 2 |
| | | Gesamt | | 1 | | | | 1 | 2 |
| | Gesamt | geschlecht | männlich | 1 | 0 | 1 | 2 | 3 | 7 |
| | | | weiblich | 1 | 3 | 5 | 0 | 2 | 11 |
| | | Gesamt | | 2 | 3 | 6 | 2 | 5 | 18 |
| nichtdeutsch | gering | geschlecht | männlich | | | 2 | | | 2 |
| | | Gesamt | | | | 2 | | | 2 |
| | hoch | geschlecht | männlich | | 0 | 1 | | | 1 |
| | | | weiblich | | 1 | 0 | | | 1 |
| | | Gesamt | | | 1 | 1 | | | 2 |
| | mittel | geschlecht | weiblich | | 1 | 1 | | 1 | 3 |
| | | Gesamt | | | 1 | 1 | | 1 | 3 |
| | Gesamt | geschlecht | männlich | | 0 | 3 | | 0 | 3 |
| | | | weiblich | | 2 | 1 | | 1 | 4 |
| | | Gesamt | | | 2 | 4 | | 1 | 7 |
| Gesamt | gering | geschlecht | männlich | 0 | 0 | 2 | | 0 | 2 |
| | | | weiblich | 1 | 2 | 5 | | 1 | 9 |
| | | Gesamt | | 1 | 2 | 7 | | 1 | 11 |
| | hoch | geschlecht | männlich | | 0 | 2 | 2 | 2 | 6 |
| | | | weiblich | | 2 | 0 | 0 | 1 | 3 |
| | | Gesamt | | | 2 | 2 | 2 | 3 | 9 |
| | mittel | geschlecht | männlich | 1 | 0 | 0 | | 1 | 2 |
| | | | weiblich | 0 | 1 | 1 | | 1 | 3 |
| | | Gesamt | | 1 | 1 | 1 | | 2 | 5 |
| | Gesamt | geschlecht | männlich | 1 | 0 | 4 | 2 | 3 | 10 |
| | | | weiblich | 1 | 5 | 6 | 0 | 3 | 15 |
| | | Gesamt | | 2 | 5 | 10 | 2 | 6 | 25 |

**Abbildung 8.18:** Output für eine erweiterte Kreuztabelle mit zwei Unterteilungen der Zeilenvariablen *geschlecht* (*GESCHL_FARBE2.SAV*)

Wenn Sie sich nicht so recht mit dem SPSS-Menü anfreunden können und lieber auf Ihre Syntaxfähigkeiten zurückgreifen, können Sie auch folgenden SPSS-Befehl im Syntaxeditor schreiben (wobei Sie bitte der Übersichtlichkeit wegen von SPSS lediglich die beobachteten absoluten Häufigkeiten anfordern):

```
CROSSTABS
    /TABLES=geschlecht BY augenfarbe BY qualifikation BY nationalität
    /CELLS=COUNT.
```

Aus der Outputdarstellung in Abbildung 8.18 wird etwa ersichtlich, dass 5 deutsche Frauen mit einer geringen Qualifikation eine braune Augenfarbe haben. Oder als anderes Beispiel: Es gibt im Datensatz 2 deutsche Männer mit hoher Qualifikation, deren Augenfarbe grau ist. Fehlen bestimmte Gruppen, weist SPSS diese nicht aus. Wie aus Abbildung 8.18 hervorgeht, sind entsprechend im Datensatz offenkundig keine Fälle für folgende Gruppen vorhanden: deutsche Männer mit geringer Qualifikation, deutsche Frauen mit mittlerer Qualifikation, nichtdeutsche Frauen mit geringer Qualifikation und nichtdeutsche Männer mit mittlerer Qualifikation.

Wie erwähnt, scheren Sie sich bitte nicht um die inhaltliche Aussagekraft dieses Beispiels – es geht hier eigentlich nur um das Prinzip!

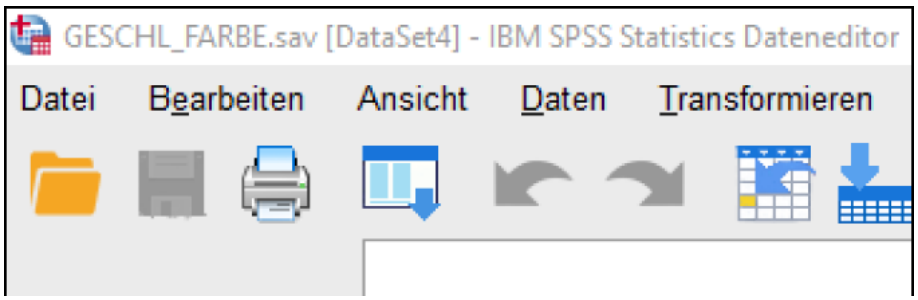
# *Assoziationen*

Statistische Zusammenhänge zwischen zwei Variablen können Sie bereits ab Nominalskalenniveau untersuchen. Man spricht in diesem Zusammenhang von **Assoziationen** oder mit einer anderen Bezeichnung: von Kontingenzen. Bei dieser schwächsten Form der statistischen Zusammenhangsanalyse geht es eigentlich nur darum zu entscheiden, ob die beiden Variablen »irgendwie« zusammenhängen. Die Art beziehungsweise Richtung und die Stärke des Zusammenhangs, die von den im nächsten Abschnitt dieses Kapitels behandelten Korrelationsmaßen berechnet werden, spielen im Rahmen der Assoziationsanalyse keine Rolle. Im Grunde genommen werden im Rahmen der Assoziationsanalyse »lediglich« mehr oder weniger auffällige Merkmalshäufungen bei bestimmten Zusammenstellungen der Variablenausprägungen untersucht.

Wenn Sie sich statistische Zusammenhänge zwischen zwei Variablen anzeigen lassen, sollten Sie sich unbedingt bewusst sein, dass hiermit keine inhaltlichen Aussagen verbunden sind. Die statistische Zusammenhangsanalyse sagt lediglich – Nomen est omen! – etwas aus über aufgrund der statistischen Datenlage gegebene Zusammenhänge, nicht aber über inhaltliche Beziehungen zwischen den Variablen! Wenn also zum Beispiel der Preis für Katzenfutter in Deutschland steigt und gleichzeitig die Nachfrage nach Uhren in Neuseeland zurückgeht, kann man zwar einen (entgegengerichteten) statistischen Zusammenhang zwischen beiden Entwicklungen berechnen, aber nur schwerlich einen inhaltlichen Bezug herstellen.

Um Ihnen zu zeigen, dass Assoziationsanalysen bereits auf der Ebene des Nominalskalenniveaus möglich sind, habe ich ein kleines Beispiel kreiert (siehe die Datendatei GESCHL_FARBE.SAV ). In diesem Beispiel werden die beiden Variablen *geschlecht* mit den Ausprägungen »männlich« und »weiblich« sowie *augenfarbe* mit den Ausprägungen »grün«, »braun«, »blau«, »grau« und »andere« (diese letztgenannte Ausprägung sozusagen als »Rest-Kategorie«) einander gegenübergestellt. Die zugehörige Datendatei ist in Abbildung 8.19 abgebildet.

Bitte geben Sie jetzt für die Kreuztabellierung von *geschlecht* (Zeilenvariable) mit *augenfarbe* (Spaltenvariable) im SPSS-Menü bei ANALYSIEREN|DESKRIPTIVE STATISTIKEN|KREUZTABELLEN|ZELLEN den Ausweis der beobachteten und der erwarteten Fallzahlen an. Alternativ können Sie folgende Syntax in den Syntaxeditor eintragen:



| | persnr | geschlecht | augenfarbe | Var | Var |
|---|---|---|---|---|---|
| 1 | 1 | männlich | grün | | |
| 2 | 2 | weiblich | blau | | |
| 3 | 3 | männlich | braun | | |
| 4 | 4 | männlich | andere | | |
| 5 | 5 | männlich | grau | | |
| 6 | 6 | weiblich | blau | | |
| 7 | 7 | männlich | grau | | |
| 8 | 8 | weiblich | grün | | |
| 9 | 9 | männlich | grün | | |
| 10 | 10 | weiblich | braun | | |
| 11 | 11 | weiblich | blau | | |
| 12 | 12 | männlich | braun | | |
| 13 | 13 | weiblich | braun | | |
| 14 | 14 | weiblich | andere | | |
| 15 | 15 | weiblich | braun | | |
| 16 | 16 | männlich | grün | | |
| 17 | 17 | weiblich | braun | | |
| 18 | 18 | weiblich | grün | | |
| 19 | 19 | weiblich | braun | | |
| 20 | 20 | weiblich | blau | | |
| 21 | 21 | männlich | braun | | |
| 22 | 22 | männlich | braun | | |
| 23 | 23 | weiblich | braun | | |
| 24 | 24 | weiblich | grün | | |
| 25 | 25 | weiblich | blau | | |
| 26 | | | | | |
| 27 | | | | | |

**Abbildung 8.19:** Datendatei *GESCHL_FARBE.SAV* mit den Variablen *geschlecht* und *augenfarbe*

```
CROSSTABS
    /TABLES=geschlecht BY augenfarbe
    /CELLS=COUNT EXPECTED.
```

Auf diese Weise erhalten Sie einen ersten, visuellen Eindruck, ob die betrachteten Variablen statistisch etwas miteinander zu tun haben oder nicht. Haben Sie etwas miteinander zu tun, dann sind die Differenzen zwischen den beobachteten und den erwarteten Fallzahlen nennenswert. Andernfalls sind die betreffenden Differenzen gering, und es kann eventuell sogar auf statistische Unabhängigkeit geschlossen werden.

In Abbildung 8.20 sind diese Differenzen zwar vorhanden, aber auf den ersten Blick auch nicht spektakulär groß. Beispielsweise haben im vorliegenden Beispielfall mehr Männer als erwartet eine grüne Augenfarbe, während gemäß Abbildung 8.20 für die Gruppe der Frauen genau Umgekehrtes gilt. Ob allerdings eine mehr oder weniger starke statistische Beziehung zwischen dem Geschlecht und der Augenfarbe vorliegt, kann zumindest ich nicht auf Anhieb erkennen. Dazu bedarf es anderer statistischer Indikatoren.

**geschlecht * augenfarbe Kreuztabelle**

| | | | augenfarbe | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | andere | blau | braun | grau | grün | Gesamt |
| geschlecht | männlich | Anzahl | 1 | 0 | 4 | 2 | 3 | 10 |
| | | Erwartete Anzahl | ,8 | 2,0 | 4,0 | ,8 | 2,4 | 10,0 |
| | weiblich | Anzahl | 1 | 5 | 6 | 0 | 3 | 15 |
| | | Erwartete Anzahl | 1,2 | 3,0 | 6,0 | 1,2 | 3,6 | 15,0 |
| Gesamt | | Anzahl | 2 | 5 | 10 | 2 | 6 | 25 |
| | | Erwartete Anzahl | 2,0 | 5,0 | 10,0 | 2,0 | 6,0 | 25,0 |

**Abbildung 8.20:** Kreuztabellen-Output mit den beobachteten und den erwarteten Fallzahlen (*GESCHL_FARBE.SAV*)

Derartige Indikatoren – man spricht dann von **Assoziationsmaßen** – gründen überwiegend auf den erwähnten Differenzen zwischen beobachteten und erwarteten absoluten Häufigkeiten, die quadriert und anschließend noch durch die jeweils erwarteten absoluten Häufigkeiten dividiert werden. Schließlich werden diese Werte noch addiert. Es ergibt sich der sogenannte **Chi-Quadrat-Kontingenzkoeffizient** $\chi^2$. Wenn dieser (näherungsweise) gleich null ist, kann auf statistische Unabhängigkeit zwischen den betrachteten Variablen geschlossen werden, andernfalls zumindest auf eine gewisse statistische Beziehung zwischen den betreffenden Variablen.

Im Beispiel der Abbildung 8.20 müssten Sie also rechnen:

$$\chi^2 = \frac{(1-0{,}8)^2}{0{,}8} + \frac{(0-2)^2}{2} + \frac{(4-4)^2}{4} + \frac{(2-0{,}8)^2}{0{,}8} + \frac{(3-2{,}4)^2}{2{,}4} + \frac{(1-1{,}2)^2}{1{,}2}$$
$$+ \frac{(5-3)^2}{3} + \frac{(6-6)^2}{6} + \frac{(0-1{,}2)^2}{1{,}2} + \frac{(3-3{,}6)^2}{3{,}6} = 6{,}667.$$

In unserem Beispiel ist der Chi-Quadrat-Kontingenzkoeffizient recht deutlich von null verschieden, sodass zumindest auf eine gewisse statistische Beziehung zwischen Geschlecht und Augenfarbe geschlossen werden kann.

Unschön am Chi-Quadrat-Kontingenzkoeffizienten ist, dass er nicht normiert ist. Viel schöner ist es, einen Indikator zu haben, der einen Wertebereich von 0 (keine statistische Beziehung) bis 1 (perfekte statistische Beziehung) aufweist. Im obigen Beispiel sollte sich ein Wert von 1 zum Beispiel dann ergeben, wenn alle Frauen beispielsweise grüne Augen und sämtliche Männer etwa blaue Augen hätten.

Um eine solche Normierung auf die Obergrenze von 1 hinzubekommen, sind verschiedene Abwandlungen des Chi-Quadrat-Kontingenzkoeffizienten entwickelt worden, wobei N im Folgenden für die gesamte Fallzahl (25 Personen in GESCHL_FARBE.SAV ), i für die Zeilenanzahl (2 Zeilen: »männlich« und »weiblich«) und j für die Spaltenzahl der Kreuztabelle (5 Spalten: »blau«, »braun«, »grün«, »grau«, »andere«) stehen:

- ✔ der **Phi-Kontingenzkoeffizient:** $\Phi = \sqrt[2]{\frac{\chi^2}{N}}$ (im obigen Beispiel also: $\sqrt[2]{\frac{6{,}667}{25}} \approx 0{,}516$),

- **Cramers** V: $V = \sqrt[2]{\frac{\chi^2}{N \cdot \text{Minimum}(i-1; j-1)}}$ (im obigen Beispiel also: $V = \sqrt[2]{\frac{6{,}667}{25 \cdot \text{Minimum}(1;4)}} = \sqrt[2]{\frac{6{,}667}{25 \cdot 1}} \approx 0{,}516$),
- der **C-Kontingenzkoeffizient**: $CC = \sqrt[2]{\frac{\chi^2}{\chi^2 + N}}$ mit einem Minimum von 0 und einem Maximum bei $\sqrt[2]{\frac{\text{Minimum}(i,j)-1}{\text{Minimum}(i,j)}}$ (im obigen Beispiel ist: $CC = \sqrt[2]{\frac{6{,}667}{6{,}667+25}} = \sqrt[2]{\frac{6{,}667}{31{,}667}} \approx 0{,}459$) und
- der – unter anderem um eine Normierung auf den Wertebereich von [0, 1] vorzunehmen – **korrigierte C-Kontingenzkoeffizient:** $CC_k = \frac{CC}{\sqrt[2]{\frac{\text{Minimum}(i,j)-1}{\text{Minimum}(i,j)}}}$ (im obigen Beispiel also: $CC_k = \frac{0{,}459}{\sqrt[2]{\frac{1}{2}}} \approx 0{,}649$).

Für nominalskalierte Merkmale wie hier (*geschlecht* und *augenfarbe*) können Sie sich in SPSS bis auf den korrigierten Kontingenzkoeffizienten alle der vorstehenden Assoziationsmaße ausgeben lassen. Dass der korrigierte C-Kontingenzkoeffizient im Übrigen nicht ausgegeben wird, ist aber nicht schlimm: Mit dem Taschenrechner ist der C-Kontingenzkoeffizient anhand der obigen Formel schnell in seine korrigierte Version umgerechnet.

Die vorstehenden Assoziationsmaße sind in SPSS zum Teil etwas anders bezeichnet:

- Den Chi-Quadrat-Kontingenzkoeffizienten finden Sie in SPSS als Pearson-Chi-Quadrat.

- ✔ Der C-Kontingenzkoeffizient heißt in SPSS einfach Kontingenzkoeffizient.

Um sich die einzelnen Assoziationsmaße nunmehr als SPSS-Output anzeigen zu lassen, müssen Sie die entsprechenden Maße in der KREUZTABELLEN-Subdialogbox STATISTIK einfach anklicken (siehe Abbildung 8.21). Damit die Kreuztabelle zwischen *geschlecht* und *augenfarbe* nicht noch einmal angezeigt wird, setzen Sie im Übrigen einfach bei der KREUZTABELLEN-Dialogbox ein Häkchen bei KEINE TABELLEN.

Alternativ hätten Sie folgendes kleines Syntaxprogramm schreiben können:

```
CROSSTABS
    /TABLES=geschlecht BY augenfarbe
    /FORMAT=NOTABLES
    /STATISTICS=CHISQ CC PHI.
```

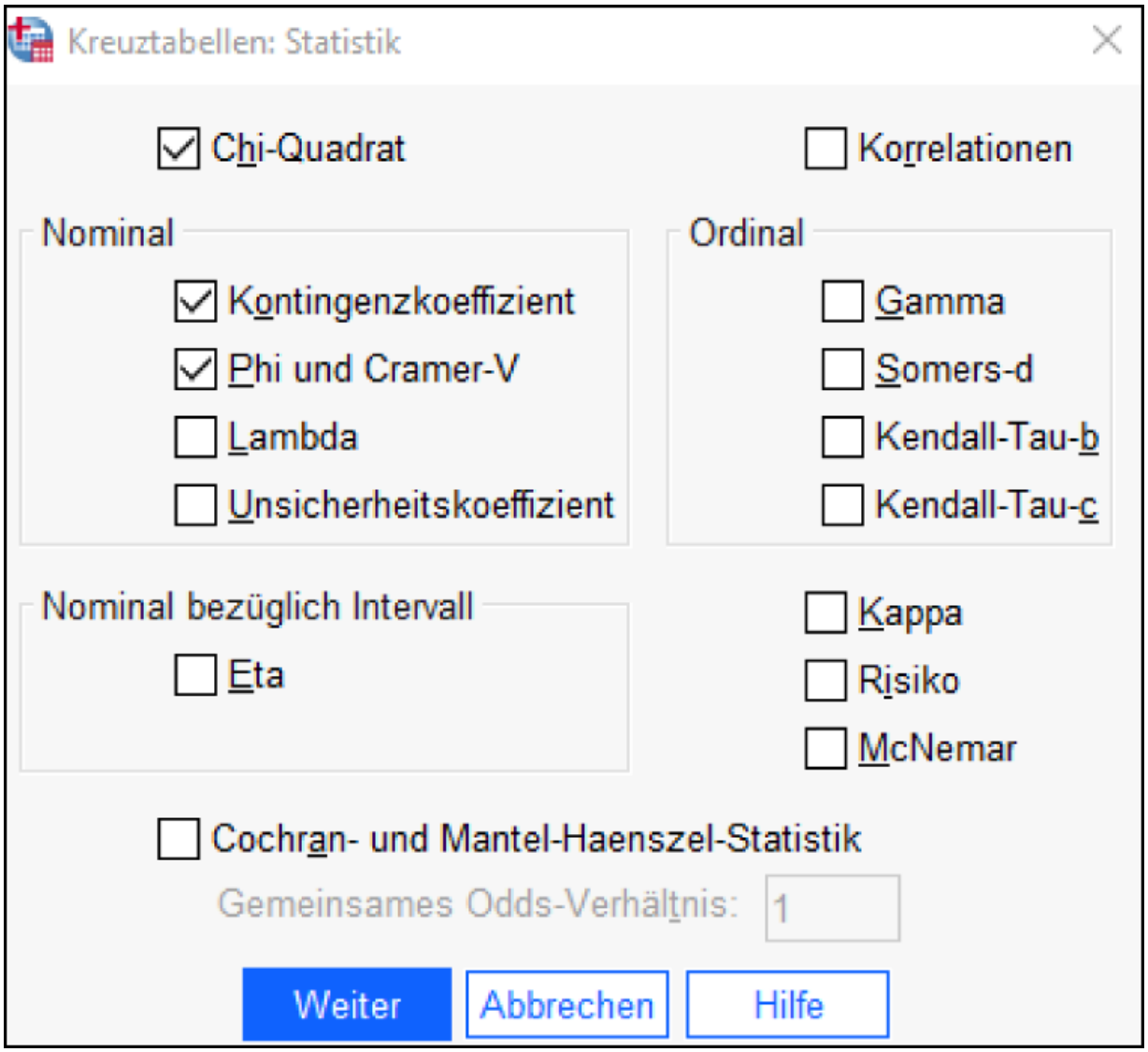



**Abbildung 8.21:** Auswahl verschiedener Assoziationsmaße (*GESCHL_FARBE.SAV*)

Mit den vorstehenden Angaben ergeben sich für unser Beispiel die in Abbildung 8.22 dargelegten Werte. Zum Glück stimmen diese Werte mit den oben händisch berechneten Werten für die einzelnen Assoziationsmaße überein. Die betreffenden Werte für die normierten Koeffizienten Phi-Kontingenzkoeffizient (0,516), Cramers V (ebenfalls 0,516) und C-Kontingenzkoeffizient (0,459) zeigen an, dass zwischen dem Geschlecht einer Person und ihrer Augenfarbe im Beispiel wohl schon ein statistischer Zusammenhang besteht, dass dieser aber keineswegs vollständig ist, vom Gefühl her eher »mittel«.

**Chi-Quadrat-Tests**

| | Wert | df | Asymptotische Signifikanz (zweiseitig) |
|---|---|---|---|
| Pearson-Chi-Quadrat | 6,667[a] | 4 | ,155 |
| Likelihood-Quotient | 9,100 | 4 | ,059 |
| Anzahl der gültigen Fälle | 25 | | |

a. 9 Zellen (90,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist ,80.

**Symmetrische Maße**

| | | Wert | Näherungsweise Signifikanz |
|---|---|---|---|
| Nominal- bzgl. Nominalmaß | Phi | ,516 | ,155 |
| | Cramer-V | ,516 | ,155 |
| | Kontingenzkoeffizient | ,459 | ,155 |
| Anzahl der gültigen Fälle | | 25 | |

**Abbildung 8.22:** SPSS-Output für verschiedene Assoziationsmaße (*GESCHL_FARBE.SAV*)

Auch bei KREUZTABELLEN können Sie das **Bootstrapping-Verfahren** anwenden, das heißt: sich für eine bestimmte Stichprobenanzahl aus dem Datensatz (hier: 1.000 Stichproben) und für die einzelnen Kennziffern wahrscheinlichkeitsbasierte (95-Prozent-)Unter- und Obergrenzen ausweisen lassen. Hierzu aktivieren Sie in der KREUZTABELLEN-Dialogbox das Bootstrapping durch einen Klick auf BOOTSTRAP und setzen dann bei BOOTSTRAPPING DURCHFÜHREN einen Haken (siehe Abbildung 8.23).

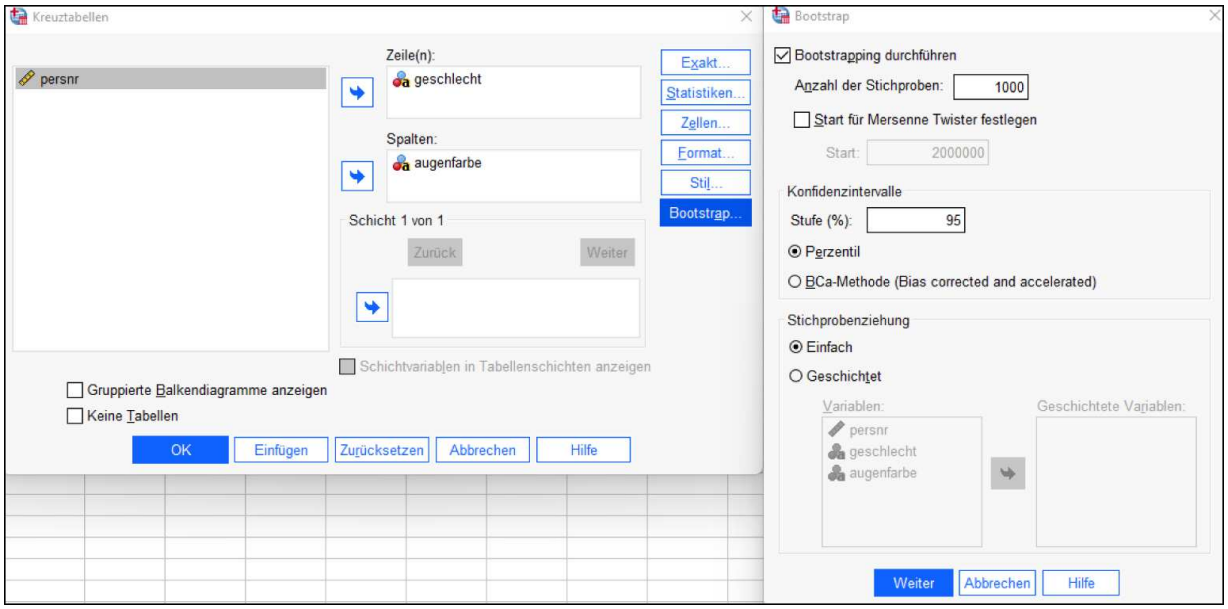



**Abbildung 8.23:** Bootstrapping für Assoziationsmaße (hier in Bezug auf *GESCHL_FARBE.SAV*)

Da Sie in der KREUZTABELLEN-Subdialogbox unter STATISTIK verschiedene Assoziationsmaße angegeben haben, werden durch das Bootstrapping, also durch relativ viele Stichprobenziehungen, bei Vorgabe einer Wahrscheinlichkeit (hier: 95 Prozent), Wertebereiche für diese Assoziationsmaße angegeben.

In unserem Beispiel ergaben sich bei meinen Bootstrapping-Berechnungen als Wertebereiche für den Phi-Kontingenzkoeffizienten und für Cramers V jeweils 0,387 als Untergrenze und 0,789 als Obergrenze. Für den C-Kontingenzkoeffizienten lautete der entsprechende Wertebereich: 0,361 als Untergrenze und 0,619 als Obergrenze (siehe Abbildung 8.24). Mit »*Verzerrung*« ist hier die Differenz zwischen dem Durchschnittswert der in der entsprechenden Zeile angegebenen statistischen Kennziffer in allen Bootstrap-Stichproben auf der einen Seite und dem Wert in der Spalte »Wert« auf der anderen Seite gemeint.

**Bootstrap für Symmetrische Maße**

| | | | Bootstrap[a] | | | |
|---|---|---|---|---|---|---|
| | | | | | 95% Konfidenzintervall | |
| | | Wert | Verzerrung | Std.-Fehler | Unterer Wert | Oberer Wert |
| Nominal- bzgl. Nominalmaß | Phi | ,516 | ,070 | ,105 | ,387 | ,789 |
| | Cramer-V | ,516 | ,070 | ,105 | ,387 | ,789 |
| | Kontingenzkoeffizient | ,459 | ,043 | ,068 | ,361 | ,619 |
| Anzahl der gültigen Fälle | | 25 | 0 | 0 | 25 | 25 |

a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

**Abbildung 8.24:** Bootstrapping-Output für verschiedene Assoziationsmaße (*GESCHL_FARBE.SAV*)

# *Korrelationen*

Während die im vorstehenden Abschnitt behandelten Assoziationsmaße bereits auf Nominalskalenniveau berechnet werden können und im Grunde genommen nur Aussagen dahin gehend ermöglichen, ob zwischen zwei Variablen »irgendwie« ein statistischer Zusammenhang besteht, ist die Lage im Fall von **Korrelationen** anders. Dort geht es nämlich sowohl um die Art beziehungsweise Richtung als auch um die Intensität des betreffenden Zusammenhangs. Dies erfordert von den miteinander korrelierten Variablen aber mindestens Ordinalskalenniveau.

Üblicherweise unterscheidet man bei den Korrelationen zwischen Rangkorrelationen einerseits und dem Korrelationskoeffizienten nach Bravais/Pearson andererseits.

Grundsätzlich fordern Sie Korrelationen in SPSS über ANALYSIEREN|KORRELATIONEN|BIVARIAT an. Der zugehörige SPSS-Befehl ist bei den Rangkorrelationen NONPAR CORR und beim Korrelationskoeffizienten nach Bravais/Pearson CORRELATIONS. Für bestimmte (Rang-)Korrelationsmaße können Sie aber auch über das KREUZTABELLEN-Menü beziehungsweise über den SPSS-Befehl CROSSTABS vorgehen, wie gleich noch näher beschrieben wird.

# *Rangkorrelationskoeffizienten*

Im Rahmen von **Rangkorrelationskoeffizienten** werden statistische Merkmalszusammenhänge für nach Rängen geordnete Variablen

untersucht. Daher müssen die betrachteten Variablen mindestens Ordinalskalenniveau aufweisen. Grundsätzlich unterscheidet man bei den Rangkorrelationskoeffizienten den Spearman'schen Rangkorrelationskoeffizienten von Kendalls Tau-Koeffizienten (beziehungsweise ähnlich wie Kendalls Tau-Koeffizient aufgebauten Maßen).

## *Rangkorrelationskoeffizient nach Spearman*

Der **Spearman'sche Rangkorrelationskoeffizient** bildet aus den Ausprägungen zweier mindestens ordinalskalierter Merkmale für beide Variablen jeweils eine Rangordnung. Wären die Variablen nicht mindestens ordinalskaliert, könnte man ja keine solche Rangordnung aufstellen – so einfach ist manchmal die Welt!

Sie können ja einfach mal den obigen Datensatz *GESCHL_FARBE.SAV* laden und versuchen, für die Stringvariablen *geschlecht* und *augenfarbe* einen Rangkorrelationskoeffizienten zu berechnen. Gehen Sie hierzu bitte über das SPSS-Menü ANALYSIEREN|KORRELATIONEN|BIVARIAT vor. Versuchen Sie, in der betreffenden Subdialogbox die beiden genannten Variablen einzugeben, werden Sie Schiffbruch erleiden! SPSS bietet Ihnen auf der linken Seite der Dialogbox gar keine dieser beiden Variablen an, sondern nur die Personennummer *persnr*, die hier aber leider ganz und gar nicht interessant ist (siehe Abbildung 8.25).

Woran liegt das? Im Unterschied zur numerischen Variablen *persnr* weisen die beiden Stringvariablen *geschlecht* und *augenfarbe* lediglich Nominalskalenniveau auf, und das reicht halt nicht, um in den elitären Kreis der Variablen aufzurücken, für die Korrelationsberechnungen möglich sind.

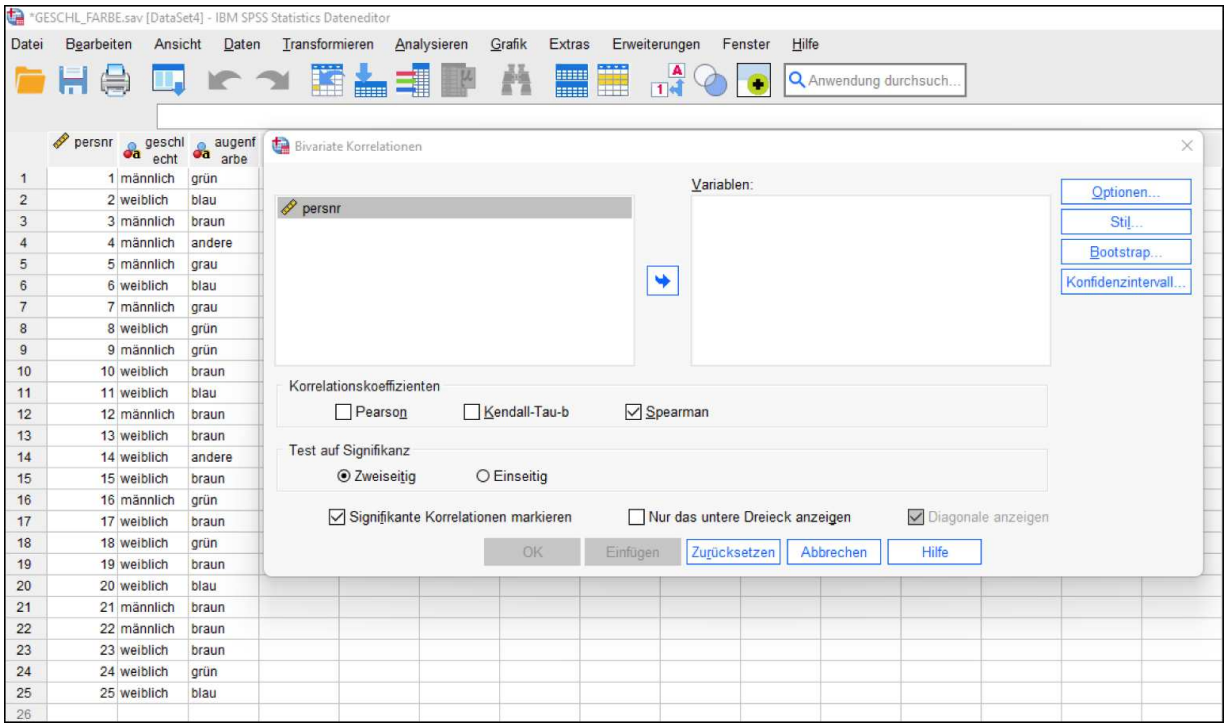



**Abbildung 8.25:** Voraussetzung von mindestens Ordinalskalenniveau bei Korrelationsberechnungen (*GESCHL_FARBE.SAV*)

Beide (nochmals: mindestens ordinalskalierten) Variablen sollten entweder beide aufsteigend (also vom kleinsten bis zum größten Wert) oder beide absteigend (also vom größten bis zum kleinsten Wert) sortiert sein. Man bildet dann für beide Variablen jeweils eine Rangfolge: Bei aufsteigender Sortierung erhält der höchste Wert den Rang 1, der zweithöchste Wert den Rang 2 und so weiter. Bei absteigender Sortierung wird demgegenüber dem niedrigsten Wert Rang 1, dem zweitniedrigsten Wert Rang 2 zugeordnet und so weiter. Diese Rangbildung weist darauf hin, dass bei Rangkorrelationskoeffizienten allein die jeweilige Rangfolge in die Berechnungen eingeht, nicht aber die Werteabstände bei den Variablen, sofern diese (natürlich) metrisch skaliert sind.

In einem nächsten Schritt werden jetzt die Unterschiede in der Rangfolge – die Differenzen der Merkmalsträger in den Rängen (nachfolgend mit $d_i$ bezeichnet) – ermittelt. Dies realisieren Sie im Rahmen des Spearman'schen Rangkorrelationskoeffizienten dadurch, dass Sie die betreffenden Differenzen zunächst quadrieren und

anschließend summieren. Das Quadrieren der Differenzen ist hier erforderlich, weil sich ansonsten die einfachen Differenzen zu 0 addieren würden und der betreffende Indikator wertlos würde.

Angenommen, Sie betrachten für drei Fälle zwei Variablen X und Y mit den X-Plätzen 1, 2 und 3 sowie den Y-Plätzen 3, 2 und 1. Dann erhalten Sie für die drei Fälle als einfache Differenzen −2, 0 und +2 beziehungsweise für deren Summe 0. Das ist höchst unglücklich, weil hier ja eigentlich eine vollständig-negative Korrelation zwischen X und Y vorliegt.

Damit der Spearman'sche Rangkorrelationskoeffizient $r_s$ innerhalb eines normierten Wertebereichs liegt (und somit Vergleiche zwischen verschiedenen Datensätzen leichter möglich werden), wird er letztlich wie folgt berechnet (mit N als gesamter Fallzahl):

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{N} d_i^2}{N \cdot \left(N^2 - 1\right)} \quad \text{mit:} \quad -1 \le r_s \le +1.$$

Erreicht $r_s$ seine Untergrenze von −1, liegt ein perfekt negativer Zusammenhang zwischen den beiden miteinander verglichenen Variablen vor. Das heißt: Die höchste Ausprägung der Variablen X ist beim betreffenden Merkmalsträger mit der niedrigsten Ausprägung der Variablen Y gekoppelt, die zweithöchste Ausprägung von X mit der zweitniedrigsten Ausprägung von Y und so weiter. Es liegt also zwischen X und Y ein gegenläufiger Verlauf der Merkmalsausprägungen beziehungsweise der Ränge vor.

Demgegenüber haben Sie es im Fall $r_s = +1$ mit einem perfekt positiven Zusammenhang zwischen X und Y zu tun. Das heißt: Die höchste Ausprägung der Variablen X ist beim betreffenden Merkmalsträger mit der höchsten Ausprägung der Variablen Y, die zweithöchste Ausprägung von X mit der zweithöchsten Ausprägung von Y verbunden und so weiter. Hier ist demnach der Verlauf der Ausprägungen von X und Y gleichgerichtet.

Ist $r_s$ gleich (beziehungsweise nahe) 0, liegt statistische Unabhängigkeit vor. Das heißt: Die Variablen X und Y haben – statistisch gesehen – nichts miteinander zu tun!

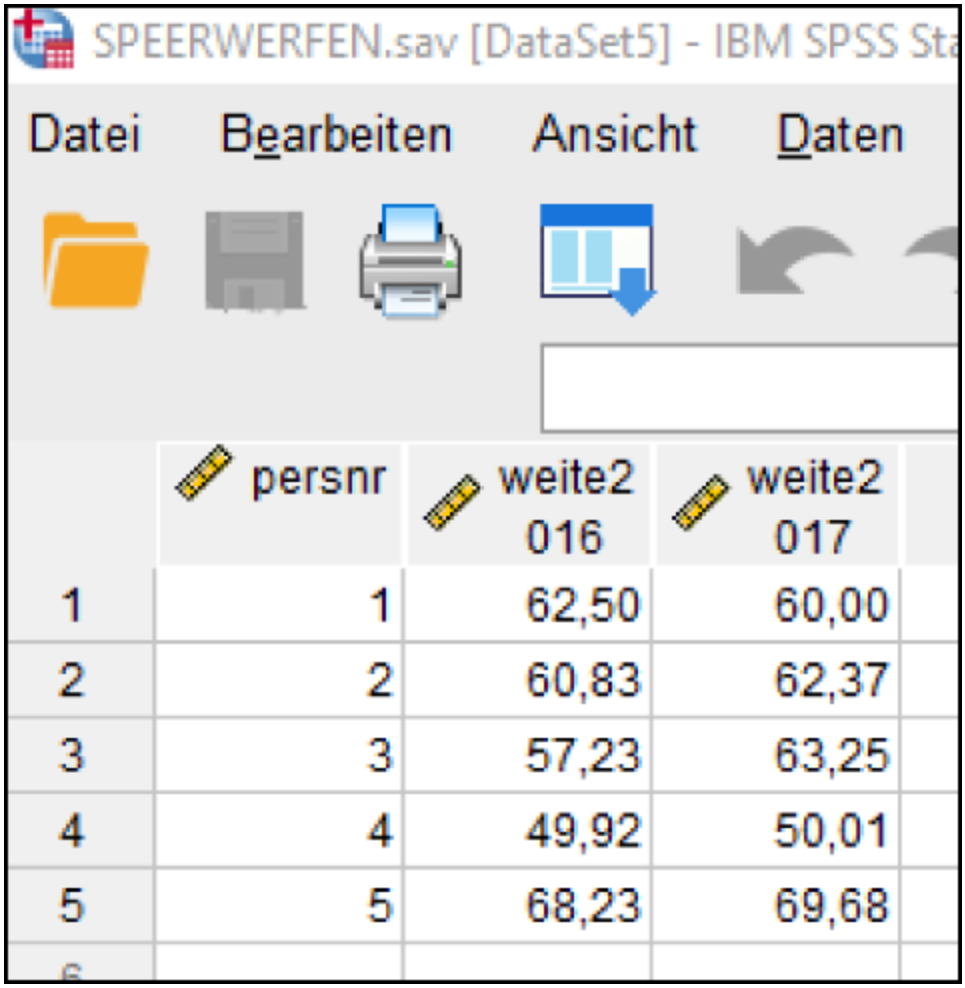
Bitte beachten Sie erneut, dass Korrelationen genauso wie Assoziationen lediglich statistische, aber nicht unbedingt inhaltliche Zusammenhänge widerspiegeln. Wenn zum Beispiel in China häufiger als in Deutschland ein Sack Reis umfällt, könnten Sie China Platz 1 und Deutschland Platz 2 bei der X-Variablen »Häufigkeit des Umfallens eines Sacks Reis« zuweisen. Auch bei der Y-Variablen »Erfolge im Tischtennis bei Weltmeisterschaften« steht China vor Deutschland auf Rang 1.

Wenn Sie nun für beide Variablen zum Beispiel den Spearman'schen Rangkorrelationskoeffizienten berechnen würden, erhielten Sie zunächst für die Differenzen $d_i$ jeweils 0, weil ja China bei beiden Variablen auf Platz 1 und Deutschland bei beiden Variablen auf Platz 2 steht. Damit ist aber auch in diesem einfachen Beispiel die quadrierte Summe der Differenzen $d_i$ gleich 0, und somit ist der Spearman'sche Rangkorrelationskoeffizient in diesem Fall gemäß der eben präsentierten Berechnungsformel gleich $+1\left(=1-\frac{6\cdot 0}{2\cdot 3}\right)$.

In einem statistischen Sinne liegt in diesem Beispielfall also eine perfekt positive, eine vollständig-gleichgerichtete (Rang-)Korrelation zwischen der X-Variablen »Häufigkeit des Umfallens eines Sacks Reis« und der Y-Variablen »Erfolge im Tischtennis bei Weltmeisterschaften« vor, ohne dass hier zwischen beiden Variablen ein sinnvoller inhaltlicher Zusammenhang herstellbar ist. Vielleicht haben Sie ja an dieser Stelle eine gute inhaltliche Erklärung – ich fürchte aber, dass Sie an dieser Herausforderung »scheitern« werden. Man spricht bei solchen statistischen, inhaltlich aber nicht sinnvollen Zusammenhängen zwischen zwei Variablen von einer Nonsenskorrelation, was ja auch

angebracht ist, wie der Nonsens des gewählten Beispiels gezeigt hat.

Lassen Sie uns nunmehr ein kleines, plausibleres Beispiel zur Rangkorrelation betrachten: Für fünf Speerwerferinnen soll in den früheren Jahren 2016 und 2017 jeweils ein Speerwurf-Wettbewerb stattgefunden haben. Dabei sollen die fünf Personen mit den Personennummern 1 bis 5 die in Abbildung 8.26 dargestellten Weiten in den Jahren 2016 und 2017 erzielt haben. Die dem Beispiel zugrunde liegende Datendatei finden Sie unter *SPEERWERFEN.SAV*.




| | persnr | weite2016 | weite2017 |
|---|---|---|---|
| 1 | 1 | 62,50 | 60,00 |
| 2 | 2 | 60,83 | 62,37 |
| 3 | 3 | 57,23 | 63,25 |
| 4 | 4 | 49,92 | 50,01 |
| 5 | 5 | 68,23 | 69,68 |

**Abbildung 8.26:** Datendatei *SPEERWERFEN.SAV*

In beiden Jahren hat gemäß Abbildung 8.26 die Speerwerferin Nummer 5 die größte Weite erzielt und die Speerwerferin Nummer 4 jeweils die geringste Weite. Platz 1 geht somit jeweils an die Speerwerferin

Nummer 5 und Platz 5 an die Speerwerferin Nummer 4. Im Einzelnen ergeben sich im Beispiel die in Tabelle 8.1 angegebenen Platzierungen.

| *Speerwerferin* | *Platz 2016* | *Platz 2017* |
|---|---|---|
| Nummer 1 | 2 | 4 |
| Nummer 2 | 3 | 3 |
| Nummer 3 | 4 | 2 |
| Nummer 4 | 5 | 5 |
| Nummer 5 | 1 | 1 |

**Tabelle 8.1:** Beispiel 1 für Rangkorrelationen

In der vorstehenden Tabelle haben Sie demnach die eigentlich metrisch skalierten Variablen *weite2016* und *weite2017* auf Ordinalskalenniveau »herunterskaliert«, also auf das bei Rangkorrelationen notwendige Skalenniveau.

Zur Berechnung des Spearman'schen Rangkorrelationskoeffizienten $r_S$ benötigen Sie die jeweiligen Platzdifferenzen $d_i$ (im Sinne der Differenz aus dem Platz 2016 und dem Platz 2017), die ganz offensichtlich folgende Ausprägungen annehmen: Speerwerferin Nummer 1: −2, Speerwerferin Nummer 2: 0, Speerwerferin Nummer 3: +2, Speerwerferin Nummer 4: 0, Speerwerferin Nummer 5: 0. Mit der Gesamtfallzahl N = 5 erhalten Sie folglich als Wert für den Spearman'schen Rangkorrelationskoeffizienten im Beispiel:

$$r_s = 1 - \frac{6 \cdot \left( (-2)^2 + 0^2 + 2^2 + 0^2 + 0^2 \right)}{5 \cdot \left( 5^2 - 1 \right)} = 1 - \frac{6 \cdot 8}{5 \cdot 24} = +0{,}600.$$

In diesem Beispiel beträgt der Spearman'sche Rangkorrelationskoeffizient also +0,600. Die Platzierungen der fünf Speerwerferinnen sind also tendenziell gleichgerichtet in beiden betrachteten Jahren 2016 und 2017, wie das positive Vorzeichen von $r_S$ anzeigt.

Dieses kleine »Speerwurf-Beispiel« war mit dem Taschenrechner noch recht leicht zu bewältigen. Haben Sie aber deutlich mehr Fälle vorliegen, wird das Ganze schon komplexer, und spätestens dann sollten Sie die Hilfe von SPSS in Anspruch nehmen.

Wie das konkret geht, demonstriere ich Ihnen am vorstehend aufgeführten »Speerwurf-Beispiel«. Sie müssen einfach über das SPSS-Menü zunächst ANALYSIEREN|KORRELATIONEN|BIVARIAT aufrufen. Anschließend fügen Sie in der sich öffnenden Dialogbox bei VARIABLEN die beiden zu analysierenden Variablen *weite2016* und *weite2017* ein und setzen bei KORRELATIONSKOEFFIZIENTEN das Häkchen bei SPEARMAN (siehe Abbildung 8.27).

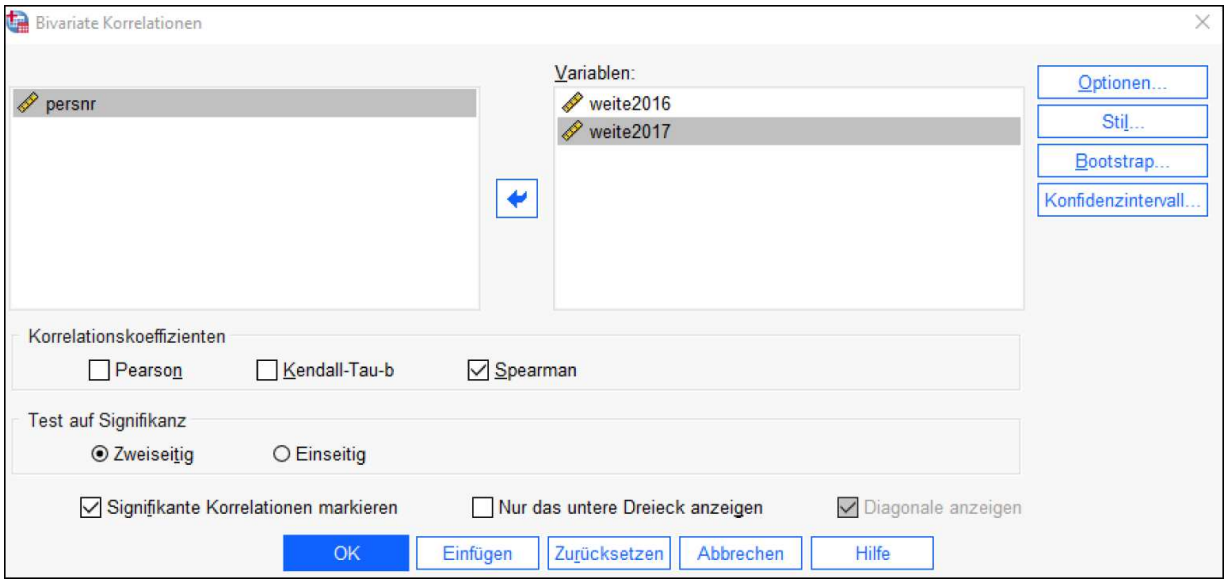



**Abbildung 8.27:** Aufruf des Spearman'schen Rangkorrelationskoeffizienten in SPSS (*SPEERWERFEN.SAV*)

Sie hätten alternativ über den SPSS-Syntaxeditor auch eingeben können:

```
NONPAR CORR
    /VARIABLES=weite2016 weite2017
    /PRINT=SPEARMAN.
```

In jedem Fall erhalten Sie das oben bereits händisch (beziehungsweise per Taschenrechner) berechnete Ergebnis für den Spearman'schen Rangkorrelationskoeffizienten in Höhe von +0,600 (siehe Abbildung 8.28).

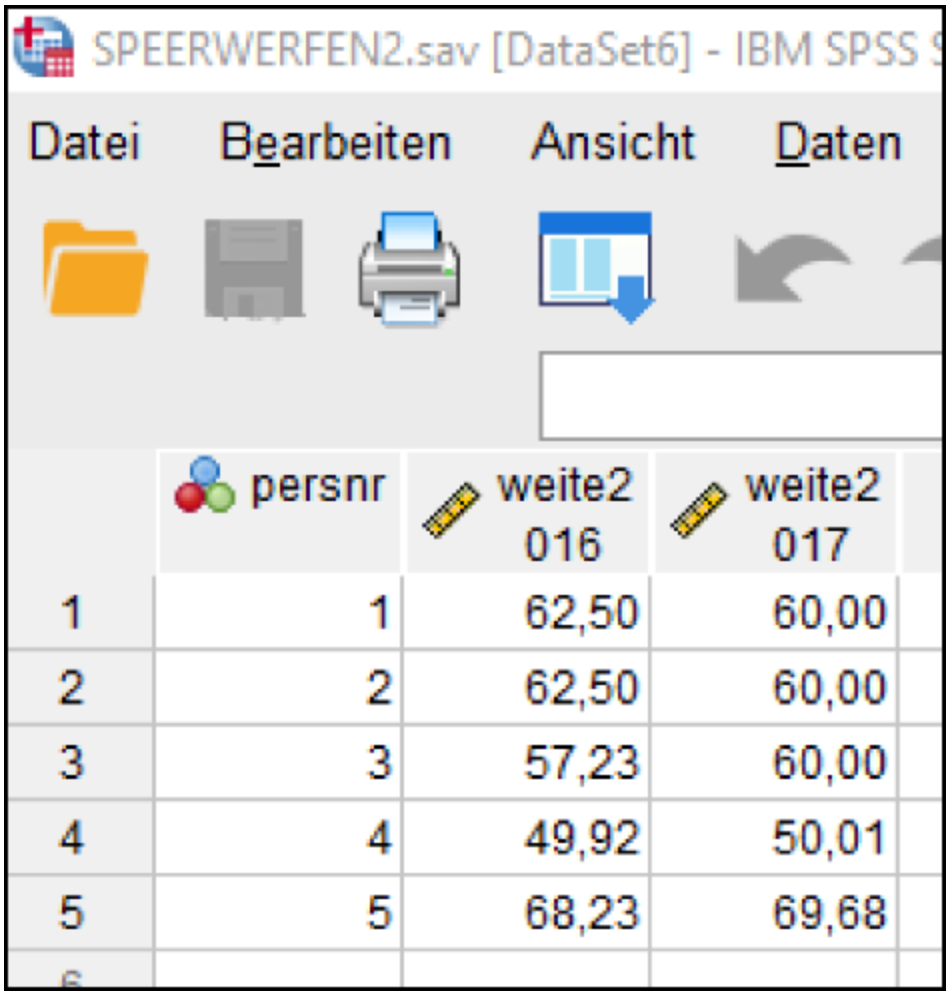
**Korrelationen**

| | | | weite2016 | weite2017 |
|---|---|---|---|---|
| Spearman-Rho | weite2016 | Korrelationskoeffizient | 1,000 | ,600 |
| | | Sig. (2-seitig) | . | ,285 |
| | | N | 5 | 5 |
| | weite2017 | Korrelationskoeffizient | ,600 | 1,000 |
| | | Sig. (2-seitig) | ,285 | . |
| | | N | 5 | 5 |

**Abbildung 8.28:** Wert des Rangkorrelationskoeffizienten – Beispiel 1 (*SPEERWERFEN.SAV*)

Etwas komplizierter gestaltet sich die Berechnung des Spearman'schen Rangkorrelationskoeffizienten, wenn Ränge mehrfach besetzt sind. Diese Mehrfachbesetzungen nennt man mit einem Fachbegriff auch **Ties**. Wenn zum Beispiel Werte bei einer Variablen doppelt vorkommen, ist die Summe der beiden infrage kommenden Ränge zu halbieren und dies den beiden betreffenden Merkmalsträgern jeweils zuzuweisen. Ein weiteres Beispiel: Kommen Werte bei einer Variablen dreimal vor, müssen Sie die betreffende Rangsumme dritteln und jedem der drei Merkmalsträger zuordnen.

Lassen Sie uns diesen Sachverhalt am »Speerwerferinnen-Beispiel« verdeutlichen. Hierzu habe ich in der Datendatei *SPEERWERFEN2.SAV* für die Speerwerferinnen Nummer 1 und Nummer 2 im Jahr 2016 jeweils eine gleiche Weite (jeweils 62,50 Meter) und für die Speerwerferinnen Nummer 1 bis 3 im Jahr 2017 ebenfalls jeweils eine gleiche Weite (jeweils 60,00 Meter) unterstellt (siehe Abbildung 8.29).




| | persnr | weite2016 | weite2017 |
|---|---|---|---|
| 1 | 1 | 62,50 | 60,00 |
| 2 | 2 | 62,50 | 60,00 |
| 3 | 3 | 57,23 | 60,00 |
| 4 | 4 | 49,92 | 50,01 |
| 5 | 5 | 68,23 | 69,68 |

**Abbildung 8.29:** Datendatei *SPEERWERFEN2.SAV* (mit Mehrfachbesetzungen)

Entsprechend ändern sich die Platzierungen in den Jahren 2016 und 2017. Im Jahr 2016 belegen die beiden Speerwerferinnen Nummer 1 und 2 jetzt jeweils den 2./3. Platz; ihnen wird also die Platzziffer $2{,}5\,(=\frac{2+3}{2})$ zugewiesen. Im Jahr 2017 haben die Werferinnen Nummer 1 bis 3 gleich weit geworfen und sind alle 2. bis 4. geworden. Ihnen wird für 2017 als (arithmetischem) Mittelwert jeweils der 3. Platz $(=\frac{2+3+4}{3})$ zugewiesen (siehe Tabelle 8.2):

| *Speerwerferin* | *Platz 2016* | *Platz 2017* |
|---|---|---|
| Nummer 1 | 2,5 | 3 |
| Nummer 2 | 2,5 | 3 |

| *Speerwerferin* | *Platz 2016* | *Platz 2017* |
|---|---|---|
| Nummer 3 | 4 | 3 |
| Nummer 4 | 5 | 5 |
| Nummer 5 | 1 | 1 |

**Tabelle 8.2:** Beispiel 2 für Rangkorrelationen

Sie erhalten nunmehr für die Differenzen $d_i$ (im Sinne der Differenz aus Platz 2016 und Platz 2017): Nummer 1: −0,5, Nummer 2: −0,5, Nummer 3: +1, Nummer 4: 0, Nummer 5: 0. Der neue Spearman'sche Rangkorrelationskoeffizient berechnet sich entsprechend gemäß der obigen Formel als:

$$r_s = 1 - \frac{6 \cdot \left( (-0{,}5)^2 + (-0{,}5)^2 + 1^2 + 0^2 + 0^2 \right)}{5 \cdot \left( 5^2 - 1 \right)} = 1 - \frac{6 \cdot 1{,}5}{5 \cdot 24} = +0{,}925.$$

Durch die Ties steigt der Spearman'sche Rangkorrelationskoeffizient also nach dieser Form der Berechnung auf +0,925. Allerdings kann die vorstehende Berechnung nur als (grobe) Näherungslösung angesehen werden. Korrekterweise muss nämlich beim Vorliegen von Ties eine andere Form der Berechnung des Spearman'schen Rangkorrelationskoeffizienten verwendet werden:

$$r_s = \frac{N^3 - N - \frac{1}{2} \cdot T_x - \frac{1}{2} \cdot T_y - 6 \cdot \sum_{i=1}^{N} d_i^2}{\sqrt[2]{\left( N^3 - N - T_x \right) \cdot \left( N^3 - N - T_y \right)}}$$

$$\text{mit}: T_x = \sum_K \left( t_{X,K}^3 - t_{X,K} \right) \text{und}: T_y = \sum_K \left( t_{Y,K}^3 - t_{Y,K} \right).$$

Hierbei kennzeichnen $t_{X,K}$ beziehungsweise $t_{Y,K}$ die Anzahl der Fälle mit gleichem Rang bei der X- und bei der Y-Variablen (K ist also hier der Index für die betreffenden Fälle). In unserem Beispielfall haben Sie zwei Fälle mit gleichem Rang bei der X-Variablen *weite2016* und drei Fälle

mit gleichem Rang bei der Y-Variablen *weite2017*. Daher ist hier $T_X = 2^3 - 2 = 6$ und $T_Y = 3^3 - 3 = 24$, sodass Sie für den Spearman'schen Rangkorrelationskoeffizienten erhalten:

$$r_s = \frac{5^3 - 5 - \frac{1}{2}\cdot 6 - \frac{1}{2}\cdot 24 - 6\cdot 1{,}5}{\sqrt[2]{\left(5^3 - 5 - 6\right)\cdot\left(5^3 - 5 - 24\right)}} \approx +0{,}918.$$

Wenn Ihnen die vorstehenden Berechnungen schon bei diesem einfachen Beispiel mit gerade einmal fünf Fällen und einem Zweier-Tie bei der X- und einem Dreier-Tie bei der Y-Variablen zu »hoch« sind und Sie dem Motto »Warum einfach, wenn es auch kompliziert geht?« so ganz und gar nichts abgewinnen können, es geht auch einfacher! Lassen Sie doch SPSS für Sie rechnen, und, voilà, Sie bekommen auf Knopfdruck und deutlich schneller, als wenn Sie sich mit der obigen Formel herumquälen, im SPSS-Ausgabefenster das betreffende Ergebnis (siehe Abbildung 8.30). Hierfür haben Sie sich ja schließlich auch ein Statistikprogramm wie SPSS angeschafft! (Es gibt im Übrigen noch eine weitere Berechnungsformel für den Spearman'schen Rangkorrelationskoeffizienten, und zwar in Anlehnung an den im nächsten Abschnitt diskutierten Bravais/Pearson'schen Korrelationskoeffizienten als Verhältnis aus der Kovarianz der Ränge von X und Y einerseits sowie dem Produkt aus den Standardabweichungen für die Ränge von X beziehungsweise von Y andererseits. Aber auch bei dieser Formel müssen Sie ziemlich viel rechnen, weshalb auch hier SPSS Ihnen die Arbeit abnehmen sollte …)

**Korrelationen**

| | | | weite2016 | weite2017 |
|---|---|---|---|---|
| Spearman-Rho | weite2016 | Korrelationskoeffizient | 1,000 | ,918* |
| | | Sig. (2-seitig) | . | ,028 |
| | | N | 5 | 5 |
| | weite2017 | Korrelationskoeffizient | ,918* | 1,000 |
| | | Sig. (2-seitig) | ,028 | . |
| | | N | 5 | 5 |

*. Die Korrelation ist auf dem 0,05 Niveau signifikant (zweiseitig).

**Abbildung 8.30:** Wert des Rangkorrelationskoeffizienten – Beispiel 2 (*SPEERWERFEN2.SAV*)

## *Kendalls Tau-Koeffizient*

Für die Berechnung von Rangkorrelationen müssen Sie nicht unbedingt auf den Spearman'schen Rangkorrelationskoeffizienten zurückgreifen. Es gibt Alternativen. Eine davon ist **Kendalls Tau-Koeffizient**. Auch er erfordert (naheliegenderweise) von den beiden Variablen, die miteinander korreliert werden sollen, mindestens Ordinalskalenniveau.

Bei Kendalls Tau-Koeffizienten werden für die beiden Variablen X und Y zwei Gruppen von Paaren gebildet (jeweils bezogen auf zwei miteinander verglichene Merkmalsträger): Befindet sich ein Merkmalsträger in beiden Merkmalsverteilungen von seinem Rang her oberhalb eines anderen Merkmalträgers, liegt ein *konkordantes Paar* vor. In unserem »Speerwerferinnen-Beispiel« aus dem vorstehenden Unterabschnitt dieses Kapitels weist zum Beispiel die Speerwerferin Nummer 5 in beiden Jahren eine bessere Platzierung als Speerwerferin Nummer 4 auf. In den Fällen hingegen, in denen sich die paarweise Rangfolge von Merkmalsträgern bei der X-Variablen von derjenigen bei der Y-Variablen unterscheidet, spricht man davon, dass ein *diskordantes Paar* vorliegt: Im »Speerwerferinnen-Beispiel« gemäß *SPEERWERFEN.SAV* (siehe Abbildung 8.26) hat 2016 die Werferin Nummer 1 zwar weiter als Werferin Nummer 2 geworfen; 2017 war dies aber genau umgekehrt.

Für die einzelnen paarweisen Vergleiche lassen sich insgesamt $\binom{N}{2} = \frac{N!}{(N-2)!\cdot 2!} = \frac{N\cdot(N-1)}{2}$ Möglichkeiten berechnen (N wieder als Symbol für die gesamte Fallzahl). Bei unseren läppischen 5 Fällen im »Speerwerferinnen-Beispiel« gibt es demnach nur insgesamt 10 Paarbeziehungen. Bei größeren Fallzahlen wird das Ganze aber recht komplex. Denken Sie zum Beispiel an einen Speerwurf-Wettbewerb – realistischer wäre vielleicht, wenn Sie in diesem Zusammenhang an einen Volkslauf oder Ähnliches denken – mit sagen wir mal 1.000 Teilnehmern, dann ergeben sich immerhin 499.500 Paarbeziehungen, und vermutlich wird Ihnen in diesem Fall der Rechenaufwand zu groß, und Sie werden an SPSS »übergeben«.

Für Kendalls Tau-Koeffizienten gibt es im Wesentlichen zwei Versionen (mit $N_c$ für die Anzahl der konkordanten Paare, $N_d$ für die Anzahl der diskordanten Paare und $\frac{N\cdot(N-1)}{2}$, wie bereits erwähnt, für die Gesamtzahl aller möglichen Paare):

1. wenn ausschließlich konkordante oder ausschließlich diskordante Paare (und keine Ties) vorliegen:

$$\tau_a = \frac{N_c - N_d}{\frac{N\cdot(N-1)}{2}}$$

2. bei einer Mischung aus konkordanten und diskordanten Paaren und eventuell bei Vorliegen von Ties ($\varsigma_X$ beziehungsweise $\varsigma_Y$):

$$\tau_b = \frac{N_c - N_d}{\sqrt[2]{(N_c + N_d + \varsigma_X)\cdot(N_c + N_d + \varsigma_Y)}};$$

Auf eine dritte, eher weniger verbreitete Version c des Kendall'schen Tau-Koeffizienten ($\tau_c$) wird im nächsten Unterabschnitt kurz eingegangen.

Lassen Sie uns nunmehr auf der Basis von *SPEERWERFEN.SAV* (keine Ties) und *SPEERWERFEN2.SAV* (ein Zweier-Tie bei der X- und ein Dreier-Tie bei der Y-Variablen) Kendalls Tau-Koeffizienten berechnen.

Beginnen Sie bitte mit *SPEERWERFEN.SAV*, also mit der Datendatei, in der keine Ties vorliegen. Sie müssen hier zunächst die insgesamt 10 Paarbeziehungen auf »konkordant« beziehungsweise auf »diskordant« abklopfen (siehe hierzu die Rangfolgen in Tabelle 8.1); es ergeben sich die in Tabelle 8.3 angegebenen Paarbeziehungen:

| ***Paarvergleich*** | ***Art der Paarbeziehung*** |
|---|---|
| Nummer 1 / Nummer 2 | Diskordant |
| Nummer 1 / Nummer 3 | Diskordant |
| Nummer 1 / Nummer 4 | Konkordant |
| Nummer 1 / Nummer 5 | Konkordant |
| Nummer 2 / Nummer 3 | Diskordant |
| Nummer 2 / Nummer 4 | Konkordant |
| Nummer 2 / Nummer 5 | Konkordant |
| Nummer 3 / Nummer 4 | Konkordant |
| Nummer 3 / Nummer 5 | Konkordant |
| Nummer 4 / Nummer 5 | Konkordant |

**Tabelle 8.3:** Beispiel 1 (keine Ties): Paarbeziehungen (*SPEERWERFEN.SAV*)

Insgesamt erhalten Sie also in diesem Beispiel ohne Ties 7 konkordante und 3 diskordante Paarbeziehungen. Da nicht ausschließlich konkordante oder diskordante Paarbeziehungen existieren, kann Kendalls Tau-Koeffizient in der Variante a nicht berechnet werden. Kendalls Tau-Koeffizient muss daher in der Variante b berechnet werden. Es ergibt sich ein Wert mit einem positiven Vorzeichen, was darauf hinweist, dass hier die konkordanten Paarbeziehungen die diskordanten Paarbeziehungen überwiegen:

$$\tau_b = \frac{7-3}{\sqrt[2]{(7+3+0)\cdot(7+3+0)}} = +0,400.$$

Wie kommen Sie jetzt zu diesem Ergebnis via SPSS? Ganz einfach: Gehen Sie bitte zunächst über das SPSS-Menü zu ANALYSIEREN|KORRELATIONEN|BIVARIAT, fügen Sie wieder in der sich öffnenden Dialogbox bei VARIABLEN die beiden zu analysierenden Variablen *weite2016* und *weite2017* ein und setzen Sie dann bei KORRELATIONSKOEFFIZIENTEN das Häkchen bei KENDALL-TAU-B (siehe Abbildung 8.31), und dann wirft Ihnen SPSS Kendalls Tau-Koeffizienten b aus (allerdings auch nur diesen).

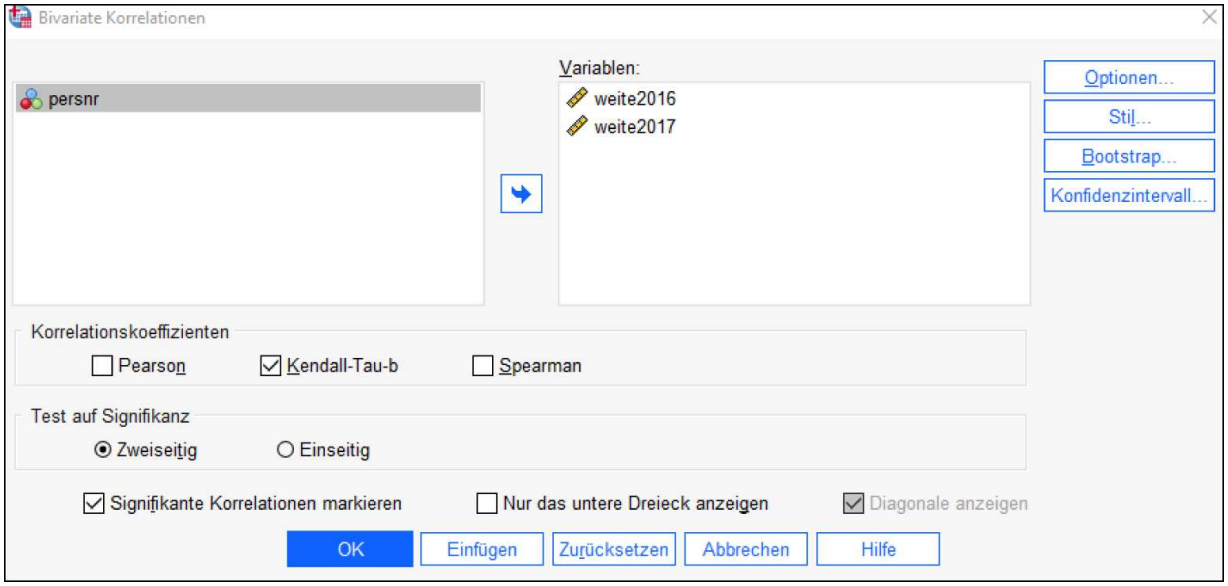



**Abbildung 8.31:** Aufruf von Kendalls Tau-Koeffizienten b in SPSS (*SPEERWERFEN.SAV*)

Die entsprechende SPSS-Syntax lautet im Übrigen:

```
NONPAR CORR
/VARIABLES=weite2016 weite2017
/PRINT=KENDALL.
```

Das erwartete Ergebnis eines Tau-Koeffizienten in Höhe von +0,400 wird auf diese Weise auch von SPSS ausgewiesen (siehe Abbildung 8.32).

Der Vollständigkeit halber lassen Sie uns auch den aus Abbildung 8.29 bekannten Fall der Datendatei *SPEERWERFEN2.SAV* mit Ties

betrachten. Auch hier müssen Sie zunächst die insgesamt 10 Paarbeziehungen auf »konkordant« beziehungsweise auf »diskordant«, jetzt aber auch noch zusätzlich auf X- und Y-Ties hin überprüfen (siehe hierzu die Rangfolgen in Tabelle 8.1); es ergeben sich (mit X = *weite2016* und Y = *weite2017*) die in Tabelle 8.4 dargestellten Paarbeziehungen.

**Korrelationen**

| | | | weite2016 | weite2017 |
|---|---|---|---|---|
| Kendall-Tau-b | weite2016 | Korrelationskoeffizient | 1,000 | ,400 |
| | | Sig. (2-seitig) | . | ,327 |
| | | N | 5 | 5 |
| | weite2017 | Korrelationskoeffizient | ,400 | 1,000 |
| | | Sig. (2-seitig) | ,327 | . |
| | | N | 5 | 5 |

**Abbildung 8.32:** Wert des Tau-b-Koeffizienten – Beispiel 1 (*SPEERWERFEN.SAV*)

| *Paarvergleich* | *Art der Paarbeziehung* |
|---|---|
| Nummer 1 / Nummer 2 | X- und Y-Tie |
| Nummer 1 / Nummer 3 | Y-Tie |
| Nummer 1 / Nummer 4 | Konkordant |
| Nummer 1 / Nummer 5 | Konkordant |
| Nummer 2 / Nummer 3 | Y-Tie |
| Nummer 2 / Nummer 4 | Konkordant |
| Nummer 2 / Nummer 5 | Konkordant |
| Nummer 3 / Nummer 4 | Konkordant |
| Nummer 3 / Nummer 5 | Konkordant |
| Nummer 4 / Nummer 5 | Konkordant |

**Tabelle 8.4:** Beispiel 2 (mit Ties): Paarbeziehungen (*SPEERWERFEN2.SAV*)

Liegt, wie im vorstehenden Beispiel bei der Paarbeziehung »Nummer 1 / Nummer 2«, sowohl bei der X- als auch bei der Y-Variablen ein Tie gleichzeitig vor, ein X/Y-Tie sozusagen, dann wird diese Paarbeziehung bei der Berechnung von Kendalls Tau-Koeffizienten *nicht* berücksichtigt!

Im Beispiel hier liegen also 0 »reine« X-Ties und 2 »reine« Y-Ties sowie 7 konkordante und 0 diskordante Paarbeziehungen vor. Kendalls Tau-Koeffizient in der Variante b lautet mit diesen Angaben:

$$\tau_b = \frac{7-0}{\sqrt[2]{(7+0+0)\cdot(7+0+2)}} \approx +0{,}882.$$

Genau dieses Ergebnis »wirft« – zum Glück! – auch SPSS »aus« (siehe Abbildung 8.33).

**Korrelationen**

| | | | weite2016 | weite2017 |
|---|---|---|---|---|
| Kendall-Tau-b | weite2016 | Korrelationskoeffizient | 1,000 | ,882* |
| | | Sig. (2-seitig) | . | ,046 |
| | | N | 5 | 5 |
| | weite2017 | Korrelationskoeffizient | ,882* | 1,000 |
| | | Sig. (2-seitig) | ,046 | . |
| | | N | 5 | 5 |

*. Die Korrelation ist auf dem 0,05 Niveau signifikant (zweiseitig).

**Abbildung 8.33:** Wert des Tau-b-Koeffizienten – Beispiel 2 (*SPEERWERFEN2.SAV*)

## *Weitere Kennziffern*

In der Literatur finden sich über die in den beiden vorstehenden Unterabschnitten dieses Kapitels behandelten Rangkorrelationskoeffizienten hinaus noch weitere entsprechende Kennziffern (unter anderem mit Kendalls Tau-Koeffizienten in der Version c eine weitere Variante des Kendall'schen Tau-Koeffizienten). Es

können etwa folgende Kennziffern angeführt werden (mit $N_c$ für die Anzahl der konkordanten Paare, $N_d$ für die Anzahl der diskordanten Paare, $\varsigma_X$ für die Anzahl der X-Ties und $\varsigma_Y$ für die Anzahl der Y-Ties sowie i für die Anzahl der unterschiedlichen X-Ausprägungen und j für die Anzahl der unterschiedlichen Y-Ausprägungen):

- ✔ der **Gamma-Index** von Goodman/Kruskal:

$$\gamma = \frac{N_c - N_d}{N_c + N_d};$$

- ✔ **Somers d-Koeffizienten:**

  a. $$d_{YX} = \frac{N_c - N_d}{N_c + N_d + \varsigma_Y};$$

  b. $$d_{XY} = \frac{N_c - N_d}{N_c + N_d + \varsigma_X};$$

  c. $$d_S = \frac{N_c - N_d}{N_c + N_d + \frac{1}{2} \cdot (\varsigma_Y + \varsigma_X)}$$

  sowie

- ✔ der bereits erwähnte **Kendall'sche Tau-Koeffizient in der Version c:**

$$\tau_c = \frac{N_c - N_d}{\frac{1}{2} \cdot N^2 \cdot \frac{Minimum(i,j) - 1}{Minimum(i,j)}}$$

Auf der Grundlage dieser Formeln können Sie mit dem Taschenrechner für das Beispiel *SPEERWERFEN2.SAV* (mit Ties; siehe Abbildung 8.29) für die einzelnen Kennziffern Werte ausrechnen. Hierbei müssen Sie im Grunde genommen nur noch bedenken, dass im Beispiel von *SPEERWERFEN2.SAV* 0 »reine« X-Ties, 2 »reine« Y-Ties, 7 konkordante und 0 diskordante Paarbeziehungen vorliegen (und die Anzahl der unterschiedlichen Werteausprägungen der X-Variablen gleich 4 und die der Y-Variablen gleich 3 ist; die Paarbeziehung »Nr.1/Nr.2« mit einem X/Y-Tie wird auch hier vernachlässigt):

$$\gamma = \frac{7-0}{7+0} = +1{,}000;$$

$$d_{YX} = \frac{7-0}{7+0+2} \approx +0{,}778;$$

$$d_{XY} = \frac{7-0}{7+0+0} = +1{,}000;$$

$$d_S = \frac{7-0}{7+0+1} = +0{,}875;$$

$$\tau_c = \frac{7-0}{0{,}5 \cdot 5^2 \cdot \frac{2}{3}} = +0{,}840.$$

Wenn Sie jetzt diese Werte mit den von SPSS »ausgeworfenen« Werten vergleichen möchten, dürfen Sie nicht über das KORRELATIONEN-, sondern müssen über das KREUZTABELLEN-Menü vorgehen, also via ANALYSIEREN|DESKRIPTIVE STATISTIKEN|KREUZTABELLEN. Dies mag Sie überraschen, lässt sich aber nicht ändern und gehört sozusagen zu den paar »Merkwürdigkeiten« von SPSS. In der KREUZTABELLEN-Subdialogbox STATISTIK auf der rechten Seite müssen Sie nun die gewünschten ordinalen Korrelationsmaße anklicken. Wenn Sie, wie in Abbildung 8.34 geschehen, dabei auch noch KORRELATIONEN und KENDALLS TAU B anklicken, können Sie sich auch über dieses KREUZTABELLEN-Menü – als Alternative zum oben genutzten KORRELATIONEN-Menü – den Spearman'schen Rangkorrelationskoeffizienten (durch das Anklicken von KORRELATIONEN) sowie den Kendall'schen Tau-Koeffizienten in der Variante b von SPSS ausweisen lassen (siehe hierzu Abbildung 8.34). Des Weiteren wird auch noch – ebenfalls via KORRELATIONEN – der Korrelationskoeffizient nach Bravais/Pearson von SPSS ausgewiesen, auf den ich aber erst im nächsten Abschnitt dieses Kapitels kommen werde.

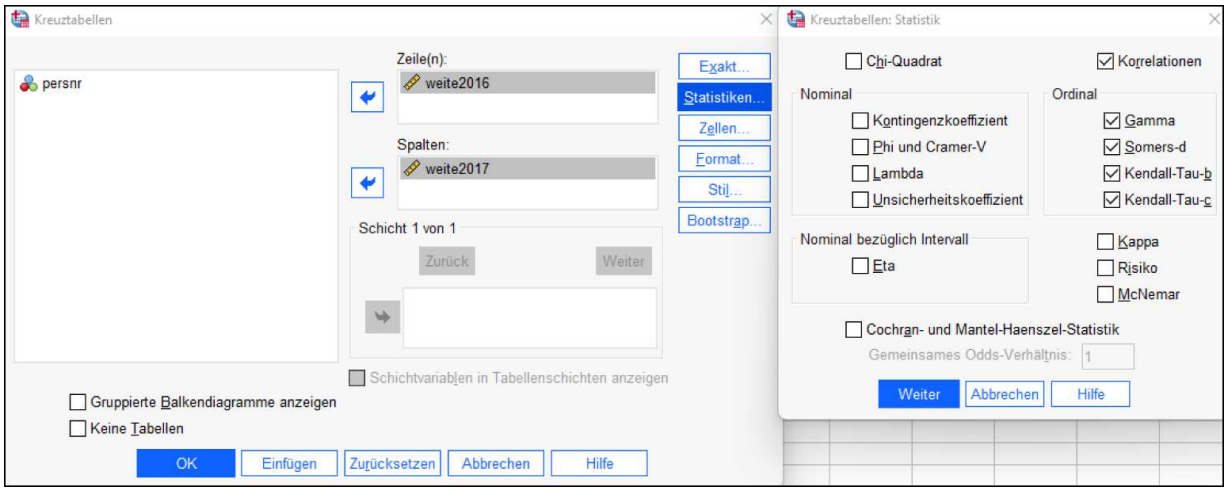



**Abbildung 8.34:** Aufruf verschiedener (weiterer) Rangkorrelationskoeffizienten in SPSS (*SPEERWERFEN2.SAV*)

Da Sie sich hier nur ein paar Korrelationskoeffizienten von SPSS ausrechnen lassen möchten, sollten Sie auf den Ausweis einer Kreuztabelle verzichten. Hierzu setzen Sie in der KREUZTABELLEN-Subdialogbox bei KEINE TABELLEN einfach ein Häkchen.

In SPSS-Syntax hätten Sie angeben müssen:

```
CROSSTABS
    /TABLES=weite2016 BY weite2017
    /FORMAT = NOTABLES
    /STATISTICS=CORR GAMMA D BTAU CTAU
    /COUNT ROUND CELL.
```

Wie aus Abbildung 8.35 ersichtlich, werden die mit dem Taschenrechner berechneten Koeffizientenwerte auch von SPSS ganz genauso berechnet. Sie können jetzt an dieser Stelle eine gute Flasche Wein öffnen, um Ihren Erfolg, richtig gerechnet zu haben, zu feiern!

Dem oberen Teil von Abbildung 8.35 können Sie dabei die verschiedenen Varianten von Somers d-Koeffizienten entnehmen. Die betreffenden Werte sind, wie ersichtlich, folgendermaßen den obigen Begrifflichkeiten zuzuordnen: »Symmetrisch« entspricht dem obigen Maß $d_S$, »*weite2016* abhängig« dem Maß $d_{XY}$ oben und »*weite2017* abhängig« dem Maß $d_{YX}$.

**Richtungsmaße**

| | | | Wert | Asymptotischer Standardfehler [a] | Näherungsweises t [b] | Näherungsweise Signifikanz |
|---|---|---|---|---|---|---|
| Ordinal- bzgl. Ordinalmaß | Somers-d | Symmetrisch | ,875 | ,109 | 3,195 | ,001 |
| | | weite2016 abhängig | 1,000 | ,000 | 3,195 | ,001 |
| | | weite2017 abhängig | ,778 | ,194 | 3,195 | ,001 |

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 | ,882 | ,110 | 3,195 | ,001 |
| | Kendall-Tau-c | ,840 | ,263 | 3,195 | ,001 |
| | Gamma | 1,000 | ,000 | 3,195 | ,001 |
| | Korrelation nach Spearman | ,918 | ,104 | 4,000 | ,028[c] |
| Intervall- bzgl. Intervallmaß | Pearson-R | ,942 | ,046 | 4,867 | ,017[c] |
| Anzahl der gültigen Fälle | | 5 | | | |

a. Die Null-Hyphothese wird nicht angenommen.

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

c. Basierend auf normaler Approximation.

**Abbildung 8.35:** Werte verschiedener Rangkorrelationskoeffizienten (und des Bravais/Pearson-Korrelationskoeffizienten) (*SPEERWERFEN2.SAV*)

# *Korrelationskoeffizient nach Bravais/Pearson*

Die »Königsklasse« der Korrelationsberechnungen bildet der **Korrelationskoeffizient nach Bravais/Pearson**. SPSS unterschlägt im Übrigen den Namen des armen Auguste Bravais und bezeichnet den betreffenden Koeffizienten einfach nur als Korrelationskoeffizienten nach Pearson.

Der Grund für die hochrangige Stellung des Korrelationskoeffizienten nach Bravais/Pearson ist, dass von den betrachteten Variablen (grundsätzlich) allesamt metrisches Skalenniveau verlangt wird. Das heißt: In den entsprechenden Korrelationsberechnungen werden nicht nur die Rangfolgenunterschiede, sondern auch noch die Abstände in den Merkmalsausprägungen berücksichtigt.

Wichtig ist, dass hier ausschließlich *lineare* Beziehungen zwischen der X- und der Y-Variablen erfasst werden.

Am besten übertragen Sie die Werte der X- und der Y-Variablen in ein Koordinatensystem, wie in Abbildung 8.36 geschehen, um einen optischen Eindruck von der Beziehung zwischen X und Y zu erhalten. Zugleich teilen Sie das betreffende Koordinatensystem anhand der jeweiligen Durchschnittswerte $\overline{X}$ bzw. $\overline{Y}$ noch einmal in vier Felder auf. Liegen – wie in Abbildung 8.36 – die einzelnen Beobachtungspunkte ausschließlich im linken unteren und im rechten oberen dieser Felder, kann man von einem klar positiven, gleichgerichteten statistischen Zusammenhang zwischen X und Y ausgehen. Befinden sich die einzelnen Beobachtungspunkte von X und Y hingegen allesamt im linken oberen und im rechten unteren Feld, kann man auf einen klar negativen, entgegengerichteten statistischen Zusammenhang zwischen X und Y schließen.

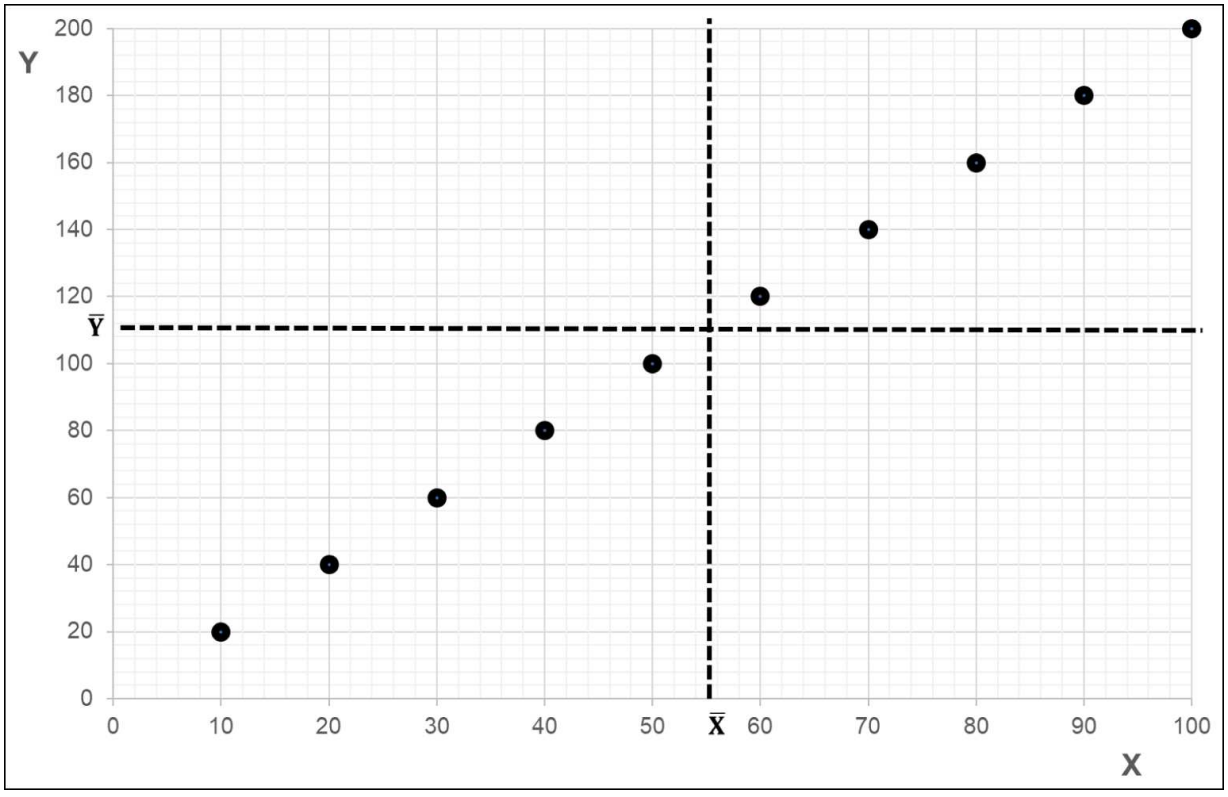



**Abbildung 8.36:** Eine perfekt-positive lineare Korrelation zwischen zwei metrischen Variablen X und Y (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und

Sozialwissenschaftler, Weinheim: Wiley, S. 153)

Je eher sich dabei die einzelnen Beobachtungspunkte auf einer Linie befinden, desto stärker ist die positive beziehungsweise negative (lineare) Korrelation. Dieser hier erwähnte Bezug auf eine Linie macht deutlich, dass es sich hier um lineare Korrelationen handelt. In Abbildung 8.36 liegt deshalb eine perfekte positive und zugleich lineare Korrelation zwischen X und Y vor, weil alle 10 Beobachtungspunkte wie »an der Schnur gezogen« auf einer Linie mit positiver Steigung liegen.

Je stärker die Beobachtungspunkte über die vier Felder streuen, die durch die Durchschnittswerte $\overline{X}$ bzw. $\overline{Y}$ aufgespannt werden, desto geringer ist die gemessene lineare Korrelation. Die Beobachtungspunkte von X und Y in Abbildung 8.37 beispielsweise legen keine lineare Korrelation zwischen X und Y nahe. Gleichwohl besteht ein klarer nichtlinearer (quadratischer) mathematischer Zusammenhang zwischen X und Y in der betreffenden Abbildung.

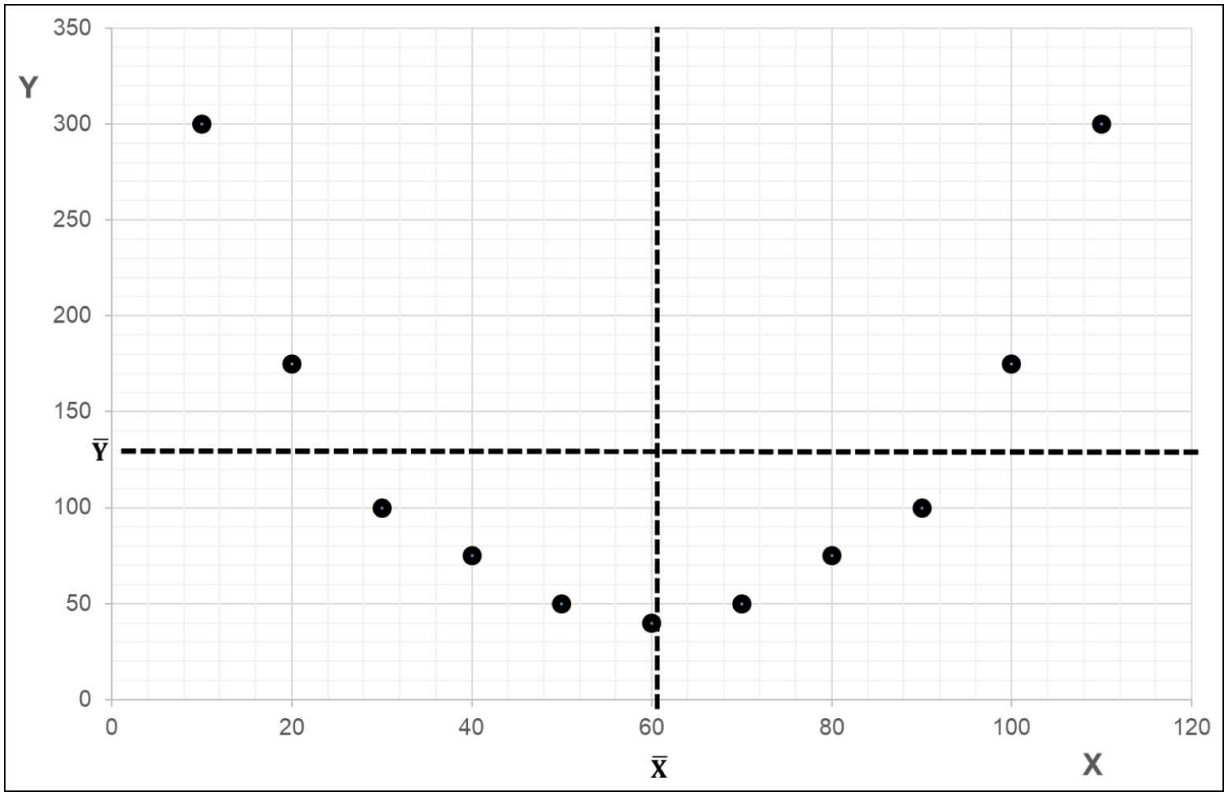


**Abbildung 8.37:** Keine lineare, aber eine nichtlineare Korrelation zwischen zwei metrischen Variablen X und Y (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 155)

> Durch das hier präsentierte Konzept der (linearen) Korrelation werden solche nichtlinearen Beziehungen zwischen X und Y nicht erfasst. Das sollten Sie unbedingt im Hinterkopf behalten.

Die linearen Korrelationsbetrachtungen in diesem Abschnitt fußen auf der sogenannten **Kovarianz** COV(X, Y) zwischen X und Y. Bei ihr geht es also um die gemeinsame Streuung zwischen den Merkmalsausprägungen von X und von Y. Ist die Kovarianz positiv, weist dies auf ein gleichgerichtetes Verhalten in den Daten von X und Y hin. Ein negativer Kovarianzwert hingegen signalisiert ein entgegengerichtetes Verhalten in den Daten von X und Y. Ist die Kovarianz schließlich gleich 0, liegt statistische (lineare) Unabhängigkeit zwischen X und Y vor:

$$COV(X,Y) = \frac{1}{N} \cdot \sum_{i=1}^{N} (X_i - \bar{X}) \cdot (Y_i - \bar{Y}).$$

Um die Kovarianzwerte auf den Wertebereich [−1; +1] zu normieren, wird die Kovarianz noch durch das Produkt aus der Standardabweichung für X ($S_x$) und der Standardabweichung für Y ($S_Y$) dividiert. Dies ergibt dann bereits den **Bravais/Pearson-Korrelationskoeffizienten** $r_{BP}$:

$$r_{BP} = \frac{COV(X,Y)}{S_X \cdot S_Y}.$$

Ist $r_{BP} = -1$, liegt ein perfekt-negativer (linearer) X-Y-Zusammenhang vor; bei $r_{BP} = +1$ hingegen spricht man von einem perfekt-positiven (linearen) Zusammenhang zwischen X und Y, und bei $r_{BP} = 0$ liegt statistische (lineare) Unabhängigkeit zwischen X und Y vor. Um das Ausmaß der jeweiligen Korrelation einordnen zu können, verwendet man bestimmte »Faustformeln«, etwa (bei Grundgesamtheitsdaten): $|r_{BP}| \leq 0{,}5$: schwache Korrelation, $0{,}5 < |r_{BP}| \leq 0{,}8$: mittlere Korrelation und

$|r_{BP}| > 0{,}8$: starke Korrelation – oder zum Beispiel (bei Stichprobendaten): $|r_{BP}| \leq 0{,}3$: schwache Korrelation, $0{,}3 < |r_{BP}| \leq 0{,}5$: mittlere Korrelation und $|r_{BP}| > 0{,}5$: starke Korrelation.

Lassen Sie uns an dieser Stelle wieder auf das Beispiel mit den fünf Speerwerferinnen aus *SPEERWERFEN2.SAV* kommen. Die Variable *weite2016* ist nachfolgend wieder die X-Variable und *weite2017* die Y-Variable. Zur Berechnung der Kovarianz führen Sie bitte mit Ihrem Taschenrechner die in Tabelle 8.5 angegebenen Berechnungen durch. Zunächst müssen Sie allerdings die Durchschnittswerte $\overline{X}$ und $\overline{Y}$ sowie die Standardabweichungen von X und Y berechnen. Für die Durchschnittswerte erhalten Sie:

$$\overline{X} = \frac{62{,}50 + 62{,}50 + 57{,}23 + 49{,}92 + 68{,}23}{5} = 60{,}076;$$

$$\overline{Y} = \frac{60{,}00 + 60{,}00 + 60{,}00 + 50{,}01 + 69{,}68}{5} = 59{,}956.$$

Die Standardabweichungen von X und Y – also $S_X$ und $S_Y$ – berechnen sich wie folgt:

$$S_X = \sqrt[2]{\frac{62{,}50^2 + 62{,}50^2 + 57{,}23^2 + 49{,}92^2 + 68{,}23^2}{5} - 60{,}076^2} \approx 6{,}156;$$

$$S_Y = \sqrt[2]{\frac{60{,}00^2 + 60{,}00^2 + 60{,}00^2 + 50{,}01^2 + 69{,}68^2}{5} - 59{,}9562^2} \approx 6{,}192.$$

So, und jetzt können Sie sich an die Berechnung der Kovarianz heranwagen. Dazu sollten Sie zunächst die nachfolgende Hilfstabelle erstellen.

| ***Speerwerferin*** | $X_i - \overline{X}$ | $Y_i - \overline{Y}$ | $(X_i - \overline{X}) \cdot (Y_i - \overline{Y})$ |
|---|---|---|---|
| Nummer 1 | 2,424 | 0,044 | 0,107 |
| Nummer 2 | 2,424 | 0,044 | 0,107 |

| *Speerwerferin* | $X_i - \bar{X}$ | $Y_i - \bar{Y}$ | $(X_i - \bar{X}) \cdot (Y_i - \bar{Y})$ |
|---|---|---|---|
| Nummer 3 | –2,846 | 0,044 | –0,125 |
| Nummer 4 | –10,156 | –9,946 | 101,012 |
| Nummer 5 | 8,154 | 9,724 | 79,289 |

**Tabelle 8.5:** Arbeitstabelle zur Berechnung der Kovarianz (in Bezug auf *SPEERWERFEN2.SAV*)

Die Kovarianz erhalten Sie ganz einfach als Summe der Werte in der letzten Spalte von Tabelle 8.5, die durch die Fallzahl N dividiert wird. In unserem Fall beträgt die Kovarianz: $\frac{180,390}{5} = 36,078.$ Damit sind Sie in der Lage, den Bravais/Pearson-Korrelationskoeffizienten $r_{BP}$ zu berechnen:

$$r_{BP} = \frac{36,078}{6,156 \cdot 6,192} \approx +0,946.$$

Diejenigen unter Ihnen, die Kapitel 7 bereits gelesen haben, werden einwenden, dass die obigen Berechnungen für die Kovarianz sowie für die Standardabweichungen von X und Y darauf basierten, dass durch die gesamte Fallzahl N dividiert wurde, dass SPSS aber jeweils durch (N-1) dividiert. Dies ist richtig, spielt hier aber keine Rolle, weil Sie beim Bravais/Pearson-Korrelationskoeffizienten den Zählerausdruck (die Kovarianz) ja durch N und auch den Nennerausdruck (das Produkt der beiden Standardabweichungen mit $\sqrt[2]{N} \cdot \sqrt[2]{N} = N$) ebenfalls durch N dividieren und sich $\frac{1}{N}$ also wegkürzt.

Bei SPSS kürzt sich $\frac{1}{N-1}$ in Zähler und Nenner beim Bravais/Pearson-Korrelationskoeffizienten weg. Es sollte also in beiden Berechnungsvarianten das Gleiche herauskommen.

Dass sich der von Ihnen per Taschenrechner ausgerechnete Korrelationskoeffizient nach Bravais/Pearson in der dritten

Nachkommastelle von dem nachfolgend präsentierten SPSS-Wert unterscheidet, hat keine inhaltlichen Gründe, sondern liegt einfach an kleinen Rundungsfehlern in den obigen Berechnungen. Die obigen Berechnungen wurden mit einer kleinen Ungenauigkeit durchgeführt, und zwar bei den Zwischenberechnungen mit jeweils lediglich drei Nachkommastellen.

Um sich in SPSS für das Beispiel *SPEERWERFEN2.SAV* den Bravais/Pearson-Korrelationskoeffizienten berechnen zu lassen, kommen Sie im SPSS-Menü über ANALYSIEREN|KORRELATION|BIVARIAT zu der in Abbildung 8.38 angegebenen Dialogbox. Dort müssen Sie jetzt lediglich unter KORRELATIONSKOEFFIZIENTEN auf PEARSON klicken. (SPSS »unterdrückt«, wie erwähnt, den armen Auguste Bravais!)

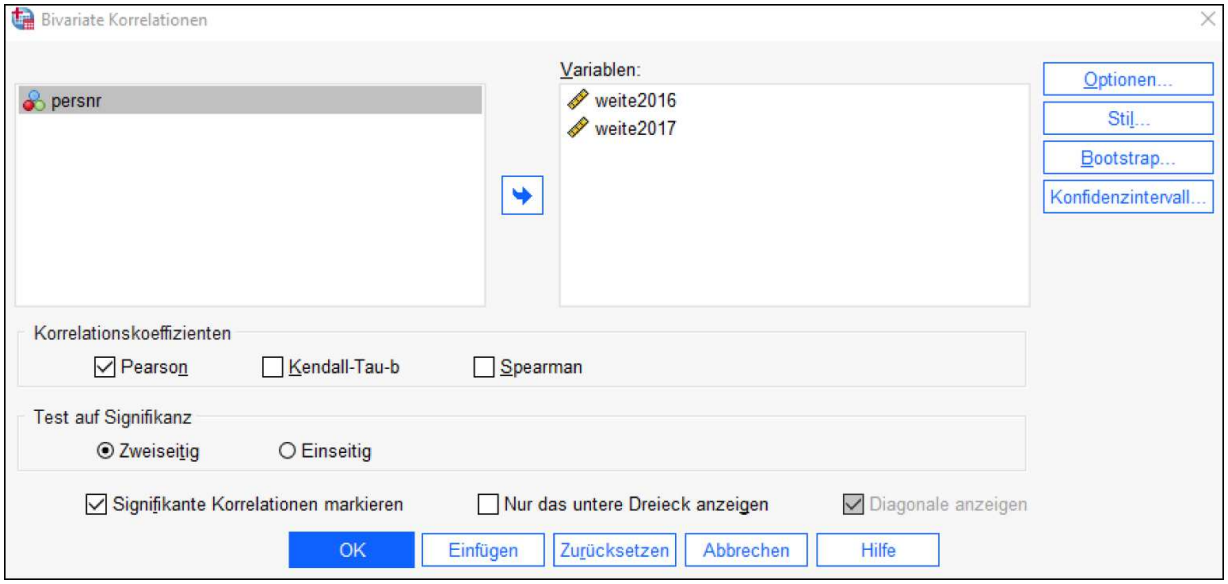



**Abbildung 8.38:** Aufruf des Bravais/Pearson-Korrelationskoeffizienten über das SPSS-KORRELATIONEN-Menü (*SPEERWERFEN2.SAV*)

Die zugehörige Syntax ist ziemlich kurz:

```
CORRELATIONS
      /VARIABLES=weite2016 weite2017.
```

Postwendend »wirft« Ihnen SPSS den gewünschten Korrelationskoeffizienten »aus«. Wie aus Abbildung 8.39 ersichtlich, beträgt er im »Speerwerferinnen-Beispiel« +0,942. SPSS gibt diesen

Wert ersichtlicherweise zweimal an, einmal in der Zeile von *weite2016* und einmal in der Zeile von *weite2017*. Dies liegt an der Darstellungsweise von SPSS, da eine Korrelationstabelle als Output von SPSS erzeugt wird. In dieser Korrelationstabelle werden die Korrelationen *weite2016/weite2016, weite2016/weite2017, weite2017/weite2016* und *weite2017/weite2017* angegeben. Natürlich sind die Korrelationen *weite2016/weite2016* und *weite2017/weite2017* immer gleich 1. Eigentlich ist ihr Ausweis durch SPSS nur »Schnickschnack« und nicht mehr. Und: Selbstverständlich sind auch die Werte für *weite2016/weite2017* und *weite2017/weite2016* gleich groß (in unserem Beispiel jeweils +0,942). Eigentlich würde auch hier ein Wert ausreichen. SPSS macht hier schon ganz schön auf »dicke Hose«, aber schaden tut's ja auch nicht, oder wie wir Hessen sagen: »Doppelt gemoppelt hält besser!«

**Korrelationen**

| | | weite2016 | weite2017 |
|---|---|---|---|
| weite2016 | Pearson-Korrelation | 1 | ,942* |
| | Sig. (2-seitig) | | ,017 |
| | N | 5 | 5 |
| weite2017 | Pearson-Korrelation | ,942* | 1 |
| | Sig. (2-seitig) | ,017 | |
| | N | 5 | 5 |

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

**Abbildung 8.39:** Ergebnis des Bravais/Pearson-Korrelationskoeffizienten nach Aufruf über das SPSS-KORRELATIONEN-Menü (*SPEERWERFEN2.SAV*)

Alternativ können Sie sich den Bravais/Pearson-Korrelationskoeffizienten – wie auch schon im vorstehenden Abschnitt dieses Kapitels dargelegt – im SPSS-Menü über

ANALYSIEREN|DESKRIPTIVE STATISTIKEN|KREUZTABELLEN ausweisen lassen (siehe Abbildung 8.40). Dazu müssen Sie in der Subdialogbox STATISTIKEN lediglich auf KORRELATIONEN klicken. Außerdem sollten Sie in der KREUZTABELLEN-Dialogbox bei KEINE TABELLEN ein Häkchen setzen, damit Ihnen SPSS keine (eventuell völlig außer Rand und Band ob ihrer Größe geratende) Kreuztabelle als Output »auswirft«.

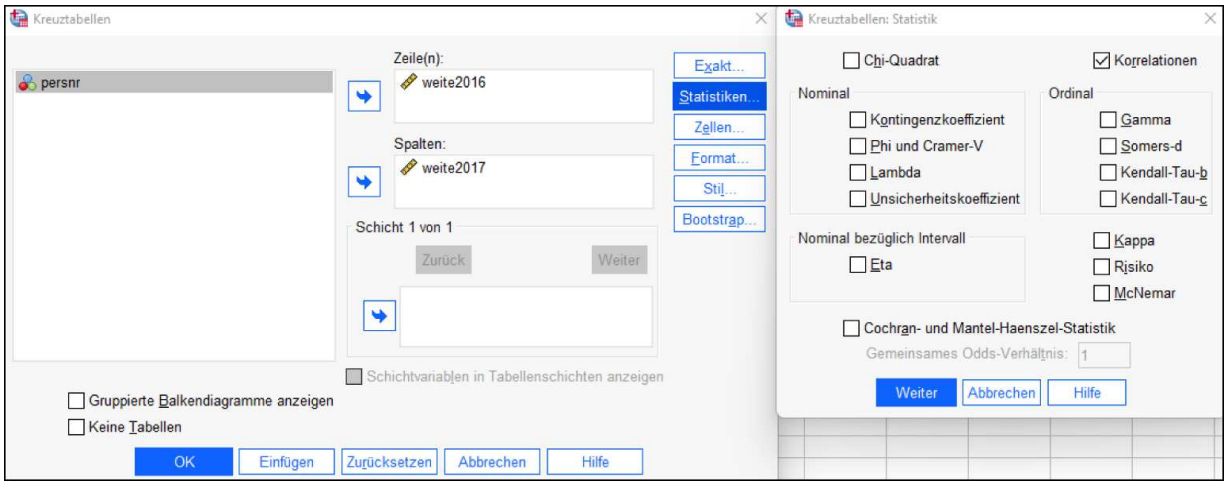



**Abbildung 8.40:** Aufruf (auch) des Bravais/Pearson-Korrelationskoeffizienten über das SPSS-KREUZTABELLEN-Menü (*SPEERWERFEN2.SAV*)

Die zugehörige SPSS-Syntax ist hier:

```
CROSSTABS
    /TABLES=weite2016 BY weite2017
    /FORMAT=NOTABLES
    /STATISTICS=CORR.
```

Auch auf diesem Wege erhalten Sie im Beispiel auf der Basis von *SPEERWERFEN2.SAV* für den Bravais/Pearson-Korrelationskoeffizienten den Wert +0,942 (siehe Abbildung 8.41). Wählen Sie den Weg über das KREUZTABELLEN-Menü, müssen Sie allerdings in Kauf nehmen, dass SPSS nicht nur den Bravais/Pearson-Korrelationskoeffizienten ausweist, sondern auch den im vorherigen Abschnitt dieses Kapitels behandelten Spearman'schen Rangkorrelationskoeffizienten. Das ist aber doch nicht weiter schlimm – oder?

**Symmetrische Maße**

| | | Wert | Asymptotischer Standardfehler [a] | Näherungsweises t [b] | Näherungsweise Signifikanz |
|---|---|---|---|---|---|
| Intervall- bzgl. Intervallmaß | Pearson-R | ,942 | ,046 | 4,867 | ,017[c] |
| Ordinal- bzgl. Ordinalmaß | Korrelation nach Spearman | ,918 | ,104 | 4,000 | ,028[c] |
| Anzahl der gültigen Fälle | | 5 | | | |

a. Die Null-Hyphothese wird nicht angenommen.

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

c. Basierend auf normaler Approximation.

**Abbildung 8.41:** Ergebnis (auch) des Bravais/Pearson-Korrelationskoeffizienten nach Aufruf über das SPSS-KREUZTABELLEN-Menü (*SPEERWERFEN2.SAV*)

Wenn Sie sich mit Korrelationskoeffizienten beschäftigen, müssen Sie immer auf das **Skalenniveau** achten: Für nominalskalierte Variablen sind keine Korrelationskoeffizienten, sondern in SPSS nur bestimmte Assoziationsmaße berechenbar, für ordinalskalierte Merkmale sollten Sie sich auf die Berechnung von Rangkorrelationskoeffizienten beschränken, und lediglich für metrische Variablen dürfen Sie den Bravais/Pearson-Korrelationskoeffizienten berechnen. Zwar weist Ihnen SPSS gegebenenfalls auch für das »falsche« Skalenniveau zum Beispiel den Bravais/Pearson-Korrelationskoeffizienten aus (etwa für eigentlich nur ordinalskalierte Variablen); in einem methodischen Sinne ist aber die Interpretation der betreffenden Werte nicht zulässig! (In der statistischen Praxis werden allerdings des Öfteren ohne schlechtes Gewissen Bravais/Pearson-Korrelationskoeffizienten für den Zusammenhang zwischen einer metrischen Variablen einerseits und einem 0/1-Dummy andererseits berechnet.)

Auch ist noch einmal darauf hinzuweisen, dass die mittels Korrelationskoeffizienten gemessenen Zusammenhänge ausschließlich statistischer Natur und inhaltlich nicht unbedingt sinnvoll sind. So wäre aus inhaltlicher Sicht der Zusammenhang zwischen der X-Variablen »Anzahl der von einer Person gekauften Quietscheentchen« und der Y-Variablen »Schuhgröße dieser

Person« doch sicherlich eine **Nonsenskorrelation** – oder haben Sie eine andere Meinung? Dann würde mich allerdings Ihre Argumentation interessieren …

Auch wichtig erscheint mir der nochmalige Hinweis, dass es bei Korrelationen nicht um **Ursache-Wirkungs-Zusammenhänge** geht. Es werden lediglich statistische Beziehungen zwischen einer X- und einer Y-Variablen berechnet, unabhängig davon, in welcher Weise eine Variable die andere genau – im Sinne von Ursache und Wirkung – beeinflusst.

Schließlich möchte ich noch kurz das Problem der **verdeckten Korrelation** beschreiben. Darunter versteht man, dass für Teilmengen der gesamten Beobachtungen durchaus statistische Zusammenhänge zwischen den Variablen X und Y bestehen können, aber nicht für die gesamte Beobachtungsmenge. Wie aus Abbildung 8.42 ersichtlich ist, besteht sowohl für die oberen eingekreisten (7) Beobachtungspunkte als auch für die unteren eingekreisten (ebenfalls 7) Beobachtungspunkte jeweils ein (leicht) positiver (linearer) statistischer Zusammenhang, nicht aber für alle (14) Beobachtungspunkte zusammen. Um solche verdeckten Korrelationen zu finden, sollten Sie – wie in Abbildung 8.42 geschehen – die einzelnen Beobachtungspunkte in ein Diagramm eintragen und dann nach derartigen Regelmäßigkeiten für Teilmengen Ausschau halten. Inhaltlich besonders interessant wird das Ganze, wenn Sie den betreffenden Teilmengen konkrete Teilgruppen zuordnen können (zum Beispiel Männer versus Frauen oder Ost- versus Westdeutschland).

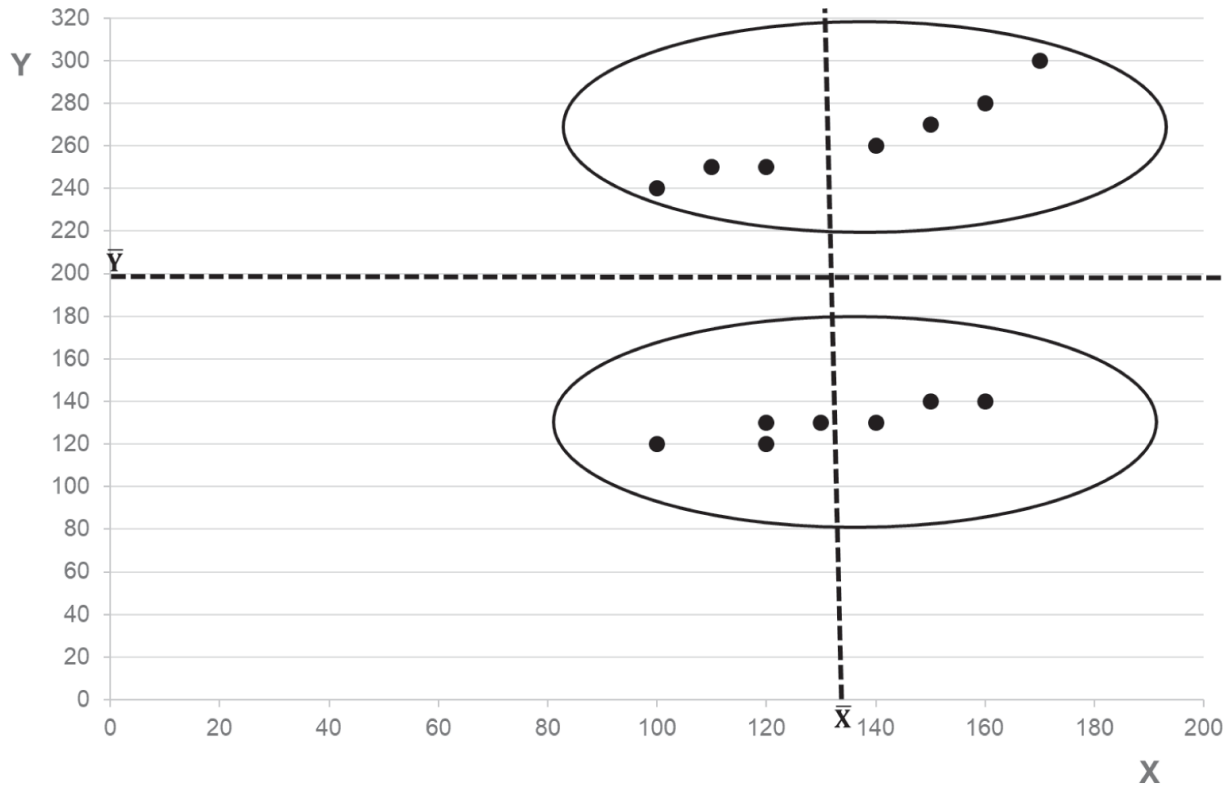



**Abbildung 8.42:** Zum Problem der verdeckten Korrelation (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 157)

# *Bootstrapping*

Mittels **Bootstrapping** können Sie für ausgewählte statistische Kennziffern bei Vorgabe eines Wahrscheinlichkeitswerts (in der Regel: 95 Prozent) Wertebereiche berechnen. Dabei zieht SPSS aus dem Datensatz eine bestimmte, von Ihnen vorgegebene Anzahl von Stichproben (im Folgenden jeweils eingestellt: 1.000 Stichproben).

Das Bootstrapping-Verfahren kann auch auf die in diesem Kapitel behandelten Assoziations- und Korrelationsmaße bezogen werden.

Lassen Sie uns mit den verschiedenen Assoziationsmaßen hier für nominalskalierte Merkmale beginnen, und zwar auf Basis der Datendatei *GESCHL_FARBE.SAV* und für den Zusammenhang zwischen den Variablen *geschlecht* und *augenfarbe*. Hierzu gehen Sie im SPSS-Menü wieder über ANALYSIEREN|DESKRIPTIVE STATISTIKEN|KREUZTABELLEN vor, wählen in der STATISTIK-

Subdialogbox auf der linken Seite die gewünschten Assoziationsmaße aus und setzen in der KREUZTABELLEN-Dialogbox bei KEINE TABELLEN ein Häkchen (siehe Abbildung 8.43).

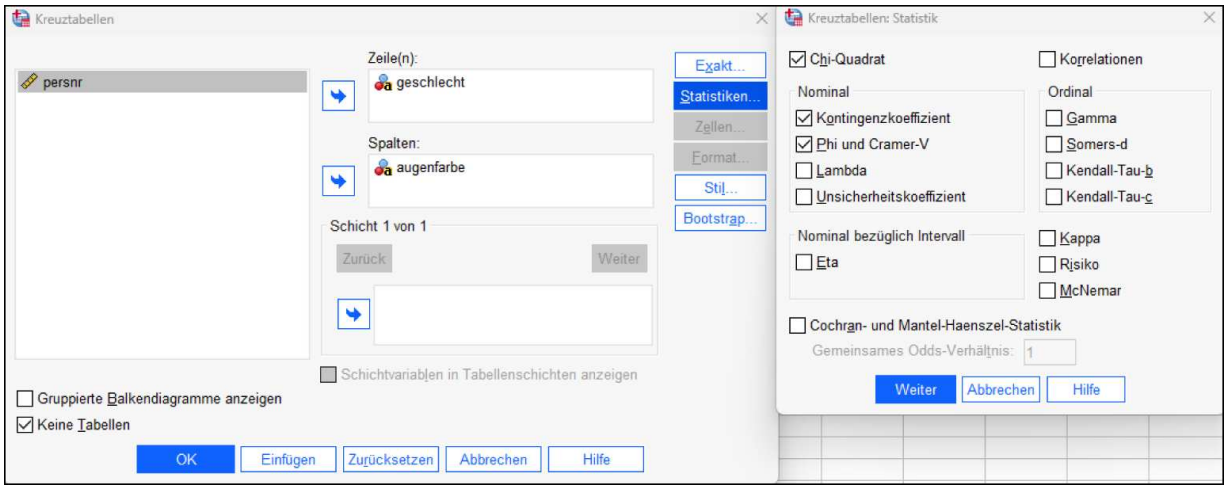



**Abbildung 8.43:** Bootstrapping bei nominalskalierten Variablen für verschiedene Assoziationsmaße (*GESCHL_FARBE.SAV*) (1)

Anschließend klicken Sie bitte in Abbildung 8.43 auf BOOTSTRAP in der KREUZTABELLEN-Subdialogbox, und es öffnet sich das in Abbildung 8.44 dargestellte BOOTSTRAP-Fenster. Dort brauchen Sie jetzt nur noch BOOTSTRAPPING DURCHFÜHREN anzuklicken (und die Anzahl der Bootstrapping-Stichproben anzugeben; hier: 1.000 Stichproben), womit Sie den Bootstrappingprozess – hier: für die ausgewählten Assoziationsmaße – in Gang setzen.

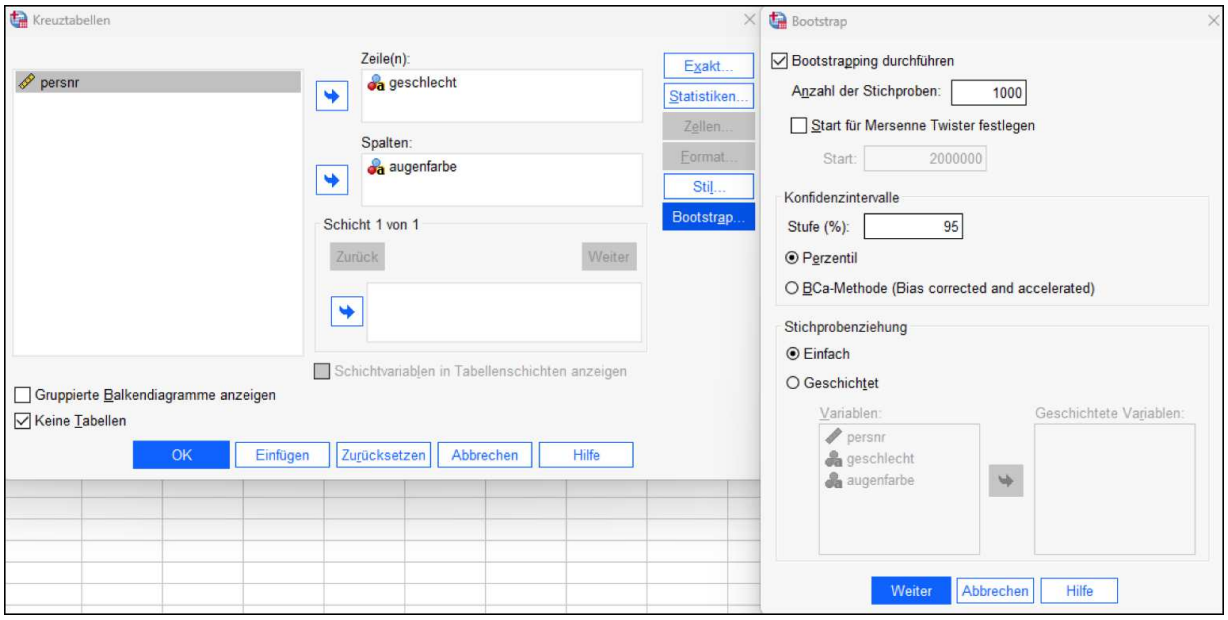



**Abbildung 8.44:** Bootstrapping bei nominalskalierten Variablen für verschiedene Assoziationsmaße (*GESCHL_FARBE.SAV*) (2)

Das ganze vorstehende Prozedere hätten Sie auch als kleines Syntaxprogramm schreiben können:

```
BOOTSTRAP
    /SAMPLING METHOD=SIMPLE
    /VARIABLES INPUT=geschlecht augenfarbe
    /CRITERIA CILEVEL=95 CITYPE=PERCENTILE NSAMPLES=1000
    /MISSING USERMISSING=EXCLUDE.
CROSSTABS
    /TABLES=geschlecht BY augenfarbe
    /FORMAT=NOTABLES
    /STATISTICS=CHISQ CC PHI
    /COUNT ROUND CELL.
```

Als Output erhalten Sie für die ausgewählten Assoziationsmaße Phi-Kontingenzkoeffizient, Cramers V und C-Kontingenzkoeffizient neben den in der Spalte »Wert« angegebenen Berechnungsergebnissen für den gesamten Datensatz auch noch in den beiden letzten Spalten die aus den Bootstrapping-Berechnungen hervorgehenden Unter- und Obergrenzen. An dieser Stelle haben meine Bootstrapping-Berechnungen gemäß Abbildung 8.45 für den Phi-Kontingenzkoeffizienten und für Cramers V jeweils als Untergrenze den Wert 0,379 und jeweils als Obergrenze den Wert 0,794 erbracht. Für den C-Kontingenzkoeffizienten habe ich hier als Untergrenze den Wert 0,354 und als Obergrenze den Wert 0,622

erhalten. (Die in Abbildung 8.45 angegebene »*Verzerrung*« bezeichnet im Übrigen die Differenz zwischen dem Durchschnittswert der in der entsprechenden Zeile angegebenen statistischen Kennziffer in allen Bootstrap-Stichproben einerseits und dem Wert in der Spalte »Wert« andererseits.)

**Bootstrap für Symmetrische Maße**

| | | | Bootstrap[a] | | | |
|---|---|---|---|---|---|---|
| | | | | | 95% Konfidenzintervall | |
| | | Wert | Verzerrung | Std.-Fehler | Unterer Wert | Oberer Wert |
| Nominal- bzgl. Nominalmaß | Phi | ,516 | ,076 | ,106 | ,379 | ,794 |
| | Cramer-V | ,516 | ,076 | ,106 | ,379 | ,794 |
| | Kontingenzkoeffizient | ,459 | ,046 | ,069 | ,354 | ,622 |
| Anzahl der gültigen Fälle | | 25 | 0 | 0 | 25 | 25 |

a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

**Abbildung 8.45:** Bootstrapping-Output bei nominalskalierten Variablen für verschiedene Assoziationsmaße (*GESCHL_FARBE.SAV*)

Bootstrapping-Berechnungen sind auch für mindestens ordinalskalierte Korrelationsbetrachtungen möglich. Hierzu müssen Sie – im »Speerwerferinnen-Beispiel« auf Basis von *SPEERWERFEN2.SAV* – in der KREUZTABELLEN-Subdialogbox die gewünschten Koeffizienten (rechts) auswählen (siehe hierzu Abbildung 8.46) und anschließend in der KREUZTABELLEN-Hauptdialogbox rechts auf BOOTSTRAP klicken und dort schließlich das Bootstrapping aktivieren.

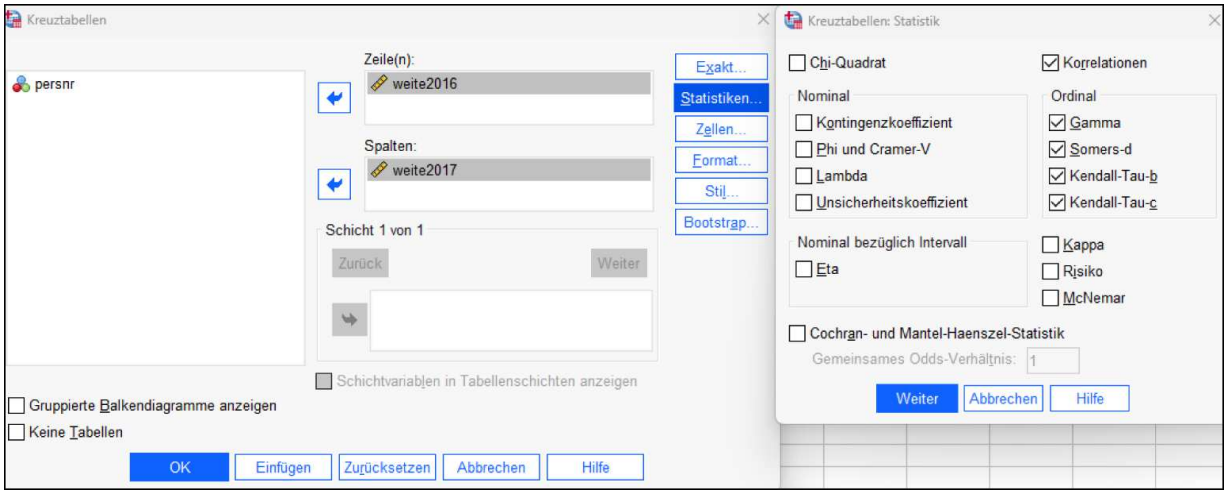



**Abbildung 8.46:** Bootstrapping bei mindestens ordinalskalierten Variablen für verschiedene Korrelationsmaße – KREUZTABELLEN-Menü (*SPEERWERFEN2.SAV*)

Sie hätten das alles aber auch als kleines SPSS-Syntaxprogramm schreiben können. Dieses Syntaxprogramm unterscheidet sich von dem obigen für die Assoziationsmaße geschriebenen im Grunde genommen nur in der Zeile /STATISTICS:

```
BOOTSTRAP
    /SAMPLING METHOD=SIMPLE
    /VARIABLES INPUT=weite2016 weite2017
    /CRITERIA CILEVEL=95 CITYPE=PERCENTILE NSAMPLES=1000
    /MISSING USERMISSING=EXCLUDE.
CROSSTABS
    /TABLES=weite2016 BY weite2017
    /FORMAT=NOTABLES
    /STATISTICS=CORR GAMMA D BTAU CTAU
    /COUNT ROUND CELL.
```

Die betreffenden Bootstrapping-Ergebnisse, die ich erhalten habe, sind in Abbildung 8.47 zusammengestellt. Aus den beiden letzten Spalten der dortigen Pivottabelle sind die jeweiligen Unter- und Obergrenzen ersichtlich, wie sie sich aus dem Bootstrapping ergeben haben.

Nicht verschwiegen werden soll, dass das KREUZTABELLEN-Bootstrapping für unser konkretes Beispiel mit lediglich 5 Fällen für *werte2016* und *werte2017* bei der Bootstrapping-Stichprobenziehung zu ein paar Warnungen und zu »lediglich« 918 Stichproben führt, was allerdings nicht dazu führt, dass das Bootstrapping abgebrochen wird. Einige der betreffenden, letztlich aber harmlosen Warnungen in unserem Beispiel sind im linken Teil von Abbildung 8.47 angegeben.

**Warnungen**

| |
|---|
| Für die Kreuztabelle von weite2016 * weite2017 für die aufgeteilte Datei $bootstrap_split=13 wurden keine Zusammenhangsmaße berechnet. Mindestens eine der Variablen in jeder Zweifach-Tabelle, anhand derer Zusammenhangsmaße berechnet werden, ist eine Konstante. |
| Für die Kreuztabelle von weite2016 * weite2017 für die aufgeteilte Datei $bootstrap_split=22 wurden keine Zusammenhangsmaße berechnet. Mindestens eine der Variablen in jeder Zweifach-Tabelle, anhand derer Zusammenhangsmaße berechnet werden, ist eine Konstante. |
| Für die Kreuztabelle von weite2016 * weite2017 für die aufgeteilte Datei $bootstrap_split=24 wurden keine Zusammenhangsmaße berechnet. Mindestens eine der Variablen in jeder Zweifach-Tabelle, anhand derer Zusammenhangsmaße berechnet werden, ist eine Konstante. |
| Für die Kreuztabelle von weite2016 * weite2017 für die aufgeteilte Datei $bootstrap_split=36 wurden keine Zusammenhangsmaße berechnet. Mindestens eine der Variablen in jeder Zweifach-Tabelle, anhand derer Zusammenhangsmaße berechnet werden, ist eine Konstante. |
| Für die Kreuztabelle von weite2016 * weite2017 für die aufgeteilte Datei $bootstrap_split=39 wurden keine Zusammenhangsmaße berechnet. Mindestens eine der Variablen in jeder Zweifach-Tabelle, anhand derer Zusammenhangsmaße berechnet werden, ist eine Konstante. |
| Für die Kreuztabelle von weite2016 * weite2017 für die aufgeteilte Datei $bootstrap_split=42 wurden keine Zusammenhangsmaße berechnet. Mindestens eine der Variablen in jeder Zweifach-Tabelle, anhand derer Zusammenhangsmaße berechnet werden, ist eine Konstante. |

**Bootstrap für Symmetrische Maße**

| | | | Bootstrap[a] | | | |
|---|---|---|---|---|---|---|
| | | | | | 95% Konfidenzintervall | |
| | | Wert | Verzerrung | Std.-Fehler | Unterer Wert | Oberer Wert |
| Ordinal- bzgl. Ordinalmaß | Kendall-Tau-b | ,882 | ,024[b] | ,102[b] | ,707[b] | 1,000[b] |
| | Kendall-Tau-c | ,840 | -,003[b] | ,136[b] | ,640[b] | ,960[b] |
| | Gamma | 1,000 | ,000[b] | ,000[b] | 1,000[b] | 1,000[b] |
| | Korrelation nach Spearman | ,918 | ,010[b] | ,087[b] | ,745[b] | 1,000[b] |
| Intervall- bzgl. Intervallmaß | Pearson-R | ,942 | ,000[b] | ,057[b] | ,810[b] | 1,000[b] |
| Anzahl der gültigen Fälle | | 5 | 0 | 0 | 5 | 5 |

a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

b. Basierend auf 918 Stichproben

**Abbildung 8.47:** Bootstrapping-Output bei mindestens ordinalskalierten Variablen für verschiedene Korrelationsmaße – KREUZTABELLEN-Menü (*SPEERWERFEN2.SAV*) (Auszug)

Ein Bootstrapping kann man für mindestens ordinalskalierte Variablen im Korrelationszusammenhang auch über das Menü ANALYSIEREN|KORRELATIONEN|BIVARIAT durchführen. Hierzu wählen Sie in der Subdialogbox BIVARIATE KORRELATIONEN – wieder für *weite2016* und *weite2017* aus dem Datensatz *SPEERWERFEN2.SAV* – die gewünschten Korrelationskoeffizienten aus und klicken anschließend auf BOOTSTRAP (siehe hierzu Abbildung 8.48), wo Sie das Bootstrapping wieder aktivieren können.

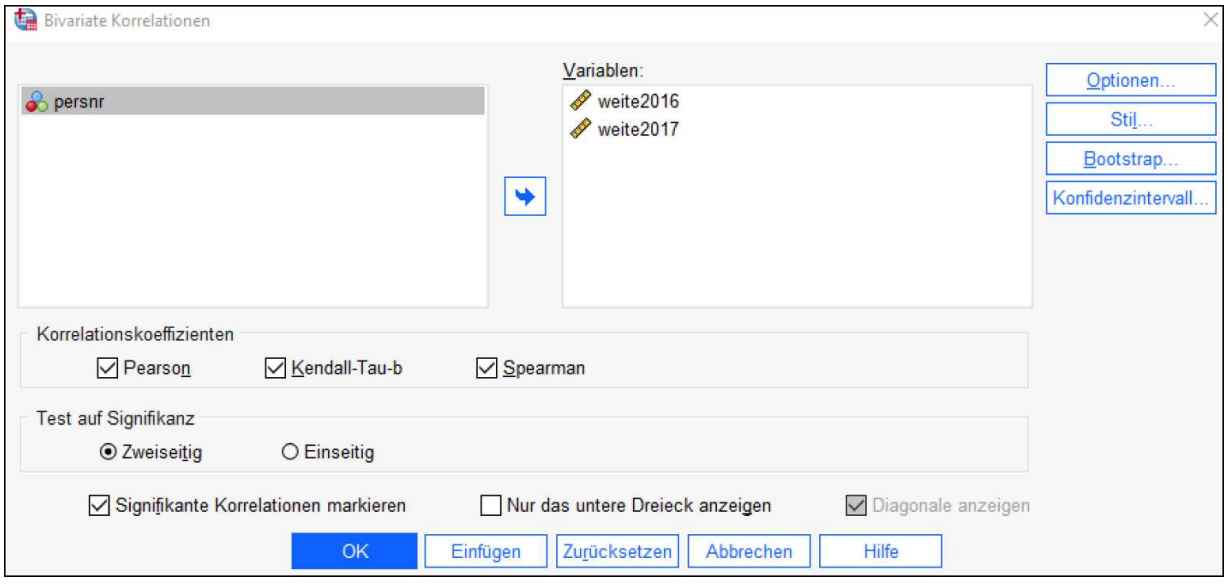


**Abbildung 8.48:** Bootstrapping bei mindestens ordinalskalierten Variablen für verschiedene Korrelationsmaße – KORRELATION-Menü (*SPEERWERFEN2.SAV*)

In diesem Fall sieht Ihre Syntax wie nachfolgend aus. Bitte beachten Sie dabei, dass Sie den BOOTSTRAP-Befehl zweimal eingeben müssen: zum einen vor dem CORRELATIONS-Befehl (für den Bravais/Pearson-Korrelationskoeffizienten) und zum anderen vor dem NONPAR-CORR-Befehl (für den Spearman'schen Rangkorrelationskoeffizienten beziehungsweise für Kendalls Tau-b-Koeffizienten):

```
BOOTSTRAP
    /SAMPLING METHOD=SIMPLE
    /VARIABLES INPUT=weite2016 weite2017
    /CRITERIA CILEVEL=95 CITYPE=PERCENTILE NSAMPLES=1000
    /MISSING USERMISSING=EXCLUDE.
CORRELATIONS
    /VARIABLES=weite2016 weite2017
    /PRINT=TWOTAIL NOSIG
    /MISSING=PAIRWISE.
BOOTSTRAP
    /SAMPLING METHOD=SIMPLE
    /VARIABLES INPUT=weite2016 weite2017
    /CRITERIA CILEVEL=95 CITYPE=PERCENTILE NSAMPLES=1000
    /MISSING USERMISSING=EXCLUDE.
NONPAR CORR
    /VARIABLES=weite2016 weite2017
    /PRINT=BOTH TWOTAIL NOSIG
    /MISSING=PAIRWISE.
```

Mit der vorstehenden Syntax erhalten Sie zwei Bootstrapping-Tabellen als Output. Der linke Teil von Abbildung 8.49 bezieht sich dabei auf den Bravais/Pearson-Korrelationskoeffizienten. Der Bootstrapping-Wertebereich ist hier sowohl bei *werte2016* als auch *werte2017* ausgewiesen (Sie wissen schon: »Doppelt gemoppelt hält besser!«), und zwar jeweils in den Zeilen »Unterer Wert« und »Oberer Wert«. Im vorliegenden Beispiel liegt der entsprechende Wertebereich für den Bravais/Pearson-Korrelationskoeffizienten zwischen 0,810 und 1,000.

Die SPSS-Darstellungsweise ist im rechten Teil von Abbildung 8.49 für den Spearman'schen Rangkorrelationskoeffizienten (von SPSS als »Spearman-Rho« bezeichnet, aber das ist nur eine andere Bezeichnung!) und für Kendalls Tau-b-Koeffizienten ganz genauso. Es ergibt sich im

Beispiel für den Spearman'schen Rangkorrelationskoeffizienten ein Bootstrapping-Wertebereich von 0,745 und 1,000 sowie für Kendalls Tau-b-Koeffizienten ein solcher zwischen 0,707 und 1,000.

**Korrelationen**

| | | | | weite2016 | weite2017 |
|---|---|---|---|---|---|
| weite2016 | Pearson-Korrelation | | | 1 | ,942* |
| | Sig. (2-seitig) | | | | ,017 |
| | N | | | 5 | 5 |
| | Bootstrap[d] | Verzerrung | | 0[e] | ,000[f] |
| | | Std.-Fehler | | 0[e] | ,055[f] |
| | | 95% Konfidenzintervall | Unterer | 1[e] | ,810[f] |
| | | | Oberer | 1[e] | 1,000[f] |
| weite2017 | Pearson-Korrelation | | | ,942* | 1 |
| | Sig. (2-seitig) | | | ,017 | |
| | N | | | 5 | 5 |
| | Bootstrap[d] | Verzerrung | | ,000[f] | 0[f] |
| | | Std.-Fehler | | ,055[f] | 0[f] |
| | | 95% Konfidenzintervall | Unterer | ,810[f] | 1[f] |
| | | | Oberer | 1,000[f] | 1[f] |

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

d. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

e. Basierend auf 991 Stichproben

f. Basierend auf 921 Stichproben

**Korrelationen**

| | | | | weite2016 | weite2017 |
|---|---|---|---|---|---|
| Kendall-Tau-b | weite2016 | Korrelationskoeffizient | | 1,000 | ,882* |
| | | Sig. (2-seitig) | | . | ,046 |
| | | N | | 5 | 5 |
| | | Bootstrap[c] Verzerrung | | ,000[d] | ,024[e] |
| | | Std.-Fehler | | ,000[d] | ,102[e] |
| | | 95% Konfidenzintervall | Unterer | 1,000[d] | ,707[e] |
| | | | Oberer | 1,000[d] | 1,000[e] |
| | weite2017 | Korrelationskoeffizient | | ,882* | 1,000 |
| | | Sig. (2-seitig) | | ,046 | . |
| | | N | | 5 | 5 |
| | | Bootstrap[c] Verzerrung | | ,024[e] | ,000[e] |
| | | Std.-Fehler | | ,102[e] | ,000[e] |
| | | 95% Konfidenzintervall | Unterer | ,707[e] | 1,000[e] |
| | | | Oberer | 1,000[e] | 1,000[e] |
| Spearman-Rho | weite2016 | Korrelationskoeffizient | | 1,000 | ,918* |
| | | Sig. (2-seitig) | | . | ,028 |
| | | N | | 5 | 5 |
| | | Bootstrap[c] Verzerrung | | ,000[d] | ,010[e] |
| | | Std.-Fehler | | ,000[d] | ,088[e] |
| | | 95% Konfidenzintervall | Unterer | 1,000[d] | ,745[e] |
| | | | Oberer | 1,000[d] | 1,000[e] |
| | weite2017 | Korrelationskoeffizient | | ,918* | 1,000 |
| | | Sig. (2-seitig) | | ,028 | . |
| | | N | | 5 | 5 |
| | | Bootstrap[c] Verzerrung | | ,010[e] | ,000[e] |
| | | Std.-Fehler | | ,088[e] | ,000[e] |
| | | 95% Konfidenzintervall | Unterer | ,745[e] | 1,000[e] |
| | | | Oberer | 1,000[e] | 1,000[e] |

*. Die Korrelation ist auf dem 0,05 Niveau signifikant (zweiseitig).

c. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

d. Basierend auf 989 Stichproben

e. Basierend auf 914 Stichproben

**Abbildung 8.49:** Bootstrapping-Output bei mindestens ordinalskalierten Variablen für verschiedene Korrelationsmaße – KORRELATION-Menü (*SPEERWERFEN2.SAV*)

# Kapitel 9

# Über Abhängigkeiten: Simple Regressionen

## IN DIESEM KAPITEL

Zum Unterschied zwischen Korrelation und Regression

Regressionsarten

Lineare Einfachregression

Lineare Mehrfachregression

Dieses Kapitel bezieht sich auf – von ihrer Grundkonzeption her – lineare Regressionsverfahren. Liegt nur eine erklärende Variable vor, spricht man von einer linearen Einfachregression. Bei mindestens zwei erklärenden Variablen haben Sie es mit einer linearen Mehrfachregression zu tun. Wichtig ist jeweils, dass die abhängige Variable eine metrisch skalierte Variable ist.

## *Zum Unterschied zwischen Korrelation und Regression*

Existiert zwischen zwei Variablen ein statistischer Zusammenhang, spricht man bei ordinalen beziehungsweise metrischen Variablen, wie in Kapitel 8 näher ausgeführt, von einer **Korrelation** zwischen den betreffenden Variablen. Dabei wird zunächst kein *Ursache-Wirkungs-Zusammenhang* formuliert.

Erst wenn ein solcher im Sinne einer abhängigen Variablen Y und einer unabhängigen Variablen X aufgestellt beziehungsweise eine mathematische Funktion zwischen den betreffenden Variablen formuliert wird, wird aus einer Korrelation eine **Regression**. Hierbei sollten sachlogische, theoretisch fundierte Überlegungen bezüglich der entsprechenden Funktion angestellt werden. Diese betreffen zum einen die Wahl der genauen mathematischen Funktionsform und zum anderen naheliegenderweise die Wahl der Einflussrichtung zwischen beiden Variablen.

Die einfache lineare Regression stellt hierbei einen linearen Ursache-Wirkungs-Zusammenhang zwischen zwei Variablen dar. Man spricht auch von einer *bivariaten Regression*. Die einfache lineare Regression dient der Ursachen- und der Wirkungsanalyse sowie der Prognose. Mitunter werden die abhängige Variable auch

»Kriteriumsvariable« und die unabhängige Variable »Prädiktorvariable« genannt. Bei mehr als einer unabhängigen Variablen handelt es sich um eine lineare Mehrfachregression.

Die in der Regressionsanalyse verwendeten Variablen sollten unbedingt in einer inhaltlichen Beziehung zueinander stehen. Ansonsten kann das Phänomen der *Nonsenskorrelation* auftreten. Mit anderen Worten: Man würde Variablen zueinander in Beziehung setzen, die wenig beziehungsweise gar nichts miteinander zu tun haben.

Gerade wenn Variablen über die Zeit hinweg erfasst werden, weisen sie häufig einen **Trend** auf. Für zwei Variablen X und Y kann man allein aufgrund dieser Tatsache immer eine (positive oder negative) statistische Korrelation ausweisen, und zwar völlig unabhängig davon, ob diese beiden Variablen in einer inhaltlichen Beziehung zueinander stehen oder nicht.

Angenommen, Sie hätten als Mitglied einer Amateur-Tischtennismannschaft in den vergangenen 20 Wochen keines Ihrer 20 Einzel verloren. Zeitgleich – das heißt: ebenfalls in den vergangenen 20 Wochen – soll die führende Volleyballmannschaft der Fidschi-Inseln ebenfalls alle ihre 20 Spiele gewonnen haben. Beide Zeitreihen weisen also einen positiven Trend auf. Statistisch ausgedrückt: Die beiden Variablen »Ihre Tischtennis-Einzel-Ergebnisse« und »Ergebnisse der führenden Volleyballmannschaft der Fidschi-Inseln« (in den vergangenen 20 Wochen) korrelieren positiv miteinander. Inhaltlich wird man allerdings ja wohl nicht ernsthaft davon ausgehen können, dass Ihre Tischtennissiegesserie etwas mit der Siegesserie des Volleyballteams auf den Fidschi-Inseln zu tun hat. Weder wird man also ernsthaft davon ausgehen können, dass Ihre Tischtenniskünste die Siegesserie der fidschianischen Volleyballmannschaft hervorgerufen haben, noch haben umgekehrt höchstwahrscheinlich die Siege des Teams von den Fidschi-Inseln Ihre Tischtenniserfolge bewirkt. Es handelt sich hier demnach um einen **Nonsenszusammenhang**!

Außerdem kommt bei der Regressionsanalyse üblicherweise erschwerend hinzu, dass man die Richtung des Ursache-Wirkungs-Zusammenhangs angeben muss. Es stellt sich im Vorhinein die Frage: Was ist Ursache, was ist Wirkung? Dies ist gerade im sozial- beziehungsweise wirtschaftswissenschaftlichen Bereich nicht einfach zu beantworten, weil dort die Variablen in der Regel in einem wechselseitigen Verhältnis zueinander stehen. Man nennt dies mit einem Fachbegriff einen *interdependenten Zusammenhang*. Beispielsweise kann man entweder der Auffassung sein, dass Armut krankmacht, oder postulieren, dass Kranksein arm macht. Im erstgenannten Fall ist »Armut« die unabhängige X-Variable und »Krankheit« die abhängige Y-Variable, im zweitgenannten Fall ist es umgekehrt: »Krankheit« ist im Sinne der Ursache die unabhängige X-Variable und »Armut« im Sinne der Wirkung die abhängige Y-Variable. Will man kein

kompliziertes Mehrgleichungsmodell mit gegenseitigen Abhängigkeiten aufstellen und stattdessen in einer einzelnen Gleichung einen X/Y-Zusammenhang formulieren, muss man sich also für eine Einflussrichtung entscheiden.

Die Ursache-Wirkungs-Thematik ist bei **Prognosen** weniger problematisch. Dies liegt daran, dass man hier üblicherweise an der zeitlichen Entwicklung einer bestimmten Variablen interessiert ist. Für diese Größe nutzt man eine oder mehrere X-Größen, die in der Vergangenheit die Entwicklung der Y-Größe möglicherweise auch nur im Sinne eines rein statistischen Zusammenhangs geprägt haben. Als X-Größe kann in diesem Zusammenhang auch allein der Faktor **Zeit** genutzt werden. Man schätzt dann mithilfe der Regressionsanalyse den zeitlichen Trend der Y-Variablen und nutzt diesen Trend dann für Prognosen.

# *Regressionsarten*

Wenn lediglich von einer X- und einer Y-Variablen die Rede ist, spricht man – wie bereits erwähnt – von einer **Einfachregression**. Verwendet man eine Funktion mit mehreren X-Variablen $Y = f(X_1, X_2, \ldots, X_K)$, handelt es sich – wie ebenfalls schon dargelegt – um eine **Mehrfachregression**, die man mitunter auch als »Multiple Regression« bezeichnet.

Je nach Skalenniveau der betrachteten Variablen und nach funktionaler Ausgestaltung können Sie vor allem folgende Regressionen voneinander unterscheiden:

- ✔ klassische lineare Regression,
- ✔ sogenannte Kurvenanpassungen (= »Linearisierungen«) bei der linearen Regression (mit 11 Regressionsmodellen in SPSS),
- ✔ Gewichtungsschätzung (WLS = Weighted Least Squares),
- ✔ Binär-logistische Regression,
- ✔ Multinomial-logistische Regression,
- ✔ Ordinale Regression,
- ✔ Probit-Regression,
- ✔ Nichtlineare Regression,
- ✔ Instrumentalvariablenmethode,
- ✔ 2-SLS-Verfahren (2 SLS = 2-Stage Least Squares) beziehungsweise 3-SLS-Verfahren (3 SLS = 3-Stage Least Squares) und
- ✔ Regression mit optimaler Skalierung.

Gegenstand dieses in die Regressionsanalyse einführenden Kapitels 9 sind die drei erstgenannten Regressionsvarianten. Die anderen Regressionsarten werden – zumindest

zum Teil – im weiterführenden, folgenden Kapitel 10 vorgestellt.

# Lineare Einfachregression

Eine einfache Möglichkeit, einen Ursache-Wirkungs-Zusammenhang zwischen zwei Variablen X und Y darzustellen, ist die **lineare Einfachregression**. Mit X wird dabei die *unabhängige Variable* (die erklärende »Ursache«) bezeichnet. Y symbolisiert die *abhängige Variable* (die zu erklärende »Wirkung«). Zwischen beiden Variablen wird im einfachen linearen Regressionsmodell folgende Geradengleichung unterstellt:

$$\hat{Y}_i = \hat{a} + \hat{b} \cdot X_i$$

Die Zeichen (Zirkumflexe) über $Y_i$, a und b weisen darauf hin, dass es sich um zu schätzende Größen handelt. In dieser Gleichung sind $\hat{a}$ und $\hat{b}$ zwei *Parameter*, die auf der Grundlage des Datenmaterials geschätzt werden. Dabei ist $\hat{b}$ die Steigung des linearen Zusammenhangs zwischen X und $\hat{Y}$, und $\hat{a}$ steht für den Ordinatenschnittpunkt, also für das absolute Glied der Gleichung. Die obige Gleichung ist deswegen linear, weil sie eine Gerade repräsentiert.

# Annahmen

Die üblichen **Voraussetzungen** der einfachen linearen Regression sind:

- ✔ typischerweise metrische Skalierung der abhängigen und der unabhängigen Variablen (aber auch 0/1-Dummies als unabhängige Variablen möglich),
- ✔ linearer Zusammenhang,
- ✔ lineare Koeffizienten,
- ✔ Zufallsstichprobe (wenn sich die Regression nicht auf die Grundgesamtheit bezieht),
- ✔ bedingter Erwartungswert des Fehlers der unabhängigen Variablen gleich 0,
- ✔ Nichtkonstanz der Ausprägungen der unabhängigen und auch der abhängigen Variablen,
- ✔ Homoskedastizität,
- ✔ unabhängige Fehlerwerte,
- ✔ (zumindest näherungsweise) Normalverteilung der Fehlerwerte (bei induktiver Anwendung der Regressionsanalyse).

## Skalierung

Grundsätzlich sollen im Rahmen der hier behandelten linearen Regressionsanalysen also sowohl die unabhängigen X-Variablen als auch die abhängige Y-Variable **metrisch skaliert** sein. Allerdings gibt es die Möglichkeit, auch nominal- beziehungsweise ordinalskalierte Variablen als unabhängige X-Variablen in die Analyse einzubeziehen. Nominalskalierte Merkmale werden in Form von **0/1-Dummies** berücksichtigt. So würde beispielsweise eine Variable *geschlecht* die Ausprägungen 0 (etwa für »weiblich«) und 1 (etwa für »männlich«) erhalten. Näheres zu einer entsprechenden Berücksichtigung nominalskalierter Merkmale folgt etwas weiter unten in diesem Kapitel.

Auch ordinalskalierte Variablen können als Dummy-Variablen berücksichtigt werden. Wollten Sie etwa eine Variable *zufriedenheit* mit den Ausprägungen »hohe Zufriedenheit«, »mittelmäßige Zufriedenheit« und »niedrige Zufriedenheit« als unabhängigen Einflussfaktor bei Ihren Regressionsberechnungen berücksichtigen, können Sie dies über zwei Dummy-Variablen *zufrieden_hoch* mit den Ausprägungen 0 für »keine hohe Zufriedenheit« und 1 für »hohe Zufriedenheit« sowie *zufrieden_mittel* mit 0 für »keine mittelmäßige Zufriedenheit« und 1 für »mittelmäßige Zufriedenheit« bewerkstelligen. Eine »niedrige Zufriedenheit« bildet hierbei jeweils die Bezugskategorie (bei *zufrieden_hoch* zusammen mit »mittelmäßiger Zufriedenheit« und bei *zufrieden_mittel* zusammen mit »hoher Zufriedenheit«). Wenn also sowohl *zufrieden_hoch* als auch *zufrieden_mittel* gleich 0 sind, ist die »niedrige Zufriedenheit« in der Regression angesprochen. Näheres zur Formulierung ordinalskalierter Merkmale als Dummy-Variablen im Rahmen der einfachen linearen Regression folgt etwas weiter unten in diesem Kapitel.

Liegen im Übrigen verhältnismäßig viele Ausprägungen von ordinalskalierten Variablen vor, können Sie sozusagen mit einem zugedrückten Auge diese Variablen (näherungsweise) als metrisch skaliert auffassen und somit sowohl als X- als auch als Y-Größen in einer linearen Regression berücksichtigen. Dies ist zum Beispiel bei einer Zufriedenheitsvariablen der Fall, wenn diese auf einer Skala von 0 (»völlig unzufrieden«) bis 10 (»völlig zufrieden«) skaliert ist. Hierbei müssen Sie aber die Annahme treffen, dass die Abstände zwischen den einzelnen Ausprägungen gleich groß sind, was ja bei einer ordinalskalierten Variablen streng genommen nicht in Ordnung ist (deswegen der Hinweis auf das »zugedrückte Auge«).

## *Linearitäten*

Bei der einfachen linearen Einfachregression liegt – Nomen est omen! – ein *linearer Zusammenhang* mit *linearen Koeffizienten* vor. Sie schätzen also bei der einfachen linearen Regression schlicht und ergreifend eine Gerade.

## *Zufallsstichprobe*

Von den obigen üblichen Annahmen einer einfachen linearen Regression ist die Annahme der **Zufallsstichprobe** besonders wichtig. Da Sie in vielen Fällen eine Regressionsbeziehung aus einer Stichprobe heraus schätzen und anschließend Ihre

Schätzung auf die Grundgesamtheit verallgemeinern möchten, muss ausgeschlossen sein, dass diese Stichprobe irgendwelche systematischen Einflüsse bei der Auswahl der ausgewählten Untersuchungseinheiten aufweist. Möchten Sie beispielsweise auf das Ausgabeverhalten in Deutschland schließen, sollten sinnvollerweise ihre Stichprobenelemente aus ganz Deutschland stammen – und nicht etwa nur aus Bayern. Aufgrund bestimmter regionaler Besonderheiten in Bayern etwa bezüglich des Bier- oder Weißwurstkonsums könnte sich an dieser Stelle ein verzerrter Schluss auf das Konsumverhalten in Deutschland insgesamt ergeben. Eine solche Verzerrung vermeiden Sie durch eine reine Zufallsauswahl (hier bezogen auf Deutschland insgesamt).

Behelfsweise können Sie durch eine Gewichtung Ihrer Fälle mittels – auf Ihre Grundgesamtheit bezogener – Hochrechnungsfaktoren Repräsentativität herstellen. Dabei müssen Sie aber Ihre Hochrechnungsfaktoren in einer Regression solcherart normieren, dass sie in der Summe Ihre Stichprobengröße ergeben. Nehmen Sie einfach mal an, Sie hätten in Ihrer Stichprobe 6 Männer und 4 Frauen, wüssten aber, dass in der Grundgesamtheit das Geschlechterverhältnis bei 1:1 liegt. Dann liegt es nahe, jedem Mann einen Gewichtungsfaktor von $\frac{5}{6} = 0{,}8\overline{3}$ und jeder Frau einen Gewichtungsfaktor von $\frac{5}{4} = 1{,}25$ zuzuweisen. Auf diese Weise werden die Geschlechteranzahlen auf 5 Männer und 5 Frauen umgewichtet; in der Summe bleiben Sie aber bei den 10 Fällen.

## Erwartungswertannahme

Die obige weitere Annahme eines **Erwartungswerts** der Abweichung des berechneten Y-Werts vom beobachteten Y-Wert bedeutet einfach nur, dass Ihre Regressionsschätzungen im Mittel mit den »wahren« Werten übereinstimmen. Dies heißt letztlich, dass eine lineare Einfachregression durch die arithmetischen Mittelwerte von X und Y verläuft. Man spricht davon, dass eine lineare Regression dort ihren **Schwerpunkt** hat.

Die vorstehende Beziehung bedeutet zugleich, dass eine Kleinst-Quadrate-Schätzung wegen der Ausreißer-Empfindlichkeit des arithmetischen Mittelwerts ebenfalls stark von Ausreißerwerten beeinflusst wird. Das sollten Sie bei der Interpretation Ihrer geschätzten Regressionsgeraden unbedingt im Hinterkopf behalten. Sie werden bald ein Beispiel hierzu kennenlernen.

## Nichtkonstanz der X- und der Y-Variablen

Dass die *X-Werte nicht konstant* sein dürfen, ist naheliegend. Ansonsten könnten Sie nicht ansatzweise eine Ursache-Wirkungs-Beziehung herstellen. Wenn nämlich die

unabhängigen X-Werte konstant wären bei variierenden Y-Werten, würde dies ja bedeuten, dass zum Beispiel ein Einkommen von jeweils 1.000 Euro als X-Wert mit Ausgaben in Höhe von 100 Euro, 200 Euro, 400 Euro und so weiter als möglichen Y-Werten gekoppelt wäre. Ein solcher X/Y-Zusammenhang erscheint zu Recht als mehr oder weniger willkürlich und darf deshalb in diesem Fall nicht als Regression geschätzt werden! Ähnlich können Sie bei einer Konstanz aller Y-Werte argumentieren.

### *Homoskedastizität*

Die drei letztgenannten obigen Annahmen klingen recht technisch und sind es im Grunde genommen auch. Mit **Homoskedastizität** ist gemeint, dass die Varianz des nicht durch die Regressionsgleichung erklärten Y-Bestandteils sich gleichmäßig um die Regressionsgerade verteilt. Es ergibt sich also bei niedrigeren X-Werten eine (in etwa) gleich hohe Streuung der Beobachtungswerte um die Regressionsgerade herum wie bei höheren X-Werten. Ist die Annahme der Homoskedastizität nicht gegeben, liegt sogenannte **Heteroskedastizität** vor.

### *Autokorrelation*

Die obige vorletzte Annahme lautet **unabhängige Fehlerwerte**. Dies kennzeichnet den Fall, dass die durch die Regression nicht erklärten Y-Bestandteile nicht voneinander abhängen. Diese Annahme ist vielfach bei Zeitreihenwerten verletzt, weil häufig über die Zeit hinweg historische Entwicklungen auf die Gegenwart »durchschlagen«. Mit einem Fachbegriff spricht man dann von **Autokorrelation**.

### *Normalverteilungsannahme*

Die obige letzte Annahme fordert eine zumindest näherungsweise **Normalverteilung** der Fehlerwerte – also der nicht erklärten Y-Bestandteile. Diese Annahme ist für die Beurteilung der geschätzten Parameterwerte der Regressionsgeraden bedeutsam, das heißt: für eine Aussage, ob die betreffenden Parameter statistisch aussagekräftig sind (technisch ausgedrückt: ob sie **signifikant** sind). Für die eigentliche Schätzung der Regressionsparameter ist die Normalverteilungsannahme hingegen nicht einschränkend.

Auf die drei letztgenannten Aspekte der Homo-/Heteroskedastizität, der Autokorrelation und der Signifikanz werde ich weiter unten in diesem Kapitel noch etwas näher eingehen.

## *Die Schätzmethode der Kleinst-Quadrate-Methode*

Ausgangspunkt einer jeden Regressionsschätzung sind vorliegende Werte für die X- und die Y-Variable, das heißt: **Beobachtungspunkte**. Deren Zusammenstellung in

einem X/Y-Diagramm bildet eine *Punktewolke* (beziehungsweise mit einem anderen Begriff: ein *Streudiagramm* oder auf Englisch: *Scatterplot*).

Nehmen Sie folgende Punktepaare an:

| $X_i$ | $Y_i$ |
|---|---|
| 0 | 10 |
| 10 | 30 |
| 20 | 40 |
| 30 | 50 |
| 40 | 80 |
| 50 | 70 |
| 60 | 60 |
| 70 | 100 |
| 80 | 120 |
| 90 | 100 |
| 100 | 130 |

**Tabelle 9.1:** X/Y-Punktepaare als Basis einer Regressionsrechnung

Durch die betreffende Punktewolke wird nun eine mathematische Funktion gelegt, die den Verlauf der Punktewolke bestmöglich widerspiegelt. Bei der einfachen linearen Regression ist dies eine Gerade, etwa die in Abbildung 9.1 angegebene durchgezogene Gerade. Üblicherweise werden zur methodischen Herleitung dieser durchgezogenen Geraden die quadrierten Abstände auf der Y-Achse zwischen dem jeweiligen Beobachtungspunkt und dem Y-Wert auf der zu schätzenden Geraden – bei gegebenem X-Wert auf der X-Achse – minimiert. Entsprechend heißt dieses Schätzverfahren **Kleinst-Quadrate-Methode** beziehungsweise auf Englisch: OLS für »Ordinary Least Squares«.

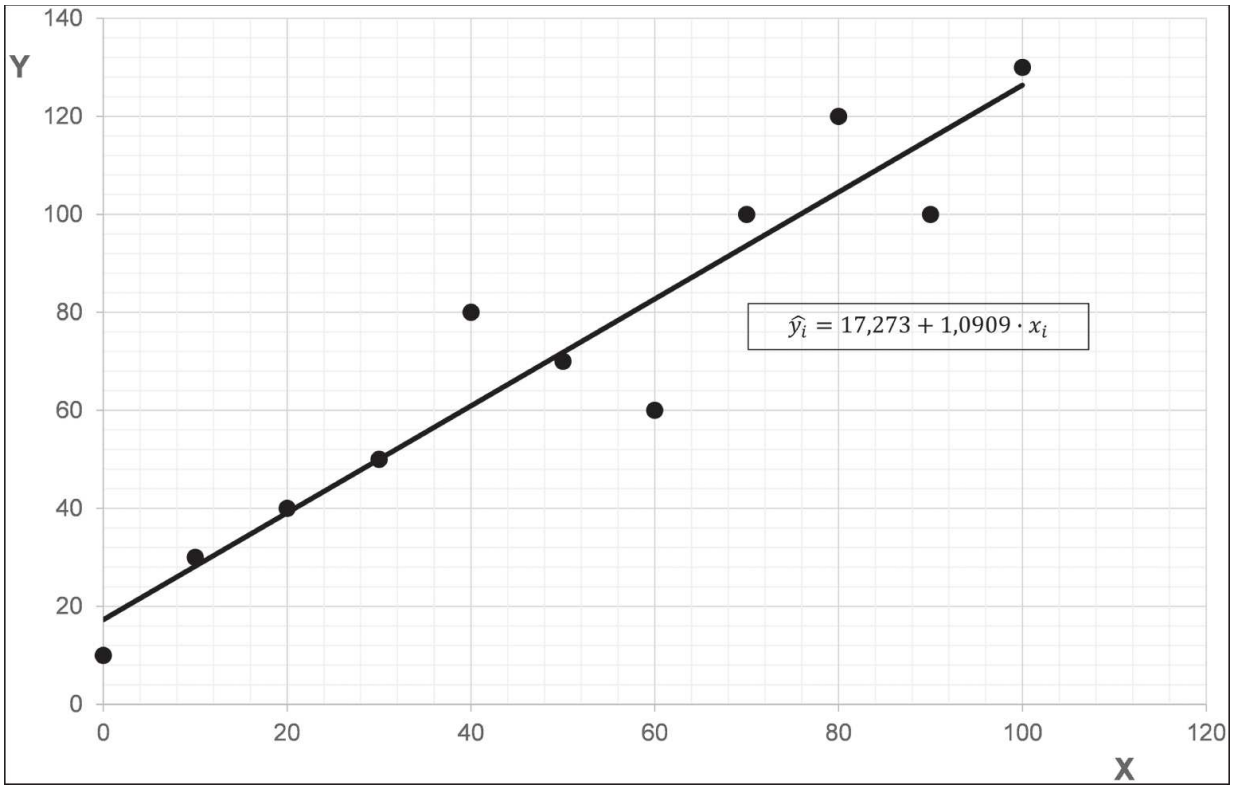



**Abbildung 9.1:** Lineare Regressionsfunktion und Punktewolke (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 164)

Dieser Minimierungsvorgang liefert am Ende Formeln zur Berechnung der Parameter $\hat{a}$ und $\hat{b}$:

$$\hat{b} = \frac{\sum_{i=1}^{N}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

$$\hat{a} = \bar{Y} - \hat{b} \cdot \bar{X}$$

In diesen Gleichungen symbolisieren $\bar{X}$ den arithmetischen Mittelwert der X-Werte und $\bar{Y}$ den arithmetischen Mittelwert der Y-Werte.

Wie das mit der Berechnung der beiden Regressionsparameter prinzipiell vonstattengeht, wird ausführlich anhand des obigen Beispiels veranschaulicht (siehe hierzu auch Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 163–167).

Erstellen Sie bitte zunächst die nachfolgende Arbeitstabelle (mit ganz unten in der letzten Zeile den Spaltensummen). Aus den beiden ersten beiden Spalten ergeben sich dabei die arithmetischen Mittelwerte $\bar{X} = \frac{550}{11} = 50$ und $\bar{Y} = \frac{790}{11} = 71{,}\overline{82}$.

| $X_i$ | $Y_i$ | $X_i - \bar{X}$ | $(X_i - \bar{X})^2$ | $Y_i - \bar{Y}$ | $(X_i - \bar{X}) \cdot (Y_i - \bar{Y})$ |
|---|---|---|---|---|---|
| 0 | 10 | –50 | 2.500 | –61,82 | 3.091,00 |
| 10 | 30 | –40 | 1.600 | –41,82 | 1.672,80 |
| 20 | 40 | –30 | 900 | –31,82 | 954,60 |
| 30 | 50 | –20 | 400 | –21,82 | 436,40 |
| 40 | 80 | –10 | 100 | 8,18 | –81,80 |
| 50 | 70 | 0 | 0 | –1,82 | 0 |
| 60 | 60 | 10 | 100 | –11,82 | –118,20 |
| 70 | 100 | 20 | 400 | 28,18 | 563,60 |
| 80 | 120 | 30 | 900 | 48,18 | 1.445,40 |
| 90 | 100 | 40 | 1.600 | 28,18 | 1.127,20 |
| 100 | 130 | 50 | 2.500 | 58,18 | 2.909,00 |
| **550** | **790** | **0** | **11.000** | **–** | **12.000** |

**Tabelle 9.2:** Arbeitstabelle für eine lineare Beispielregression

Als Steigung der Regression erhalten Sie folglich:

$$\hat{b} = \frac{12.000}{11.000} = 1,\overline{09}$$ (in Abbildung 9.1: gerundet 1,0909)

und der Ordinatenschnittpunkt (das heißt: der Schnittpunkt mit der Y-Achse) bestimmt sich in diesem Beispiel wie folgt:

$$\hat{a} = \frac{790}{11} = \frac{\frac{12.000}{11.000} \cdot 550}{11} = 17,\overline{27}$$ (in Abbildung 9.1: gerundet: 17,273)

# *Eine lineare Regression in SPSS schätzen*

So, nun lassen Sie uns endlich mit einer SPSS-Anwendung loslegen!

## *Scatterplots*

Zu diesem Zweck öffnen Sie bitte die Datendatei *LINEAREREGRESSION_GERADE1.SAV*. In diesem Zusammenhang soll eine lineare Einfachregression zwischen *ausgaben* als Y-Variable und *einkommen* als X-Variable erstellt werden. Um sich einen ersten Überblick über diesen Zusammenhang zu verschaffen, empfiehlt es sich – und dies gilt generell! – zunächst ein Diagramm – einen sogenannten **Scatterplot** (also auf Deutsch: ein Streudiagramm) – zu zeichnen

(beziehungsweise von SPSS zeichnen zu lassen). Dies führen Sie am besten über GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM durch, wobei Sie in unserem Fall *ausgaben* als Y-Variable und *einkommen* als X-Variable definieren. Offenbaren sich dabei zum Beispiel nichtlineare Zusammenhänge, so sind diese prinzipiell via Transformation in lineare Zusammenhänge zwischen X und Y überführbar, wie gleich noch gezeigt wird. Allerdings ist dabei zu beachten, dass die Linearität der Koeffizienten nicht verloren geht.

Sollten Sie eine frühere SPSS-Version als SPSS 29 verwenden, müssen Sie an dieser Stelle aus dem SPSS-Menü einen leicht anderen Weg gehen: zum Beispiel in SPSS 24 GRAFIK|ALTE DIALOGFELDER|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM.

Die SPSS-Syntax für den genannten Scatterplot lautet im Übrigen:

```
GRAPH
    /SCATTERPLOT(BIVAR)=einkommen WITH ausgaben
    /MISSING=LISTWISE.
```

Mit dem letztgenannten Unterbefehl MISSING = LISTWISE werden Fälle mit einem fehlenden Wert bei mindestens einer der berücksichtigten Variablen von den weiteren Berechnungen komplett ausgeschlossen.

Im vorliegenden Beispielfall erhalten Sie den in Abbildung 9.2 angegebenen Scatterplot.

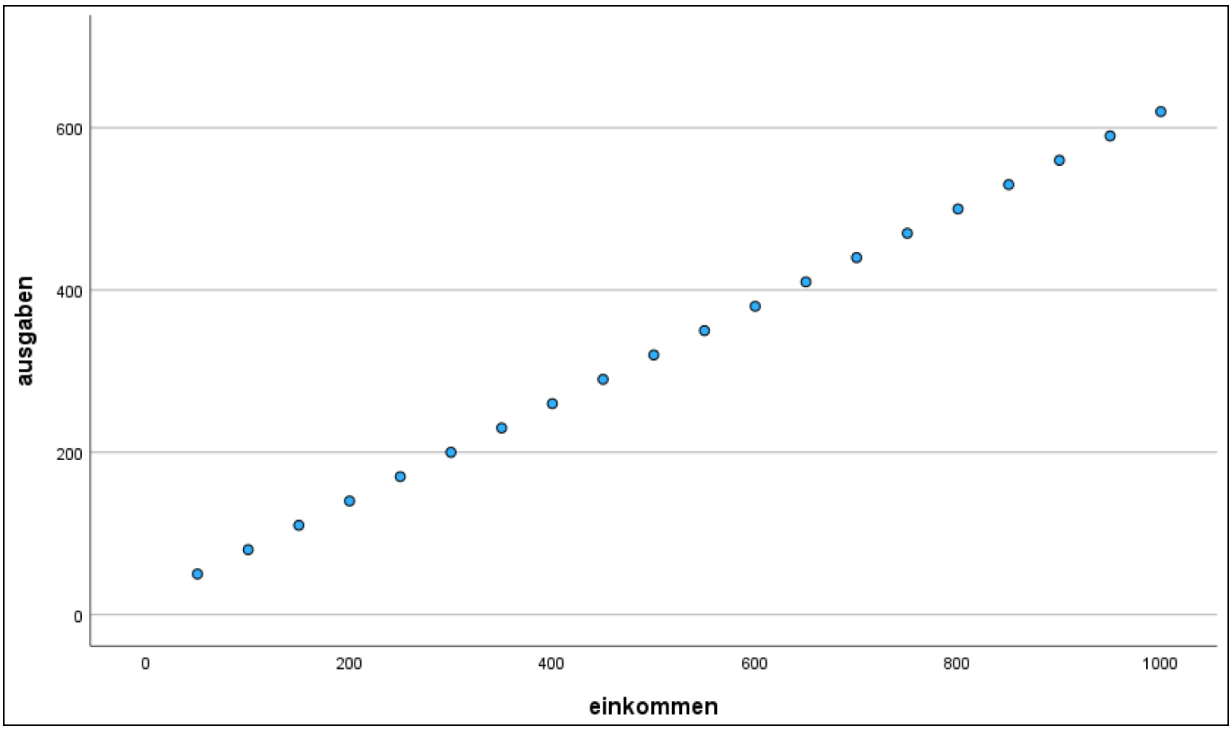



**Abbildung 9.2**: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE1.SAV*

## *OLS-Schätzung*

Dieser Scatterplot deutet auf eine klare lineare Abhängigkeit zwischen *ausgaben* und *einkommen* hin. Mittels ANALYSIEREN|REGRESSION|LINEAR schätzen Sie daher bitte mal zwischen diesen beiden Variablen eine lineare Einfachregression (siehe Abbildung 9.3); das heißt in Syntaxschreibweise:

```
REGRESSION
    /MISSING LISTWISE
    /STATISTICS COEFF OUTS R ANOVA
    /CRITERIA=PIN(.05) POUT(.10)
    /NOORIGIN
    /DEPENDENT ausgaben
    /METHOD=ENTER einkommen.
```

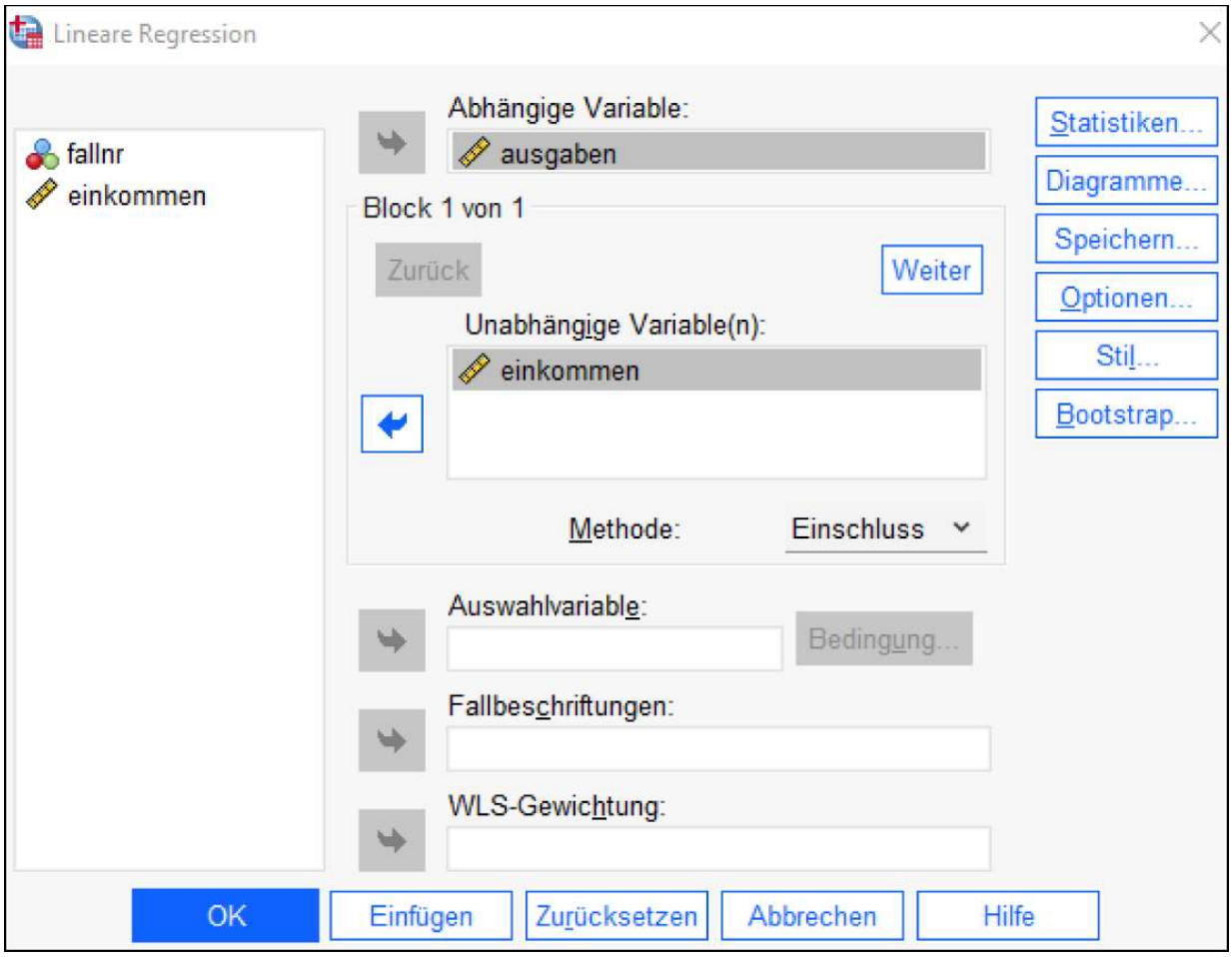



**Abbildung 9.3:** Lineare Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE1.SAV*

Es ergibt sich (unter anderem) der in Abbildung 9.4 dargelegte SPSS-Output.

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten: Regressionskoeffizient B | Nicht standardisierte Koeffizienten: Std.-Fehler | Standardisierte Koeffizienten: Beta | T | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Konstante) | 20,000 | ,000 | | . | . |
| | einkommen | ,600 | ,000 | 1,000 | . | . |

a. Abhängige Variable: ausgaben

**Abbildung 9.4:** Ergebnis der linearen Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE1.SAV* (Auszug)

Im vorliegenden Beispiel handelt es sich um einen perfekten linearen Zusammenhang zwischen der unabhängigen X-Variablen (*einkommen*) und der abhängigen Y-Variablen

(*ausgaben*). Gemäß Abbildung 9.4 lautet die Schätzgleichung für diesen Zusammenhang:

$$\hat{Y}_i = 20 + 0{,}6 \cdot X_i.$$

Das konstante Glied in Höhe von +20 ergibt sich bei einem X-Wert in Höhe von 0. Die Steigung beträgt hier 0,6. Das heißt: Erhöht sich das Einkommen um einen Euro, erhöhen sich die privaten Konsumausgaben um 0,60 Euro. Die Schätzergebnisse können nur innerhalb des vorgegebenen X-Wertebereichs – hier zwischen 50 Euro (Fall Nummer 1) und 1.000 Euro (Fall Nummer 20) – sinnvoll interpretiert werden. Damit entzieht sich streng genommen auch das absolute Glied von +20 einer Interpretation, da ein X-Wert (= ein Einkommen) von 0 im Datensatz nicht beobachtet worden ist.

# *Beeinflussung der linearen Regressionsgeraden durch Ausreißerwerte*

Dass die Regressionsschätzungen bei einer linearen Einfachregression stark von **Ausreißerwerten** beeinflusst werden (worauf in diesem Kapitel bereits hingewiesen worden ist), zeigt sich bei einer entsprechenden Schätzung auf Basis des Datensatzes *LINEAREREGRESSION_GERADE2.SAV*. Der betreffende Scatterplot weist für die Variable *ausgaben* bei Fall Nummer 20 einen Ausreißerwert auf (50.000 Euro; siehe Abbildung 9.5 – erneut über GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM mit *ausgaben* als Y-Variablen und *einkommen* als X-Variablen erstellt).

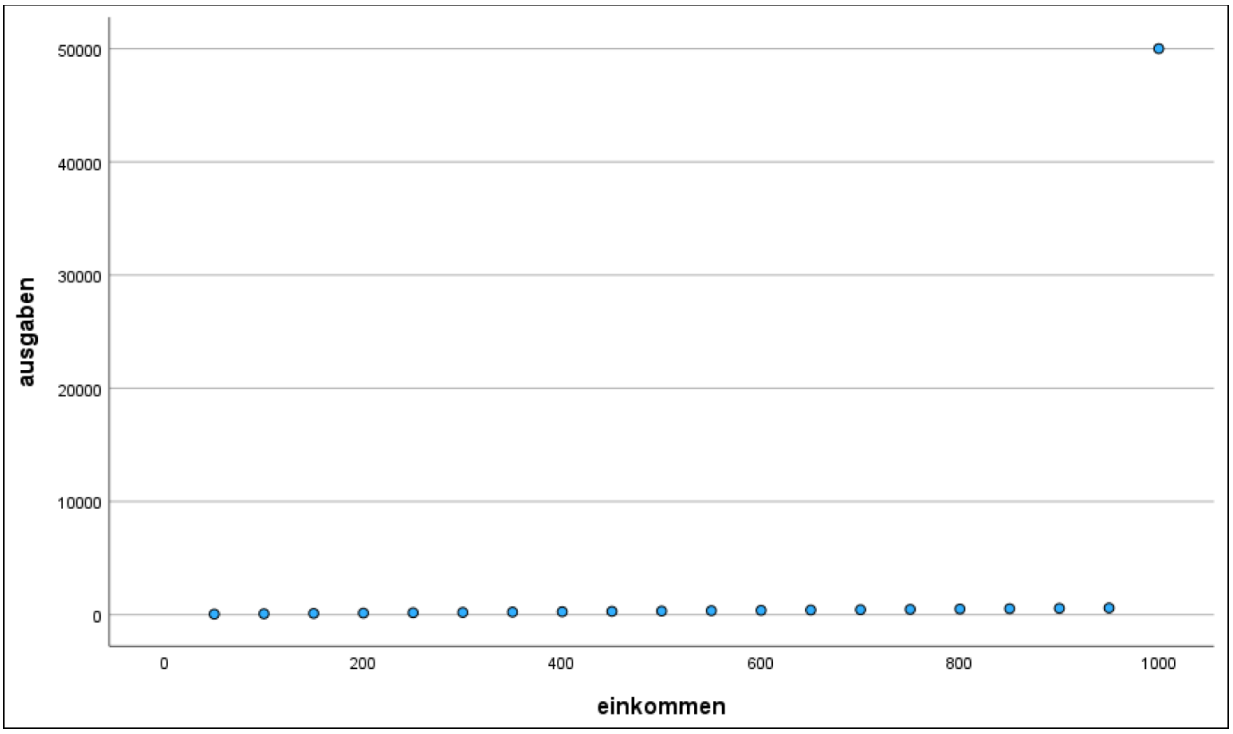



**Abbildung 9.5**: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE2.SAV*

Wie Abbildung 9.6 zeigt, verändert sich die Regressionsgerade allein durch diesen einen Ausreißerwert grundlegend, und zwar zu:

$$\hat{Y}_i = -4.918 + 14{,}709 \cdot X_i$$

**Koeffizienten**[a]

| Modell | | Nicht standardisierte Koeffizienten: Regressionskoeffizient B | Nicht standardisierte Koeffizienten: Std.-Fehler | Standardisierte Koeffizienten: Beta | T | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Konstante) | -4918,000 | 4878,860 | | -1,008 | ,327 |
| | einkommen | 14,709 | 8,146 | ,392 | 1,806 | ,088 |

a. Abhängige Variable: ausgaben

**Abbildung 9.6**: Ergebnis der linearen Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE2.SAV*

Es ist zu erkennen, dass sich das absolute Glied geradezu dramatisch von +20 zu −4.918 und auch die Steigung von +0,6 zu +14,709 durch einen einzigen Ausreißerwert verändert hat. In diesem Fall ist zu überlegen, ob dieser Ausreißerwert nicht eventuell aus dem Datensatz herausgenommen werden sollte. Unter inhaltlichen Gesichtspunkten erscheinen Güterausgaben in Höhe von 50.000 Euro bei einem Einkommen von 1.000 Euro zwar nicht völlig abwegig (sie könnten aus Vermögensauflösungen stammen),

sind aber angesichts des üblichen Ausgabenverhaltens, demzufolge Ausgaben aus dem laufenden Einkommen finanziert werden (und deshalb kleiner als beziehungsweise gleich dem Einkommen sind), sehr kritisch zu hinterfragen. Daher wäre hier meines Erachtens in der Tat zu überlegen, den betreffenden »Ausreißer« bei den Schätzungen *nicht* zu berücksichtigen.

# Keine Variation der abhängigen Variablen

Auch die nächsten Beispiele illustrieren, dass Sie den linearen Regressionsansatz keinesfalls »auf Knopfdruck« in SPSS durchführen sollten. Öffnen Sie bitte zunächst den Datensatz *LINEAREREGRESSION_GERADE3.SAV*. In diesem Datensatz sind die Ausgaben (in der Variablen *ausgaben*) für alle 20 Untersuchungseinheiten konstant (jeweils 200 Euro). Dies geht aus dem Scatterplot in Abbildung 9.7 hervor, der wieder über GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM mit *ausgaben* als Y-Variablen und *einkommen* als X-Variablen erstellt worden ist.

In diesem Fall reagiert SPSS auf eine entsprechende Regressionsanforderung nach dem üblichen Schema über ANALYSIEREN|REGRESSION|LINEAR mit der unabhängigen Größe *einkommen* und der abhängigen Variablen *ausgaben* allergisch. Es verweigert die Berechnung, und es erscheint die in Abbildung 9.8 wiedergegebene Warnung im Sinne einer Fehlermeldung.

# Bereichsspezifische Konstanz der abhängigen Variablen

Auf den ersten Blick ähnlich gelagert scheint der in Abbildung 9.9 dargestellte Fall (auf Basis der Datendatei *LINEAREREGRESSION_GERADE4.SAV*). Der Scatterplot ist auch hier wieder via GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM mit *ausgaben* als Y-Variablen und *einkommen* als X-Variablen erzeugt worden. Bis zu einem Einkommen von 450 Euro lautet das Ausgabenniveau durchgängig 200 Euro; ab 500 Euro liegt es demgegenüber konstant bei 500 Euro.

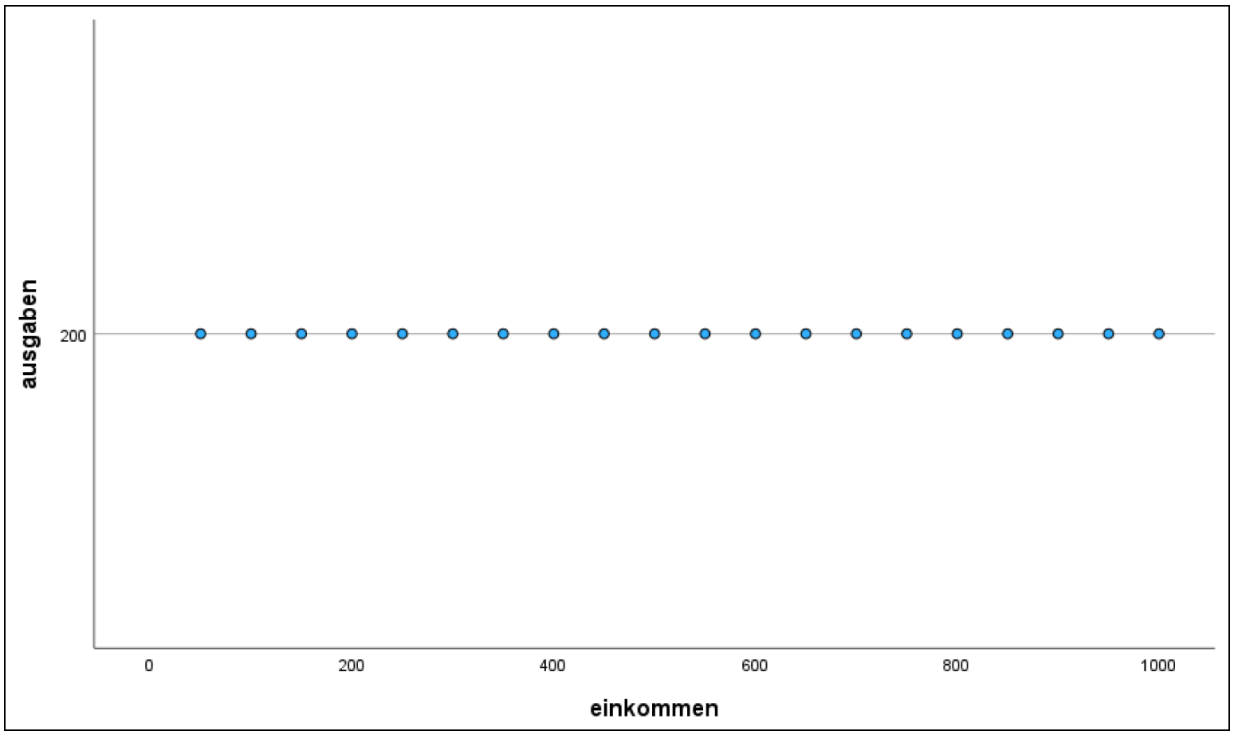



**Abbildung 9.7:** Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE3.SAV*

**Warnungen**

Die abhängige Variable ausgaben wurde gelöscht. Statistiken können nicht berechnet werden.

**Abbildung 9.8:** Ergebnis der linearen Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE3.SAV*

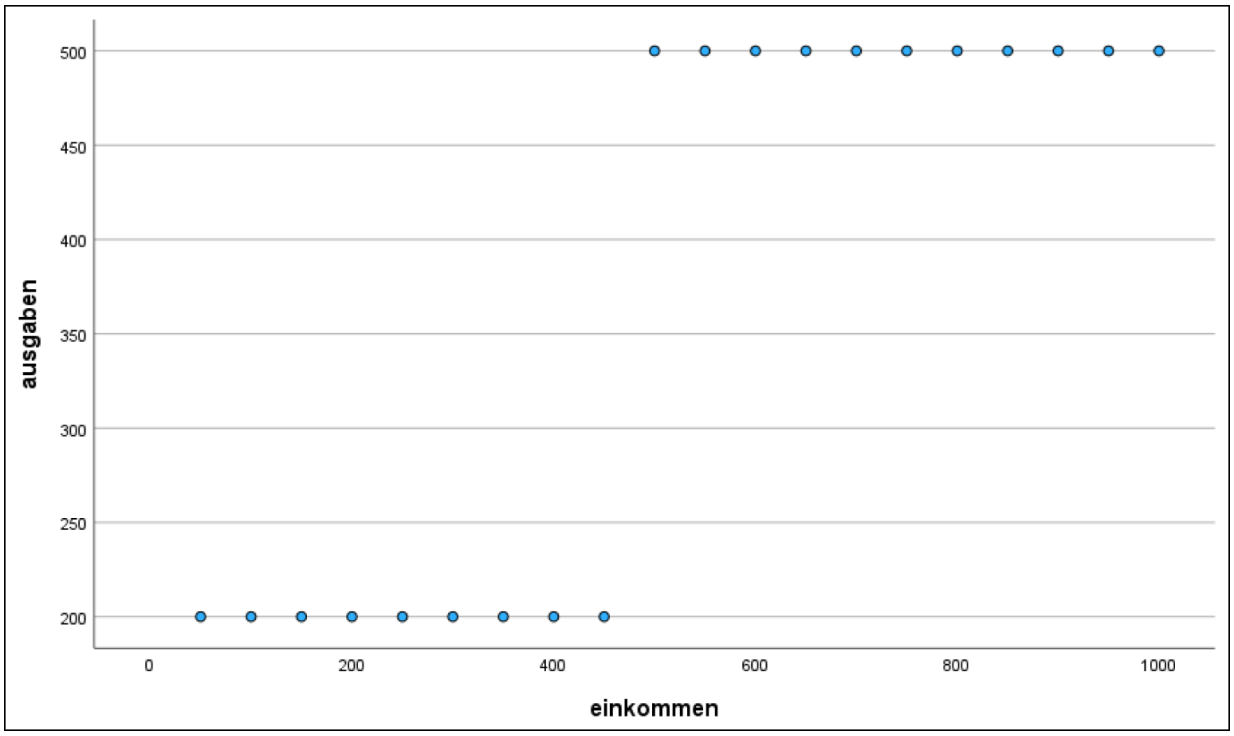



**Abbildung 9.9:** Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE4.SAV*

Das lineare Regressionsmodell reagiert auf die **Sprungstelle** bei den Ausgaben (das heißt auf den Ausgabensprung von konstant 200 Euro im Einkommensbereich von 50 bis 450 Euro auf konstant 500 Euro im Einkommensbereich von 500 bis 1.000 Euro) so, als handele es sich um eine lineare Änderung zwischen den beiden Einkommensbereichen. Es wird folglich eine lineare Regressionsbeziehung berechnet. Das entsprechende SPSS-Ergebnis findet sich in Abbildung 9.10. Es erscheint aber, vorsichtig ausgedrückt, sehr fraglich, ob der ausgewiesene (und über ANALYSIEREN|REGRESSION|LINEAR angeforderte) Regressionszusammenhang mit einer Konstanten in Höhe von +130,526 und einer Steigung in Höhe von +0,447 für den gesamten X-Wertebereich von 50 bis 1.000 Euro die Datenlage angemessen widerspiegelt.

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | Regressionskoeffizient B | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | 130,526 | 36,950 | | 3,532 | ,002 |
| | einkommen | ,447 | ,062 | ,863 | 7,240 | <,001 |

a. Abhängige Variable: ausgaben

**Abbildung 9.10:** Ergebnis der linearen Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE4.SAV*

# *Entgegengesetzte allgemeine Regression im Vergleich zu bereichsspezifischen Regressionen*

Vielleicht noch etwas »verrückter« ist die Lage auf Basis der Datendatei *LINEAREREGRESSION_GERADE5.SAV*. Der mittels GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM mit *ausgaben* als Y-Variable und *einkommen* als X-Variable erzeugte Scatterplot in Abbildung 9.11 zeigt, dass sich das Ausgabeverhalten für vier Einkommensbereiche voneinander unterscheidet, und zwar wird für die Einkommensbereiche von 50 bis 200 Euro, von 250 bis 450 Euro, von 500 bis 750 Euro und von 800 bis 1.000 Euro jeweils innerhalb dieser Bereiche ein steigender Verlauf der Beobachtungspunkte evident.

Wenn Sie jetzt vermuten, dass eine Regressionsberechnung über den gesamten Einkommensbereich ebenfalls eine steigende Gerade ergibt, liegen Sie leider komplett falsch! Wie Abbildung 9.12 verdeutlicht, folgt nämlich ein genau umgekehrtes Ergebnis: ein sinkender Verlauf der (erneut mittels ANALYSIEREN|REGRESSION|LINEAR geschätzten) Regressionsgeraden mit einem konstanten Glied in Höhe von +152,368 und einer Steigung von −0,126. In einer solchen Situation sollten Sie offenkundig anhand des Scatterplots für die unterschiedlichen X-Wertebereiche am besten *separate* Regressionen schätzen.

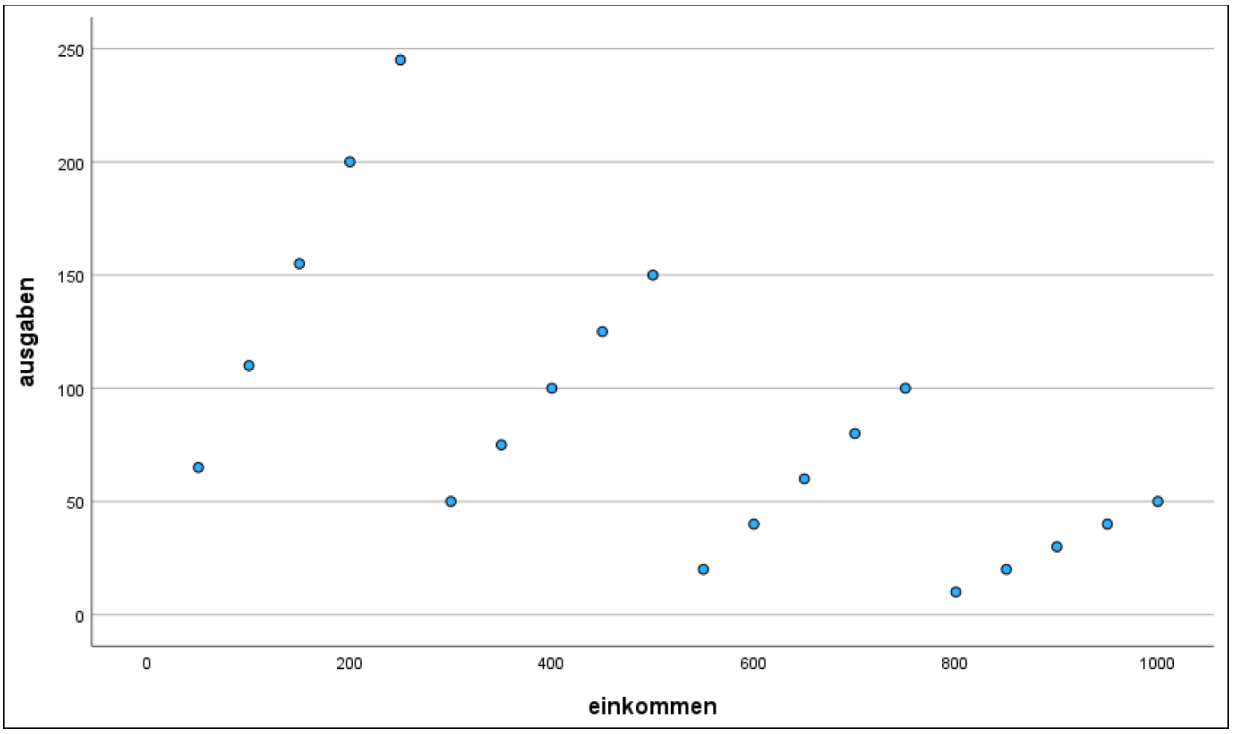



**Abbildung 9.11:** Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE5.SAV*

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | Regressionskoeffizient B | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | 152,368 | 24,231 | | 6,288 | <,001 |
| | einkommen | -,126 | ,040 | -,592 | -3,113 | ,006 |

a. Abhängige Variable: ausgaben

**Abbildung 9.12:** Ergebnis der linearen Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE5.SAV*

# *Zur Aussagekraft einer Regressionsbeziehung*

Um die Güte einer geschätzten Regressionsbeziehung beurteilen zu können, bedarf es verschiedener Kriterien. Im Folgenden werden davon beschrieben:

- ✔ das Bestimmtheitsmaß,
- ✔ die sogenannte ANOVA und
- ✔ die Signifikanz der Regressionskoeffizienten.

Lassen Sie mich mit der Konzeption des Bestimmtheitsmaßes beginnen.

## *Bestimmtheitsmaß*

Je enger der Zusammenhang zwischen Y und X in einer Regression ist, das heißt: Je enger die Punktewolke um die Regressionsfunktion herum liegt, desto höher ist der Erklärungsgehalt einer linearen Regression. Dieser Erklärungsgehalt wird durch das **Bestimmtheitsmaß** B gemessen. Dieses ist bei der linearen Einfachregression das Quadrat des Bravais/Pearson-Korrelationskoeffizienten (siehe Kapitel 8):

$$B = r_{BP}^2.$$

Etwas ausführlicher ist das Bestimmtheitsmaß folgendermaßen definiert:

$$B = 1 - \frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2}.$$

Es spiegelt somit das Verhältnis aus der durch die Regressionsgleichung für die Beziehung zwischen X und Y erklärten Varianz (Zähler) einerseits und der Gesamtvarianz für die betreffende Beziehung (Nenner) andererseits wider. Wenn durch das Regressionsmodell nichts erklärt wird, hat B seine Untergrenze von 0. Erklärt die Regressionsbeziehung hingegen die Gesamtvarianz vollständig, so hat B seine Obergrenze von 1 inne. Da in der Realität typischerweise eine Variable Y nicht vollständig durch eine einzelne andere Variable X erklärt wird, weist B in der Praxis üblicherweise einen Wert kleiner 1 auf.

Es gelten zum Beispiel folgende »Faustregeln« (bei Grundgesamtheitsdaten): $B \leq 0{,}25$: schwache lineare Anpassung, $0{,}25 < B \leq 0{,}64$: mittlere lineare Anpassung und $B > 0{,}64$: starke lineare Anpassung – oder (bei Stichprobendaten): $B \leq 0{,}09$: schwache lineare Anpassung, $0{,}09 < B \leq 0{,}25$: mittlere lineare Anpassung und $B > 0{,}25$: starke lineare Anpassung.

Zur Berechnung des Bestimmtheitsmaßes für den aus Abbildung 9.2 bekannten Regressionszusammenhang können Sie wieder eine Arbeitstabelle aufstellen (siehe in diesem Zusammenhang auch Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 170). In Tabelle 9.3 sind die Summen in der letzten Zeile durch Fettdruck hervorgehoben.

Mit diesen Informationen können Sie das Bestimmtheitsmaß berechnen. Im Beispiel liegt es recht hoch:

$$B = 1 - \frac{1.472{,}7272}{14.563{,}6379} \approx 0{,}8989.$$

SPSS weist das Bestimmtheitsmaß standardmäßig vor der Darstellung der Regressionsparameter aus. Wählen Sie hierzu die Datendatei

*LINEAREREGRESSION_GERADE6.SAV* aus. Den zugehörigen Scatterplot erhalten Sie wieder über GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM. Er ist in Abbildung 9.13 dargestellt. Man erkennt einen positiven Zusammenhang zwischen *einkommen* und *ausgaben*, der allerdings nicht perfekt linear ist.

Lassen Sie sich jetzt wieder mittels ANALYSIEREN|REGRESSION|LINEAR von SPSS die Regressionsgerade zwischen *einkommen* als unabhängige X-Variable und *ausgaben* als abhängige Y-Variable ausrechnen, erhalten Sie zugleich das Bestimmtheitsmaß (siehe hierzu Abbildung 9.14). Zunächst wird im SPSS-Output angegeben, welche Variablen als erklärende Größe(n) in die Regressionsgleichung aufgenommen worden sind. Im vorliegenden Fall ist das – wenig überraschend – die Variable *einkommen*. Gleich danach sind unter der Überschrift »Modellzusammenfassung« der Bravais/Pearson-Korrelationskoeffizient unter der Bezeichnung R mit einem Wert von 0,952, das Bestimmtheitsmaß – bekanntlich bei der linearen Einfachregression das Quadrat des Bravais/Pearson-Korrelationskoeffizienten und daher von SPSS als »R-Quadrat« bezeichnet – mit einem Wert von 0,907 und das sogenannte *korrigierte Bestimmtheitsmaß* mit einem Wert von 0,902 angegeben. Das Bestimmtheitsmaß ist in diesem Beispiel also sehr hoch, was auf eine hohe Erklärungskraft des vorliegenden Zusammenhangs zwischen *einkommen* und *ausgaben* hinweist.

| $X_i$ | $Y_i$ | $\hat{Y}_1 = 17{,}2727 + 1{,}0909 \cdot X_1$ | $(Y_i - \hat{Y}_1)^2$ | $(Y_i - \bar{Y})^2$ mit: $\hat{Y} = \frac{790}{11} \approx 71{,}8182$ |
|---|---|---|---|---|
| 0 | 10 | 17,2727 | 52,8926 | 3.821,4899 |
| 10 | 30 | 28,1817 | 3,3058 | 1.748,7629 |
| 20 | 40 | 39,0907 | 0,8264 | 1.012,3979 |
| 30 | 50 | 49,9997 | 0,0000 | 476,0339 |
| 40 | 80 | 60,9087 | 364,4628 | 66,9419 |
| 50 | 70 | 71,8177 | 3,3058 | 3,3059 |
| 60 | 60 | 82,7267 | 516,5289 | 139,6699 |
| 70 | 100 | 93,6357 | 40,4959 | 794,2139 |
| 80 | 120 | 104,5447 | 238,8430 | 2.321,4859 |
| 90 | 100 | 115,4537 | 238,8430 | 794,2139 |
| 100 | 130 | 126,3627 | 13,2231 | 3.385,1219 |
| **550** | **790** | **–** | **1.472,7273** | **14.563,6379** |

**Tabelle 9.3:** Arbeitstabelle für die Berechnung des Bestimmtheitsmaßes

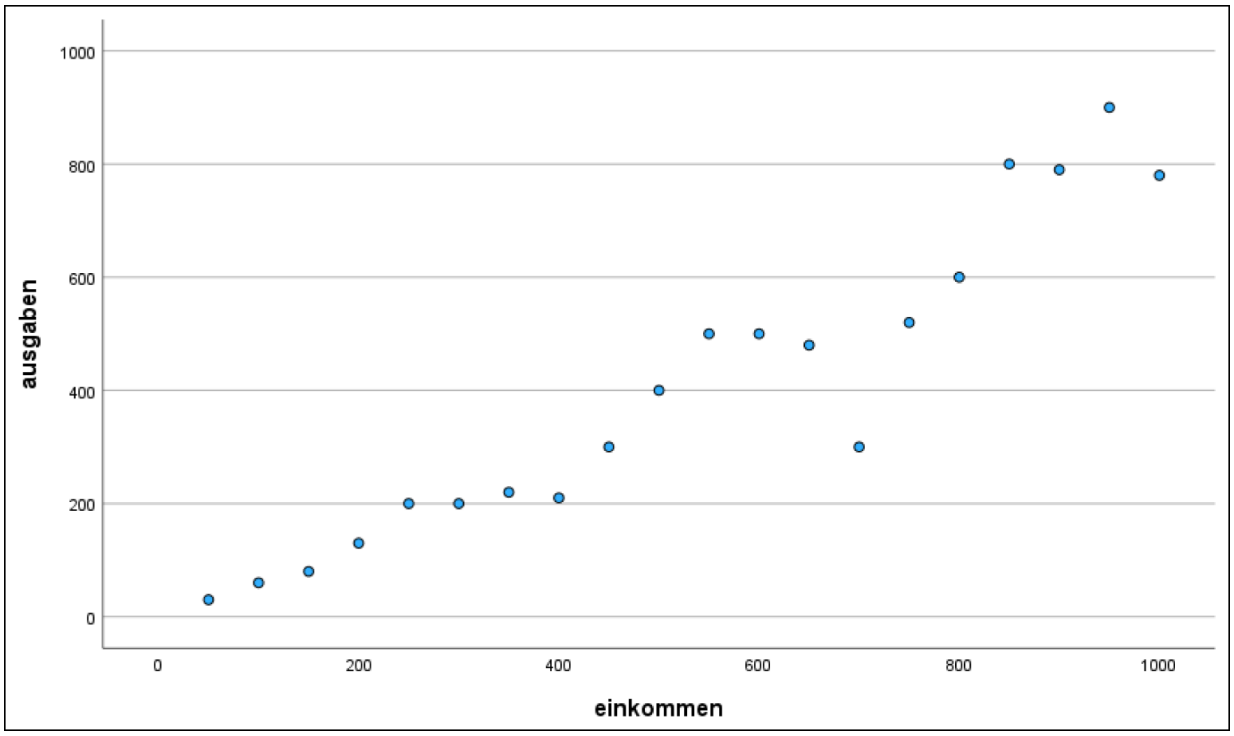



**Abbildung 9.13:** Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE6.SAV*

Das korrigierte Bestimmtheitsmaß $B_{korr}$ ist im Übrigen folgendermaßen definiert:

$$B_{korr} = 1 - (1 - B) \cdot \frac{N - 1}{N - K - 1}.$$

Hierin bezeichnen N die Anzahl der Beobachtungen und K die Anzahl der unabhängigen Variablen. Im Unterschied zu B kann $B_{korr}$ auch negative Werte annehmen. Das korrigierte Bestimmtheitsmaß ist außer in dem Fall, in dem B = 1 ist, kleiner als B. Es spielt in erster Linie bei der unten näher erörterten multiplen Regression eine Rolle, nicht so sehr hier bei der einfachen linearen Regression, was auch in den geringen Unterschieden zwischen B und $B_{korr}$ im Beispiel deutlich wird. Im Beispiel berechnet sich $B_{korr}$ folgendermaßen:

$$B_{korr} = 1 - (1 - 0{,}907) \cdot \frac{20 - 1}{20 - 1 - 1} = 0{,}902.$$

## *ANOVA*

Bei den SPSS-Regressionsangaben folgt dann – wie aus Abbildung 9.14 zu entnehmen ist – eine Rubrik **ANOVA**. Hierbei steht ANOVA für den englischen Ausdruck »ANalysis Of VAriance«. In dieser Rubrik wird Auskunft über die durch die Regression erklärte beziehungsweise nicht erklärte Varianz sowie über die Gesamtvarianz des Regressionsmodells gegeben. Besonders interessant sind dabei die

Spalten »F« und »Sig.«. In der Spalte »F« wird ein sogenannter F-Wert angegeben, der für die Modellgüte steht und der hinsichtlich seiner *Signifikanz* (das heißt: hinsichtlich seiner statistischen Aussagekraft) in der Spalte »Sig.« beurteilt wird.

In unserem Beispiel wird der F-Wert von 175,190 als hochsignifikant eingestuft, da unter »Sig.« eine Irrtumswahrscheinlichkeit nahe 0 angegeben ist. Inhaltlich besehen ist die Angabe einer geringen Irrtumswahrscheinlichkeit Ausdruck einer hohen Signifikanz. Das heißt hier: Die berechnete Regressionsgerade wird auch über den F-Wert als Modell mit einem hohen Erklärungsgrad eingestuft. Im Zusammenhang mit dem *statistischen Testen* in Teil IV wird noch genauer auf derartige Aspekte eingegangen.

## *Signifikanz der Regressionskoeffizienten*

Letzteres gilt auch bezüglich der statistischen Beurteilung der geschätzten Regressionsparameter, die in Abbildung 9.14 ganz am Ende stehen. Auch bei diesen kann mittels statistischer Tests (auf Basis der t-Verteilung; daher hier auch die Spalte »T«) deren Signifikanz überprüft werden. Findet sich unter »Sig.« hierbei ein hoher Wert für die *Überschreitungs-* beziehungsweise *Irrtumswahrscheinlichkeit* (als äußerste Grenze wird in der Regel 0,10 beziehungsweise 10 Prozent verwendet, meistens aber schon 0,05 beziehungsweise 5 Prozent), werden Sie die Nullhypothese (»Parameterwert = 0«) nicht verwerfen, und der zugehörige Parameter ist nicht signifikant (von 0 verschieden). (Mehr zu derartigen Tests finden Sie in Teil IV.)

Im Beispiel ergibt sich für die Konstante ein Wert von -56,158, der allerdings aufgrund einer **Signifikanzangabe** von 0,170 in »Sig.« nicht als statistisch aussagekräftig eingestuft wird (er wäre es erst ab einem Signifikanzniveau von 0,170 beziehungsweise 17 Prozent); das heißt: Er wird nicht als signifikant angesehen. Demgegenüber kann der Steigungsparameter in Höhe von 0,869 angesichts der Signifikanzangabe von »< 0,001« in »Sig.« als hochsignifikant eingestuft werden.

**Regression**

**Aufgenommene/Entfernte Variablen[a]**

| Modell | Aufgenommene Variablen | Entfernte Variablen | Methode |
|---|---|---|---|
| 1 | einkommen[b] | . | Einschluß |

a. Abhängige Variable: ausgaben

b. Alle gewünschten Variablen wurden eingegeben.

**Modellzusammenfassung**

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|---|---|---|---|---|
| 1 | ,952[a] | ,907 | ,902 | 84,497 |

a. Einflußvariablen : (Konstante), einkommen

**ANOVA[a]**

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 1255085,865 | 1 | 1255085,865 | 175,790 | <,001[b] |
| | Nicht standardisierte Residuen | 128514,135 | 18 | 7139,674 | | |
| | Gesamt | 1383600,000 | 19 | | | |

a. Abhängige Variable: ausgaben

b. Einflußvariablen : (Konstante), einkommen

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | Regressionskoeffizient B | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | -56,158 | 39,251 | | -1,431 | ,170 |
| | einkommen | ,869 | ,066 | ,952 | 13,259 | <,001 |

a. Abhängige Variable: ausgaben

**Abbildung 9.14:** Vollständiges Ergebnis der linearen Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE6.SAV*

In Tabelle 9.4 sind als Übersicht die Regressionsergebnisse aller vorstehenden Regressionen angegeben. Dabei basiert die in der ersten Zeile angegebene Regression auf einer perfekten Punktewolke und ist daher mit einem Bestimmtheitsmaß von 1,000 verbunden. In einem solchen Fall weist SPSS, wie aus Tabelle 9.4 ersichtlich ist, bei den Regressionsparametern keine Signifikanzniveaus aus.

Der in der zweiten Zeile genannte Fall ist durch einen (extremen) Ausreißerwert gekennzeichnet. Dies führt dazu, dass sich die Werte der Regressionsparameter sehr stark (und beim absoluten Glied sogar bezüglich des Vorzeichens) verändern. Diese Werte sind entweder gar nicht (absolutes Glied) oder nur (ganz) schwach signifikant (Steigungsparameter); auch der Wert des Bestimmtheitsmaßes sinkt deutlich auf jetzt nur noch 0,153.

In der dritten Zeile der folgenden Tabelle ist der Fall angegeben, bei dem die abhängige Variable *ausgaben* keine Variation aufweist. Hier ist keine Regressionsberechnung möglich gewesen.

Im vierten Fall mit jeweils konstanten Ausgabenwerten innerhalb zweier Einkommensbereiche resultiert eine positiv geneigte Gerade mit positivem Ordinatenschnittpunkt. Die Regressionsparameter sind beide hochsignifikant, und auch das Bestimmtheitsmaß mit einem erklärten Varianzanteil in Höhe von fast 75 Prozent ist recht hoch. Diese Befunde weisen auf die Problematik hin, dass man sich nicht nur auf die Bestimmtsheitsmaß- beziehungsweise auf die Signifikanzangaben bei der Schätzung einer Regressionsgeraden verlassen darf, weil in diesem vierten Fall die Y-Variable im Grunde genommen ja auch keine ausreichende Variation aufweist.

Die letztgenannte Signifikanzproblematik kann auch auf den fünften Fall in Tabelle 9.4 bezogen werden. Bei diesem Fall ergibt sich bekanntlich trotz positiver Zusammenhänge zwischen *ausgaben* und *einkommen* in vier Einkommensbereichen ein negativer Gesamtzusammenhang. Durch das Bestimmtheitsmaß werden 35 Prozent der Gesamtvarianz erklärt, und die beiden geschätzten Regressionsparameter sind beide hochsignifikant, obwohl der negative Gesamtzusammenhang die im Grunde genommen positive Beziehung zwischen *ausgaben* und *einkommen* in den vier Einkommensbereichen ganz und gar nicht angibt!

Der in der letzten Zeile von Tabelle 9.4 wiedergegebene lineare Regressionszusammenhang schließlich zeigt ein sehr hohes Bestimmtheitsmaß in Höhe von 90,7 Prozent erklärter Varianz und einen hochsignifikanten Steigungsparameter – bei allerdings nicht-signifikantem absoluten Glied.

Die in Tabelle 9.4 enthaltenen Bestimmtheitsmaße können hier problemlos miteinander verglichen werden, weil in allen durchgeführten linearen Regressionen die abhängige Variable jeweils inhaltlich gleich ist (jeweils *ausgaben*). So kann etwa gesagt werden, dass das Bestimmtheitsmaß für den Fall in der vierten Zeile von Tabelle 9.4 mit 0,744 mehr als doppelt so hoch wie im fünften Fall von Tabelle 9.4 mit einem Bestimmtheitsmaß von 0,350 ist.

Hätte man die abhängige Variable vor Durchführung der Regressionsberechnungen zum Beispiel logarithmiert (wie dies nachfolgend zum Teil gemacht wird), dann hätte das für eine derartige Regressionsberechnung ermittelte Bestimmtheitsmaß, streng genommen, nicht mit den Werten des

Bestimmtheitsmaßes für Regressionsberechnungen mit *ausgaben* als abhängiger Variable verglichen werden dürfen!

| *Datendatei* | *Absolutes Glied (in Klammern Signifikanzniveau)* | *Steigung (in Klammern Signifikanzniveau)* | *Bestimmtheitsmaß* |
|---|---|---|---|
| *LINEAREREGRESSION_GERADE1.SAV* | 20,000 (.) | 0,600 (.) | 1,000 |
| *LINEAREREGRESSION_GERADE2.SAV* | –4.918,000 (0,327) | 14,709 (0,088) | 0,153 |
| *LINEAREREGRESSION_GERADE3.SAV* | – | – | – |
| *LINEAREREGRESSION_GERADE4.SAV* | 130,526 (0,002) | 0,447 (< 0,001) | 0,744 |
| *LINEAREREGRESSION_GERADE5.SAV* | 152,368 (< 0,001) | –0,126 (0,006) | 0,350 |
| *LINEAREREGRESSION_GERADE6.SAV* | –56,158 (0,170) | 0,869 (< 0,001) | 0,907 |

**Tabelle 9.4:** Übersicht über die Signifikanz der geschätzten linearen Regressionen zwischen *ausgaben* und *einkommen* auf Basis der Datendateien *LINEAREREGRESSION_GERADE1.SAV* bis *LINEAREREGRESSION_GERADE6.SAV* (gemäß jeweiligem SPSS-Output)

# Quasilineare Regression

Zeigt sich in einem Scatterplot eine nichtlineare Beziehung zwischen der X- und der Y-Variablen, können Sie durch Umformungen (**Transformationen**) das lineare Regressionsmodell auch auf eine **nichtlineare Gleichung** anwenden. Man spricht dann von einer **quasilinearen Regression**.

Im Folgenden wird auf derartige nichtlineare Zusammenhänge und ihre Berücksichtigung im Modellrahmen der linearen Regression eingegangen.

# Quadratische Funktion

Ein Beispiel für eine Abweichung von der üblichen Geradenkonzeption stellt ein Polynom zweiten Grades dar, das heißt: eine **quadratische Regressionsfunktion**:

$$\hat{Y}_i = \hat{a} + \hat{b} \cdot X_i + \hat{c} \cdot X_i^2.$$

Lassen Sie uns in diesem Zusammenhang den Scatterplot für die Beziehung zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_QUADRAT1.SAV* betrachten (siehe Abbildung 9.15). Der Scatterplot wurde dabei wieder über GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES

STREUDIAGRAMM erzeugt. Er deutet auf einen quadratischen Zusammenhang – zumindest näherungsweise – hin.

Um einen entsprechenden quadratischen Regressionszusammenhang schätzen zu können, bilden Sie einfach eine neue X-Variable für den quadratischen Term in der obigen Gleichung, und zwar:

```
COMPUTE einkommen2 = einkommen**2.
EXECUTE.
```

Im nächsten Schritt fordern Sie über ANALYSIEREN|REGRESSION|LINEAR eine lineare Regressionsbeziehung an, jetzt aber (auf Basis des entsprechend präparierten Datensatzes *LINEAREREGRESSION_QUADRAT1A.SAV*) zwischen *ausgaben* als Y-Variable sowie *einkommen* und *einkommen2* als X-Variablen (siehe Abbildung 9.16).

Es ergibt sich dann der in Abbildung 9.17 dargestellte Output. Das (hier nicht abgebildete) Bestimmtheitsmaß von 0,943 deutet darauf hin, dass die Beziehung zwischen *ausgaben* sowie *einkommen* und *einkommen2* ziemlich gut in der gewählten quadratischen Form dargestellt werden kann. Auch die einzelnen Regressionskoeffizienten gelten als signifikant, wie aus der Spalte »Sig.« hervorgeht. Im Einzelnen ist folgende Gleichung zwischen *ausgaben* (Y) und *einkommen* beziehungsweise *einkommen2* (X) geschätzt worden, wobei der Parameterwert für den quadratischen X-Ausdruck nur sehr gering ist (und in SPSS, das standardmäßig nur drei Nachkommastellen ausweist, erst erscheint, wenn Sie auf die betreffende Zelle der zugehörigen Pivottabelle klicken, wie in Abbildung 9.17 dargelegt):

$$\hat{Y}_i = 425{,}815 - 0{,}653 \cdot X_i + 0{,}000297 \cdot X_i^2.$$

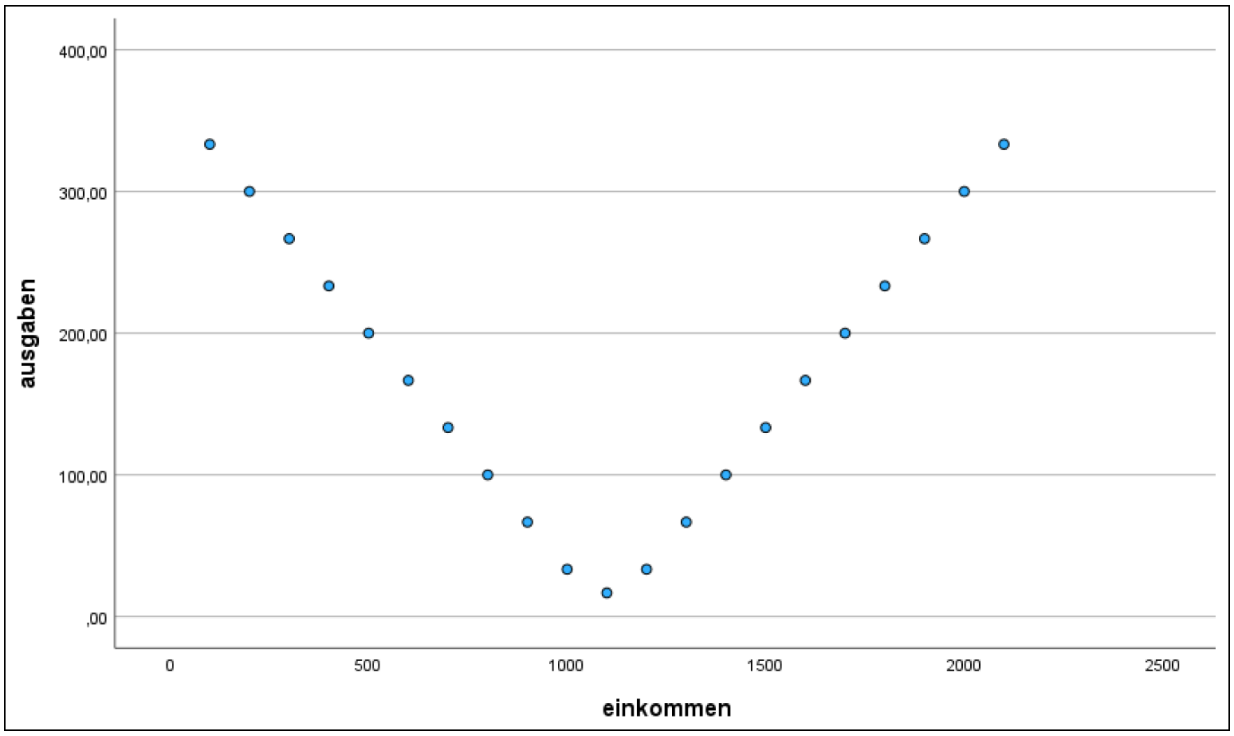



**Abbildung 9.15:** Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_QUADRAT1.SAV*

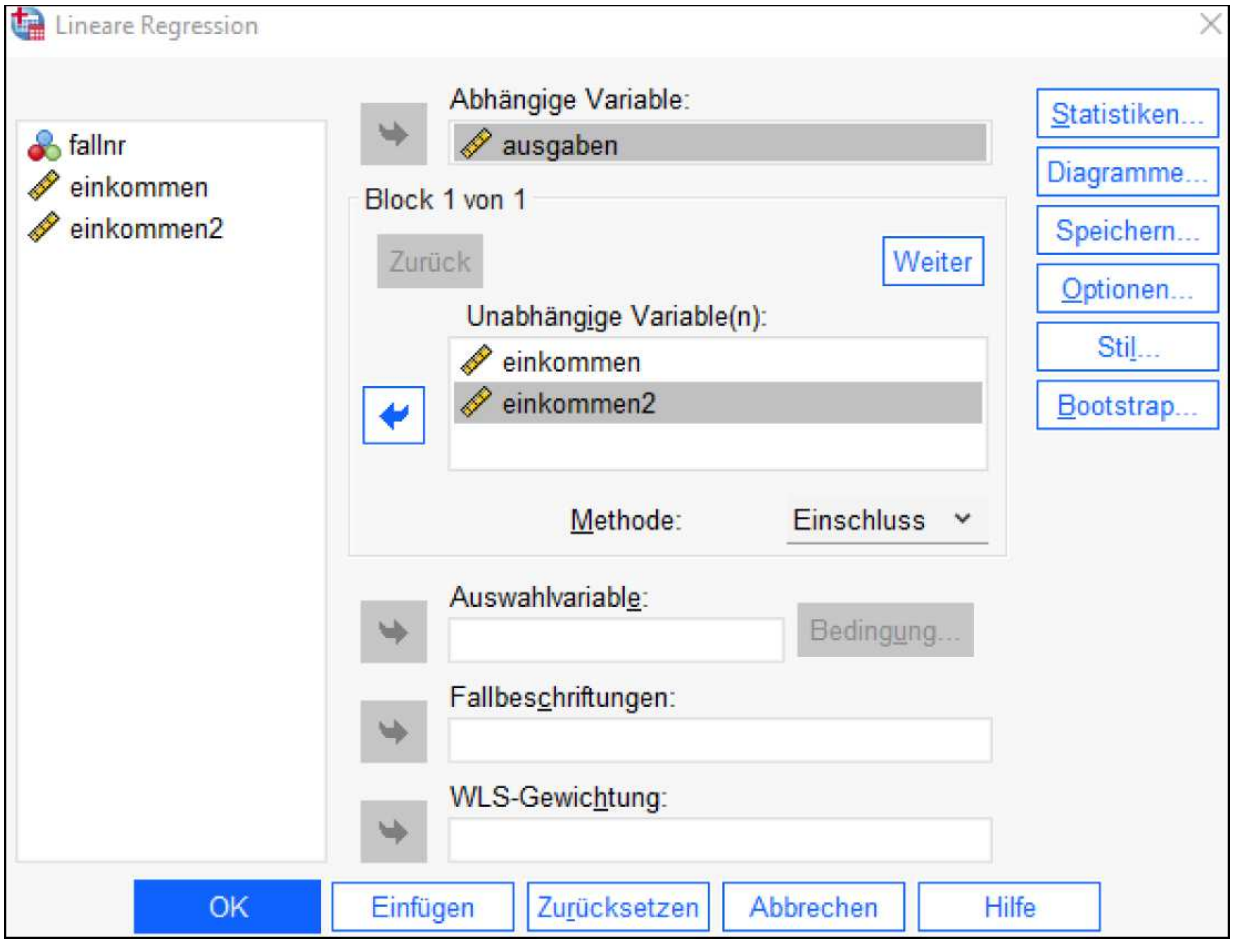



**Abbildung 9.16:** Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben* sowie *einkommen* und *einkommen2* in *LINEAREREGRESSION_QUADRAT1A.SAV*

Eine quadratische Gleichung ist ein Polynom zweiten Grades. Was so furchtbar kompliziert klingt, meint nichts anderes, als dass der höchste X-Exponent gleich 2 ist. Für alle anderen Polynome kann der vorstehend skizzierte Lösungsweg zur Schätzung einer linearen Regressionsbeziehung ebenfalls beschritten werden. Um zum Beispiel ein Polynom dritten Grades $\hat{Y}_i = \hat{a} + \hat{b} \cdot X_i + \hat{c} \cdot X_i^2 + \hat{d} \cdot X_i^3$ zu schätzen, müssen Sie entsprechend zusätzlich zu oben einfach nur noch eine weitere Variable *einkommen3* (= *einkommen***3) berücksichtigen.

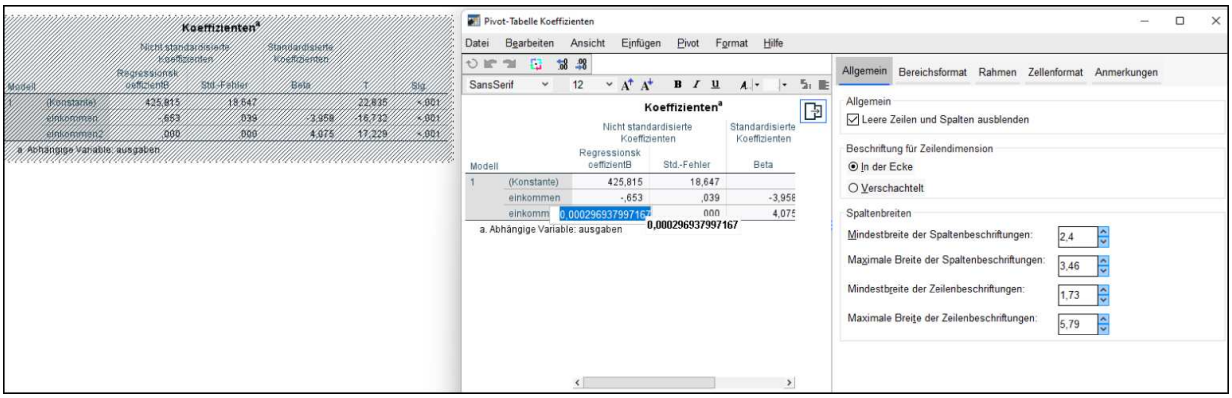



**Abbildung 9.17:** Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* sowie *einkommen* und *einkommen2* in *LINEAREREGRESSION_QUADRAT1.SAV* (Auszug)

# *Potenzfunktion*

Eine **Potenzfunktion** hat als Regressionsbeziehung grundsätzlich folgendes Aussehen:

$$\hat{Y}_i = \hat{a} \cdot X_i^{\hat{b}}.$$

Diese Gleichung lässt sich nach (bezogen auf die verwendete Basis: mehr oder weniger beliebiger) Logarithmierung beider Gleichungsseiten (hier nach natürlicher Logarithmierung) in die Grundform einer linearen Regressionsfunktion überführen:

$$\ln(\hat{Y}_i) = \ln(\hat{a}) + \hat{b} \cdot \ln(X_i).$$

Bei der Logarithmierung ist natürlich zu beachten, dass der Logarithmus einer nichtpositiven Zahl nicht definiert ist.

Auf der Basis der Datendatei *LINEAREREGRESSION_POTENZ1.SAV* ist die Punktewolke in Abbildung 9.18 gegeben (die hier als Scatterplot über GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM generiert wurde).

Auf Basis dieser Datendatei werden zunächst einmal die logarithmierte abhängige Variable *ausgaben_ln* und die logarithmierte unabhängige Variable *einkommen_ln* berechnet (der neue Datensatz ist für Sie in *LINEAREREGRESSION_POTENZ1A.SAV* bereits aufbereitet worden):

```
COMPUTE ausgaben_ln = ln(ausgaben).
COMPUTE einkommen_ln = ln(einkommen).
EXECUTE.
```

Für diese beiden Variablen wird anschließend via ANALYSIEREN|REGRESSION|LINEAR eine lineare Regressionsbeziehung geschätzt (siehe Abbildung 9.19).

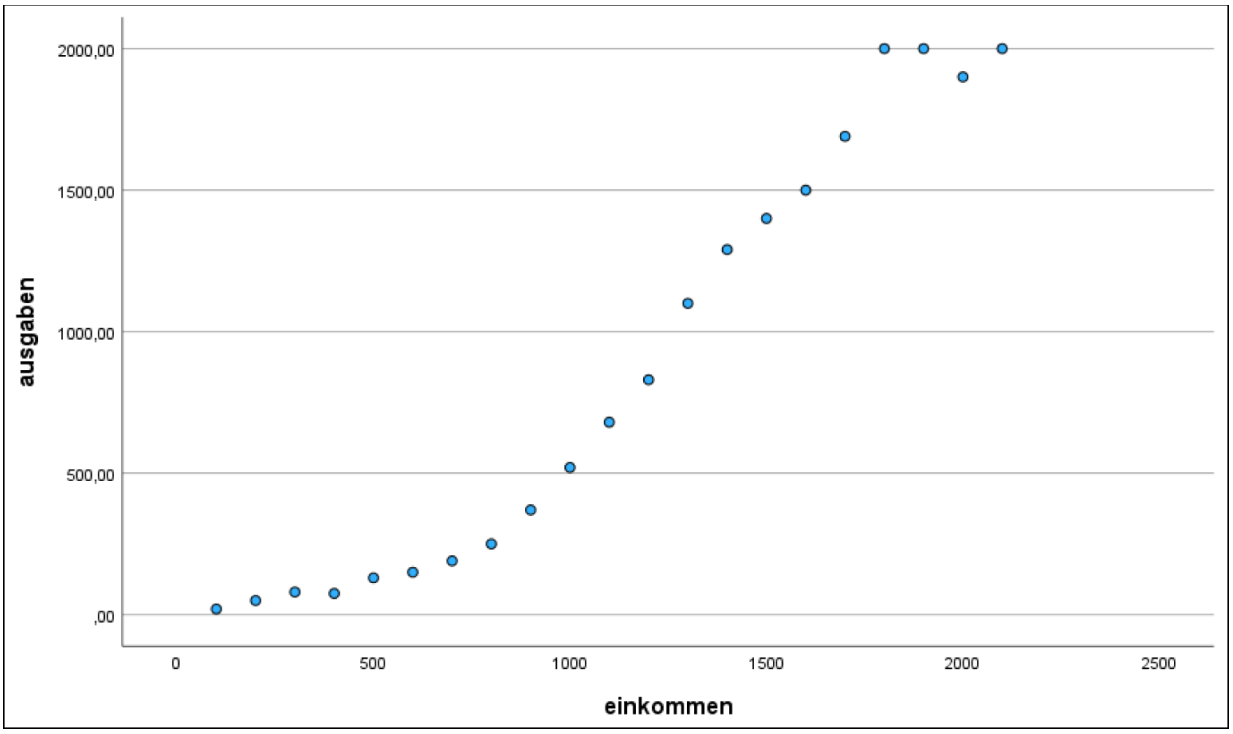



**Abbildung 9.18:** Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_POTENZ1.SAV*

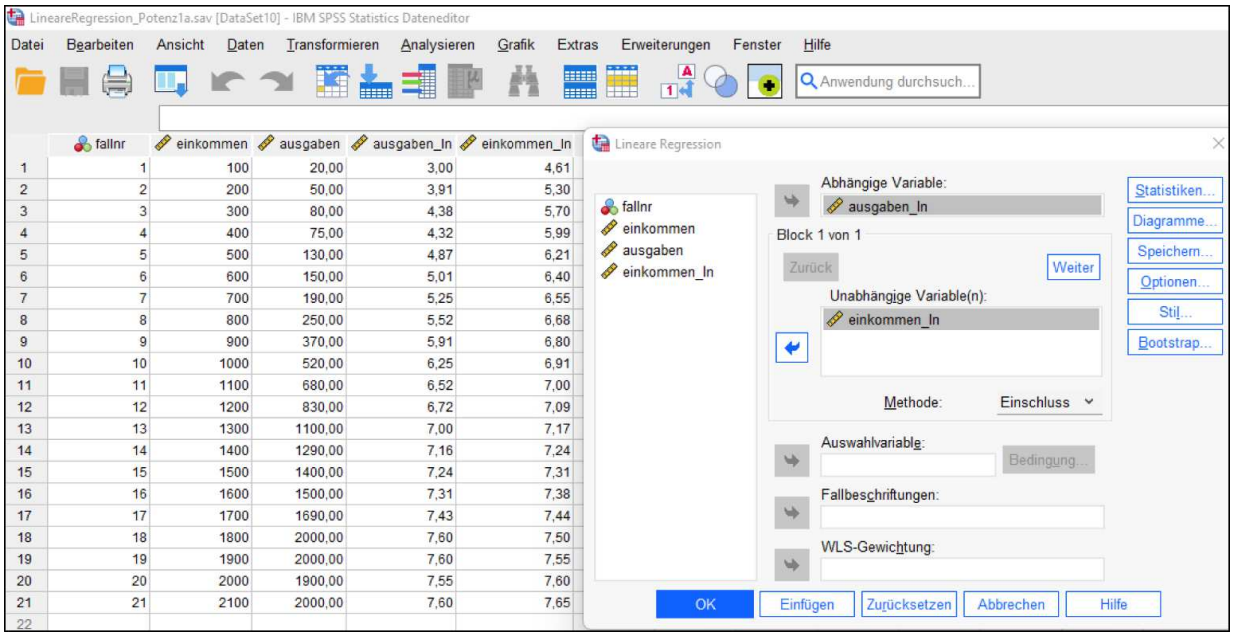



**Abbildung 9.19:** Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben_ln* und *einkommen_ln* in *LINEAREREGRESSION_POTENZ1A.SAV*

Das Ergebnis der Regressionsschätzung ist in Abbildung 9.20 dargelegt. Die geschätzte Gleichung lautet demnach:

$$\ln(\hat{Y}_i) = -5{,}384 + 1{,}698 \cdot \ln(X_i).$$

Inhaltlich bedeutet dies: Erhöht sich das Einkommen um ein Prozent, so erhöhen sich die Ausgaben um circa 1,7 Prozent.

**Regression**

[DataSet10] C:\SPSS-Dateien\LineareRegression_Potenz1a.sav

**Aufgenommene/Entfernte Variablen[a]**

| Modell | Aufgenommene Variablen | Entfernte Variablen | Methode |
|---|---|---|---|
| 1 | einkommen_ln[b] | . | Einschluß |

a. Abhängige Variable: ausgaben_ln

b. Alle gewünschten Variablen wurden eingegeben.

**Modellzusammenfassung**

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|---|---|---|---|---|
| 1 | ,978[a] | ,956 | ,954 | ,30497 |

a. Einflußvariablen : (Konstante), einkommen_ln

**ANOVA[a]**

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 38,536 | 1 | 38,536 | 414,333 | <,001[b] |
| | Nicht standardisierte Residuen | 1,767 | 19 | ,093 | | |
| | Gesamt | 40,303 | 20 | | | |

a. Abhängige Variable: ausgaben_ln

b. Einflußvariablen : (Konstante), einkommen_ln

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | RegressionskoeffizientB | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | -5,384 | ,568 | | -9,474 | <,001 |
| | einkommen_ln | 1,698 | ,083 | ,978 | 20,355 | <,001 |

a. Abhängige Variable: ausgaben_ln

**Abbildung 9.20:** Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben_ln* und *einkommen_ln* in *LINEAREREGRESSION_POTENZ1A.SAV*

Um die eigentliche (Potenz-)Funktionsform zu erhalten, setzen Sie bitte zunächst $\ln(\hat{a}) = -5{,}384$ und $\hat{b} = 1{,}698.$ Jetzt müssen Sie nur noch beachten, dass Sie ja nicht an der Schätzung des (natürlichen) Logarithmus von $\hat{a}$, sondern an $\hat{a}$ direkt interessiert sind. Über exp(-5,384) erhalten Sie für $\hat{a}$ dann den Wert 0,0046, (exp(…) verwenden Sie hier, weil wir ja auf den natürlichen Logarithmus ln(…) Bezug genommen hatten), sodass Ihre geschätzte Regressionsbeziehung letztlich lautet:

$$\hat{Y}_i = 0{,}0046 \cdot X_i^{1{,}698}.$$

# *Exponentialfunktion*

Bei einer **Exponentialfunktion** ist folgende Regressionsbeziehung zu schätzen:

$$\hat{Y}_i = \hat{a} \cdot \hat{b}^{X_i}.$$

Logarithmieren Sie beide Seiten, erhalten Sie auch hier die Grundform einer linearen Regressionsfunktion:

$$\ln(\hat{Y}_i) = \ln(\hat{a}) + X_i \cdot \ln(\hat{b}).$$

Öffnen Sie bitte in diesem Zusammenhang die Datendatei *LINEAREREGRESSION_EXPONENTIAL1.SAV*. Für diesen Datensatz ergibt sich in Bezug auf die Variablen *ausgaben* und *einkommen* der in Abbildung 9.21 dargestellte Scatterplot (der im Übrigen dem Scatterplot auf Basis von *LINEAREREGRESSION_POTENZ1.SAV* in Abbildung 9.18 nicht unähnlich ist; die Werte von *ausgaben* sind in *LINEAREREGRESSION_POTENZ1.SAV* etwas größer als die von *ausgaben* in *LINEAREREGRESSION_EXPONENTIAL1.SAV*).

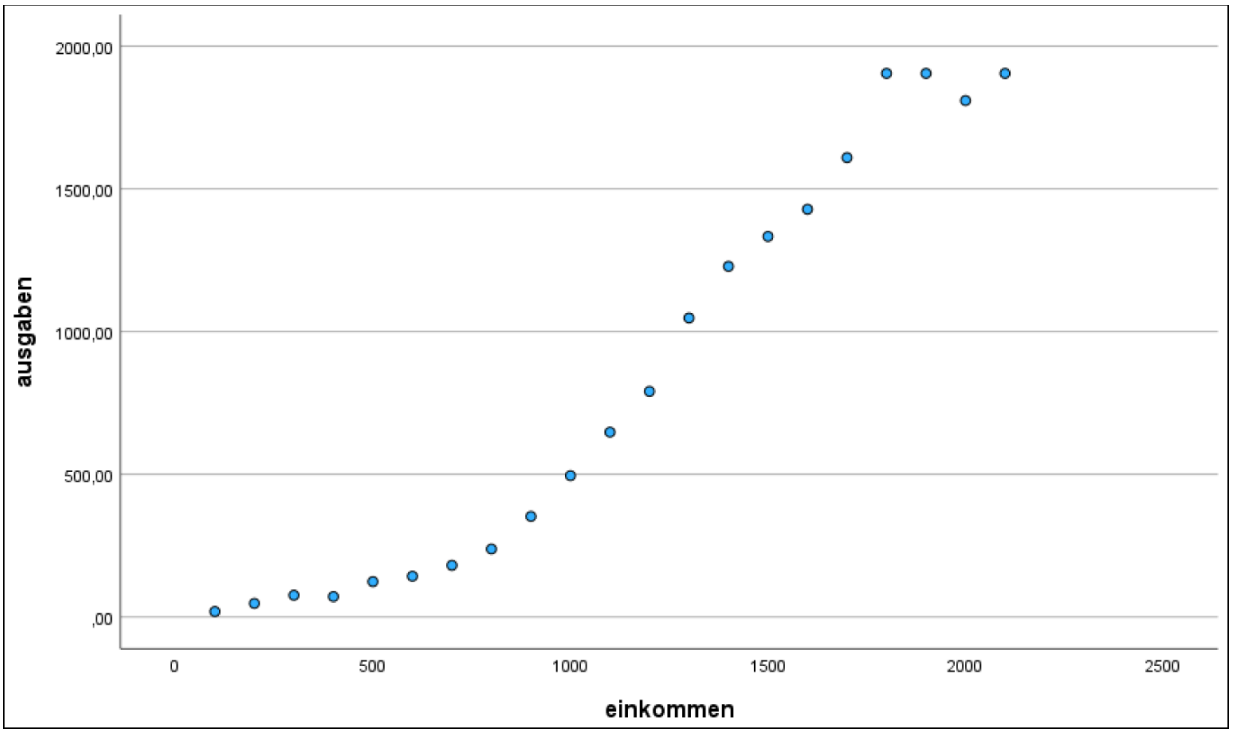



**Abbildung 9.21:** Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_EXPONENTIAL1.SAV*

Auf der Grundlage dieser Datendatei wird die logarithmierte abhängige Variable *ausgaben_ln* berechnet (die neue Datendatei habe ich Ihnen als *LINEAREREGRESSION_EXPONENTIAL1A.SAV* zur Verfügung gestellt):

```
COMPUTE ausgaben_ln = ln(ausgaben).
EXECUTE.
```

Für *ausgaben_ln* und *einkommen* wird dann mittels ANALYSIEREN|REGRESSION|LINEAR eine lineare Regressionsbeziehung geschätzt (siehe auch Abbildung 9.22):

$$\ln\left(\hat{Y}_i\right) = 3,626 + 0,002 \cdot X_i .$$

Die vorstehende Gleichung drückt inhaltlich Folgendes aus: Erhöht sich das Einkommen (die X-Variable) um eine (Geld-)Einheit, dann erhöhen sich die Ausgaben (die Y-Variable) um 0,2 Prozent.

**Aufgenommene/Entfernte Variablen[a]**

| Modell | Aufgenommene Variablen | Entfernte Variablen | Methode |
|---|---|---|---|
| 1 | einkommen[b] | . | Einschluß |

a. Abhängige Variable: ausgaben_ln

b. Alle gewünschten Variablen wurden eingegeben.

**Modellzusammenfassung**

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|---|---|---|---|---|
| 1 | ,965[a] | ,931 | ,927 | ,38232 |

a. Einflußvariablen : (Konstante), einkommen

**ANOVA[a]**

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 37,526 | 1 | 37,526 | 256,737 | <,001[b] |
| | Nicht standardisierte Residuen | 2,777 | 19 | ,146 | | |
| | Gesamt | 40,303 | 20 | | | |

a. Abhängige Variable: ausgaben_ln

b. Einflußvariablen : (Konstante), einkommen

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | Regressionskoeffizient B | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | 3,626 | ,173 | | 20,961 | <,001 |
| | einkommen | ,002 | ,000 | ,965 | 16,023 | <,001 |

a. Abhängige Variable: ausgaben_ln

**Abbildung 9.22:** Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben_ln* und *einkommen* in *LINEAREREGRESSION_EXPONENTIAL1A.SAV*

Nun müssen Sie $\ln(\hat{a}) = 3{,}626$ und $\ln(\hat{b}) = 0{,}002$ nach $\hat{a}$ und $\hat{b}$ auflösen, und zwar über die Verwendung von exp (…): $\hat{a} = \exp(3{,}626) = 37{,}562$ und $\hat{b} = \exp(0{,}002) = 1{,}002,$ sodass die geschätzte Regressionsbeziehung folgendermaßen aussieht:

$$Y_i = 37{,}562 \cdot 1{,}002^{X_i}.$$

# *Logistische Funktion*

Beispielsweise bei Sättigungseffekten im Rahmen der Konsumnachfrage bietet es sich an, eine obere Schranke K für die abhängige Variable zu modellieren (im Beispiel: bei einem bestimmten Konsumniveau für ein Gut). Dabei ist diese Obergrenze für $\hat{Y}_i$ von ihrem Wert her vorzugeben. Man kann derartige Zusammenhänge gut durch eine **logistische Funktion** abbilden:

$$\hat{Y}_i = \frac{K}{1 + e^{\hat{a} + \hat{b} \cdot X_i}} \quad \text{mit: } \hat{b} < 0 \text{ und e (Euler'sche Zahl) = 2,71828…}$$

Bitte verwechseln Sie die hier präsentierte Regressionsschätzung nicht mit der in Kapitel 10 vorgestellten binär-logistischen Regressionsschätzung! Während bei der hier relevanten Schätzung einer logistischen Funktion die abhängige Y-Variable eine metrisch skalierte Variable ist, ist die abhängige Y-Variable bei der binär-logistischen Regression ein 0/1-Dummy.

Die eben dargelegte Gleichung kann folgendermaßen in Richtung einer einfachen linearen Regressionsbeziehung umgeformt werden:

$$\hat{Y}_i = \frac{K}{1 + e^{\hat{a} + \hat{b} \cdot X_i}}$$

$$\Leftrightarrow 1 + e^{\hat{a} + \hat{b} \cdot X_i} = \frac{K}{\hat{Y}_i}$$

$$\Leftrightarrow e^{\hat{a} + \hat{b} \cdot X_i} = \frac{K}{\hat{y}_i} - 1$$

$$\Leftrightarrow \ln\left(e^{\hat{a} + \hat{b} \cdot X_i}\right) = \ln\left(\frac{K}{\hat{Y}_i} - 1\right)$$

$$\Leftrightarrow \ln\left(\frac{K}{\hat{Y}_i} - 1\right) = \hat{a} + \hat{b} \cdot X_i.$$

Hierbei wird demnach $\hat{Y}_i$ durch $\ln\left(\frac{K}{\hat{Y}_i} - 1\right)$ ersetzt.

Weil der Logarithmus von nichtpositiven Zahlen nicht definiert ist, ist darauf zu achten, dass das Verhältnis aus der gesetzten Grenze K und den Y-Werten nicht kleiner gleich 1 ist. Daher sollten Sie die Grenze K höher als den größten Beobachtungswert für Y ansetzen, sofern es sich bei diesem nicht um einen Ausreißerwert handelt (weil Sie diesen Wert dann gegebenenfalls (auch) aus Ihren Betrachtungen ausschließen sollten).

Mithilfe der Datendatei *LINEAREREGRESSION_LOGISTISCH1.SAV* sieht der dank GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM ermittelte Scatterplot so aus, dass eine Sättigung bei einem Ausgabenniveau von 600 Euro stattfindet (siehe Abbildung 9.23). K wird daher im Folgenden gleich 600 gesetzt.

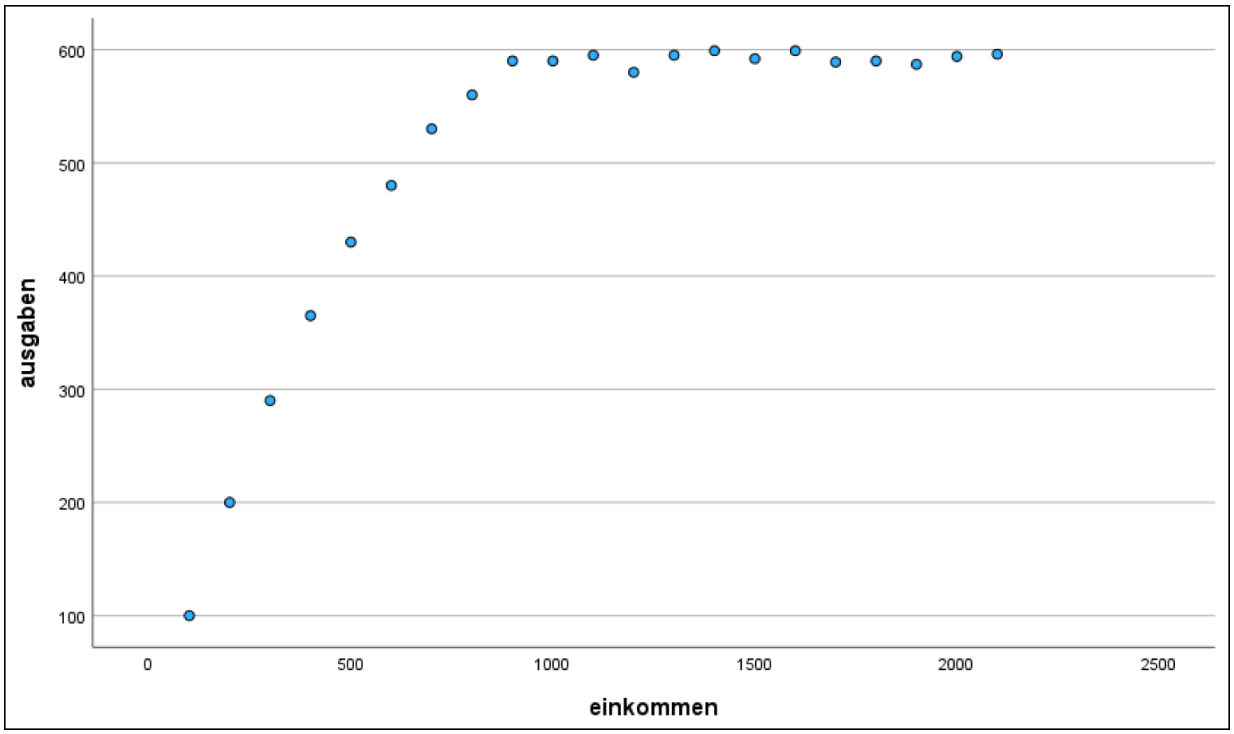



**Abbildung 9.23:** Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_LOGISTISCH1.SAV*

Berechnen Sie nun in dieser Datendatei als neue abhängige Variable die Variable *ausgaben_logistisch*, wobei der neue Datensatz für Sie als *LINEAREREGRESSION_LOGISTISCH1A.SAV* aufbereitet worden ist:

```
COMPUTE ausgaben_logistisch = ln((600/ausgaben)-1).
EXECUTE.
```

Zwischen *ausgaben_logistisch* und *einkommen* wird anschließend über ANALYSIEREN|REGRESSION|LINEAR die klassische lineare Regression

durchgeführt. Diese ergibt im Beispiel folgende Regressionsparameter: $\hat{a} = 0{,}272$ und $\hat{b} = -0{,}003$ (siehe Abbildung 9.24).

Damit sieht in diesem Beispiel die geschätzte Regressionsbeziehung folgendermaßen aus:

$$\hat{Y}_i = \frac{600}{1 + e^{0{,}272 - 0{,}003 \cdot X_i}}.$$

# *Lineare und quasilineare Regressionen gleichzeitig*

SPSS bietet für insgesamt 11 Funktionsformen lineare Regressionsschätzungen an, darunter die vorstehend behandelten Arten der Geradengleichung, der quadratischen Funktion, der Potenz-, der Exponential- und der logistischen Funktion. Sie müssen in diesem Zusammenhang nur ANALYSIEREN|REGRESSION|KURVENANPASSUNG auswählen.

**Aufgenommene/Entfernte Variablen[a]**

| Modell | Aufgenommene Variablen | Entfernte Variablen | Methode |
|---|---|---|---|
| 1 | einkommen[b] | . | Einschluß |

a. Abhängige Variable: ausgaben_logistisch

b. Alle gewünschten Variablen wurden eingegeben.

**Modellzusammenfassung**

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|---|---|---|---|---|
| 1 | ,839[a] | ,704 | ,688 | 1,25793 |

a. Einflußvariablen : (Konstante), einkommen

**ANOVA[a]**

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 71,503 | 1 | 71,503 | 45,187 | <,001[b] |
| | Nicht standardisierte Residuen | 30,066 | 19 | 1,582 | | |
| | Gesamt | 101,569 | 20 | | | |

a. Abhängige Variable: ausgaben_logistisch

b. Einflußvariablen : (Konstante), einkommen

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | RegressionskoeffizientB | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | ,272 | ,569 | | ,477 | ,639 |
| | einkommen | -,003 | ,000 | -,839 | -6,722 | <,001 |

a. Abhängige Variable: ausgaben_logistisch

**Abbildung 9.24**: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben_logistisch* und *einkommen* in *LINEAREREGRESSION_LOGISTISCH1A.SAV*

Um diese Möglichkeit der alternativen Regressionsschätzung zu illustrieren, beziehe ich mich an dieser Stelle auf die Datendatei *LINEAREREGRESSION_DIVERSE1.SAV*. Über GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM gewinnen Sie dabei zwischen *ausgaben* und *einkommen* den Scatterplot gemäß Abbildung 9.25.

So, und jetzt öffnen Sie bitte mittels ANALYSIEREN|REGRESSION|KURVENANPASSUNG die Subdialogbox, mit deren Hilfe Sie verschiedene alternative Funktionsformulierungen auswählen können. Klicken Sie doch einfach mal alle Funktionsformen an und geben Sie bei der logistischen Regression die Obergrenze von 400 an (siehe Abbildung 9.26).

Die zugehörige Syntax ist (mit der Sie auch noch in der letzten Befehlszeile ein Übersichtsdiagramm anfordern):

```
CURVEFIT
    /VARIABLES=ausgaben WITH einkommen
    /CONSTANT
    /MODEL=LINEAR LOGARITHMIC INVERSE QUADRATIC CUBIC COMPOUND POWER S GROWTH
EXPONENTIAL LGSTIC
    /UPPERBOUND=400
    /PLOT FIT.
```

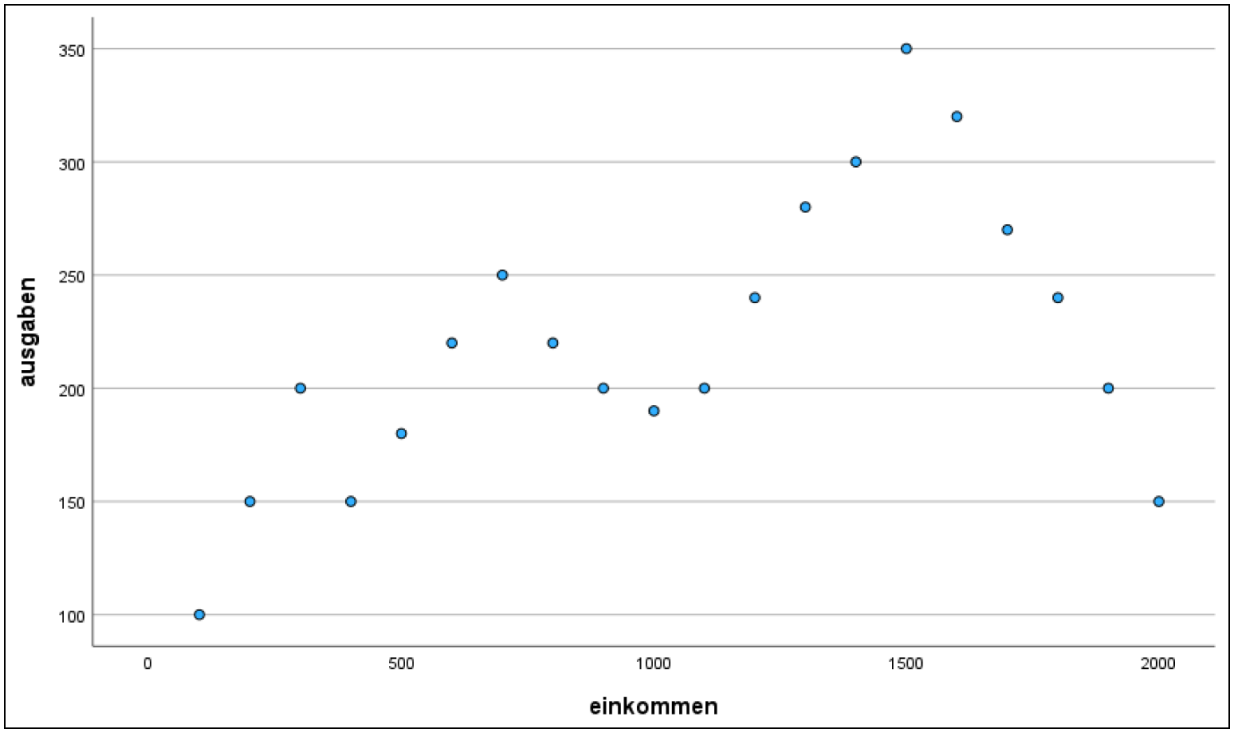



**Abbildung 9.25**: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_DIVERSE1.SAV*

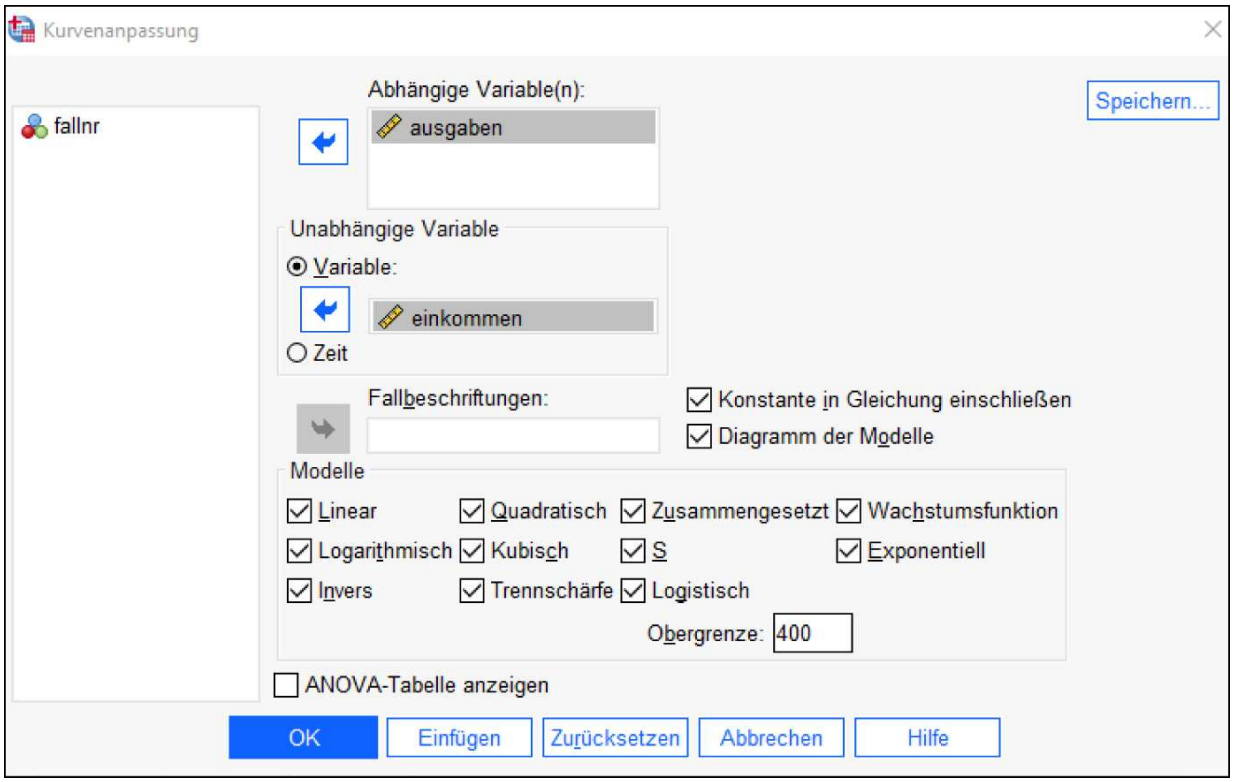



**Abbildung 9.26:** Anfordern verschiedener linearer Regressionsbeziehungen zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_DIVERSE1.SAV*

Wenn Sie nun diesen Befehl ausführen beziehungsweise den entsprechenden Menüpunkt aufrufen, erhalten Sie einen umfangreichen SPSS-Output (siehe Abbildung 9.27 und Abbildung 9.28). Zunächst wird in Abbildung 9.27 in Form einer »Modell-Beschreibung« jede der 11 Varianten aufgeführt. Danach und nach einer »Zusammenfassung der Fallverarbeitung« und nach einer »Zusammenfassung der Variablenverarbeitung« erhalten Sie die geschätzten Koeffizienten der einzelnen Gleichungen. Abschließend werden in Abbildung 9.28 die betreffenden Schätzfunktionen für den zugrunde liegenden Datensatz noch einmal grafisch dargestellt – vor allem im Vergleich zur Lage der ursprünglichen Datenpunkte. Von den geschätzten Funktionen weist im Übrigen die kubische Funktion (das heißt: ein Polynom dritten Grades) eine besonders hohe Erklärungskraft auf, was angesichts des zyklischen Verlaufs der Punktewolke nicht überrascht (wenngleich die Bestimmtheitsmaße nur bedingt miteinander vergleichbar sind, weil als abhängige Variable nicht immer *ausgaben* gewählt worden ist, sondern zum Beispiel *ausgaben_ln*).

**Modell-Beschreibung**

| | | |
|---|---|---|
| Modell | | MOD_1 |
| Abhängige Variable | 1 | ausgaben |
| Gleichung | 1 | Linear |
| | 2 | Logarithmisch |
| | 3 | Invers |
| | 4 | Quadratisch |
| | 5 | Kubisch |
| | 6 | Zusammengesetzt[a] |
| | 7 | Netzstrom[a] |
| | 8 | S[a] |
| | 9 | Wachstumsfunktion[a] |
| | 10 | Exponentiell[a] |
| | 11 | Logistisch[a,b] |
| Unabhängige Variable | | einkommen |
| Konstante | | Eingeschlossen |
| Variable zur Beschriftung der Beobachtungen in Diagrammen | | Nicht angegeben |
| Toleranz für die Aufnahme von Termen in Gleichungen | | ,0001 |

a. Das Modell erfordert, dass alle nicht-fehlenden Werte positiv sind.

b. Für alle abhängigen Variablen wird die theoretische Obergrenze auf 400 gesetzt.

**Zusammenfassung der Variablenverarbeitung**

| | | Variablen | |
|---|---|---|---|
| | | Abhängige Variable ausgaben | Unabhängige Variable einkommen |
| Anzahl der positiven Werte | | 20 | 20 |
| Anzahl der Nullen | | 0 | 0 |
| Anzahl der negativen Werte | | 0 | 0 |
| Anzahl der fehlenden Werte | Benutzerdefiniert fehlend | 0 | 0 |
| | Systemdefiniert fehlend | 0 | 0 |

**Zusammenfassung der Fallverarbeitung**

| | N |
|---|---|
| Gesamtzahl der Fälle | 20 |
| Ausgeschlossene Fälle[a] | 0 |
| Prognostizierte Fälle | 0 |
| Neu erstellte Fälle | 0 |

a. Fälle mit einem fehlenden Wert in einer Variablen werden aus der Analyse ausgeschlossen.

**Modellzusammenfassung und Parameterschätzer**

Abhängige Variable: ausgaben

| | Modellübersicht | | | | | Parameterschätzungen | | | |
|---|---|---|---|---|---|---|---|---|---|
| Gleichung | R-Quadrat | F | df1 | df2 | Sig. | Konstante | b1 | b2 | b3 |
| Linear | ,281 | 7,029 | 1 | 18 | ,016 | 161,684 | ,056 | | |
| Logarithmisch | ,411 | 12,573 | 1 | 18 | ,002 | -111,176 | 49,342 | | |
| Invers | ,388 | 11,392 | 1 | 18 | ,003 | 251,836 | -17419,560 | | |
| Quadratisch | ,548 | 10,309 | 2 | 17 | ,001 | 79,798 | ,279 | ,000 | |
| Kubisch | ,659 | 10,330 | 3 | 16 | <,001 | 152,561 | -,092 | ,000 | -1,370E-7 |
| Zusammengesetzt | ,293 | 7,465 | 1 | 18 | ,014 | 158,627 | 1,000 | | |
| Netzstrom | ,485 | 16,957 | 1 | 18 | <,001 | 37,524 | ,257 | | |
| S | ,532 | 20,430 | 1 | 18 | <,001 | 5,531 | -97,988 | | |
| Wachstumsfunktion | ,293 | 7,465 | 1 | 18 | ,014 | 5,067 | ,000 | | |
| Exponentiell | ,293 | 7,465 | 1 | 18 | ,014 | 158,627 | ,000 | | |
| Logistisch | ,272 | 6,729 | 1 | 18 | ,018 | ,004 | ,999 | | |

Die unabhängige Variable ist einkommen.

**Abbildung 9.27:** Ergebnisse verschiedener linearer Regressionsbeziehungen zwischen *ausgaben* sowie *einkommen* in *LINEAREREGRESSION_DIVERSE1.SAV* (1)

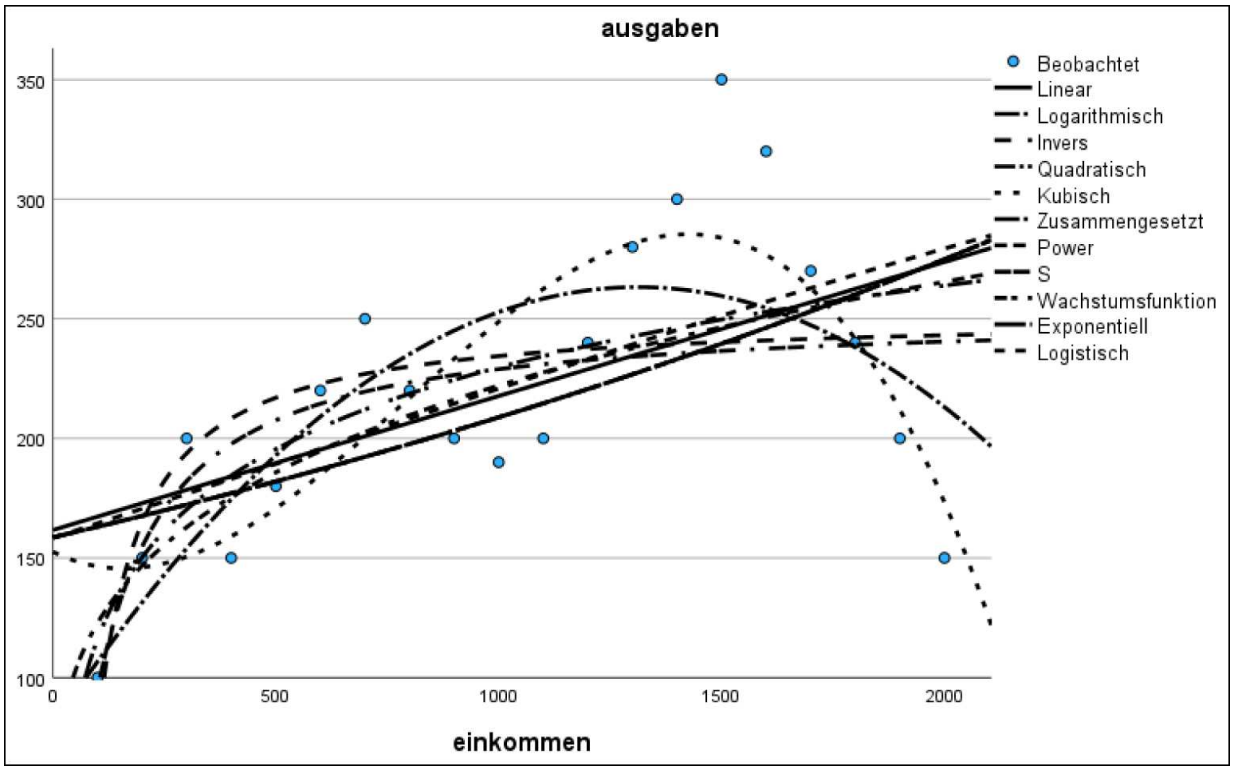



**Abbildung 9.28:** Ergebnisse verschiedener linearer Regressionsbeziehungen zwischen *ausgaben* sowie *einkommen* in *LINEAREREGRESSION_DIVERSE1.SAV* (2)

# *Modifikationen/Erweiterung der linearen Regressionsanalyse*

Die vorstehend betrachteten Regressionstechniken können in verschiedener Hinsicht modifiziert beziehungsweise erweitert werden. Im Folgenden werden als entsprechende Themen die Erweiterung um 0/1-Dummies als unabhängige X-Variablen, eine Zeitvariable als unabhängige X-Variable mit dem zugehörigen Problem der Autokorrelation, Intervallschätzungen für die Regressionskoeffizienten sowie die Residuenanalyse und hier speziell die Heteroskedastizitätsproblematik angesprochen.

# *0/1-Dummies als unabhängige Variablen*

Sie können in einer Regressionsberechnung auch nominal- und ordinalskalierte Variablen als unabhängige Variablen (also auf der »rechten Seite« der Regressionsgleichung) berücksichtigen, indem Sie sie als **0/1-Dummy-Variablen** ausgestalten. Beim Geschlecht würden Sie etwa der Gruppe der Frauen den Variablenwert 0 und der Gruppe der Männer den Variablenwert 1 zuweisen.

## *Grundsätzliches/Interpretationen*

Ein 0/1-Dummy kann in SPSS ganz leicht mithilfe des COMPUTE-Befehls gebildet werden, im Beispiel des Datensatzes *LINEAREREGRESSION_DUMMY.SAV* für eine Altersabgrenzung in unter 40-Jährige (Ausprägung »1«) und in 40-Jährige und Ältere (Ausprägung »0«) auf Basis der ursprünglichen Variablen *alter*:

```
COMPUTE alter_dummy = 0.
COMPUTE alter_dummy = alter < 40.
EXECUTE.
```

In der zweiten Zeile setzen Sie mit dem dortigen Befehl die Dummy-Variable auf den Wert »1«. Alternativ hätten Sie an Syntax auch eingeben können:

```
COMPUTE alter_dummy = 0.
IF (alter < 40) alter_dummy = 1.
EXECUTE.
```

Im Beispiel erhalten also alle Personen im Alter von 40 und mehr Lebensjahren den Wert 0 in der Dummy-Variablen *alter_dummy*. Demgegenüber ist einem Alter kleiner 40 Lebensjahren mittels des jeweils zweiten oben stehenden Befehls in *alter_dummy* der Wert 1 zugeordnet.

Der Zusammenhang zwischen *ausgaben* und *alter* in *LINEAREREGRESSION_DUMMY.SAV* sieht als Scatterplot, der mittels GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM erzeugt worden ist, wie in Abbildung 9.29 aus. Ziehen Sie nunmehr bei 40 Lebensjahren in der nachfolgenden Abbildung gedanklich einen

senkrechten Strich, sind alle Punkte links von diesem Strich den Personen unter 40 Jahren zugeordnet, alle Punkte auf dem Strich beziehungsweise rechts von diesem Strich hingegen Personen im Alter von 40 und mehr Jahren. Anhand der nachfolgenden Punktewolke ist ersichtlich, dass die jüngeren Personen (unter 40 Jahren) tendenziell niedrigere Ausgaben als die älteren Personen (40 Jahre und mehr) haben.

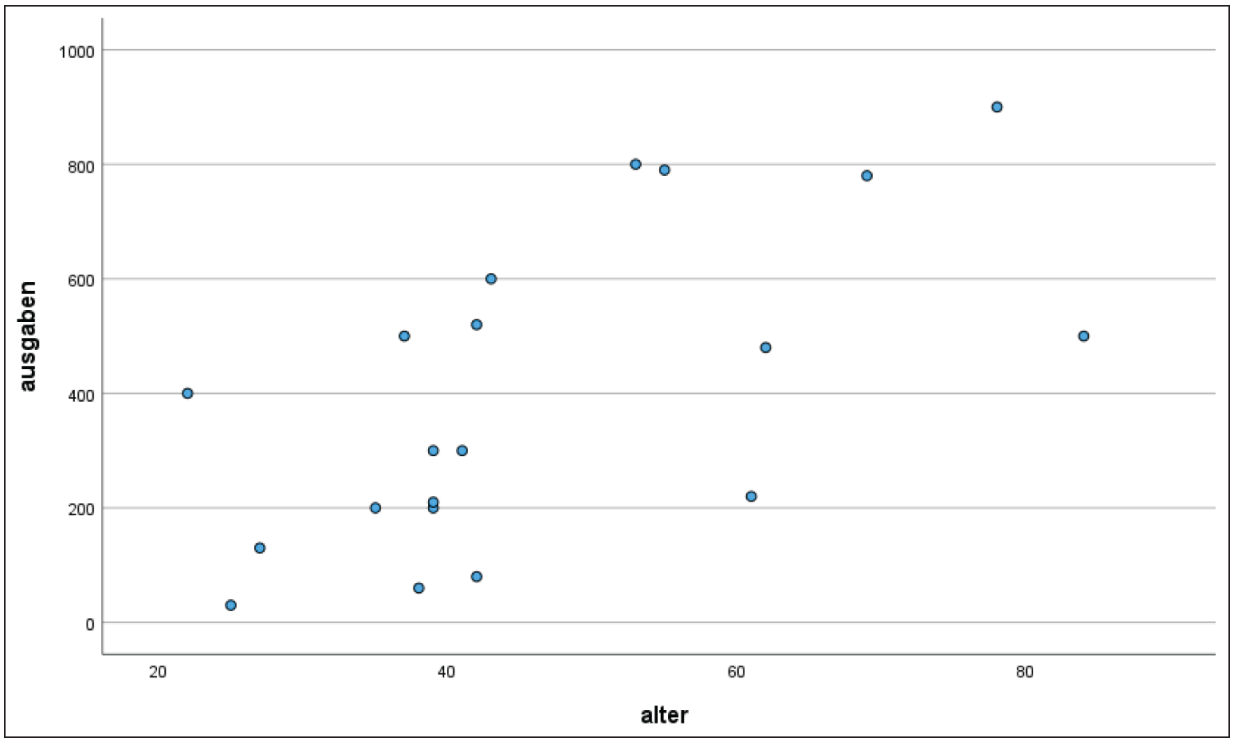



**Abbildung 9.29:** Scatterplot zwischen *ausgaben* und *alter* in der Datendatei *LINEAREREGRESSION_DUMMY.SAV*

Der letztgenannte Zusammenhang müsste sich in einer Regressionsbeziehung zwischen *ausgaben* und *alter_dummy* widerspiegeln. Zu diesem Zweck ist über ANALYSIEREN|REGRESSION|LINEAR zwischen diesen beiden Variablen in der aktualisierten Datendatei *LINEAREREGRESSION_DUMMY1.SAV* eine entsprechende lineare Regression geschätzt worden (mit *ausgaben* als abhängiger Variable und *alter_dummy* als unabhängiger Variable; siehe Abbildung 9.30). Die betreffende SPSS-Schätzung ergibt ein absolutes Glied von 542,727 und einen der Dummyvariablen zugeordneten Parameter von -317,172. Inhaltlich bedeutet dies, dass bei einem Dummywert in Höhe von 0, das heißt für Personen im Alter von 40 und mehr Jahren, die durchschnittlichen Ausgaben bei 542,73 Euro und bei einem Dummywert in Höhe von 1, das heißt für Personen im Alter unter 40 Jahren, um 317,17 Euro niedriger, also bei 225,56 Euro, liegen. Eine 0/1-Dummyvariable wirkt sich demzufolge auf die Höhe des absoluten Glieds aus. Dass es sich hierbei um Durchschnittswerte, in diesem Fall der Ausgaben, handelt, ergibt sich daraus, dass die mittels OLS-Verfahren geschätzte Regressionsgerade mittlere Zusammenhänge zwischen den an der Regression beteiligten Variablen widerspiegelt.

**Aufgenommene/Entfernte Variablen[a]**

| Modell | Aufgenommene Variablen | Entfernte Variablen | Methode |
|---|---|---|---|
| 1 | alter_dummy[b] | . | Einschluß |

a. Abhängige Variable: ausgaben

b. Alle gewünschten Variablen wurden eingegeben.

**Modellzusammenfassung**

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|---|---|---|---|---|
| 1 | ,600[a] | ,360 | ,324 | 221,816 |

a. Einflußvariablen : (Konstante), alter_dummy

**ANOVA[a]**

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 497959,596 | 1 | 497959,596 | 10,121 | ,005[b] |
| | Nicht standardisierte Residuen | 885640,404 | 18 | 49202,245 | | |
| | Gesamt | 1383600,000 | 19 | | | |

a. Abhängige Variable: ausgaben

b. Einflußvariablen : (Konstante), alter_dummy

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | RegressionskoeffizientB | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | 542,727 | 66,880 | | 8,115 | <,001 |
| | alter_dummy | -317,172 | 99,699 | -,600 | -3,181 | ,005 |

a. Abhängige Variable: ausgaben

**Abbildung 9.30:** Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* und *alter_dummy* in *LINEAREREGRESSION_DUMMY1.SAV*

Zur Überprüfung der vorstehenden Aussagen können Sie sich mittels ANALYSIEREN|MITTELWERTE UND PROPORTIONEN VERGLEICHEN|MITTELWERTE und *ausgaben* als abhängiger Variable und *alter_dummy* als SCHICHT die entsprechenden arithmetischen Mittelwerte »auswerfen« lassen. In der Tat liefert eine derartige SPSS-Berechnung die obigen Durchschnittswerte in Höhe von 542,73 Euro für die 40-Jährigen und Älteren und in Höhe von 225,56 Euro für die unter 40-Jährigen (siehe Abbildung 9.31).

**Bericht**

ausgaben

| alter_dummy | Mittelwert | N | Std.-Abweichung |
|---|---|---|---|
| ,00 | 542,73 | 11 | 264,200 |
| 1,00 | 225,56 | 9 | 153,143 |
| Insgesamt | 400,00 | 20 | 269,854 |

**Abbildung 9.31:** Ausgaben-Durchschnittswerte für zwei Altersgruppen auf Basis von *LINEAREREGRESSION_DUMMY.SAV*

Die hier durchgeführte Regression mit ausschließlich einem 0/1-Dummy als unabhängiger X-Variable hat in erster Linie didaktische Gründe. In der Praxis finden Sie 0/1-Dummies in der Regel in Kombination mit metrischen Größen als weiteren unabhängigen X-Variablen vor.

## *Mehrere Dummy-Variablen*

Im vorstehenden Beispiel haben Sie lediglich zwei Altersklassen voneinander unterschieden. Natürlich können Sie aber mehr Altersklassen in Ihre Berechnungen einbeziehen. Nehmen Sie bitte mal an, dass Sie auf Basis von *LINEAREREGRESSION_DUMMY.SAV* einen Regressionszusammenhang zwischen den Ausgaben als abhängiger Größe und dem Einkommen sowie dem Alter als unabhängige Variablen herstellen möchten. Okay, es handelt sich hier eigentlich schon um eine multiple Regression mit mehreren unabhängigen X-Variablen. Lassen Sie mich aber bitte aus didaktischen Gründen diesen Fall bereits an dieser Stelle erklären.

Nehmen Sie des Weiteren an, dass Sie drei Altersklassen voneinander unterscheiden möchten: unter 40 Jahre, 40 Jahre bis unter 60 Jahre sowie 60 Jahre und älter. Sie könnten nun für jede Altersgruppe drei getrennte Regressionen im Beispiel zwischen *ausgaben* und *einkommen* durchführen. Einfacher geht es aber mit einer einzigen Regression, in der Sie die Altersklassen als 0/1-Dummies berücksichtigen. Wenn Sie zum Beispiel die unter 40-Jährigen als Referenzgruppe, das heißt als Bezugsgruppe, annehmen, müssen Sie in diesem Fall zwei 0/1-Dummies bilden, und zwar für die 40- bis unter 60-Jährigen und für die 60-Jährigen und Älteren jeweils mit dem Wert 1, wenn die betreffende Altersgruppe gemeint ist:

```
COMPUTE alter_40_59 = 0.
IF (alter >= 40 and alter < 60) alter_40_59 = 1.
COMPUTE alter_60plus = 0.
```

```
IF (alter >= 60) alter_60plus = 1.
EXECUTE.
```

Im Folgenden wird exemplarisch (in der aktualisierten Datendatei *LINEAREREGRESSION_DUMMY1A.SAV*) eine Regressionsbeziehung zwischen *ausgaben* als Y-Variable sowie *alter_40_59* und *alter_60plus* als X-Variablen über ANALYSIEREN|REGRESSION|LINEAR in die Tat umgesetzt (siehe Abbildung 9.32).

Welche Koeffizientenwerte dürfen Sie für *alter_40_59* und *alter_60plus* erwarten? Wenn Sie sich noch einmal den Scatterplot zwischen *ausgaben* und *alter* in Abbildung 9.29 anschauen, sind positive Regressionsparameter für die beiden genannten 0/1-Dummies zu erwarten, da sich aus Abbildung 9.29 für *ausgaben* und *alter* ein tendenziell positiver Zusammenhang darstellt. Um dies zu erkennen, müssen Sie in der obigen Abbildung 9.29 einfach bei 40 Jahren und jetzt auch bei 60 Jahren gedanklich jeweils eine senkrechte Linie ziehen.

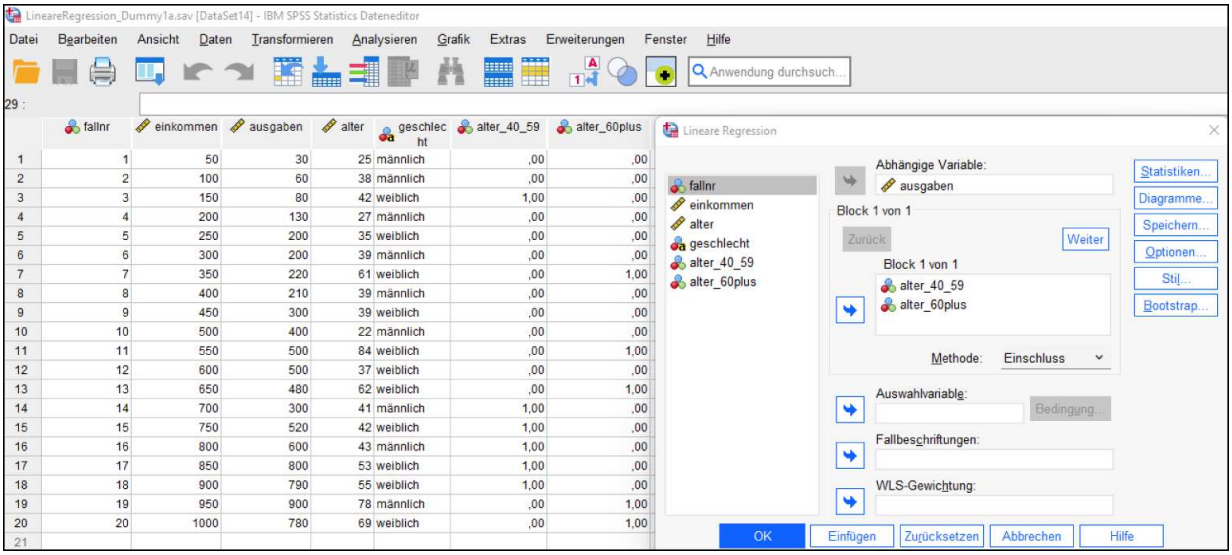



**Abbildung 9.32:** Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben* sowie *alter_40_59* und *alter_60plus* in *LINEAREREGRESSION_DUMMY1A.SAV*

Wie Sie wissen, wirken sich die Parameterwerte von 0/1-Dummies auf die Höhe des konstanten Glieds der Regressionsgleichung aus. Sie stellen für die unterschiedenen Gruppen die arithmetischen Mittelwerte der abhängigen Variablen (hier: für *ausgaben*) dar.

Auch hier wird darauf hingewiesen, dass die hier durchgeführte Regression mit ausschließlich 0/1-Dummies als unabhängigen X-Variablen in erster Linie didaktische Gründe hat. In der Praxis finden Sie 0/1-Dummies in der Regel in Kombination mit metrischen Größen als weiteren unabhängigen X-Variablen vor.

Mit folgendem kleinen Syntaxprogramm ermitteln Sie zunächst die betreffenden arithmetischen Mittelwerte für die drei unterschiedenen Altersklassen, und zwar

weiterhin auf Basis von *LINEAREREGRESSION_DUMMY.SAV*:

```
TEMPORARY.
SELECT IF (alter < 40).
DESCRIPTIVES VARIABLES = ausgaben
/STATISTICS = mean.
TEMPORARY.
SELECT IF (alter >= 40 and alter < 60).
DESCRIPTIVES VARIABLES = ausgaben
/STATISTICS = mean.
TEMPORARY.
SELECT IF (alter >= 60).
DESCRIPTIVES VARIABLES = ausgaben
/STATISTICS = mean.
```

SPSS liefert Ihnen als Ergebnis folgende arithmetischen Ausgaben-Mittelwerte:

Unter 40 Jahre: 225,56 Euro, 40 Jahre bis unter 60 Jahre: 515,00 Euro und 60 Jahre und älter: 576,00 Euro.

Mit steigender Altersklasse steigen also im Beispiel erwartungsgemäß die arithmetischen Ausgaben-Mittelwerte. Dies muss sich auch in den Regressionsberechnungen widerspiegeln, und in der Tat tut es dies auch (siehe Abbildung 9.33): Die Konstante reflektiert die Durchschnittsausgaben für die Referenzgruppe der unter 40-Jährigen in Höhe von 225,56 Euro, für die 40-Jährigen bis unter 60-Jährigen ergeben sich aus der Regressionsgleichung Durchschnittsausgaben in Höhe von 225,56 Euro + 289,44 Euro = 515,00 Euro, und für die 60-Jährigen und Älteren betragen die Durchschnittsausgaben auf Basis der betreffenden Regressionsgleichung: 225,56 Euro + 350,44 Euro = 576,00 Euro. Das sind genau die mittels der DESCRIPTIVES-Prozedur für die drei Altersgruppen ermittelten arithmetischen Mittelwerte.

## *Interaktionsterme mit 0/1-Dummies*

Wenn Sie den Verdacht haben, dass in *LINEAREREGRESSION_DUMMY1A.SAV* das Ausgabeverhalten mit bestimmten Alters- und Geschlechtsausprägungen in sozusagen systematischem Zusammenhang steht, können Sie **Interaktionsterme** in Ihr Regressionsmodell einbauen. Dies bewerkstelligen Sie durch die Multiplikation der betreffenden Variablen, in diesem Fall der relevanten Alters- mit der Geschlechtsvariablen.

Bilden Sie daher zunächst einmal die Geschlechtsvariable *geschl_dummy* und (nochmals) die Altersvariable *alter_dummy*:

```
IF (geschlecht = 'männlich') geschl_dummy = 0.
IF (geschlecht = 'weiblich') geschl_dummy = 1.
EXECUTE.
COMPUTE alter_dummy = 0.
COMPUTE alter_dummy = alter < 40.
EXECUTE.
```

Im nächsten Schritt bilden Sie die Interaktions-Variable *alter_geschl*:

```
COMPUTE alter_geschl = geschl_dummy * alter_dummy.
EXECUTE.
```

Diese neue Variable *alter_geschl* nimmt den Wert 1 nur dann an, wenn sowohl *geschl_dummy* als auch *alter_dummy* gleich 1 sind (das heißt: »weiblich« und »Unter 40 Jahre alt«). Ansonsten weist *alter_geschl* die Ausprägung 0 auf.

Am besten speichern Sie jetzt erst einmal den veränderten Datensatz als *LINEAREREGRESSION_DUMMY2.SAV* ab. Dieser Datensatz enthält nunmehr folgende Variablen: *fallnr, ausgaben, einkommen, alter, geschlecht, alter_40_59, alter_60plus, geschl_dummy* und *alter_dummy*.

**Regression**

**Aufgenommene/Entfernte Variablen[a]**

| Modell | Aufgenommene Variablen | Entfernte Variablen | Methode |
|---|---|---|---|
| 1 | alter_60plus, alter_40_59[b] | . | Einschluß |

a. Abhängige Variable: ausgaben

b. Alle gewünschten Variablen wurden eingegeben.

**Modellzusammenfassung**

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|---|---|---|---|---|
| 1 | ,606[a] | ,367 | ,293 | 226,935 |

a. Einflußvariablen : (Konstante), alter_60plus, alter_40_59

**ANOVA[a]**

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 508107,778 | 2 | 254053,889 | 4,933 | ,020[b] |
| | Nicht standardisierte Residuen | 875492,222 | 17 | 51499,542 | | |
| | Gesamt | 1383600,000 | 19 | | | |

a. Abhängige Variable: ausgaben

b. Einflußvariablen : (Konstante), alter_60plus, alter_40_59

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | RegressionskoeffizientB | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | 225,556 | 75,645 | | 2,982 | ,008 |
| | alter_40_59 | 289,444 | 119,605 | ,504 | 2,420 | ,027 |
| | alter_60plus | 350,444 | 126,578 | ,577 | 2,769 | ,013 |

a. Abhängige Variable: ausgaben

**Abbildung 9.33:** Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* sowie *alter_40_59* und *alter_60plus* in *LINEAREREGRESSION_DUMMY1A.SAV*

Mit der Interaktions-Variablen *alter_geschl* können Sie nun via ANALYSIEREN|REGRESSION|LINEAR eine neue Regression durchführen (siehe Abbildung 9.34).

Das zugehörige Regressionsergebnis findet sich in Abbildung 9.35. Es zeigt einen positiven Interaktionsterm, und zwar in der Weise, dass in unserem Fall die Kombination »weiblich« und »Unter 40 Jahre alt« zu höheren Ausgaben führt. Leider ist der zugehörige Regressionsparameter – ebenso wie im Übrigen der Geschlechts-Dummy – nicht signifikant. Dies weist darauf hin, dass eine solche Interaktion hier im Beispiel keine allzu große statistische Aussagekraft hat.

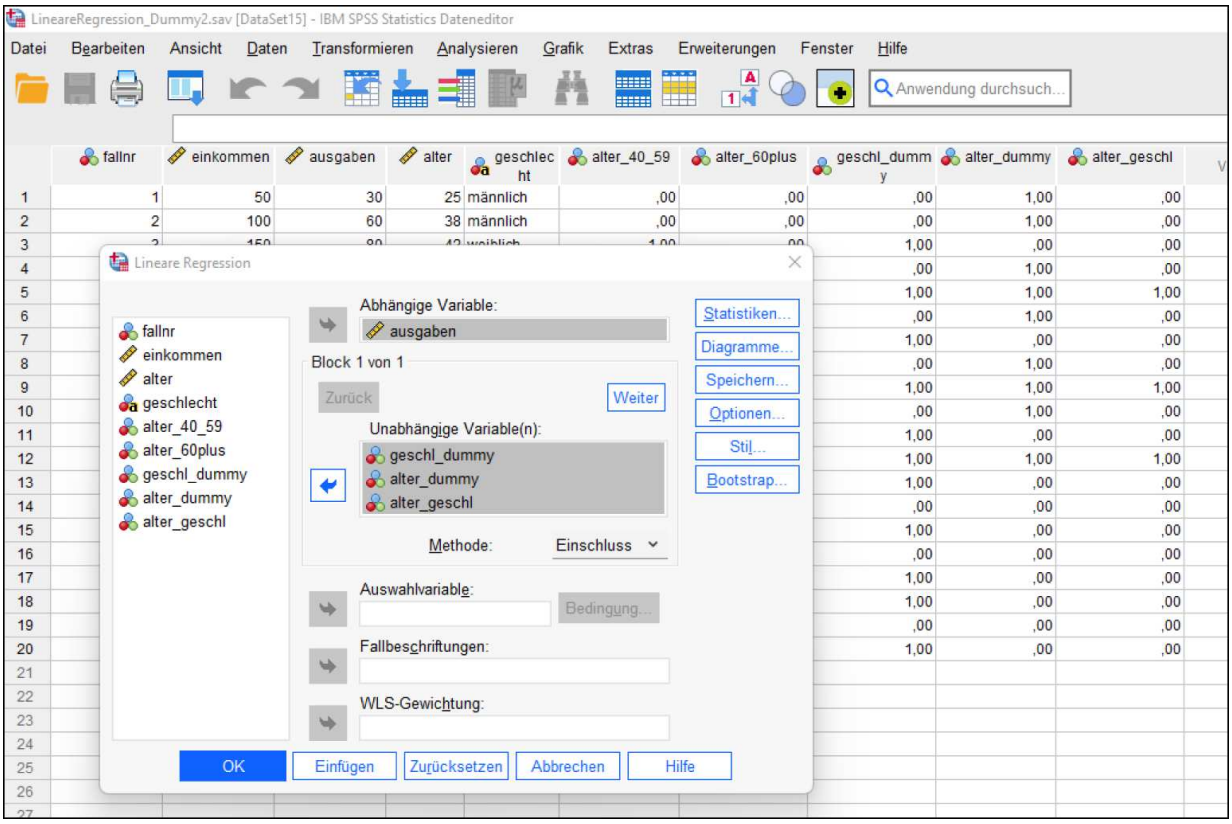



**Abbildung 9.34:** Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben* sowie *geschl_dummy*, *alter_dummy* und *alter_geschl* in *LINEAREREGRESSION_DUMMY2.SAV*

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten: Regressionskoeffizient B | Nicht standardisierte Koeffizienten: Std.-Fehler | Standardisierte Koeffizienten: Beta | T | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Konstante) | 600,000 | 130,690 | | 4,591 | <,001 |
| | geschl_dummy | -78,750 | 153,248 | -,149 | -,514 | ,614 |
| | alter_dummy | -428,333 | 160,062 | -,810 | -2,676 | ,017 |
| | alter_geschl | 240,417 | 221,596 | ,326 | 1,085 | ,294 |

a. Abhängige Variable: ausgaben

**Abbildung 9.35:** Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* sowie *geschl_dummy*, *alter_dummy* und *alter_geschl* in *LINEAREREGRESSION_DUMMY2.SAV*

# *Zeitreihenwerte als abhängige Variable*

Sie können die Regressionsanalyse auch auf reine **Zeitreihendaten** anwenden, indem Sie den Faktor Zeit als »erklärende« X-Größe verwenden. Als Zeit können Sie zum Beispiel Jahresangaben verwenden. In *LINEAREREGRESSION_ZEIT1.SAV* finden Sie für eine Ausgaben-Zeitreihe sowohl (ältere) Jahresangaben in der Variablen *jahr* als auch standardisierte Zeitangaben in der Variablen *time*, wobei *time* aus *jahr* wie folgt berechnet wird (und demnach das Jahr 1996 zu 0 wird, das Jahr 1997 zu 1, das Jahr 1998 zu 2 und so weiter):

```
COMPUTE time = jahr -1996.
EXECUTE.
```

Üblicherweise standardisiert man Jahresangaben in der vorstehend beschriebenen Art und Weise, um Zeitreihen mit unterschiedlichen Jahresangaben sozusagen besser miteinander vergleichen zu können.

## *Eine »reine« Zeitreihenbetrachtung*

Der Zusammenhang zwischen *ausgaben* und *jahr* ist als Scatterplot mittels GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM in Abbildung 9.36 dargestellt. Er weist auf einen positiven Zusammenhang, also auf einen positiven zeitlichen Trend hin.

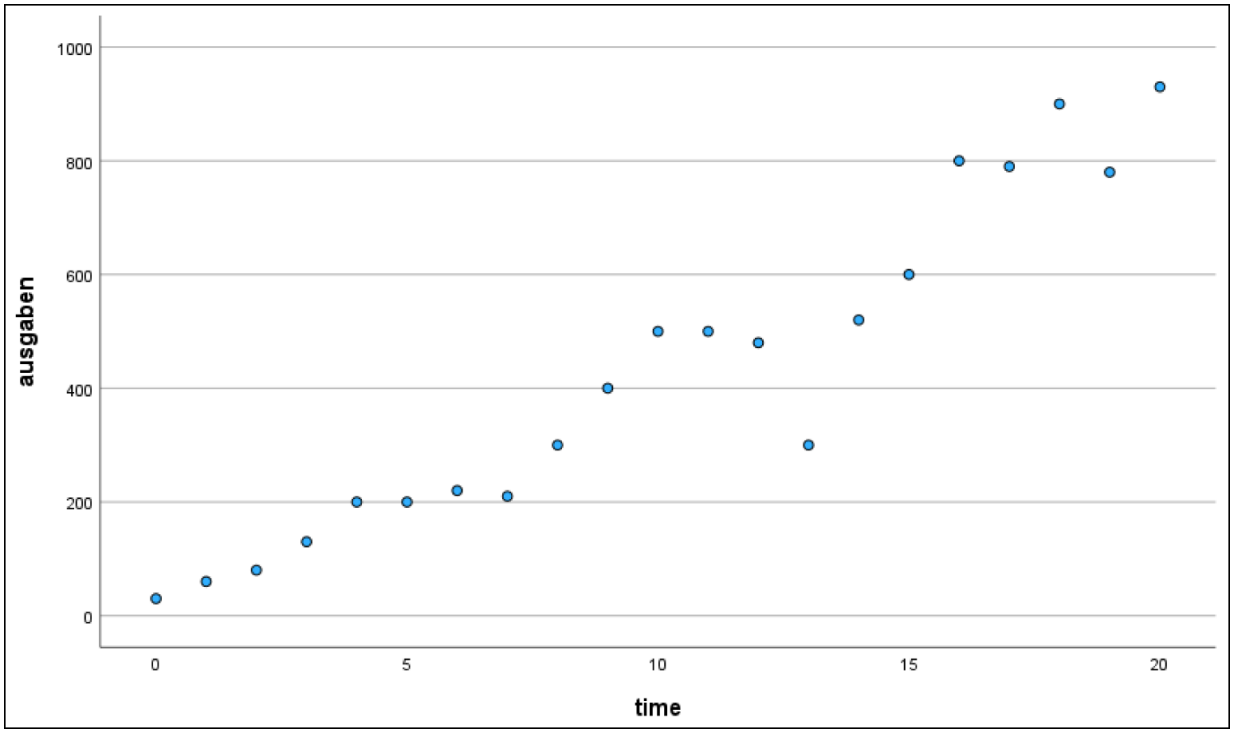



**Abbildung 9.36:** Scatterplot zwischen *ausgaben* und *jahr* in der Datendatei *LINEAREREGRESSION_ZEIT1.SAV*

Wenn Sie jetzt über ANALYSIEREN|REGRESSION|LINEAR eine Regression zwischen *ausgaben* und *time* durchführen, tun Sie dies, um einen (linearen) **Trend** für die Ausgabenentwicklung zu ermitteln, das heißt zum Beispiel: um auf Basis der Regressionsgleichung in die Zukunft gerichtete **Prognosen** abgeben zu können. Aus gleich noch näher erläuterten Gründen klicken Sie bitte in der Subdialogbox STATISTIKEN auch noch DURBIN-WATSON an, genauso wie dies in Abbildung 9.37 dargestellt ist.

Auf Basis von *LINEAREREGRESSION_ZEIT1.SAV* ergibt sich der in Abbildung 9.38 dargelegte Regressionsoutput. Wie ersichtlich, zeigt sich ein positiver Trend, was an dem positiven Koeffizienten von *time* abzulesen ist. Von einem auf das andere Jahr steigen die Ausgaben (im Durchschnitt) gemäß der Regressionsgleichung um +44,40 Euro. Nicht plausibel an der vorgenommenen Schätzung ist allerdings das negative absolute Glied von −18,79 Euro, denn das bedeutet technisch, dass im Jahr 1996 (bei *time* = 0) die Ausgaben negativ gewesen wären. Das absolute Glied ist indes statistisch nicht signifikant.

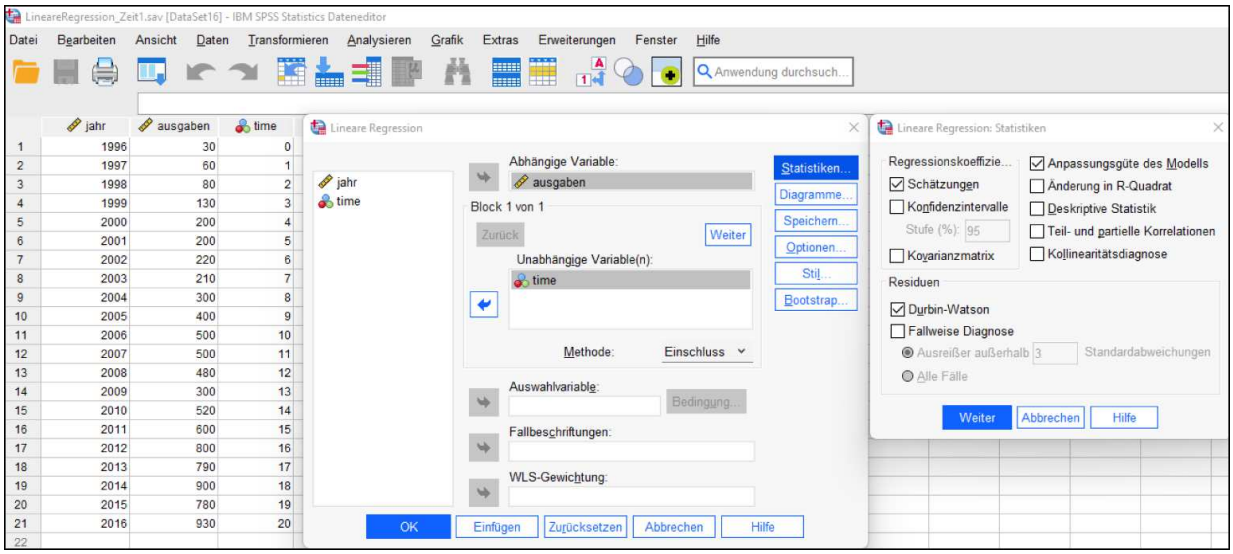



**Abbildung 9.37**: Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben* und *time* in *LINEAREREGRESSION_ZEIT1.SAV*

Die Regressionsgleichung

$$ausgaben = -18{,}79 + 44{,}40 \cdot time$$

kann dennoch für Prognosen verwendet werden. Sind Sie zum Beispiel an dem geschätzten Ausgabenwert für das Jahr 2023 interessiert, müssen Sie *time* auf 2023 − 1996 = 27 setzen, und Sie erhalten für *ausgaben* als Schätzwert 1.180,01 Euro.

## *Berechnung mehrerer Trendfunktionen*

Da ein Trend verschiedene funktionale Ausprägungen aufweisen kann: linear, quadratisch und so weiter, empfiehlt sich neben der Betrachtung eines linearen Trends

die Betrachtung anderer Trendfunktionen, und zwar über ANALYSIEREN|KURVENANPASSUNG (siehe Abbildung 9.39).

In unserem Beispiel gibt es allerdings das Problem, dass die Variable *time* bei 0 anfängt und daher bestimmte Trendvarianten, die *time* logarithmieren oder durch *time* dividieren, nicht berechnet werden könnten. Deswegen ist eine Variable *time2* berechnet worden, die beim Wert 1 anfängt und den nachfolgenden Schätzungen zugrunde liegt (der aktualisierte Datensatz befindet sich in *LINEAREREGRESSION_ZEIT1A.SAV*):

```
COMPUTE time2 = time + 1.
EXECUTE.
```

Im vorliegenden Fall erhalten Sie dann den in Abbildung 9.40 und Abbildung 9.41 wiedergegebenen Outputauszug. Die Bestimmtheitsmaße zeigen, dass der bereits geschätzte lineare Trend recht gut »passt«, dass aber zum Beispiel ein quadratischer oder ein kubischer Trend (also Trendfunktionen zweiter beziehungsweise dritter Ordnung) noch ein bisschen besser »passen« (wobei allerdings die ausgewiesenen Bestimmtheitsmaße hier nur bedingt miteinander verglichen werden können, weil als abhängige Variable nicht immer *ausgaben* genutzt worden ist, sondern etwa auch *ausgaben_ln*).

**Regression**

**Aufgenommene/Entfernte Variablen[a]**

| Modell | Aufgenommene Variablen | Entfernte Variablen | Methode |
|---|---|---|---|
| 1 | time[b] | . | Einschluß |

a. Abhängige Variable: ausgaben

b. Alle gewünschten Variablen wurden eingegeben.

**Modellzusammenfassung[b]**

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers | Durbin-Watson-Statistik |
|---|---|---|---|---|---|
| 1 | ,959[a] | ,919 | ,915 | 83,666 | 1,305 |

a. Einflußvariablen : (Konstante), time

b. Abhängige Variable: ausgaben

**ANOVA[a]**

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 1518124,805 | 1 | 1518124,805 | 216,877 | <,001[b] |
| | Nicht standardisierte Residuen | 132999,004 | 19 | 6999,948 | | |
| | Gesamt | 1651123,810 | 20 | | | |

a. Abhängige Variable: ausgaben

b. Einflußvariablen : (Konstante), time

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | RegressionskoeffizientB | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | -18,788 | 35,248 | | -,533 | ,600 |
| | time | 44,403 | 3,015 | ,959 | 14,727 | <,001 |

a. Abhängige Variable: ausgaben

**Residuenstatistik[a]**

| | Minimum | Maximum | Mittelwert | Std.-Abweichung | N |
|---|---|---|---|---|---|
| Nicht standardisierter vorhergesagter Wert | -18,79 | 869,26 | 425,24 | 275,511 | 21 |
| Nicht standardisierte Residuen | -258,446 | 119,541 | ,000 | 81,547 | 21 |
| Standardisierter vorhergesagter Wert | -1,612 | 1,612 | ,000 | 1,000 | 21 |
| Standardisierte Residuen | -3,089 | 1,429 | ,000 | ,975 | 21 |

a. Abhängige Variable: ausgaben

**Abbildung 9.38**: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* und *time* in *LINEAREREGRESSION_ZEIT1.SAV*

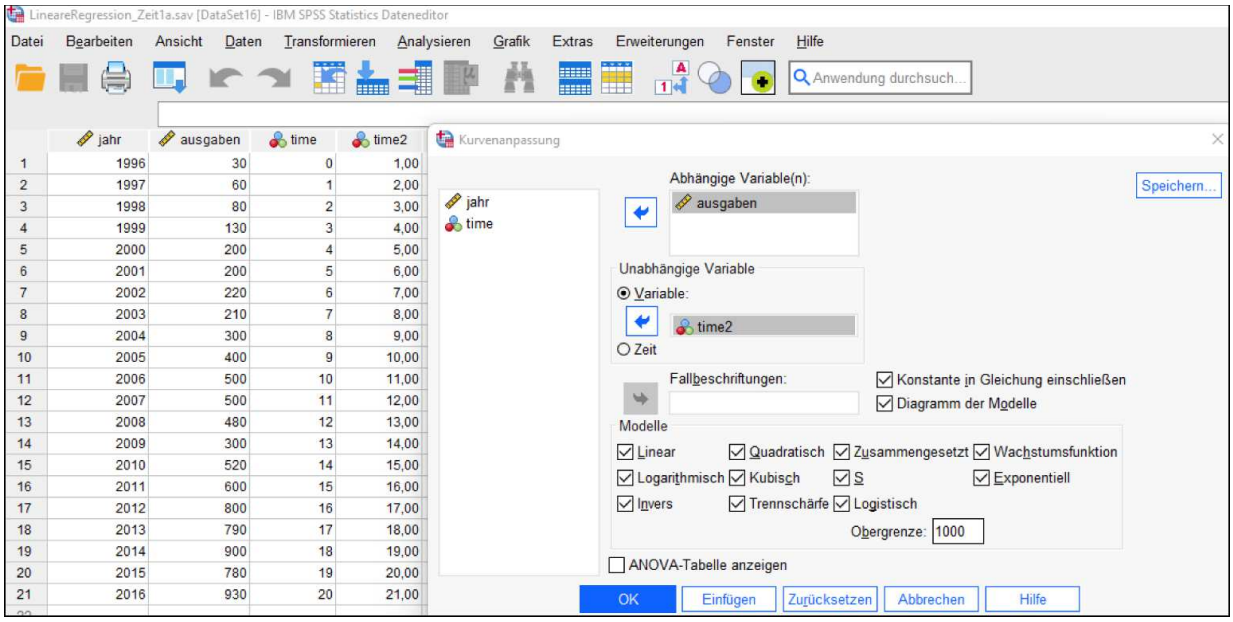



**Abbildung 9.39:** Anfordern verschiedener Trendfunktionen für *ausgaben* in *LINEARE_REGRESSION_ZEIT1A.SAV*

**Modellzusammenfassung und Parameterschätzer**

Abhängige Variable: ausgaben

| | Modellübersicht | | | | | Parameterschätzungen | | | |
|---|---|---|---|---|---|---|---|---|---|
| Gleichung | R-Quadrat | F | df1 | df2 | Sig. | Konstante | b1 | b2 | b3 |
| Linear | ,919 | 216,877 | 1 | 19 | <,001 | -63,190 | 44,403 | | |
| Logarithmisch | ,740 | 54,074 | 1 | 19 | <,001 | -228,036 | 302,307 | | |
| Invers | ,389 | 12,089 | 1 | 19 | ,003 | 566,784 | -815,411 | | |
| Quadratisch | ,931 | 121,802 | 2 | 18 | <,001 | 15,218 | 23,948 | ,930 | |
| Kubisch | ,932 | 77,376 | 3 | 17 | <,001 | -8,338 | 35,478 | -,350 | ,039 |
| Zusammengesetzt | ,866 | 123,036 | 1 | 19 | <,001 | 65,172 | 1,152 | | |
| Netzstrom | ,962 | 485,147 | 1 | 19 | <,001 | 26,791 | 1,131 | | |
| S | ,753 | 57,883 | 1 | 19 | <,001 | 6,379 | -3,723 | | |
| Wachstumsfunktion | ,866 | 123,036 | 1 | 19 | <,001 | 4,177 | ,141 | | |
| Exponentiell | ,866 | 123,036 | 1 | 19 | <,001 | 65,172 | ,141 | | |
| Logistisch | ,923 | 226,309 | 1 | 19 | <,001 | ,023 | ,780 | | |

Die unabhängige Variable ist time2.

**Abbildung 9.40:** Ergebnis der Schätzung verschiedener Trendfunktionen für *ausgaben* in *LINEARE_REGRESSION_ZEIT1.SAV* (1)

## *Zum Problem der Autokorrelation*

Gerade bei Zeitreihenwerten tritt oft das Problem der **Autokorrelation** auf. Dies bedeutet, dass die Residuen der OLS-Schätzung, also die Differenzen aus beobachteten und geschätzten Y-Werten, miteinander korrelieren, das heißt statistisch voneinander abhängen. Wie eingangs zu diesem Kapitel bei den Annahmen der linearen Regressionsanalyse erwähnt, sollten Sie bei Ihren entsprechenden Schätzungen darauf achten, dass in Ihrem Datensatz möglichst keine Autokorrelation existiert.

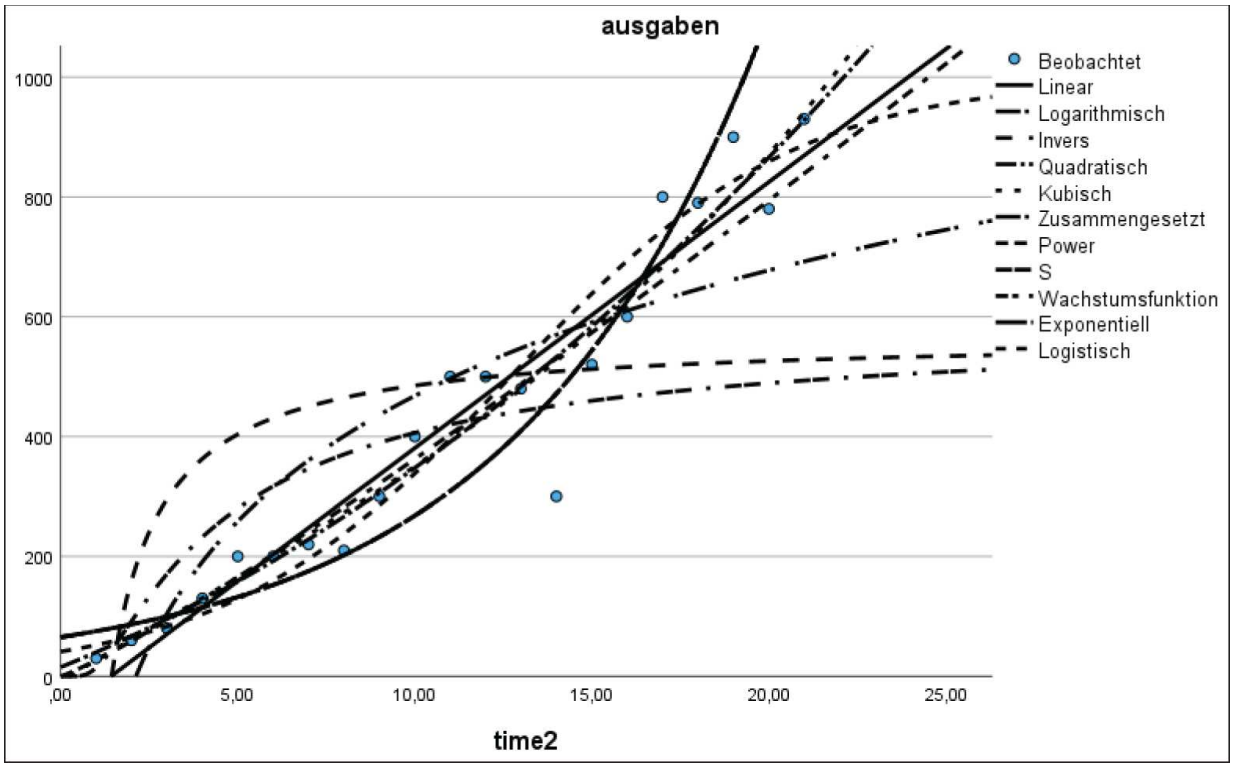



**Abbildung 9.41:** Ergebnis der Schätzung verschiedener Trendfunktionen für *ausgaben* in *LINEARE_REGRESSION_ZEIT1.SAV* (2)

Bei Zeitreihenwerten tritt dieser Fall deshalb auf, weil historische Entwicklungen die Gegenwart und damit auch die in der Gegenwart beobachteten Y-Werte prägen. Wenn die beobachteten Y-Werte in Ihrem Datensatz nach aneinander angrenzenden Gebietseinheiten angeordnet wären, kann im Übrigen auch Autokorrelation auftreten, weil auch wegen kultureller und wirtschaftlicher Abhängigkeiten zwischen Nachbarregionen gegenseitige Abhängigkeiten in den beobachteten Y-Werten vorliegen können.

In SPSS können Sie mittels des **Durbin/Watson-Tests** überprüfen, ob Autokorrelation erster Ordnung vorliegt (allerdings dann halt nicht, ob Autokorrelation einer höheren Ordnung existiert). Mit einer Autokorrelation erster Ordnung ist gemeint, dass zwischen genau zwei benachbarten Y-Werten beziehungsweise zwischen genau zwei benachbarten Residuen Autokorrelation existiert. Hierzu hatten Sie ja in Abbildung 9.37 für den »reinen« Zeitreihenfall den Durbin/Watson-Test in SPSS angefordert.

Für den Durbin/Watson-Test gilt als Faustregel, dass bei einem Wert nahe 2 (etwas genauer: zwischen circa 1,6 und circa 2,4) keine Autokorrelation erster Ordnung vorliegt. Demgegenüber wird bei Werten nahe 0 auf starke positive und bei Werten nahe 4 auf starke negative Autokorrelation erster Ordnung geschlossen.

Wie sieht es nun aber etwas konkreter bei Zusammenhängen zwischen zwei Variablen bezüglich der Autokorrelation aus, die über die Zeit hinweg betrachtet worden sind? Zu diesem Zweck öffnen Sie bitte *LINEAREREGRESSION_ZEIT2.SAV* und fordern über

ANALYSIEREN|REGRESSION|LINEAR und durch Anklicken des Durbin/Watson-Tests in der Subdialogbox STATISTIKEN eine lineare Regression zwischen *ausgaben* und *einkommen* an.

In unserem Fall liegt der Durbin/Watson-Wert mit 1,305 im Bereich positiver Autokorrelation erster Ordnung. Allein schon durch das Vorliegen eines Trends bei »reinen« Zeitreihenbetrachtungen erhalten Sie hier Autokorrelation. Im Ergebnis wird also eine positive Autokorrelation erster Ordnung ausgewiesen (siehe Abbildung 9.42).

**Modellzusammenfassung[b]**

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers | Durbin-Watson-Statistik |
|---|---|---|---|---|---|
| 1 | ,959[a] | ,919 | ,915 | 83,666 | 1,305 |

a. Einflußvariablen : (Konstante), einkommen

b. Abhängige Variable: ausgaben

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | RegressionskoeffizientB | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | -63,190 | 37,859 | | -1,669 | ,111 |
| | einkommen | ,888 | ,060 | ,959 | 14,727 | <,001 |

a. Abhängige Variable: ausgaben

**Abbildung 9.42:** Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_ZEIT2.SAV* (Auszug)

Wie kann man nun das Problem der Autokorrelation beseitigen beziehungsweise zumindest abmildern? Eine in der Praxis vielfach angewandte Lösung des Problems besteht darin, die Variablenwerte zu transformieren. Beispielsweise könnten Sie in unserem Beispiel die Differenz aus zwei aufeinanderfolgenden Werten für *ausgaben* und *einkommen* bilden. Weil es in SPSS etwas schwierig ist, aufeinanderfolgende Fälle miteinander zu verknüpfen, habe ich dies in EXCEL entsprechend vorbereitet (in der Datei *DIFFERENZEN1.XLSX* in Form der beiden Variablen *ausg_diff* und *eink_diff*). Diese Datei habe ich anschließend nach SPSS transportiert und dort unter *LINEAREREGRESSION_ZEIT3.SAV* abgespeichert.

Wenn Sie jetzt auf die Idee kommen, zwischen *ausg_diff* und *eink_diff* eine lineare Regression durchführen zu wollen, liegen Sie leider falsch. Die Variable *eink_diff* weist nämlich ausschließlich den Wert 50 auf und hat somit nicht im Mindesten Varianz, was aber für Regressionsberechnungen notwendig ist. Aus diesem Grund wird im Folgenden über ANALYSIEREN|REGRESSION|LINEAR eine andere lineare

Regression berechnet werden, und zwar eine solche zwischen der »Differenzen-Variablen« *ausg_diff* als Y-Variablen und *einkommen* als X-Variablen (siehe Abbildung 9.43).

Wie aus Abbildung 9.44 hervorgeht, verändert sich der Durbin/Watson-Wert von zuvor 1,305 in der neuen Regression zu 2,517. Dieser Wert liegt zumindest nicht allzu weit entfernt vom Obergrenzenwert 2,4, für den keine Autokorrelation ersten Grades angenommen wird. Durch die Differenzenbildung ist somit die positive Autokorrelation ersten Grades zumindest (markant) vermindert worden. Allerdings sind die neuen Regressionsparameter nicht signifikant, und das korrigierte Bestimmtheitsmaß zur Beurteilung der Anpassungsgüte der betreffenden linearen Parameter ist sogar negativ, sodass die betreffende berechnete Regressionsbeziehung im Grunde genommen leider keine (statistische) Aussagekraft besitzt. (Allerdings geht es hier ja um bivariate Regressionen, und daher sollte man das *korrigierte* Bestimmtheitsmaß nicht überinterpretieren.)

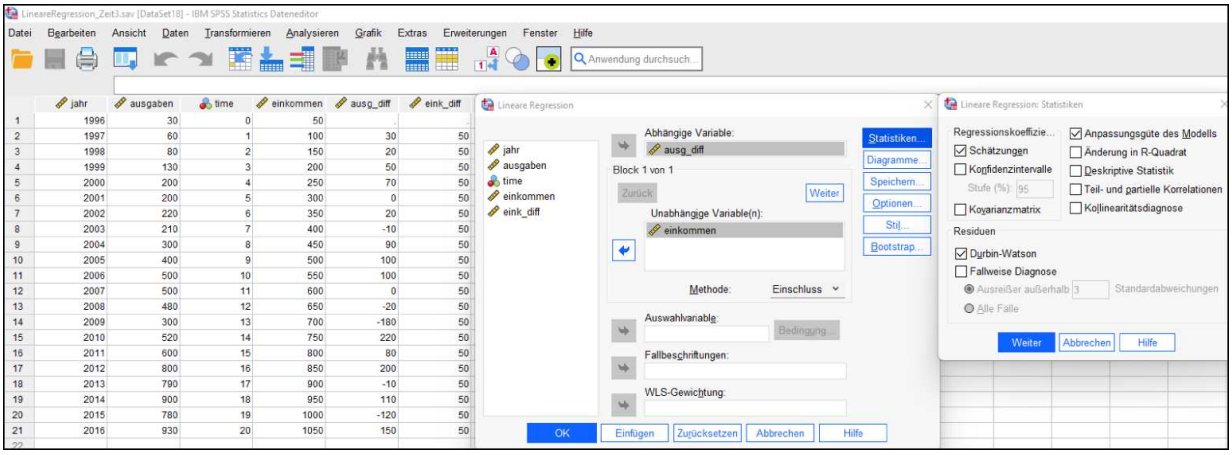



**Abbildung 9.43:** Anfordern einer linearen Regressionsbeziehung zwischen *ausg_diff* und *einkommen* in *LINEAREREGRESSION_ZEIT3.SAV*

**Modellzusammenfassung**[b]

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers | Durbin-Watson-Statistik |
|---|---|---|---|---|---|
| 1 | ,107[a] | ,011 | -,043 | 97,614 | 2,517 |

a. Einflußvariablen : (Konstante), einkommen

b. Abhängige Variable: ausg_diff

**Koeffizienten**[a]

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | Regressionskoeffizient B | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | 25,113 | 48,697 | | ,516 | ,612 |
| | einkommen | ,035 | ,076 | ,107 | ,457 | ,653 |

a. Abhängige Variable: ausg_diff

**Abbildung 9.44:** Ergebnis einer linearen Regressionsbeziehung zwischen *ausg_diff* und *einkommen* in *LINEAREREGRESSION_ZEIT3.SAV* (Auszug)

## *Zur Schätzung von Konfidenzintervallen*

Führen Sie über ANALYSIEREN|REGRESSION|LINEAR eine lineare Regression durch, können Sie in der Subdialogbox STATISTIKEN insofern über die Voreinstellungen hinausgehen, als Sie zum Beispiel KONFIDENZINTERVALLE anfordern. Dadurch erhalten Sie Wertebereiche, sogenannte **Konfidenzintervalle**, für die geschätzten Regressionsparameter.

Wählen Sie doch bitte einfach *LINEAREREGRESSION_GERADE6.SAV* als Datendatei für die Regression zwischen *ausgaben* (Y-Variable) und *einkommen* (X-Variable) aus. Wenn Sie jetzt in der oben genannten Subdialogbox STATISTIKEN bei KONFIDENZINTERVALLE ein Häkchen setzen, wie in Abbildung 9.45 gezeigt, fordern Sie für die beiden Regressionskoeffizienten – hier bei Vorgabe einer 95-Prozent-»Vertrauenswahrscheinlichkeit« – nicht mehr nur jeweils einen konkreten Schätzwert an, sondern um diesen Schätzwert herum jeweils einen Wertebereich mit einer Unter- und einer Obergrenze.

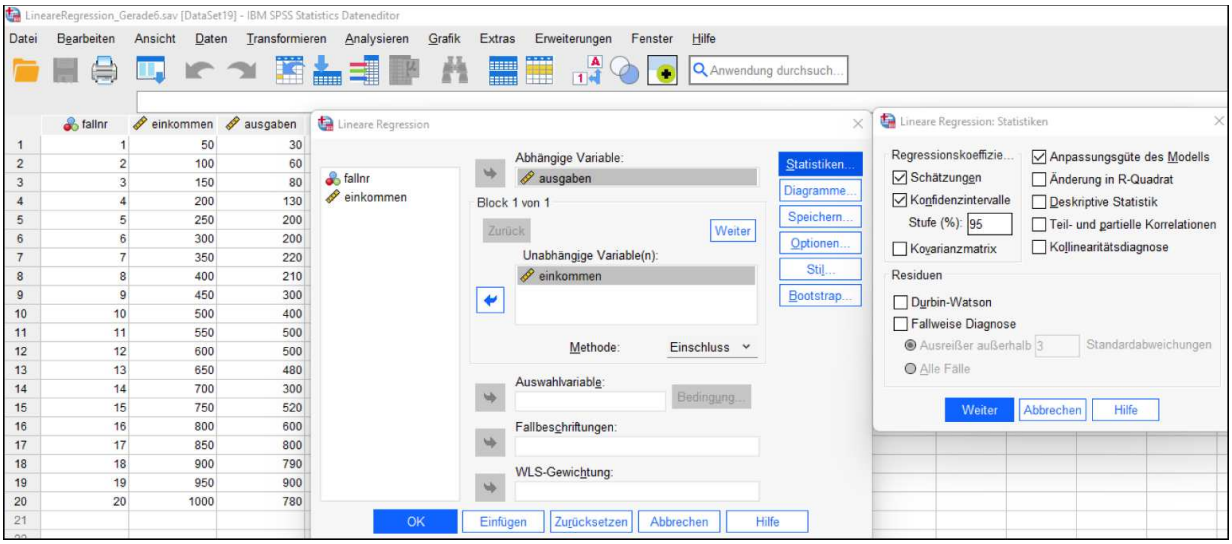



**Abbildung 9.45**: Anfordern von Konfidenzintervallen für die geschätzten Regressionsparameter einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_GERADE6.SAV*

Hätten Sie nur eine Punktschätzung für die beiden Regressionsparameter angefordert, hätten Sie −56,16 Euro für die Konstante und +0,869 für die Steigung erhalten. Dadurch, dass Sie aber auch 95-Prozent-Konfidenzintervalle für diese beiden Parameter angefordert haben, »wirft« SPSS Ihnen jeweils einen Wertebereich »aus«. Mit 95-prozentiger Wahrscheinlichkeit liegt das konstante Glied der Regressionsgleichung gemäß Abbildung 9.46 zwischen −138,62 Euro und 26,31 Euro und die Steigung zwischen +0,731 und +1,007.

**Koeffizienten**[a]

| Modell | | Nicht standardisierte Koeffizienten: RegressionskoeffizientB | Nicht standardisierte Koeffizienten: Std.-Fehler | Standardisierte Koeffizienten: Beta | T | Sig. | 95,0% Konfidenzintervalle für B: Untergrenze | 95,0% Konfidenzintervalle für B: Obergrenze |
|---|---|---|---|---|---|---|---|---|
| 1 | (Konstante) | -56,158 | 39,251 | | -1,431 | ,170 | -138,622 | 26,306 |
| | einkommen | ,869 | ,066 | ,952 | 13,259 | <,001 | ,731 | 1,007 |

a. Abhängige Variable: ausgaben

**Abbildung 9.46**: Konfidenzintervalle für die geschätzten Regressionsparameter als Ergebnis einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_GERADE6.SAV* (Ergebnisauszug)

Ergänzend zur »reinen« Schätzung von Konfidenzintervallen über den gesamten Datensatz hinweg (mittels der Konfidenzintervallangabe) können Sie mittels **Bootstrapping** ein Konfidenzintervall für die Regressionsparameter konstruieren. Hierzu zieht SPSS aus Ihrer Datenbasis eine bestimmte Anzahl von Stichproben (standardmäßig: 1.000 Stichproben) und erstellt auf dieser Grundlage die jeweiligen Konfidenzintervalle.

Sie müssen bei ANALYSIEREN|REGRESSION|LINEAR einfach nur die Subdialogbox BOOTSTRAP ansteuern und dort BOOTSTRAPPING

DURCHFÜHREN anklicken. Dies ist in Abbildung 9.47 erneut für eine lineare Regression zwischen *ausgaben* (Y-Variable) und *einkommen* (X-Variable) auf Basis von *LINEAREREGRESSION_GERADE6.SAV* dargestellt.

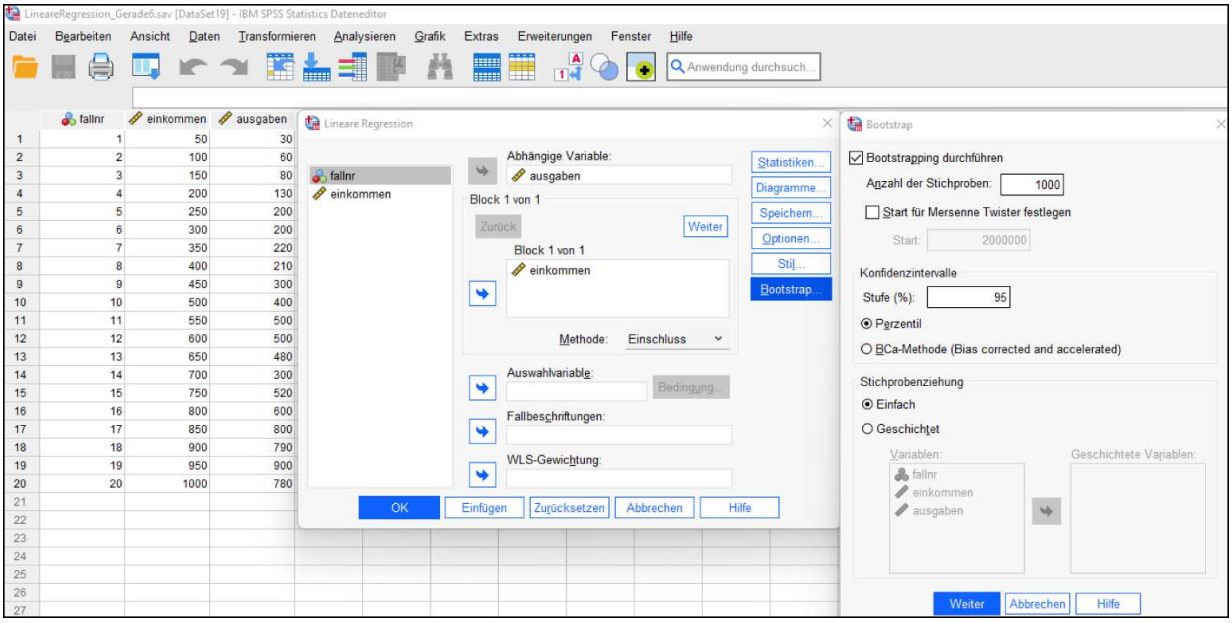



**Abbildung 9.47:** BOOTSTRAP-Anfordern von Konfidenzintervallen für die geschätzten Regressionsparameter einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_GERADE6.SAV*

Ich habe eine entsprechende BOOTSTRAP-Berechnung durchgeführt (siehe Abbildung 9.48). In diesem Fall haben die verschiedenen Zufallsziehungen zu folgenden Konfidenzintervallen (jeweils auf dem 95-Prozent-Niveau) geführt: bei der Konstanten zwischen −111,03 Euro und −13,97 Euro sowie bei der Steigung zwischen +0,729 und +0,984.

**Bootstrap für Koeffizienten**

| Modell | | Regressionskoeffizient B | Bootstrap[a] Verzerrung | Std.-Fehler | Sig. (2-seitig) | 95% Konfidenzintervall Unterer | Oberer |
|---|---|---|---|---|---|---|---|
| 1 | (Konstante) | -56,158 | -,305 | 23,962 | ,040 | -111,034 | -13,969 |
| | einkommen | ,869 | -,002 | ,062 | <,001 | ,729 | ,984 |

a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

**Abbildung 9.48:** BOOTSTRAP-Konfidenzintervalle für die geschätzten Regressionsparameter als Ergebnis einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_GERADE6.SAV* (Ergebnisauszug)

# *Residuenanalyse*

Mit der sogenannten **Residuenanalyse** können Sie die Annahmen der linearen Regressionsanalyse in Bezug auf die Restgrößen der Schätzungen, die *Residuen*, also in Bezug auf die Abweichungen der geschätzten von den beobachteten Y-Werten, einer genaueren Betrachtung unterziehen. Im Folgenden wird konkret auf die Heteroskedastizitätsproblematik und die Normalverteilungsannahme der Residuen eingegangen.

## *Zum Problem der Heteroskedastizität*

Bezüglich der oben genannten **Homoskedastizitäts**-Annahme wird also hier zunächst der gegenteilige (und unerwünschte) Fall der Heteroskedastizität anhand der Residuenverteilung überprüft. Unter Homoskedastizität (welch furchtbares Wort!) versteht man, dass die Streuung – die Varianz – der Residuen um die Regressionsgerade herum (möglichst) gleich ist beziehungsweise keinen auffälligen Mustern folgt. Entsprechend meint **Heteroskedastizität**, dass solche auffälligen Muster vorliegen.

Lassen Sie uns in diesem Zusammenhang auf *LINEAREREGRESSION_HETEROSKEDASTIZITÄT.SAV* als Datendatei beziehen. Führen Sie zunächst einmal über ANALYSIEREN|REGRESSION|LINEAR eine einfache lineare Regression zwischen *ausgaben* und *einkommen* durch, wobei Sie aber jetzt in der Subdialogbox SPEICHERN die einfachen Residuen durch Anklicken von RESIDUEN|NICHT STANDARDISIERT (siehe Abbildung 9.49) als neue Variable in Ihren Datensatz einbauen. Diese Variable wird von SPSS als *res_1* bezeichnet (siehe Abbildung 9.50).

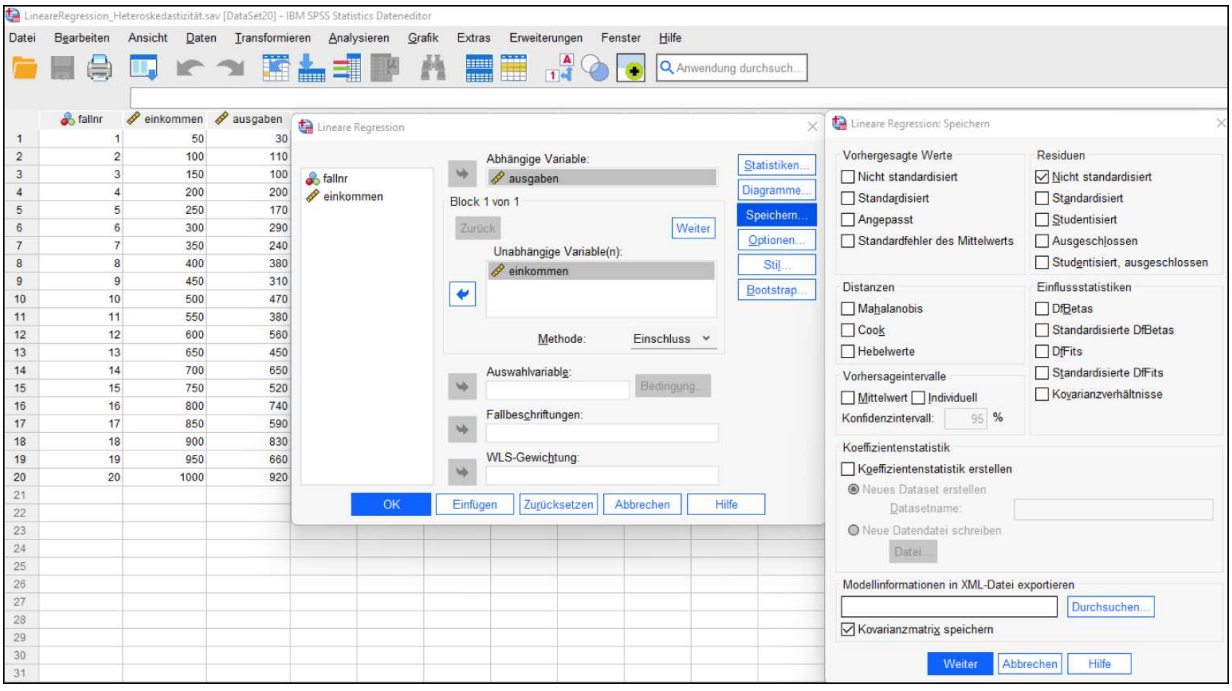
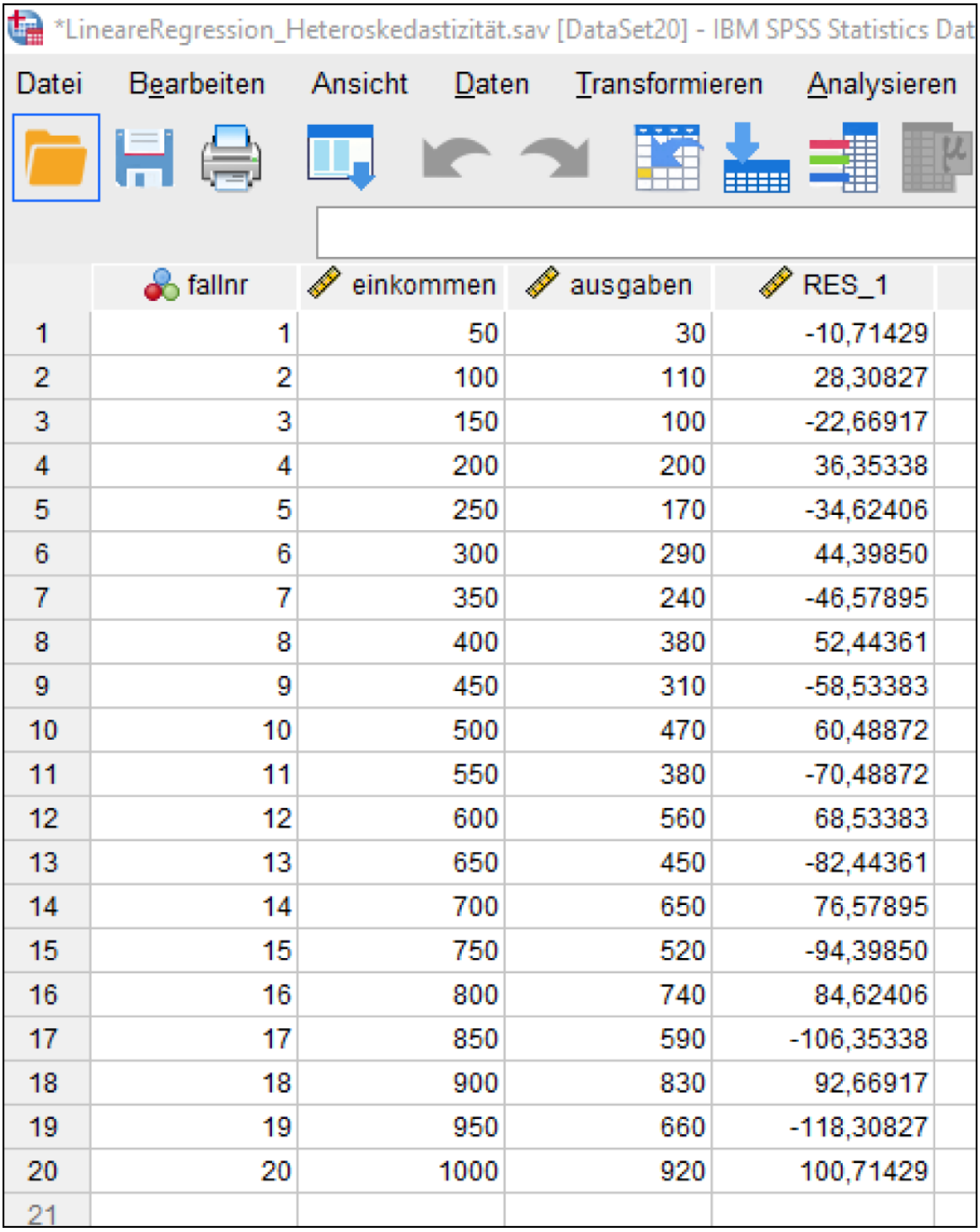



**Abbildung 9.49:** Anfordern der einfachen Residuen bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT.SAV*

Über GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM können Sie nun den Zusammenhang zwischen den einfachen Residuen (Variable *res_1*) und der unabhängigen Variablen *einkommen* veranschaulichen. Wie Sie Abbildung 9.51 entnehmen können, zeigt sich als Streudiagramm ein trichterförmiges Bild. Ein solches Muster ist typisch für das Vorliegen von Heteroskedastizität, heißt es doch hier, dass die Varianz, also vereinfacht die Streuung, der Residuen mit steigendem Wert der unabhängigen Variablen *einkommen* wächst. Dies ist hier sogar inhaltlich einigermaßen plausibel. Bei geringen Einkommen hält sich die Ausgabenstreuung deshalb in engen Grenzen, weil die betreffenden Konsumenten ihr Einkommen überwiegend in Ausgaben (vor allem für Grundbedarfsgüter) umsetzen (müssen), während mit steigendem Einkommen die Wahlfreiheit der Konsumenten zwischen der Anschaffung von (teuren) Konsumgütern auf der einen Seite und der Bildung von Ersparnissen (zulasten der Konsumausgaben) steigt.




| | fallnr | einkommen | ausgaben | RES_1 |
|---|---|---|---|---|
| 1 | 1 | 50 | 30 | -10,71429 |
| 2 | 2 | 100 | 110 | 28,30827 |
| 3 | 3 | 150 | 100 | -22,66917 |
| 4 | 4 | 200 | 200 | 36,35338 |
| 5 | 5 | 250 | 170 | -34,62406 |
| 6 | 6 | 300 | 290 | 44,39850 |
| 7 | 7 | 350 | 240 | -46,57895 |
| 8 | 8 | 400 | 380 | 52,44361 |
| 9 | 9 | 450 | 310 | -58,53383 |
| 10 | 10 | 500 | 470 | 60,48872 |
| 11 | 11 | 550 | 380 | -70,48872 |
| 12 | 12 | 600 | 560 | 68,53383 |
| 13 | 13 | 650 | 450 | -82,44361 |
| 14 | 14 | 700 | 650 | 76,57895 |
| 15 | 15 | 750 | 520 | -94,39850 |
| 16 | 16 | 800 | 740 | 84,62406 |
| 17 | 17 | 850 | 590 | -106,35338 |
| 18 | 18 | 900 | 830 | 92,66917 |
| 19 | 19 | 950 | 660 | -118,30827 |
| 20 | 20 | 1000 | 920 | 100,71429 |
| 21 | | | | |

**Abbildung 9.50**: Neue Variable *res_1* für die einfachen Residuen bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT.SAV*

Ein ähnliches Bild wie in Abbildung 9.51 erhalten Sie alternativ, wenn Sie bei ANALYSIEREN|REGRESSION|LINEAR in der Subdialogbox DIAGRAMME als Y-Variable die standardisierten Residuen **zresid* und als X-Variable die standardisierten vorhergesagten Werte für die Ausgabenvariable **zpred* eingeben (siehe hierzu

Abbildung 9.52). SPSS standardisiert im Übrigen sowohl die Residuen als auch die Regressionswerte der abhängigen Variablen durch Division durch den sogenannten Standardfehler der Schätzung. Zugegebenermaßen etwas technisch ausgedrückt stellt er die Quadratwurzel aus der Fehlervarianz (also der Varianz der Residuen) dar (wobei zudem nicht durch die gesamte Fallzahl, sondern durch die Fallzahl minus 2 Fällen dividiert wird).

Diese Standardisierung ändert aber nichts an den Grundaussagen an dieser Stelle. Auch für diese Form der Darstellung erhalten Sie in unserem Beispiel eine trichterförmige Abbildung (siehe Abbildung 9.53). Dies zeigt natürlich auch hier Heteroskedastizität an.

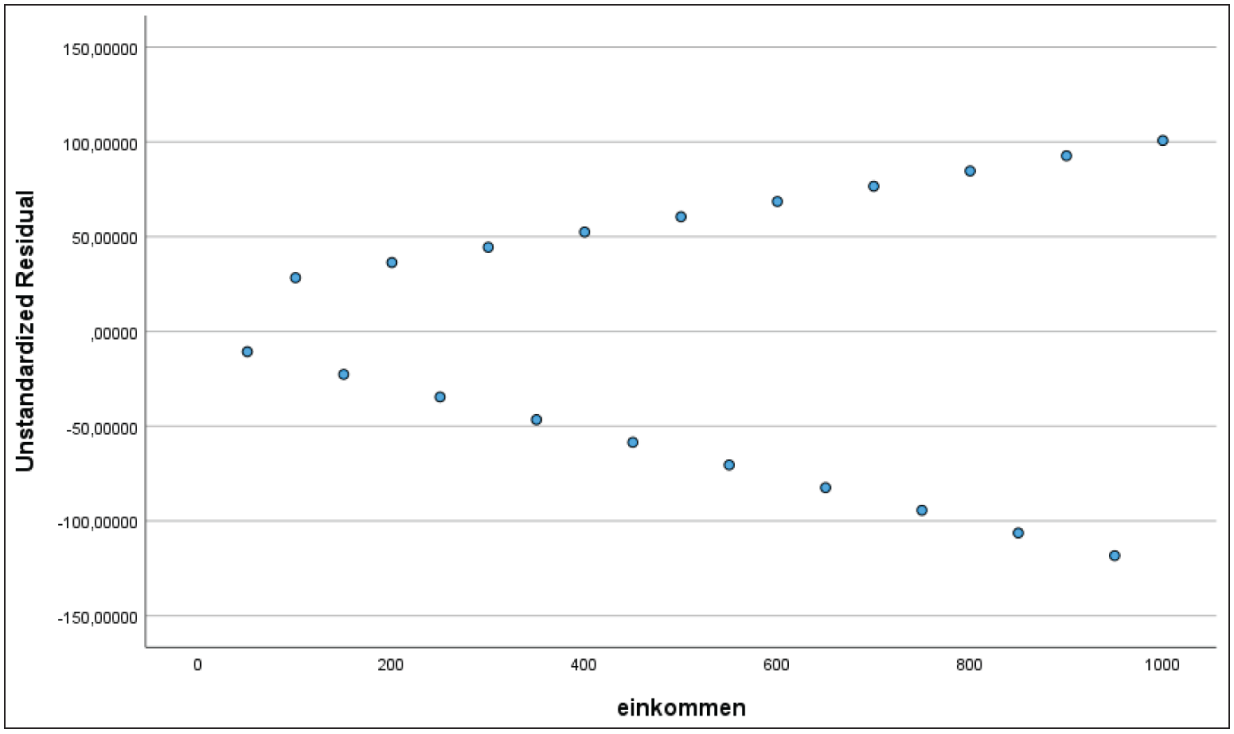



**Abbildung 9.51:** Streudiagramm für die einfachen Residuen in Abhängigkeit von *einkommen* bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT.SAV*

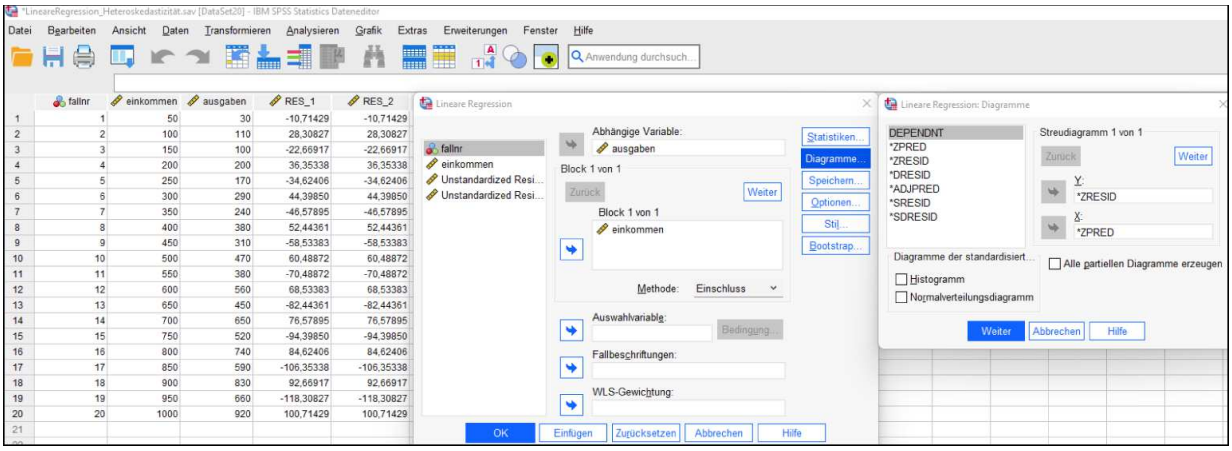



**Abbildung 9.52:** Standardisierte vorhergesagte Werte der abhängigen Variablen bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT.SAV*

Um festzustellen, ob bei Ihren Regressionsschätzungen Heteroskedastizität vorliegt, können Sie über die eben aufgezeigten grafischen Möglichkeiten hinaus in SPSS auch verschiedene Tests durchführen. Aus Platzgründen muss hier der Verweis auf den Breusch/Pagan-, den Cook/Weisberg-, den White- und den Levene-Test auf Heteroskedastizität genügen.

Eine zur Verminderung der Heteroskedastizitätsproblematik vielfach angewandte Variablentransformation stellt die Logarithmierung sowohl der unabhängigen als auch der abhängigen Variablen dar. Zu diesem Zweck bilden Sie in der Datenbasis *LINEAREREGRESSION_HETEROSKEDASTIZITÄT.SAV* bitte folgende neue Variable (wobei Sie ersichtlicherweise hier den natürlichen Logarithmus verwenden – Sie hätten aber auch jeden anderen Logarithmus verwenden können; die neue Datendatei ist im Übrigen in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT1.SAV* abgespeichert):

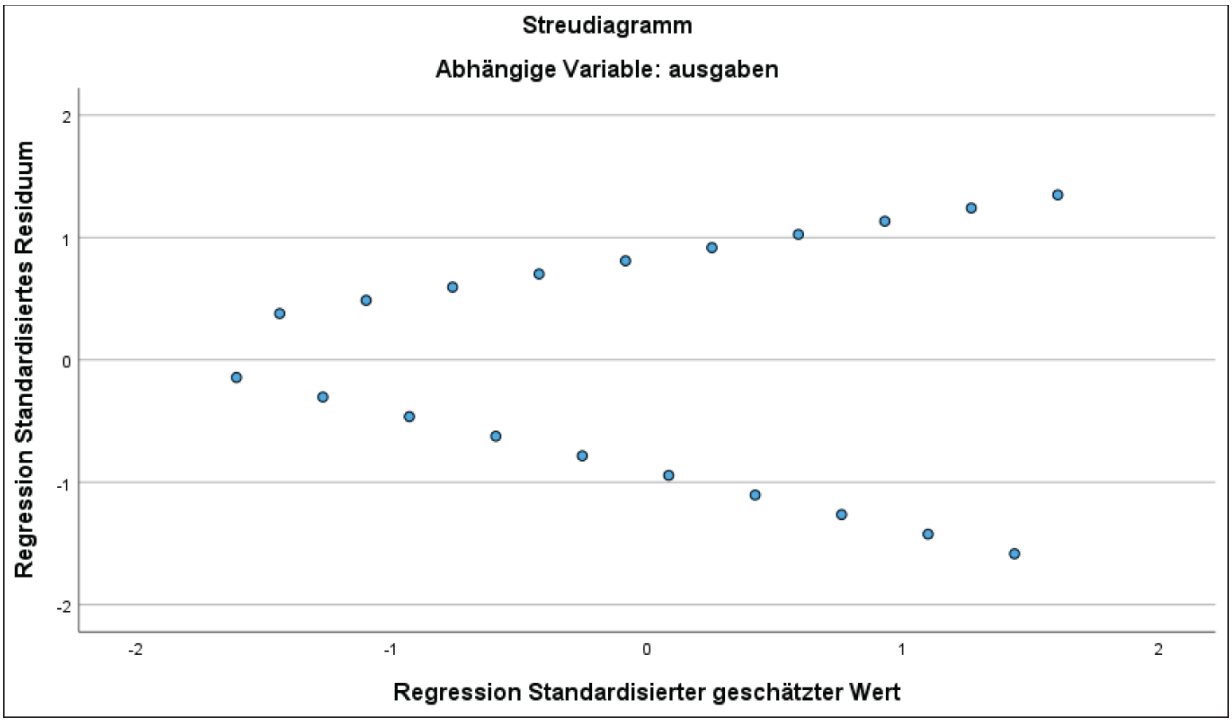



**Abbildung 9.53:** Streudiagramm für die standardisierten Residuen in Abhängigkeit von den standardisierten geschätzten Ausgabenwerten bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT.SAV*

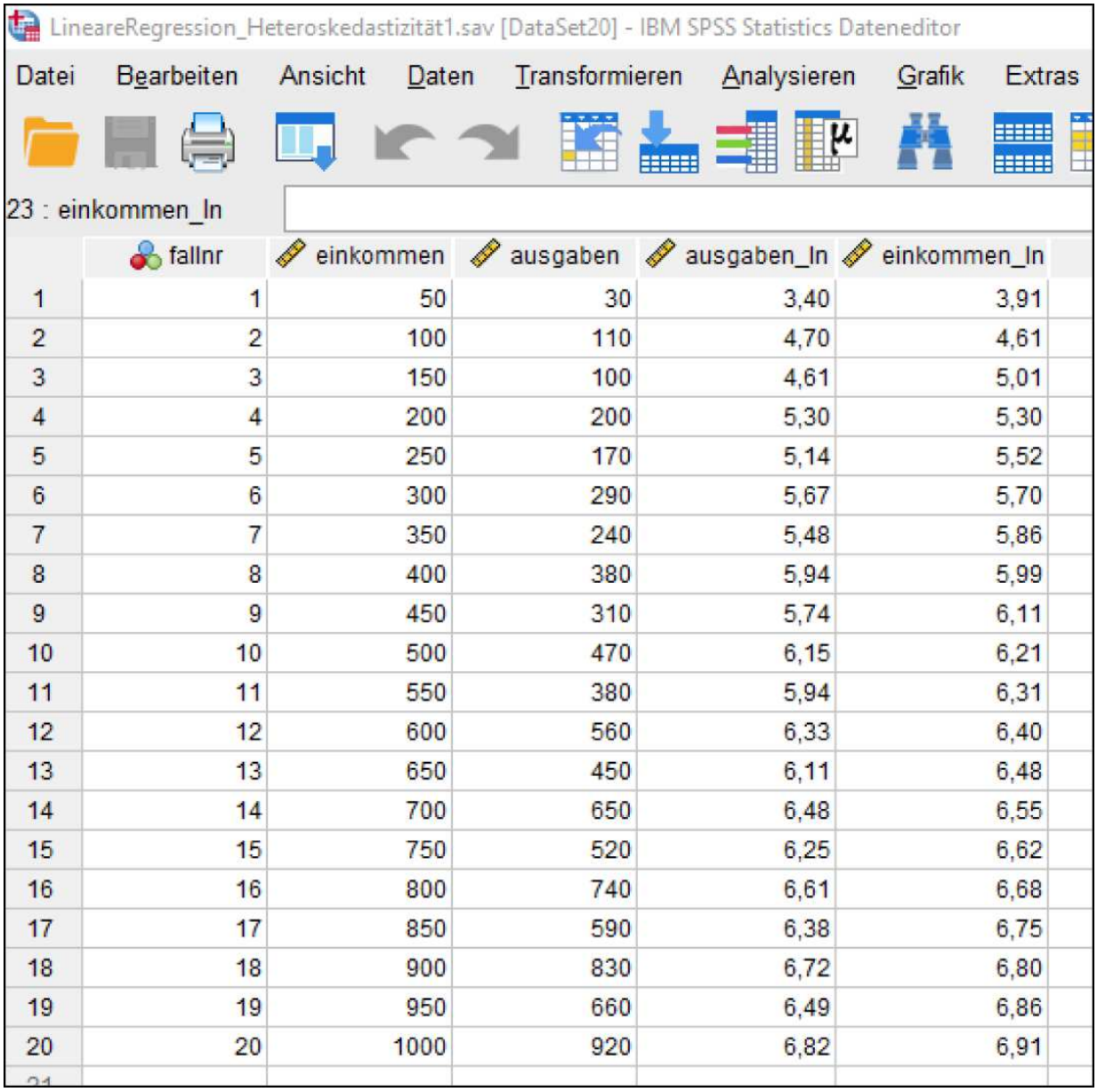
```
COMPUTE ausgaben_ln = ln(ausgaben).
COMPUTE einkommen_ln = ln(einkommen).
EXECUTE.
```

Die Werteausprägungen von *ausgaben_ln* und von *einkommen_ln* gehen aus Abbildung 9.54 hervor.

Anschließend führen Sie mit ANALYSIEREN|REGRESSION|LINEAR eine lineare Regression zwischen *ausgaben_ln* (Y-Variable) und *einkommen_ln* (X-Variable) durch und geben in der Subdialogbox DIAGRAMME als dortige Y-Variable die standardisierten Residuen **zresid* und als dortige X-Variable die standardisierten vorhergesagten Werte für die Ausgabenvariable **zpred* ein. Durch die Logarithmierungen haben Sie die Varianzunterschiede bei den Residuen verringert. Gleichwohl – und da muss man ehrlich bleiben – weist auch noch das Streudiagramm in Abbildung 9.55 ein trichterförmiges Muster auf. Wenngleich nicht mehr ganz so ausgeprägt vorhanden (der »Trichter« hat sich etwas verengt), existiert weiterhin in unserem Beispiel Heteroskedastizität.

Eine präzisere, aber leider auch etwas komplexere Alternative zur Vermeidung beziehungsweise zur Verminderung von Heteroskedastizität stellt eine **gewichtete Regressionsschätzung** dar. Sie wird oft auch als *WLS-Ansatz* bezeichnet (WLS = Weighted Least Squares). Durch die Gewichtung beider Regressionsseiten soll die

Fehlerstreuung »homogenisiert« werden. Hierzu ist im Idealfall die Heteroskedastizitätsstruktur bekannt.




| | fallnr | einkommen | ausgaben | ausgaben_ln | einkommen_ln |
|---|---|---|---|---|---|
| 1 | 1 | 50 | 30 | 3,40 | 3,91 |
| 2 | 2 | 100 | 110 | 4,70 | 4,61 |
| 3 | 3 | 150 | 100 | 4,61 | 5,01 |
| 4 | 4 | 200 | 200 | 5,30 | 5,30 |
| 5 | 5 | 250 | 170 | 5,14 | 5,52 |
| 6 | 6 | 300 | 290 | 5,67 | 5,70 |
| 7 | 7 | 350 | 240 | 5,48 | 5,86 |
| 8 | 8 | 400 | 380 | 5,94 | 5,99 |
| 9 | 9 | 450 | 310 | 5,74 | 6,11 |
| 10 | 10 | 500 | 470 | 6,15 | 6,21 |
| 11 | 11 | 550 | 380 | 5,94 | 6,31 |
| 12 | 12 | 600 | 560 | 6,33 | 6,40 |
| 13 | 13 | 650 | 450 | 6,11 | 6,48 |
| 14 | 14 | 700 | 650 | 6,48 | 6,55 |
| 15 | 15 | 750 | 520 | 6,25 | 6,62 |
| 16 | 16 | 800 | 740 | 6,61 | 6,68 |
| 17 | 17 | 850 | 590 | 6,38 | 6,75 |
| 18 | 18 | 900 | 830 | 6,72 | 6,80 |
| 19 | 19 | 950 | 660 | 6,49 | 6,86 |
| 20 | 20 | 1000 | 920 | 6,82 | 6,91 |

**Abbildung 9.54**: Um zwei logarithmierte Variablen erweiterter Datensatz *LINEAREREGRESSION_HETEROSKEDASTIZITÄT1.SAV*

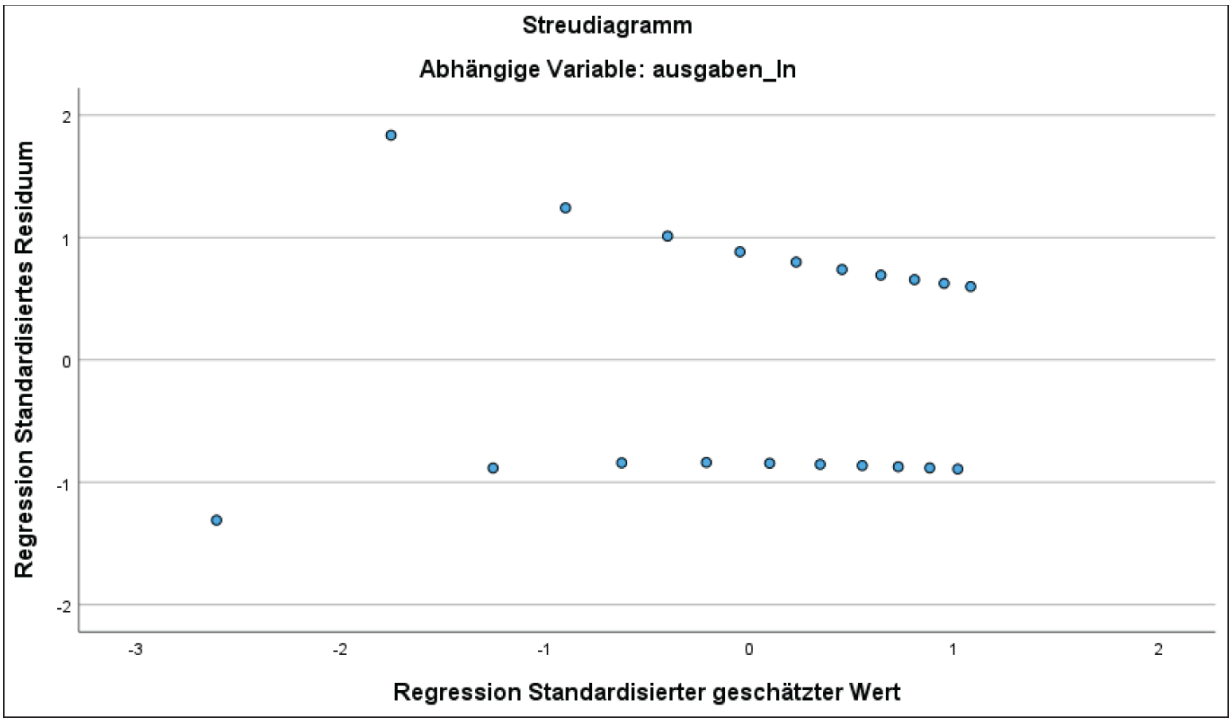



**Abbildung 9.55**: Streudiagramm für die standardisierten Residuen in Abhängigkeit von den standardisierten geschätzten Ausgabenwerten bei einer linearen Regression zwischen *ausgaben_ln* und *einkommen_ln* in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT1.SAV*

Ich habe im Folgenden für eine WLS-Schätzung auf Basis von *LINEAREREGRESSION_HETEROSKEDATIZITÄT.SAV* einfach mal beide Seiten der Regressionsgleichung durch die Werte der X-Variablen (*einkommen*) dividiert (inhaltlich betrachtet, sozusagen die Bedeutung der Einkommensvariablen umgekehrt zu ihrer Höhe angesetzt). Dies ist hier die beschriebene Gewichtung beider Regressionsseiten. Die Gewichtungsvariable habe ich mit *x_invers* bezeichnet (der Bequemlichkeit wegen finden Sie den um die nachfolgende Variable veränderten Datensatz in *LINEAREREGRESSION_HETEROSKEDATIZITÄT2.SAV*):

```
COMPUTE x_invers=1/einkommen.
EXECUTE.
```

Anschließend habe ich den WLS-Ansatz über ANALYSIEREN|REGRESSION|GEWICHTUNGSSCHÄTZUNG aufgerufen (siehe Abbildung 9.56).

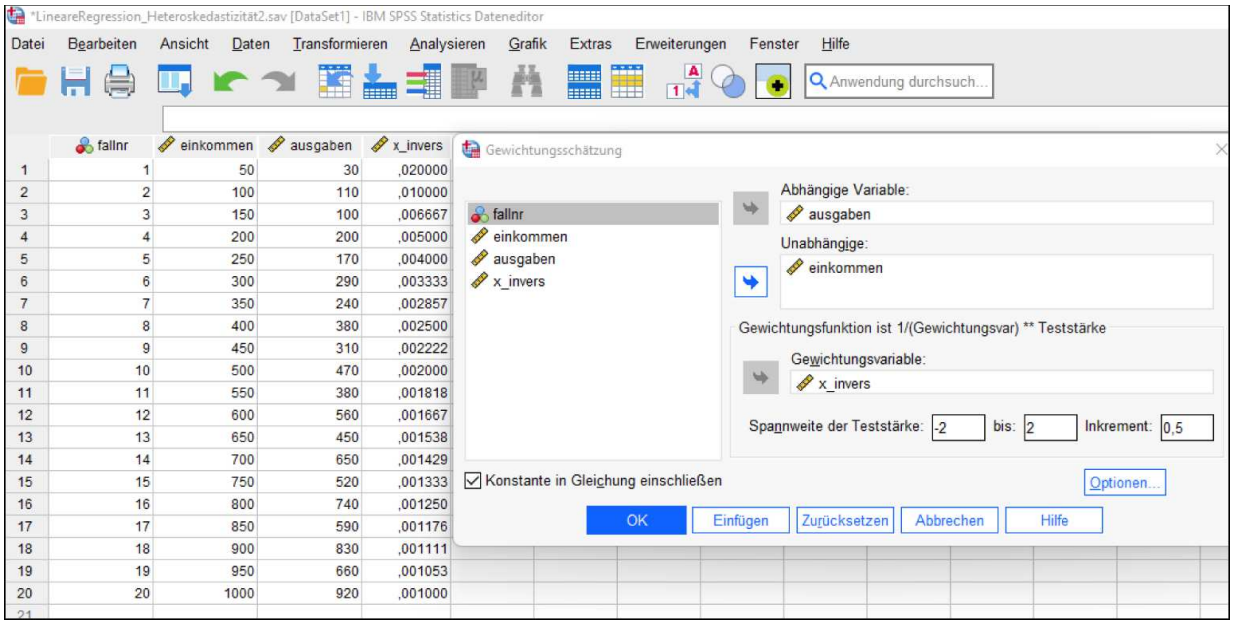



**Abbildung 9.56**: Aufruf einer WLS-Schätzung für den Zusammenhang zwischen *ausgaben* und *einkommen* mit der Gewichtungsvariablen *x_invers* auf Basis von *LINEAREREGRESSION_HETEROSKEDASTIZITÄT2.SAV*

Alternativ zur Bezugnahme auf das SPSS-Menü hätten Sie an Syntax eingeben können:

```
TSET NEWVAR=NONE.
WLS ausgaben WITH einkommen
    /SOURCE x_invers
    /POWER -2 TO 2 BY 0.5
    /CONSTANT
    /PRINT BEST.
```

Im vorliegenden Beispiel erhalten Sie die nachfolgende WLS-Schätzung für den Zusammenhang zwischen *ausgaben* und *einkommen* (siehe Abbildung 9.57).

**Statistiken des besten Modells**

| Modellbeschreibung | | |
|---|---|---|
| Abhängige Variable | | ausgaben |
| Unabhängige Variablen | 1 | einkommen |
| Gewichtung | Quelle | x_invers |
| | Wert der Teststärke | -1,500 |
| Modell: MOD_1. | | |

| Modellzusammenfassung | |
|---|---|
| Multiples R | ,978 |
| R-Quadrat | ,957 |
| Korrigiertes R-Quadrat | ,954 |
| Standardfehler des Schätzers | ,652 |
| Wert der Log-Likelihood-Funktion | -109,260 |

**ANOVA**

| | Quadratsumme | Freiheitsgrade | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|
| Regression | 169,173 | 1 | 169,173 | 397,830 | <,001 |
| Residuum | 7,654 | 18 | ,425 | | |
| Insgesamt | 176,827 | 19 | | | |

**Koeffizienten**

| | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | | |
|---|---|---|---|---|---|---|
| | B | Std.-Fehler | Beta | Std.-Fehler | t | Sig. |
| (Konstante) | -2,513 | 11,033 | | | -,228 | ,822 |
| einkommen | ,826 | ,041 | ,978 | ,049 | 19,946 | <,001 |

**Abbildung 9.57:** Ergebnis einer WLS-Schätzung für den Zusammenhang zwischen *ausgaben* und *einkommen* mit der Gewichtungsvariablen *x_invers* auf Basis von *LINEAREREGRESSION_HETEROSKEDASTIZITÄT2.SAV*

## *Zur Normalverteilungsannahme der Residuen*

Für verschiedene Tests wird eine **Normalverteilung** der Residuen unterstellt. Diese Annahme ist allerdings weniger gewichtig als zum Beispiel die Homoskedastizitätsannahme, weil sich im Unterschied zum Vorliegen von

Heteroskedastizität beim Abweichen von der Normalverteilungsannahme keine verzerrten Schätzparameter ergeben.

Nähere Ausführungen zur Normalverteilung folgen in Teil IV. Dort wird auf entsprechende Normalverteilungstests eingegangen. Dies werden dann dort der Chi-Quadrat-Verteilungstest, der Kolmogorov/Smirnov- und der Shapiro/Wilk-Test sein.

Sie können im Rahmen einer linearen Regression mittels ANALYSIEREN|REGRESSION|LINEAR in der Subdialogbox DIAGRAMME via Augenschein beurteilen, ob die Residuen zumindest näherungsweise normalverteilt sind. Führen Sie auf Basis von *LINEAREREGRESSION_GERADE6.SAV* im Folgenden eine Regression zwischen *ausgaben* (Y-Variable) und *einkommen* (X-Variable) durch. Hierbei klicken Sie in der Subdialogbox DIAGRAMME sowohl auf HISTOGRAMM als auch auf NORMALVERTEILUNGSDIAGRAMM (siehe Abbildung 9.58), was im Übrigen aber nur funktioniert, wenn Sie die Bootstrapping-Einstellung deaktiviert haben.

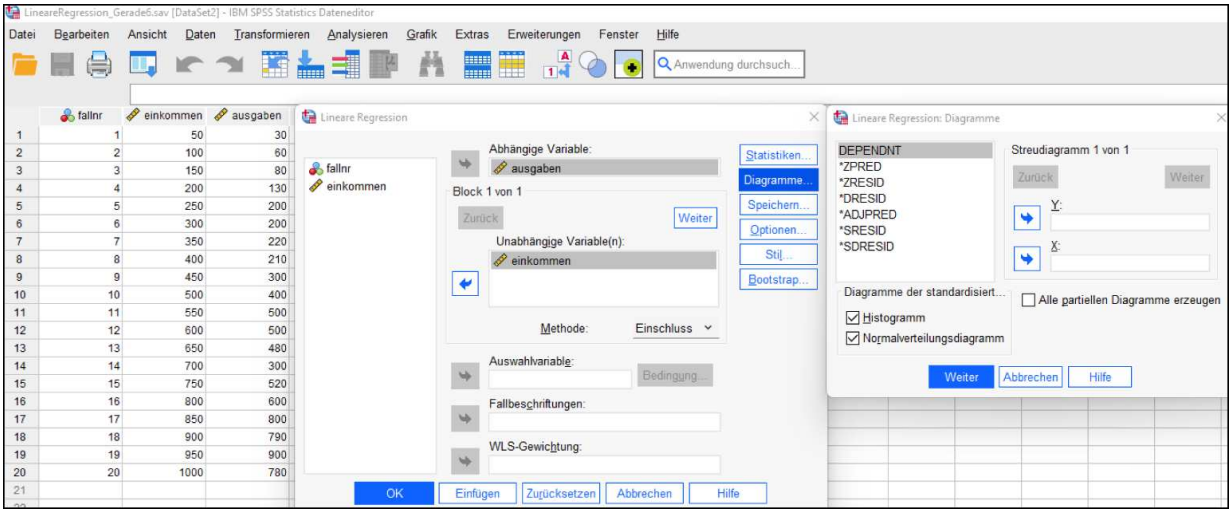



**Abbildung 9.58:** Anfordern der Prüfung auf Normalverteilung in Bezug auf die Residuen bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_GERADE6.SAV*

Auf der Grundlage der vorgenannten Anforderung liefert SPSS Ihnen unter anderem die beiden in Abbildung 9.59 enthaltenen Bilder. Dabei »wirft« SPSS Ihnen das Residuen-*Histogramm* zusammen mit einer (theoretischen) Normalverteilungskurve »aus«. Anhand des linken Bilds in Abbildung 9.59 kann man zwar sicherlich nicht auf eine »hundertprozentige« Normalverteilung der (standardisierten) Residuen in unserem Beispiel schließen. Gleichwohl sieht das Histogramm auch nicht gänzlich anders aus als die auch »Glockenkurve« genannte Normalverteilungskurve (wobei allerdings hier vier Histogramm-Balken diesbezüglich nur vage Aussagen zulassen).

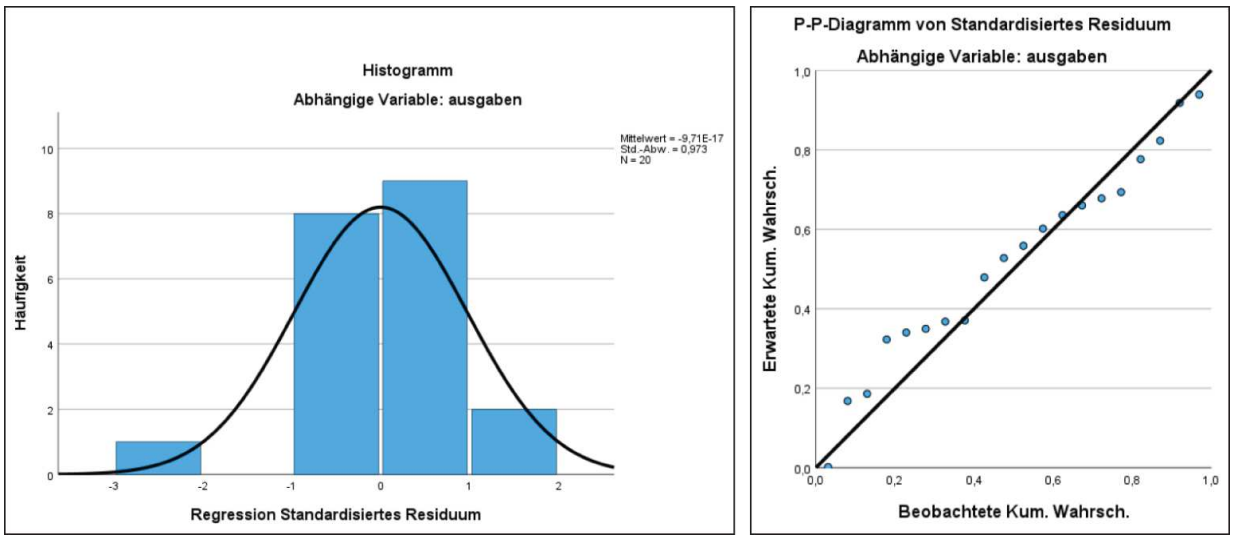



**Abbildung 9.59:** Histogramm und Normalverteilung(-skurve) in Bezug auf die Residuen bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_GERADE6.SAV*

Analoges kommt aus dem rechten Teil von Abbildung 9.59 zum Ausdruck. Dort ist das sogenannte **P/P-Diagramm** abgebildet. Ein P/P-Diagramm stellt die beobachteten den erwarteten standardisierten Residuen gegenüber (jeweils in Form kumulierter Wahrscheinlichkeiten). Im Normalverteilungsfall müssten die Werte auf der Diagonalen liegen. In unserem Beispiel liegen die einzelnen Punkte zumindest nicht allzu weit weg von dieser Diagonalen.

# *Multiple lineare Regression*

Vom Prinzip her funktioniert die **multiple lineare Regression** wie die einfache lineare Regression. Entsprechend stellt sie einen gerichteten linearen Zusammenhang zwischen einer abhängigen und – jetzt hier: – *mehreren* unabhängigen Variablen dar. Auch sie dient der Ursachen- und der Wirkungsanalyse sowie der Prognose. Wichtig ist, dass auch bei ihr der zugrunde liegende kausale Zusammenhang in jedem Fall theoretisch gut zu begründen ist. Die abhängige Variable wird auch »Kriteriumsvariable« und die unabhängigen Variablen werden auch »Prädiktorvariablen« genannt.

## *Annahmen*

Über die bereits bei der linearen Einfachregression genannten **Voraussetzungen**:

- ✔ linearer Zusammenhang,
- ✔ lineare Koeffizienten,
- ✔ Zufallsstichprobe,
- ✔ bedingter Erwartungswert der Fehler der unabhängigen Variablen gleich 0,
- ✔ Nichtkonstanz der Ausprägungen der unabhängigen und der abhängigen Variablen,

- ✔ Homoskedastizität,
- ✔ unabhängige Fehlerwerte (das heißt: keine Autokorrelation) und
- ✔ (zumindest näherungsweise) Normalverteilung der Fehlerwerte

hinaus ist bei der linearen Mehrfachregression insbesondere noch als Voraussetzung zu nennen:

- ✔ möglichst keine Multikollinearität.

Mit fehlender Multikollinearität ist gemeint, dass die unabhängigen Variablen möglichst nicht allzu stark miteinander zusammenhängen (korrelieren) und daher möglichst große eigene Beiträge zur Erklärung der abhängigen Variablen liefern. Gedulden Sie sich noch ein bisschen – bald werde ich ausführlich auf das Multikollinearitätsproblem eingehen.

# *Schätzgleichung*

Bei einer **linearen multiplen Regression** wird von folgender Gleichung ausgegangen, die wie bei der einfachen linearen Regression mit der **Kleinst-Quadrate-Methode** geschätzt wird:

$$\hat{Y}_i = \hat{a}_0 + \hat{a}_1 \cdot X_{1i} + \hat{a}_2 \cdot X_{2i} + \ldots + \hat{a}_k \cdot X_{Ki} + e_i.$$

In dieser Gleichung dienen K X-Variablen der Bestimmung der einen Y-Variablen. Die Koeffizienten dieser K X-Variablen sind zu schätzen, was mit dem »Dach« über ihrer Bezeichnung symbolisiert wird. In den $e_i$ sind alle durch den Zusammenhang zwischen Y und $X_1$, $X_2$, …, $X_K$ nicht erklärten Einflüsse gebündelt. Man fasst diese $e_i$-Größen auch unter dem Begriff **Störterm** (alternativ: *Störgröße*) zusammen.

Da eine Regressionsgleichung wie die vorstehende mit Störeinflüssen behaftet ist, dürfen Sie eine geschätzte Regressionsgleichung nicht einfach mathematisch nach einer der unabhängigen Größen umformen und diese Variable auf diese Weise sozusagen zur abhängigen Variablen machen. Diese Aussage trifft im Übrigen nicht nur auf die multiple, sondern auch schon auf die bivariate Regression zu!

Die Interpretation eines geschätzten Koeffizienten (für eine metrische Variable) erfolgt ceteris paribus (= bei Konstanthaltung der Werte aller anderen X-Variablen), das heißt: Eine Veränderung zum Beispiel der Variablen $X_1$ (zum Beispiel: Einkommen) um eine Einheit führt zu einer Veränderung der Y-Variablen (zum Beispiel: Ausgaben) um $\hat{a}_1$-Einheiten. Symbolisieren $X_1$ beispielsweise das Einkommen und Y die Ausgaben und beläuft sich der geschätzte Parameterwert $\hat{a}_1$ auf +0,750, können Sie die Aussage

treffen, dass eine Erhöhung des Einkommens um 1 Euro bei Konstanthaltung der Werte der anderen Variablen $X_2$, $X_3$, …, $X_K$ die Ausgaben um 75 Cent steigert.

Prinzipiell können auch bei der linearen Mehrfachregression nichtlineare Variablenzusammenhänge in lineare transformiert werden, zum Beispiel dadurch, dass Sie die abhängige Größe und/oder die unabhängigen Variablen logarithmieren, quadrieren oder dergleichen.

## *Aussagekraft der Schätzung*

Für die einzelnen Regressionsparameter wird mittels **statistischer Tests** deren **Signifikanz** überprüft (siehe hierzu auch Teil IV). Es wird auf der Grundlage einer Testverteilung (konkret: auf Basis der sogenannten t-Verteilung) die Nullhypothese, dass ein bestimmter Parameter $a_i = 0$ ist, gegen die Alternativhypothese $a_i \neq 0$ getestet. Wird diese Nullhypothese bei einem bestimmten Signifikanzniveau α (zum Beispiel bei einem Prozent als sogenannter Irrtumswahrscheinlichkeit) verworfen, gilt der betreffende Parameter bei diesem Niveau von α als signifikant (von 0 verschieden). SPSS gibt an dieser Stelle die Signifikanz der Parameterwerte direkt aus.

Wie bei einer Einfachregression kann auch bei einer Mehrfachregression der Erklärungsgrad der Regressionsbeziehung durch das Bestimmtheitsmaß ausgedrückt werden, und zwar als Relation zwischen der durch die Regression erklärten Varianz und der Gesamtvarianz (als Summe aus erklärter und unerklärter Varianz). Man nennt diesen Indikator im Fall einer linearen Mehrfachregression **multiples Bestimmtheitsmaß**.

Da dieser Indikator umso größer wird, desto größer die Zahl der unabhängigen Variablen ist, wird anstelle des einfachen Bestimmtheitsmaßes B häufig das **korrigierte Bestimmtheitsmaß** $B_{korr}$ verwendet:

$$B_{korr} = 1 - (1 - B) \cdot \frac{N - 1}{N - K - 1}.$$

Mit N wird die Anzahl der Beobachtungen und mit K die Anzahl der unabhängigen Variablen bezeichnet. Im Unterschied zu B kann $B_{korr}$ auch negative Werte annehmen. Das korrigierte Bestimmtheitsmaß ist außer in dem Fall, in dem B = 1 ist, kleiner als B.

## *Beispiel*

Im Folgenden wird eine lineare Mehrfachregression auf Basis der Datendatei *LINEAREREGRESSION_MULTIPEL.SAV* mit *ausgaben* als abhängiger Y-Variable und mit *einkommen, haushaltsgroesse* sowie *vermögen* als unabhängigen X-Variablen durchgeführt, und zwar wieder über ANALYSIEREN|REGRESSION|LINEAR bzw. über folgende Syntax:

```
REGRESSION
    /MISSING LISTWISE
    /STATISTICS COEFF OUTS R ANOVA
```

```
/DEPENDENT ausgaben
/METHOD=ENTER einkommen hhgroesse vermögen.
```

Um den Daten in *LINEAREREGRESSION_MULTIPEL.SAV* zumindest eine gewisse inhaltliche Bedeutung zu geben, kann man sich vorstellen, dass es sich bei den Variablen des genannten Datensatzes um die Durchschnittswerte der Ausgaben, der Einkommen, der Haushaltsgrößen und der Vermögen in einer bestimmten Region im Zeitverlauf handelt.

## *Bivariate Korrelationen*

Zunächst werden für diese Regression die **bivariaten Korrelationszusammenhänge** zwischen *ausgaben* einerseits sowie *einkommen, hhgroesse* beziehungsweise *vermögen* andererseits überprüft. Die entsprechenden, in Abbildung 9.60 festgehaltenen bivariaten Zusammenhänge sind über GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM erzeugt worden. Die drei folgenden Korrelationsbilder zeigen im Beispiel jeweils einen positiven Zusammenhang zwischen der Y-Größe *ausgaben* und der jeweiligen X-Variablen, wobei per Augenmaß die für eine lineare Regression zentrale Linearitätsannahme jeweils zumindest plausibel erscheint.

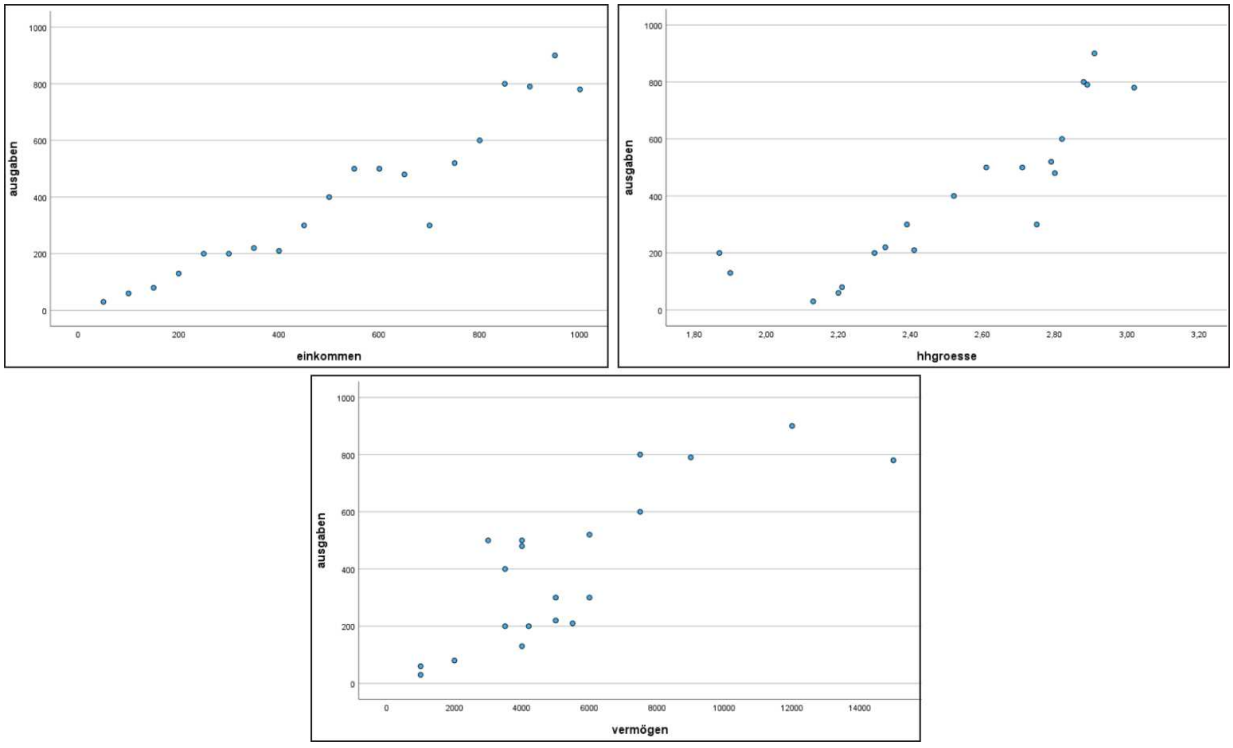



**Abbildung 9.60:** Bivariate Korrelationen zwischen *ausgaben* einerseits sowie *einkommen*, *hhgroesse* und *vermögen* andererseits in *LINEAREREGRESSION_MULTIPEL.SAV*

## *Methoden*

Bei der eben angegebenen Syntax wurde als Methode ENTER angegeben. Dies bedeutet, dass alle von Ihnen als unabhängige Variablen benannten Größen in der

Regression berücksichtigt werden (sofern SPSS nicht das Ausmaß der unten näher behandelten Multikollinearität bei einer bestimmten Variablen als zu hoch einstuft und diese Variable dann bei der Regressionsschätzung weglässt). In der betreffenden Subdialogbox müssen Sie bei METHODE lediglich EINSCHLUSS auswählen (das ist die Menü-Entsprechung für den ENTER-Befehl). Wie Sie Abbildung 9.61 entnehmen können, gibt es aber noch weitere Möglichkeiten, auf die im Folgenden jeweils eingegangen wird.

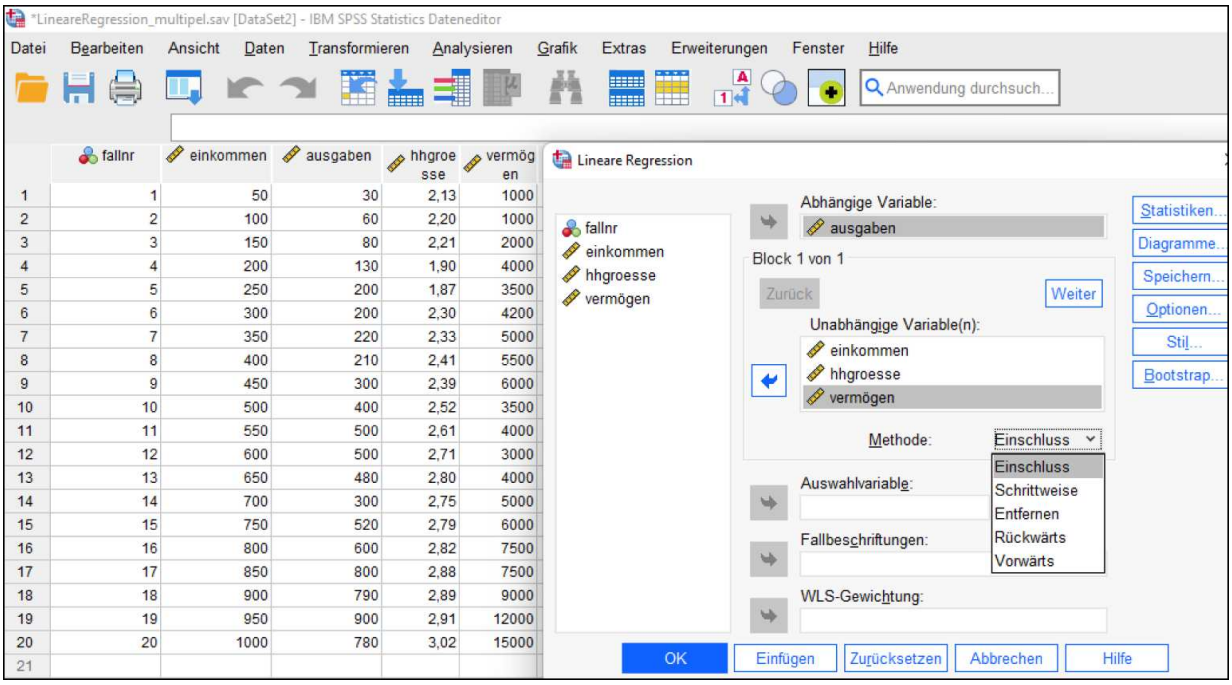



**Abbildung 9.61**: Methoden zur Berechnung einer linearen Mehrfachregression am Beispiel von *LINEAREREGRESSION_MULTIPEL.SAV*

Lassen Sie uns erst mal bei **EINSCHLUSS** (als Syntax, wie bereits geschrieben: ENTER) bleiben. Mit dieser Eingabe erhalten Sie in unserem Beispiel den in Abbildung 9.62 dargestellten Output, der, wie bereits beschrieben, alle drei unabhängigen Variablen umfasst. Hierbei ist aber nur der Regressionskoeffizient von *einkommen* statistisch signifikant. Letzteres erkennen Sie anhand der Spalte »Sig.«, in der für insbesondere *hhgroesse* und *vermögen* die in dieser Spalte angegebenen Wahrscheinlichkeiten deutlich über gängigen Überschreitungs- beziehungsweise Irrtumswahrscheinlichkeiten liegen.

Da im Übrigen die eigentlichen Parameterwerte wesentlich durch die jeweiligen Werteeinheiten beeinflusst werden, ist die Betrachtung der normierten Größen der standardisierten Parameter interessant. Deren Verhältnisse spiegeln die relative Bedeutung der einzelnen Variablen wider. Aus Abbildung 9.62 geht zum Beispiel in diesem Zusammenhang hervor, dass in der vorliegenden Regression auf *ausgaben* – in absoluten Beträgen der Parameterwerte – der Variablen *einkommen* (0,990) die größte Bedeutung vor *hhgroesse* (0,079) und *vermögen* (0,044) zukommt.

Verwenden Sie stattdessen als Methode **SCHRITTWEISE** (als Syntax: STEPWISE), werden die unabhängigen Variablen schrittweise in den Schätzansatz aufgenommen. Einfach ausgedrückt wird zunächst die unabhängige Variable, die am stärksten mit der abhängigen Variablen korreliert, in den Schätzansatz aufgenommen, dann gegebenenfalls die am zweitstärksten mit der abhängigen Variablen korrelierende unabhängige Variable und so weiter. Ob eine unabhängige Variable in das Schätzmodell aufgenommen wird, hängt davon ab, ob sich hierdurch die Modellgüte in Form des Bestimmtheitsmaßes signifikant erhöht oder nicht. Außerdem wird bei jedem (Prüf-)Schritt getestet, ob die am wenigsten »brauchbare« unabhängige Variable weggelassen werden kann oder nicht. Im vorliegenden Beispiel werden die Variablen *hhgroesse* und *vermögen* aus der Regressionsgleichung entfernt, und es wird folglich nur eine Regression zwischen *ausgaben* und *einkommen* von SPSS durchgeführt (siehe Abbildung 9.63). Dieses Ergebnis hatte sich bereits bei der EINSCHLUSS-Methode durch die fehlende Signifikanz der Parameterwerte für *hhgroesse* und *vermögen* angedeutet.

**Regression**

**Aufgenommene/Entfernte Variablen[a]**

| Modell | Aufgenommene Variablen | Entfernte Variablen | Methode |
|---|---|---|---|
| 1 | vermögen, hhgroesse, einkommen[b] | . | Einschluß |

a. Abhängige Variable: ausgaben

b. Alle gewünschten Variablen wurden eingegeben.

**Modellzusammenfassung**

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|---|---|---|---|---|
| 1 | ,953[a] | ,909 | ,892 | 88,650 |

a. Einflußvariablen : (Konstante), vermögen, hhgroesse, einkommen

**ANOVA[a]**

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 1257859,300 | 3 | 419286,433 | 53,353 | <,001[b] |
| | Nicht standardisierte Residuen | 125740,700 | 16 | 7858,794 | | |
| | Gesamt | 1383600,000 | 19 | | | |

a. Abhängige Variable: ausgaben

b. Einflußvariablen : (Konstante), vermögen, hhgroesse, einkommen

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | Regressionskoeffizient B | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | 63,378 | 360,612 | | ,176 | ,863 |
| | einkommen | ,903 | ,279 | ,990 | 3,243 | ,005 |
| | hhgroesse | -61,928 | 183,068 | -,079 | -,338 | ,740 |
| | vermögen | ,003 | ,012 | ,044 | ,291 | ,775 |

a. Abhängige Variable: ausgaben

**Abbildung 9.62:** Ergebnis einer Regression zwischen *ausgaben* sowie *einkommen*, *hhgroesse* und *vermögen* in *LINEAREREGRESSION_MULTIPEL.SAV*: EINSCHLUSS beziehungsweise ENTER

**Koeffizienten**[a]

| Modell | | Nicht standardisierte Koeffizienten: Regressionskoeffizient B | Nicht standardisierte Koeffizienten: Std.-Fehler | Standardisierte Koeffizienten: Beta | T | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Konstante) | -56,158 | 39,251 | | -1,431 | ,170 |
| | einkommen | ,869 | ,066 | ,952 | 13,259 | <,001 |

a. Abhängige Variable: ausgaben

**Ausgeschlossene Variablen**[a]

| Modell | | Beta In | T | Sig. | Partielle Korrelation | Kollinearitätsstatistik: Toleranz |
|---|---|---|---|---|---|---|
| 1 | hhgroesse | -,109[b] | -,532 | ,601 | -,128 | ,127 |
| | vermögen | ,066[b] | ,502 | ,622 | ,121 | ,312 |

a. Abhängige Variable: ausgaben

b. Einflußvariablen im Modell: (Konstante), einkommen

**Abbildung 9.63:** Ergebnis einer Regression zwischen *ausgaben* sowie *einkommen*, *hhgroesse* und *vermögen* in *LINEAREREGRESSION_MULTIPEL.SAV*: SCHRITTWEISE beziehungsweise STEPWISE (Ergebnisauszug)

Bei **ENTFERNEN** (als Syntax: REMOVE) werden unabhängige Variablen im Block von der Regression ausgeschlossen, wenn SPSS sie insgesamt (als Block) nicht als signifikant ansieht. In unserem Beispiel führt dies zu dem ungünstigen Ergebnis, dass alle als unabhängige Variablen vorgesehenen Größen von der Regression ausgeschlossen werden, also auch der gemäß AUSSCHLUSS-Methode ja signifikante Parameter für *einkommen* (siehe den letzten Iterations- beziehungsweise Berechnungsschritt Nummer 2 in Abbildung 9.64).

Bei **RÜCKWÄRTS** (als Syntax: BACKWARD) werden zunächst – analog zu ENTER – alle unabhängigen Variablen im Schätzmodell berücksichtigt. Anschließend wird jeweils schrittweise geprüft, ob – vereinfacht formuliert – die unabhängige Variable mit der jeweils geringsten Korrelation zur abhängigen Größe im Modell verbleibt oder nicht. Für unser Beispiel sind die betreffenden Ergebnisse in Abbildung 9.65 angegeben. Wie ersichtlich, beginnt die RÜCKWÄRTS-Methode im vorliegenden Beispiel mit allen unabhängigen Variablen, schließt dann aber im zweiten Schritt *vermögen* und im dritten Schritt *hhgroesse* von den Regressionsberechnungen aus.

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten: Regressionskoeffizient B | Nicht standardisierte Koeffizienten: Std.-Fehler | Standardisierte Koeffizienten: Beta | T | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Konstante) | 63,378 | 360,612 | | ,176 | ,863 |
| | einkommen | ,903 | ,279 | ,990 | 3,243 | ,005 |
| | hhgroesse | -61,928 | 183,068 | -,079 | -,338 | ,740 |
| | vermögen | ,003 | ,012 | ,044 | ,291 | ,775 |
| 2 | (Konstante) | 400,000 | 60,341 | | 6,629 | <,001 |

a. Abhängige Variable: ausgaben

**Ausgeschlossene Variablen[a]**

| Modell | | Beta In | T | Sig. | Partielle Korrelation | Kollinearitätsstatistik: Toleranz |
|---|---|---|---|---|---|---|
| 2 | einkommen | ,952[b] | 13,259 | <,001 | ,952 | 1,000 |
| | hhgroesse | ,876[b] | 7,700 | <,001 | ,876 | 1,000 |
| | vermögen | ,811[b] | 5,876 | <,001 | ,811 | 1,000 |

a. Abhängige Variable: ausgaben

b. Einflußvariable: (Konstante)

**Abbildung 9.64:** Ergebnis einer Regression zwischen *ausgaben* sowie *einkommen*, *hhgroesse* und *vermögen* in *LINEAREREGRESSION_MULTIPEL.SAV*: ENTFERNEN beziehungsweise REMOVE (Ergebnisauszug)

Die letzte Methode, in SPSS eine multiple Regression zu berechnen, ist **VORWÄRTS** (als Syntax: FORWARD): Die unabhängigen Variablen werden hier also auch gegebenenfalls schrittweise in das Regressionsmodell aufgenommen, und zwar – wiederum vereinfacht ausgedrückt – in der Reihenfolge der Korrelation mit der abhängigen Größe. Sie erhalten in unserem Fall das gleiche Ergebnis wie bei STEPWISE in Abbildung 9.63 (und daher wurde hier für den FORWARD-Fall auf eine eigene Abbildung verzichtet).

Bis auf ENTER sind alle vorstehend beschriebenen Methoden zur Schätzung einer linearen Mehrfachregression rein statistische Methoden, da hier ausschließlich nach statistischen Kriterien von SPSS entschieden wird, ob eine unabhängige Variable in das Schätzmodell aufgenommen wird oder nicht. Es sollte aber immer geprüft werden, ob die berücksichtigten Variablen auch unter inhaltlichen Gesichtspunkten plausibel sind!

# *Multikollinearität*

Neben der Verletzung der Modellannahmen bezüglich der Störgröße (das heißt etwa bei Vorliegen von *Autokorrelation* oder von *Heteroskedastizität*) stellt sich bei einer Mehrfachregression noch ein weiteres Problem. Gemeint ist – auch in Bezug auf die Aussagekraft des multiplen (korrigierten) Bestimmtheitsmaßes – das Problem der **Multikollinearität**. Mitunter wird einfach auch nur von Kollinearität gesprochen.

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | |
|---|---|---|---|---|---|---|
| | | Regressionskoeffizient B | Std.-Fehler | Beta | T | Sig. |
| 1 | (Konstante) | 63,378 | 360,612 | | ,176 | ,863 |
| | einkommen | ,903 | ,279 | ,990 | 3,243 | ,005 |
| | hhgroesse | -61,928 | 183,068 | -,079 | -,338 | ,740 |
| | vermögen | ,003 | ,012 | ,044 | ,291 | ,775 |
| 2 | (Konstante) | 109,885 | 314,538 | | ,349 | ,731 |
| | einkommen | ,962 | ,187 | 1,055 | 5,132 | <,001 |
| | hhgroesse | -85,242 | 160,161 | -,109 | -,532 | ,601 |
| 3 | (Konstante) | -56,158 | 39,251 | | -1,431 | ,170 |
| | einkommen | ,869 | ,066 | ,952 | 13,259 | <,001 |

a. Abhängige Variable: ausgaben

**Ausgeschlossene Variablen[a]**

| Modell | | Beta In | T | Sig. | Partielle Korrelation | Kollinearitätsstatistik Toleranz |
|---|---|---|---|---|---|---|
| 2 | vermögen | ,044[b] | ,291 | ,775 | ,073 | ,252 |
| 3 | vermögen | ,066[c] | ,502 | ,622 | ,121 | ,312 |
| | hhgroesse | -,109[c] | -,532 | ,601 | -,128 | ,127 |

a. Abhängige Variable: ausgaben

b. Einflußvariablen im Modell: (Konstante), hhgroesse, einkommen

c. Einflußvariablen im Modell: (Konstante), einkommen

**Abbildung 9.65:** Ergebnis einer Regression zwischen *ausgaben* sowie *einkommen*, *hhgroesse* und *vermögen* in *LINEAREREGRESSION_MULTIPEL.SAV*: RÜCKWÄRTS beziehungsweise BACKWARD (Ergebnisauszug)

## *Begriff*

Allgemein wird bei Multikollinearität der Erklärungsgehalt eines Regressionsmodells sozusagen »künstlich« dadurch erhöht, dass die unabhängigen Variablen miteinander (stark) korrelieren. Beispielsweise kann man sich vorstellen, dass die gemessene Armutsquote als Y-Variable sowohl in Abhängigkeit von der Geschlechterausprägung »Weibliche Bezugsperson im Haushalt« als auch in Abhängigkeit vom Haushaltstyp »Alleinerziehend« als X-Variablen steigt. Dabei ist der ermittelte Gesamteffekt vermutlich nach oben verzerrt, weil in der Regel den Haushalten der Alleinerziehenden Frauen vorstehen. Aus diesem Grund werden durch »Alleinerziehend« zumindest teilweise die gleichen armutserhöhenden Effekte abgebildet, die bereits in »Weiblicher Haushaltsvorstand« enthalten sind.

Man kann dies mithilfe eines *Venn-Diagramms* veranschaulichen: In Abbildung 9.66 ist der Erklärungsbeitrag von drei erklärenden Größen $X_1$, $X_2$ und $X_3$ hinsichtlich Y in Kreisform dargestellt. Offenkundig überlappen sich die betreffenden Erklärungsbeiträge. Je größer diese Überlappungen – das heißt: die Multikollinearitäten – zwischen den Erklärungsgrößen sind, desto verzerrter ist der in der Regressionsbeziehung angegebene eigenständige Erklärungsbeitrag einer X-Größe in Bezug auf die zu erklärende Größe Y.

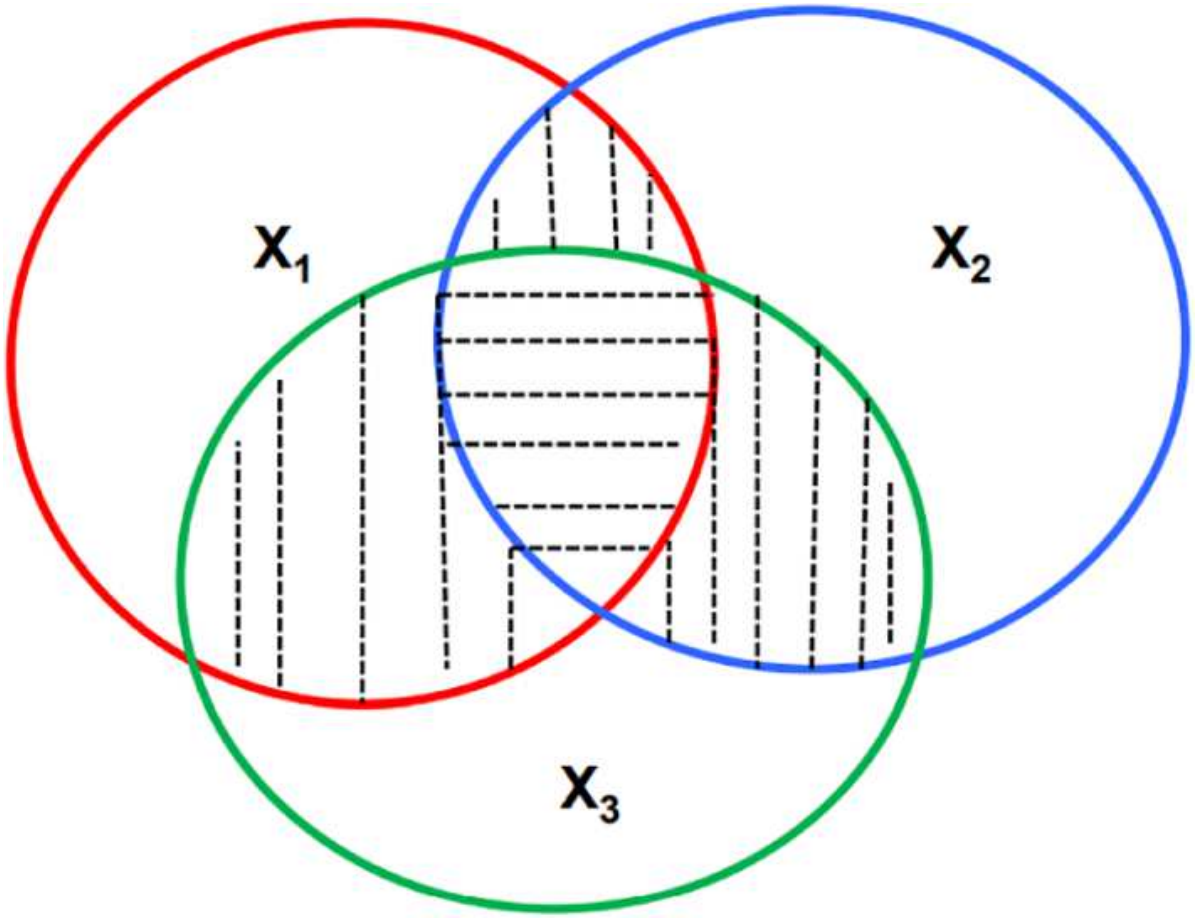


**Abbildung 9.66:** Zur Multikollinearitätsproblematik (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 335)

## *Überprüfung von Multikollinearität*

Eine Multikollinearitätsprüfung kann über den sogenannten **Toleranzwert** $T_j = 1 - r_j^2$ (mit $r_j^2$ als Bestimmtheitsmaß für die Regression der betreffenden X-Variablen j auf alle anderen X-Variablen) beziehungsweise als Kehrwert des Toleranzwerts: als **Varianz-Inflations-Faktor** $VIF_j = \frac{1}{1 - r_j^2}$ erfolgen. Im Sinne einer »Faustregel« spricht man bei $T_j$-Werten größer 0,2 beziehungsweise bei VIF-Werten kleiner 5,0 davon, dass keine große Multikollinearität vorliegt. (Manche Statistiker sind schon mit einem VIF-Wert kleiner 10,0 zufrieden.)

## *SPSS-Beispiel*

In SPSS können Sie sich eine Beurteilung des Ausmaßes der vorliegenden Multikollinearitäten recht einfach ausgeben lassen, wenn Sie – weiterhin bei Verwendung von *LINEAREREGRESSION_MULTIPEL.SAV* – unter ANALYSIEREN|REGRESSION|LINEAR|STATISTIKEN bei KOLLINEARITÄTSDIAGNOSE (und bei TEIL- UND PARTIELLE KORRELATIONEN) ein Häkchen setzen (siehe Abbildung 9.67).

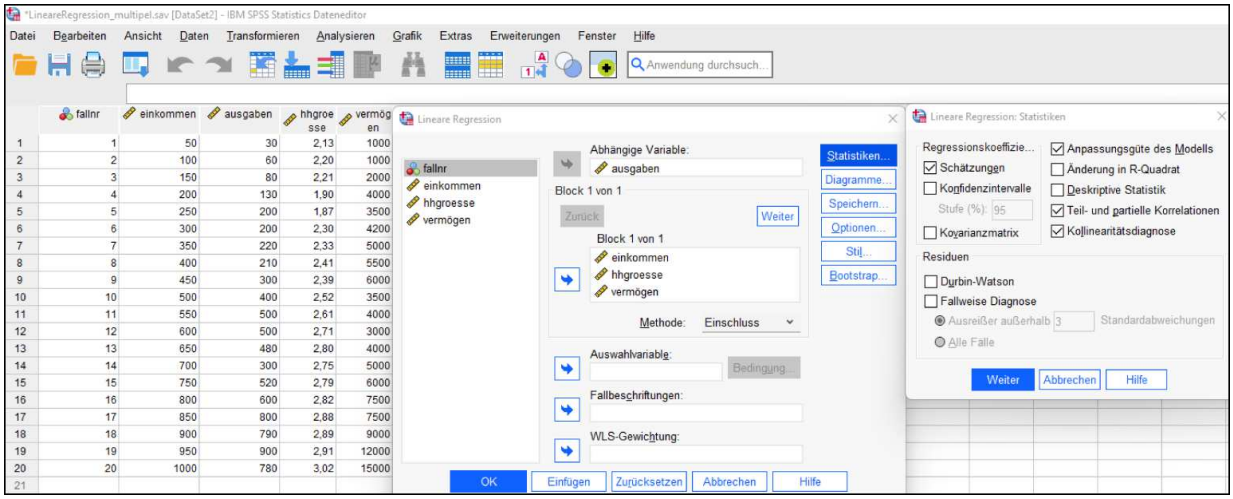



**Abbildung 9.67:** Anfordern einer Kollinearitätsdiagnose für eine Regression zwischen *ausgaben* sowie *einkommen*, *hhgroesse* und *vermögen* in *LINEAREREGRESSION_MULTIPEL.SAV*

Die zugehörige Syntax lautet:

```
REGRESSION
    /MISSING LISTWISE
    /STATISTICS COEFF OUTS R ANOVA COLLIN TOL ZPP
    /CRITERIA=PIN(.05) POUT(.10)
    /NOORIGIN
```

```
/DEPENDENT ausgaben
/METHOD=ENTER einkommen hhgroesse vermögen.
```

Im vorliegenden Beispiel erhalten Sie den in Abbildung 9.68 angegebenen SPSS-Output. Wie aus den beiden letzten Spalten der oberen Pivottabelle in dieser Abbildung ersichtlich, liegen – mit Ausnahme von *vermögen* – die VIF-Werte (deutlich) oberhalb beziehungsweise die Toleranzwerte (deutlich) unterhalb der oben genannten Grenzwerte von 5,0 (VIF) beziehungsweise 0,2 (Toleranz), was jeweils auf Multikollinearität hindeutet.

Falls der von SPSS als weiterer Multikollinearitätsindikator ausgewiesene (und für die mathematischen Spezialisten unter Ihnen: aus den sogenannten Eigenwerten abgeleitete) *Konditionsindex* einen Wert größer 30 aufweist, deutet auch dies auf starke Multikollinearität hin. Letzteres ist aus dem unteren Teil von Abbildung 9.68 in unserem Beispiel in der vierten Zeile ersichtlich (mit einem Wert von 59,344). Die in der gleichen Zeile ausgewiesenen *Varianzanteile* geben an, zwischen welchen unabhängigen X-Variablen kollineare Beziehungen existieren, und zwar gilt dies bei Varianzanteilen größer 0,5. In unserem Beispiel wird in der vierten Zeile mit Ausnahme von *vermögen* jeweils ein höherer Wert als 0,5 ausgewiesen.

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | | Korrelationen | | | Kollinearitätsstatistik | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Regressionskoeffizient B | Std.-Fehler | Beta | T | Sig. | Nullter Ordnung | Partiell | Teil | Toleranz | VIF |
| 1 | (Konstante) | 63,378 | 360,612 | | ,176 | ,863 | | | | | |
| | einkommen | ,903 | ,279 | ,990 | 3,243 | ,005 | ,952 | ,630 | ,244 | ,061 | 16,415 |
| | hhgroesse | -61,928 | 183,068 | -,079 | -,338 | ,740 | ,876 | -,084 | -,025 | ,103 | 9,714 |
| | vermögen | ,003 | ,012 | ,044 | ,291 | ,775 | ,811 | ,073 | ,022 | ,252 | 3,967 |

a. Abhängige Variable: ausgaben

**Kollinearitätsdiagnose[a]**

| Modell | Dimension | Eigenwert | Konditionsindex | Varianzanteile | | | |
|---|---|---|---|---|---|---|---|
| | | | | (Konstante) | einkommen | hhgroesse | vermögen |
| 1 | 1 | 3,745 | 1,000 | ,00 | ,00 | ,00 | ,00 |
| | 2 | ,210 | 4,225 | ,01 | ,01 | ,00 | ,12 |
| | 3 | ,044 | 9,256 | ,00 | ,17 | ,00 | ,66 |
| | 4 | ,001 | 59,344 | ,99 | ,82 | 1,00 | ,22 |

a. Abhängige Variable: ausgaben

**Abbildung 9.68:** Ergebnis einer Kollinearitätsdiagnose für eine Regression zwischen *ausgaben* sowie *einkommen*, *hhgroesse* und *vermögen* in *LINEAREREGRESSION_MULTIPEL.SAV* (Ergebnisauszug)

## *Verminderung von Multikollinearität*

Multikollinearität können Sie dadurch vermindern, dass Sie unabhängige Variablen aus Ihren Regressionsberechnungen ausschließen, so wie dies oben bei den Regressionsmethoden STEPWISE, BACKWARD, FORWARD und REMOVE ja von der Idee her geschehen ist beziehungsweise bei ENTER von SPSS auch automatisch betrieben wird.

Es geht hier nicht darum, Multikollinearität vollständig zu beseitigen. Würden die einzelnen unabhängigen X-Variablen überhaupt nicht miteinander korrelieren, könnten Sie anstelle einer linearen Mehrfachregression mit K X-Variablen nämlich genauso gut K lineare Einfachregressionen schätzen. Anders formuliert: Die lineare Mehrfachregression hat in einem solchen Fall keinen inhaltlichen Mehrwert gegenüber der linearen Einfachregression (beziehungsweise gegenüber K linearen Einfachregressionen). Daher ist vom Grundsatz her bei einer linearen Mehrfachregression sozusagen ein mittleres Multikollinearitätsausmaß erwünscht.

Eine andere Möglichkeit, Multikollinearität zu verringern, besteht in der Zusammenfassung unabhängiger Variablen. Zu diesem Zweck werden nachfolgend zunächst die beiden Variablen *einkommen* und *vermögen* durch eine (Indikator-)Variable *wealth* (im entsprechend aktualisierten Datensatz *LINEAREREGRESSION_MULTIPEL1.SAV*) repräsentiert:

```
COMPUTE wealth = (einkommen + (vermögen/10))/2.
EXECUTE.
```

Die Größe *wealth* ist hierbei – inhaltlich mehr oder weniger willkürlich – als Durchschnitt aus Einkommens- und (angepasstem) Vermögenswert gebildet worden. Es geht mir hier in erster Linie ja um die Darstellung des Prinzips der Variablenzusammenfassung zur Multikollinearitätsreduktion und weniger um inhaltliche Präzision.

Für die lineare Regression zwischen *ausgaben* als Y-Variable sowie *wealth* und *hhgroesse* als X-Variablen erhalten Sie die neue Kollinearitätsdiagnose in Abbildung 9.69. Gegenüber den in Abbildung 9.68 dargestellten Befunden haben sich die Toleranz- und VIF-Werte verbessert, und zwar in Richtung des Überschreitens (VIF-Angaben) beziehungsweise des Unterschreitens (Toleranzwerte) der kritischen Multikollinearitätsgrenzen. So liegt jeweils der Toleranzwert mit 0,299 knapp oberhalb des Schwellenwerts von 0,2, und der VIF-Wert ist jeweils mit 3,349 knapp niedriger als der Schwellenwert von 5,0. Allerdings liegt der Konditionsindex (in der vierten Zeile) mit 30,271, wenn auch nur knapp, weiterhin oberhalb des kritischen Werts von 30. Insgesamt sind die eben erwähnten Verbesserungen zwar keineswegs optimal, aber doch immerhin schon mal ein bisschen was …

**Koeffizienten**[a]

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | | Korrelationen | | | Kollinearitätsstatistik | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Regressionskoeffizient B | Std.-Fehler | Beta | T | Sig. | Nullter Ordnung | Partiell | Teil | Toleranz | VIF |
| 1 | (Konstante) | -603,826 | 248,050 | | -2,434 | ,026 | | | | | |
| | wealth | ,535 | ,136 | ,610 | 3,932 | ,001 | ,916 | ,690 | ,333 | ,299 | 3,349 |
| | hhgroesse | 284,766 | 120,809 | ,365 | 2,357 | ,031 | ,876 | ,496 | ,200 | ,299 | 3,349 |

a. Abhängige Variable: ausgaben

**Kollinearitätsdiagnose**[a]

| Modell | Dimension | Eigenwert | Konditionsindex | Varianzanteile | | |
|---|---|---|---|---|---|---|
| | | | | (Konstante) | wealth | hhgroesse |
| 1 | 1 | 2,855 | 1,000 | ,00 | ,01 | ,00 |
| | 2 | ,141 | 4,494 | ,02 | ,33 | ,00 |
| | 3 | ,003 | 30,271 | ,98 | ,66 | 1,00 |

a. Abhängige Variable: ausgaben

**Abbildung 9.69**: Ergebnis einer Kollinearitätsdiagnose für eine Regression zwischen *ausgaben* sowie *wealth* und *hhgroesse* in *LINEAREREGRESSION_MULTIPEL1.SAV* (Ergebnisauszug)

Ein noch besseres Resultat erzielen Sie, wenn Sie die laut den obigen Befunden ja vergleichsweise wenig korrelierte Variable *vermögen* in den Regressionsberechnungen belassen und dafür die Variablen *einkommen* und *hhgroesse* zu einer Variablen zusammenfassen. Lassen Sie uns aus den beiden letztgenannten Variablen eine Variable *einkommen_prokopf* bilden. Inhaltlich handelt es sich hierbei um das Pro-Kopf-Einkommen (zum Beispiel im Sinne eines durchschnittlichen Pro-Kopf-Einkommens für eine bestimmte Region; der neue Datensatz lautet im Übrigen *LINEAREREGRESSION_MULTIPEL2.SAV*):

```
COMPUTE einkommen_prokopf = einkommen/hhgroesse.
EXECUTE.
```

Mit dieser neuen Variablen führen Sie bitte in *LINEAREREGRESSION_MULTIPEL2.SAV* über ANALYSIEREN|REGRESSION|LINEAR eine lineare Mehrfachregression durch, jetzt zwischen *ausgaben* als abhängiger Variable sowie *einkommen_prokopf* und *vermögen* als unabhängigen Variablen. Machen Sie dabei erneut in der Subdialogbox STATISTIKEN bei KOLLINEARITÄTSDIAGNOSE (und bei TEIL- UND PARTIELLE KORRELATIONEN) ein Häkchen. Die Toleranzwerte der Variablen *einkommen_prokopf* und *vermögen* sind jeweils oberhalb der kritischen Toleranzgrenze von 0,2, und die betreffenden VIF-Werte sind jeweils niedriger als der kritische VIF-Grenzwert von 5,0. Des Weiteren ist, wie aus dem unteren Teil von Abbildung 9.70 zu erkennen ist, der Konditionsindex mit einem Wert in Höhe von 8,747 (letzte Zeile) deutlich unter dem Grenzwert von 30. Mit dem neuen Schätzansatz ist die Multikollinearität also deutlich vermindert worden. Der Wermutstropfen hierbei ist allerdings, dass der Regressionsparameter von *vermögen* weiterhin statistisch nicht signifikant ist, Sie also in diesem (konstruierten) Beispiel nicht von einem signifikanten Einfluss der Vermögens- auf die Ausgabenhöhe ausgehen dürfen.

**Koeffizienten**[a]

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | | | Korrelationen | | | Kollinearitätsstatistik | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | RegressionskoeffizientB | Std.-Fehler | Beta | T | Sig. | Nullter Ordnung | Partiell | Teil | Toleranz | VIF |
| 1 | (Konstante) | -130,529 | 51,376 | | -2,541 | ,021 | | | | | |
| | einkommen_prokopf | 2,422 | ,409 | ,837 | 5,920 | <,001 | ,940 | ,821 | ,480 | ,329 | 3,038 |
| | vermögen | ,010 | ,011 | ,125 | ,884 | ,389 | ,811 | ,210 | ,072 | ,329 | 3,038 |

a. Abhängige Variable: ausgaben

**Kollinearitätsdiagnose**[a]

| Modell | Dimension | Eigenwert | Konditionsindex | Varianzanteile (Konstante) | einkommen_prokopf | vermögen |
|---|---|---|---|---|---|---|
| 1 | 1 | 2,807 | 1,000 | ,02 | ,01 | ,01 |
| | 2 | ,157 | 4,234 | ,68 | ,01 | ,21 |
| | 3 | ,037 | 8,747 | ,30 | ,98 | ,78 |

**Abbildung 9.70:** Ergebnis einer Kollinearitätsdiagnose für eine Regression zwischen *ausgaben* sowie *einkommen_prokopf* und *vermögen* in *LINEAREREGRESSION_MULTIPEL2.SAV* (Ergebnisauszug)

# Kapitel 10

# Komplexere Regressionen

## IN DIESEM KAPITEL

Logit-Regression

Ordinale Regression

Multinomiale Regression

In diesem Kapitel wird auf im Vergleich zu Kapitel 9 weitere, (aus der Vielzahl der von SPSS angebotenen Ansätze) ausgewählte Regressionsansätze eingegangen. Bei ihnen wird die abhängige Regressionsvariable nicht mehr unbeschränkt als metrische Variable modelliert, sondern als binäre (also als 0/1-)Variable beziehungsweise als ordinale oder als multinomiale Variable. In späteren Kapiteln wird noch auf weitere spezielle Regressionsarten eingegangen wie die Cox-Regression (Kapitel 12) oder die robuste Regression (Kapitel 16).

## *Binär-logistische Regression (Logit-Ansatz)*

Bei der binär-logistischen Regression, die auch als **Logit-Analyse** bezeichnet wird, lautet die Frage: Besteht zwischen mehreren unabhängigen und einer binären abhängigen Variablen ein Zusammenhang? Auch hier ist der betreffende kausale Zusammenhang auf jeden Fall theoretisch gut zu begründen!

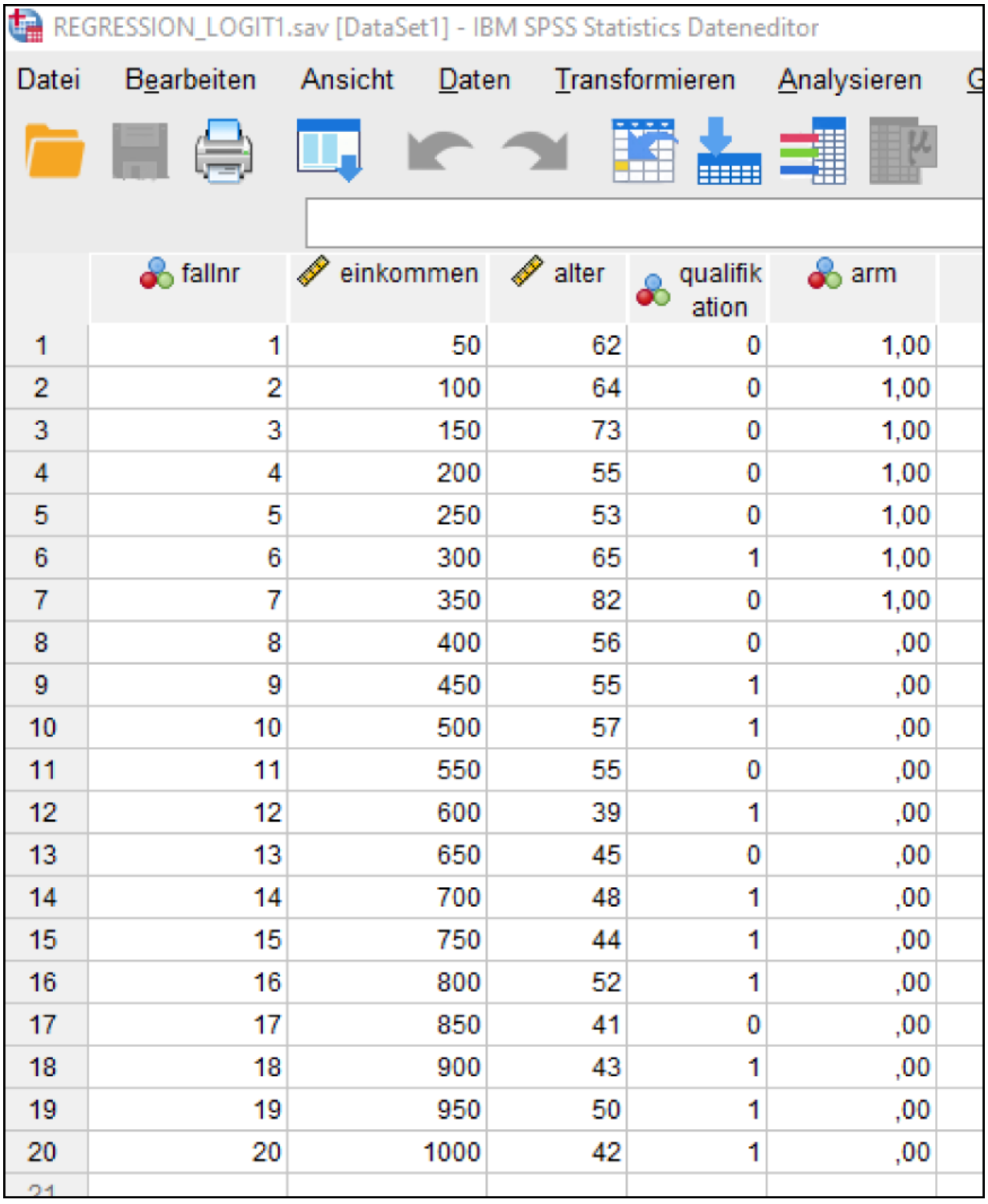
Bei einer Logit-Regression wird also die abhängige Variable als 0/1-Dummy modelliert. Wenn Ihre abhängige Variable nicht als 0/1-Dummy vorliegt, müssen Sie sie durch eine Variablentransformation zu einem solchen »machen«.

Bitte verwechseln Sie die hier behandelte Logit-Regression nicht mit der in Kapitel 9 präsentierten quasilinearen logistischen Regressionsschätzung! Während bei der dort relevanten Schätzung einer logistischen Funktion die abhängige Y-Variable eine metrisch skalierte Variable ist, ist hier bei der binär-logistischen Regression die abhängige Y-Variable ein 0/1-Dummy.

Unter Bezugnahme auf die Datendatei *REGRESSION_LOGIT.SAV* soll es Ihre Aufgabe sein, zunächst einmal aus den dortigen Einkommensangaben in der Variablen *einkommen* eine Unterscheidung in arme und nicht-arme Menschen vorzunehmen. Dabei soll ein Einkommenswert in Höhe von 400 Euro die Unterscheidungsgrenze (inhaltlich: die Armutsgrenze) darstellen. Den Zustand der Armut modellieren Sie mit der Ausprägung 1 und den Zustand der Nichtarmut mit der Ausprägung 0 bei der neuen Variablen *arm*:

```
COMPUTE arm = 0.
IF (einkommen < 400) arm = 1.
EXECUTE.
```

Ihr Datensatz müsste nun so aussehen, wie in Abbildung 10.1 dargelegt. Speichern Sie an dieser Stelle am besten *REGRESSION_LOGIT.SAV* mit der neuen Variablen *arm* ab; ich habe dies in *REGRESSION_LOGIT1.SAV* getan.




| | fallnr | einkommen | alter | qualifikation | arm |
|---|---|---|---|---|---|
| 1 | 1 | 50 | 62 | 0 | 1,00 |
| 2 | 2 | 100 | 64 | 0 | 1,00 |
| 3 | 3 | 150 | 73 | 0 | 1,00 |
| 4 | 4 | 200 | 55 | 0 | 1,00 |
| 5 | 5 | 250 | 53 | 0 | 1,00 |
| 6 | 6 | 300 | 65 | 1 | 1,00 |
| 7 | 7 | 350 | 82 | 0 | 1,00 |
| 8 | 8 | 400 | 56 | 0 | ,00 |
| 9 | 9 | 450 | 55 | 1 | ,00 |
| 10 | 10 | 500 | 57 | 1 | ,00 |
| 11 | 11 | 550 | 55 | 0 | ,00 |
| 12 | 12 | 600 | 39 | 1 | ,00 |
| 13 | 13 | 650 | 45 | 0 | ,00 |
| 14 | 14 | 700 | 48 | 1 | ,00 |
| 15 | 15 | 750 | 44 | 1 | ,00 |
| 16 | 16 | 800 | 52 | 1 | ,00 |
| 17 | 17 | 850 | 41 | 0 | ,00 |
| 18 | 18 | 900 | 43 | 1 | ,00 |
| 19 | 19 | 950 | 50 | 1 | ,00 |
| 20 | 20 | 1000 | 42 | 1 | ,00 |

**Abbildung 10.1:** Um die Variable *arm* erweiterter neuer Datensatz *REGRESSION_LOGIT1.SAV*

Im Rahmen eines Logit-Modells ist – wie erwähnt – die abhängige Variable ein *0/1-Dummy*: Das zugrunde liegende Ereignis tritt entweder

ein oder nicht. Es wird die Wahrscheinlichkeit geschätzt, dass das betreffende Ereignis eintritt. Als Schätzverfahren wird die *Maximum-Likelihood-Methode* genutzt:

$$p(Y=1)=\frac{1}{1+e^{-z}} \Rightarrow z_i=\hat{\beta}_0+\hat{\beta}_1\cdot X_{1i}+\hat{\beta}_2\cdot X_{2i}+\ldots+\hat{\beta}_k\cdot X_{Ki}+\varepsilon_i.$$

Gemäß der vorstehenden Gleichung wird die Wahrscheinlichkeit p geschätzt, dass die Variable Y den Wert 1 annimmt (also das fragliche Ereignis eintritt), und zwar bei bestimmten Wertekonstellationen der erklärenden X-Variablen. Dabei wird die Variable Y in eine logistisch verteilte Variable z schätztechnisch überführt.

Bei der Maximum-Likelihood-Methode lautet die Entscheidungsregel, dass man jene Schätzfunktion bezüglich des interessierenden Grundgesamtheitsparameters auswählen soll, die bei dem beobachteten Stichprobenergebnis die größte Wahrscheinlichkeit hat.

Die geschätzten Koeffizienten eines Logit-Modells können in dem Sinne interpretiert werden, dass sie die Wahrscheinlichkeit für das Auftreten des zugrunde liegenden Ereignisses verändern. Um die Koeffizienten leicht interpretieren zu können, müssen Sie die geschätzten Werte potenzieren – man bildet dann mit einem Fachbegriff die sogenannten Odd ratios – und danach mit 100 Prozent vergleichen. Ein potenzierter Koeffizient in Höhe von 1,25 zum Beispiel besagt, dass ein Anstieg der betreffenden unabhängigen Variablen um eine Einheit die Wahrscheinlichkeit, dass das Y-Ereignis eintritt, um 25 Prozent (125-100) erhöht. Demgegenüber müssten Sie einen potenzierten Koeffizienten in Höhe von zum Beispiel 0,90 dahin gehend interpretieren, dass eine Erhöhung der zugeordneten unabhängigen Variablen um eine Einheit die Wahrscheinlichkeit, dass das Y-Ereignis eintritt, um 10 Prozent vermindert (90-100).

Die Modellgüte kann bei der logistischen Regression (unter anderem) via – in Analogie zur linearen Regression entwickelter – **Bestimmtheitsmaße** beurteilt werden, die mitunter auch als *Pseudo-Bestimmtheitsmaße* bezeichnet werden. Grob ausgedrückt, ergibt sich ein solches Maß aus dem relativen Vergleich der im Rahmen der Maximum-

Likelihood-Schätzung berechneten sogenannten *Likelihood-Funktion* bei völliger Unabhängigkeit aller Variablen (sogenanntes Nullmodell; $L_0$) einerseits und bei Berücksichtigung des Zusammenhangs zwischen der abhängigen und den unabhängigen Variablen ($L_1$) andererseits. Je geringer dieses Verhältnis ist, desto besser ist die Erklärungskraft des letztlich geschätzten Regressionsmodells gegenüber dem Ausgangsmodell, also gegenüber dem Nullmodell. Aus dieser Grundüberlegung heraus ist das **Bestimmtheitsmaß nach Cox/Snell** $\left(r_{CS}^2\right)$ kreiert worden (mit N als Fallzahl):

$$r_{CS}^2 = 1 - \left(\frac{L_0}{L_1}\right)^{\frac{2}{N}}.$$

Da dieses (Pseudo-)Bestimmtheitsmaß nie den Wert 1 für eine perfekte Modellgüte erreichen kann, ist von Nagelkerke entsprechend ein auf diesen oberen Grenzwert normiertes Bestimmtheitsmaß entwickelt worden. Das **Bestimmtheitsmaß nach Nagelkerke** $\left(r_N^2\right)$ hat folgende Definition:

$$r_N^2 = \frac{r_{CS}^2}{1 - L_0^{\frac{2}{N}}}, 0 \le r_N^2 \le 1.$$

Eine weitere, von SPSS (unter anderem) im Rahmen von Logit-Schätzungen ausgewiesene Bestimmtheitsmaß-Variante ist das (Pseudo-)**Bestimmtheitsmaß von McFadden** $\left(r_{McF}^2\right)$, das allerdings wie das Bestimmtheitsmaß nach Cox/Snell daran krankt, dass der obere Grenzwert von 1 nicht erreicht werden kann (ln steht im Übrigen nachfolgend für den natürlichen Logarithmus zur Basis e mit e = 2,71828…):

$$r^2_{McF} = 1 - \frac{\ln(L_1)}{\ln(L_0)}$$

beziehungsweise in korrigierter Form unter Berücksichtigung der Anzahl der unabhängigen Variablen (= K):

$$r^2_{McF,korr} = 1 - \frac{\ln(L_1) - K}{\ln(L_0)}.$$

Lassen Sie uns nun auf der Basis von *REGRESSION_LOGIT1.SAV* eine Logit-Regression zwischen *arm* einerseits sowie *alter* und *qualifikation* andererseits schätzen. Im SPSS-Menü gelangen Sie über ANALYSIEREN|REGRESSION|BINÄR LOGISTISCH zur logistischen Regression.

Sie haben grundsätzlich (das heißt: bei mehreren X-Variablen) – ähnlich wie bei einer linearen Mehrfachregression – die Wahl zwischen verschiedenen Berechnungsmethoden:

- **EINSCHLUSS** (als Syntax: ENTER): Es werden alle unabhängigen Variablen bei der Regressionsberechnung berücksichtigt.
- **SCHRITTWEISE** (als Syntax: STEPWISE): Hier werden die unabhängigen Variablen schrittweise in den Schätzansatz aufgenommen. Es wird zunächst die unabhängige Variable, die am stärksten mit der abhängigen Variablen korreliert, in den Schätzansatz aufgenommen, dann (gegebenenfalls) die am zweitstärksten mit der abhängigen Variablen korrelierende unabhängige Variable und so weiter. Ob eine unabhängige Variable in das Schätzmodell aufgenommen wird, hängt davon ab, ob sich hierdurch die Modellgüte in Form des (Pseudo-)Bestimmtheitsmaßes signifikant erhöht oder nicht. Außerdem wird bei jedem (Prüf-)Schritt getestet, ob die am wenigsten »brauchbare« unabhängige Variable weggelassen werden kann oder nicht.
- **ENTFERNEN** (als Syntax: REMOVE): Hier werden unabhängige Variablen im Block von der Regression ausgeschlossen, wenn SPSS

sie insgesamt (als Block) nicht als signifikant ansieht.

- **RÜCKWÄRTS** (als Syntax: BACKWARD): Hier werden zunächst – analog zu ENTER – alle unabhängigen Variablen im Schätzmodell berücksichtigt. Anschließend wird jeweils sequenziell geprüft, ob die unabhängige Variable mit der jeweils geringsten Korrelation zur abhängigen Größe im Modell verbleibt oder nicht.
- **VORWÄRTS** (als Syntax: FORWARD): Diese Methode funktioniert im Prinzip wie STEPWISE: Die unabhängigen Variablen werden hier also auch (gegebenenfalls) sequenziell in das Regressionsmodell aufgenommen, und zwar in der (absteigenden) Reihenfolge der Korrelation mit der abhängigen Größe.

Bei den nachfolgenden Berechnungen für die Wahrscheinlichkeit, dass eine Person arm ist in Abhängigkeit vom Alter und vom Qualifikationsniveau (*qualifikation* = 1, wenn hoch qualifiziert, *qualifikation* = 0, wenn nicht hoch qualifiziert), ist als Berechnungsmethode EINSCHLUSS unterstellt (siehe auch Abbildung 10.2).

Die entsprechende Syntax sieht so aus:

```
LOGISTIC REGRESSION VARIABLES arm
    /METHOD=ENTER alter qualifikation
    /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

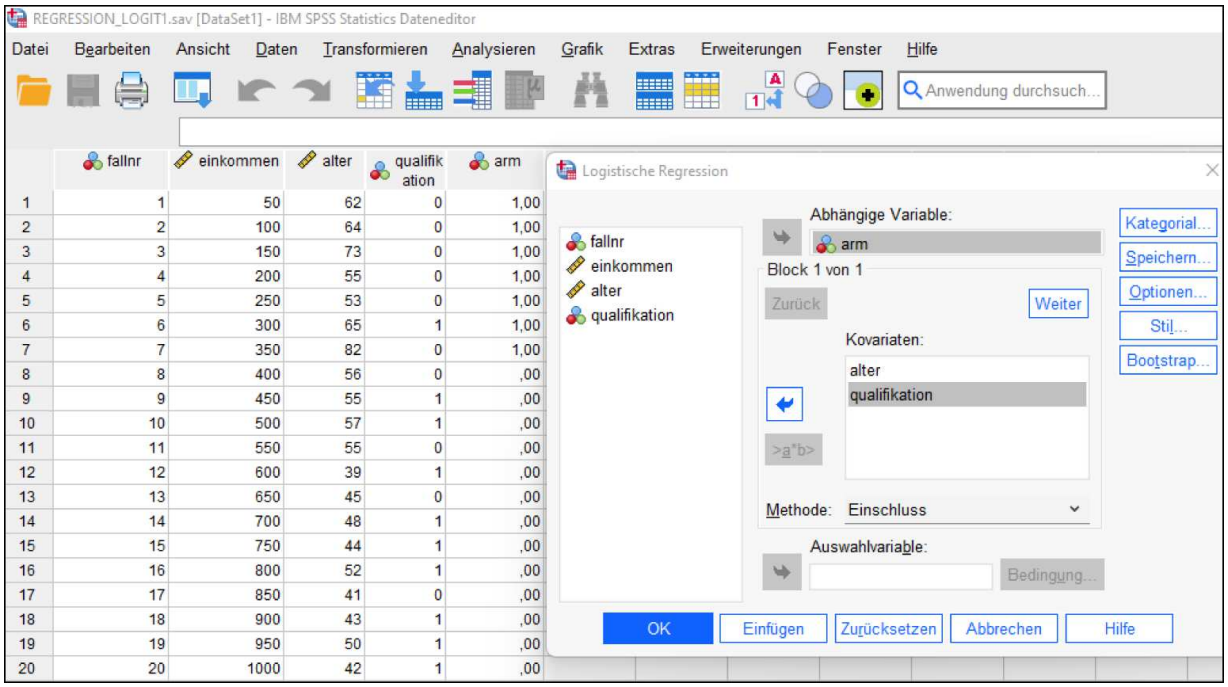



**Abbildung 10.2:** Anfordern einer Logit-Regressionsschätzung zwischen *arm* sowie *alter* und *qualifikation* in der Datendatei *REGRESSION_LOGIT1.SAV*

Lassen Sie diese Syntax »laufen«, erhalten Sie in unserem Fall unter anderem den in Abbildung 10.3 dargestellten Output. Mit einem Wert von 0,813 für Nagelkerkes Bestimmtheitsmaß ist die betreffende Logit-Schätzung als ziemlich ordentlich einzustufen (siehe oberer Teil von Abbildung 10.3). Die geschätzten Regressionsparameter sind allerdings selbst auf einem Signifikanzniveau von 10 Prozent nicht signifikant (siehe unterer Teil von Abbildung 10.3). Dabei sind die Werte der Koeffizienten in der Spalte »RegressionskoeffizientB« beziehungsweise in der Spalte »Exp(B)« angegeben. Anhand dieser letztgenannten Spalte »Exp(B)« kann für die Variable *alter* (indirekt) interpretiert werden, dass ein zusätzliches Lebensjahr die Wahrscheinlichkeit, dem Armutsbereich anzugehören (wenn also *arm* = 1 ist), um (159,4-100) Prozent = 59,4 Prozent erhöht. Für die Variable *qualifikation* gilt hingegen, dass eine hohe Qualifikation (wenn also *qualifikation* = 1 ist) das Armutsrisiko um (100-40,0) Prozent = 60,0 Prozent senkt. Allerdings ist der betreffende Parameter von qualifikation ja leider im vorliegenden Beispiel überhaupt nicht signifikant, das heißt: nicht statistisch aussagekräftig.

**Modellzusammenfassung**

| Schritt | -2 Log-Likelihood | Cox & Snell R-Quadrat | Nagelkerkes R-Quadrat |
|---|---|---|---|
| 1 | 8,035[a] | ,591 | ,813 |

a. Schätzung beendet bei Iteration Nummer 8, weil die Parameterschätzer sich um weniger als ,001 änderten.

**Variablen in der Gleichung**

| | | Regressionskoeffizient B | Standardfehler | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Schritt 1[a] | alter | ,466 | ,288 | 2,630 | 1 | ,105 | 1,594 |
| | qualifikation | -3,212 | 2,562 | 1,571 | 1 | ,210 | ,040 |
| | Konstante | -25,510 | 15,850 | 2,590 | 1 | ,108 | ,000 |

a. In Schritt 1 eingegebene Variablen: alter, qualifikation.

**Abbildung 10.3:** Ergebnis einer Logit-Regressionsschätzung zwischen *arm* und *alter* in der Datendatei *REGRESSION_LOGIT1.SAV* (Ergebnisauszug)

# *Ordinale Regression*

Bei der **ordinalen Regression** kann die abhängige Variable mehr als zwei Ausprägungen annehmen. Dabei sind die Y-Ausprägungen ordinal skaliert, untereinander also in einer Rangordnung.

Öffnen Sie in diesem Zusammenhang den Datensatz *REGRESSION_ORDINAL.SAV*. Um hier jetzt eine ordinale abhängige Variable zu kreieren, sollen aus den Einkommensangaben vier Einkommensbereiche gebildet werden:

```
IF (einkommen < 250) bereich = 1.
IF (einkommen >= 250 and einkommen < 500) bereich = 2.
IF (einkommen >= 500 and einkommen < 750) bereich = 3.
IF (einkommen >= 750) bereich = 4.
EXECUTE.
```

Die Einkommensbereiche sind insofern hierarchisch geordnet, als natürlich die Einkommen im Bereich 1 am niedrigsten und im Bereich 4 am höchsten sind.

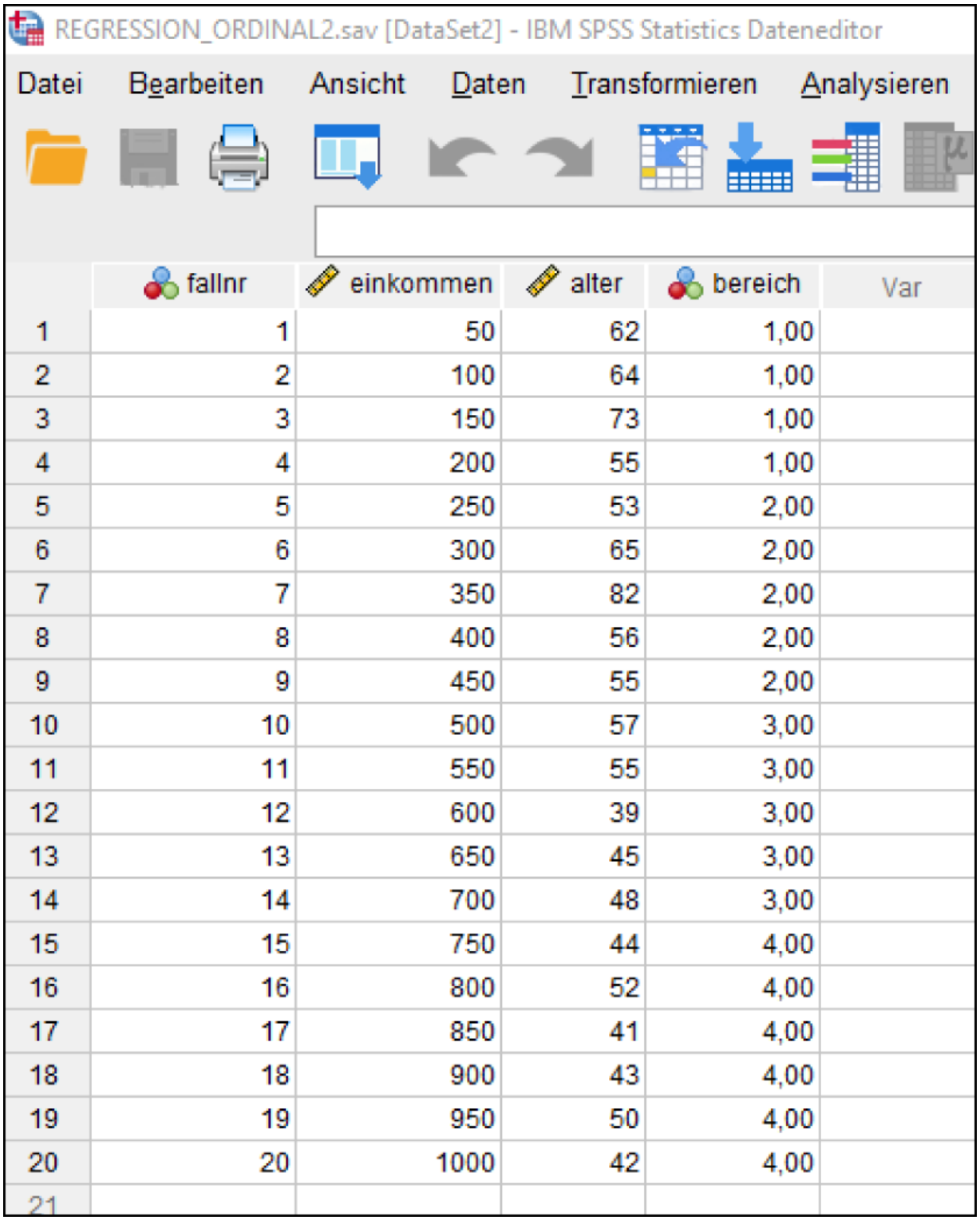
Sie müssten nunmehr den in Abbildung 10.4 dargelegten Datensatz auf Ihrem Bildschirm sehen. Am besten speichern Sie diesen Datensatz mit

der neuen Variablen *bereich* ab – ich habe dies im Übrigen unter *REGRESSION_ORDINAL2.SAV* getan.

Für diesen neuen Datensatz schätzen Sie im nächsten Schritt eine ordinale Regression zwischen *bereich* als Y-Variable und *alter* als X-Variable. Da es sich bei Alter um eine metrische Variable handelt, tragen Sie diese bitte hier in der Rubrik KOVARIATE(N) ein. Wäre im Datensatz zum Beispiel noch eine kategoriale Variable wie *geschlecht* vorhanden gewesen, hätten Sie diese Variable zusätzlich noch unter FAKTOR(EN) eintragen können. Auf diese Art und Weise hätten Sie eine nach der (Gruppierungs-)Variablen *geschlecht* verfeinerte Regressionsschätzung erhalten. Im hier vorliegenden, vergleichsweise einfachen Fall wählen Sie im SPSS-Menü bitte zunächst ANALYSIEREN|REGRESSION|ORDINAL aus (siehe Abbildung 10.5) beziehungsweise geben an Syntax ein:

```
PLUM bereich WITH alter
    /CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5) PCONVERGE
(1.0E-6) SINGULAR(1.0E-8)
    /LINK=LOGIT
    /PRINT=FIT PARAMETER SUMMARY.
```

Mit den vorstehenden Setzungen erhalten Sie unter anderem ein Nagelkerke-Bestimmtheitsmaß in Höhe von 0,469. Des Weiteren sind die Parameterschätzungen für die Konstante sowie für den Altersparameter von SPSS angegeben worden, und zwar für die Einkommensbereiche 1 bis 3; der Bereich 4 ist, wie aus Abbildung 10.6 ersichtlich wird, als Referenzbereich ausgewählt worden. Die bereichsspezifischen Parameterwerte sind daher in Relation zu Bereich 4 (das heißt zum obersten Einkommensbereich) zu sehen. Die »Lage«-Angabe zeigt über das negative Vorzeichen an, dass aus Sicht der Referenzkategorie – das heißt aus Sicht des obersten Einkommensbereichs – in den unteren drei Bereichen tendenziell höhere Lebensalter vorliegen.




| | fallnr | einkommen | alter | bereich | Var |
|---|---|---|---|---|---|
| 1 | 1 | 50 | 62 | 1,00 | |
| 2 | 2 | 100 | 64 | 1,00 | |
| 3 | 3 | 150 | 73 | 1,00 | |
| 4 | 4 | 200 | 55 | 1,00 | |
| 5 | 5 | 250 | 53 | 2,00 | |
| 6 | 6 | 300 | 65 | 2,00 | |
| 7 | 7 | 350 | 82 | 2,00 | |
| 8 | 8 | 400 | 56 | 2,00 | |
| 9 | 9 | 450 | 55 | 2,00 | |
| 10 | 10 | 500 | 57 | 3,00 | |
| 11 | 11 | 550 | 55 | 3,00 | |
| 12 | 12 | 600 | 39 | 3,00 | |
| 13 | 13 | 650 | 45 | 3,00 | |
| 14 | 14 | 700 | 48 | 3,00 | |
| 15 | 15 | 750 | 44 | 4,00 | |
| 16 | 16 | 800 | 52 | 4,00 | |
| 17 | 17 | 850 | 41 | 4,00 | |
| 18 | 18 | 900 | 43 | 4,00 | |
| 19 | 19 | 950 | 50 | 4,00 | |
| 20 | 20 | 1000 | 42 | 4,00 | |
| 21 | | | | | |

**Abbildung 10.4:** Um die Variable *bereich* erweiterter neuer Datensatz *REGRESSION_ORDINAL2.SAV*

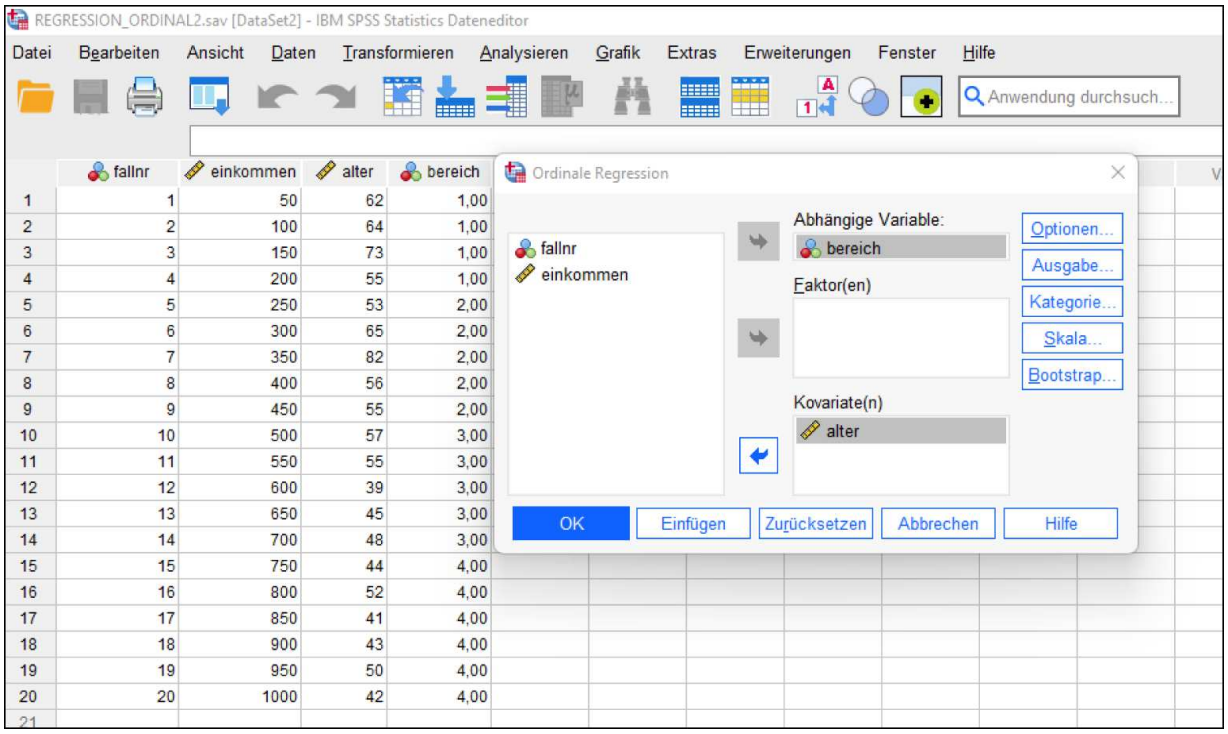



**Abbildung 10.5:** Anfordern einer ordinalen Regressionsschätzung zwischen *bereich* und *alter* in der Datendatei *REGRESSION_ORDINAL2.SAV*

**Pseudo R-Quadrat**

| | |
|---|---|
| Cox und Snell | ,439 |
| Nagelkerke | ,469 |
| McFadden | ,210 |

Verknüpfungsfunktion: Logit.

**Parameterschätzer**

| | | Schätzer | Standard Fehler | Wald | Freiheitsgrade | Sig. | Konfidenzintervall 95% Untergrenze | Konfidenzintervall 95% Obergrenze |
|---|---|---|---|---|---|---|---|---|
| Schwelle | [bereich = 1,00] | -10,536 | 3,312 | 10,119 | 1 | ,001 | -17,028 | -4,044 |
| | [bereich = 2,00] | -8,640 | 2,996 | 8,315 | 1 | ,004 | -14,513 | -2,767 |
| | [bereich = 3,00] | -6,984 | 2,786 | 6,283 | 1 | ,012 | -12,445 | -1,523 |
| Lage | alter | -,156 | ,055 | 8,043 | 1 | ,005 | -,263 | -,048 |

Verknüpfungsfunktion: Logit.

**Abbildung 10.6:** Ergebnis einer ordinalen Regressionsschätzung zwischen *bereich* und *alter* in der Datendatei *REGRESSION_ORDINAL2.SAV* (Ergebnisauszug)

# *Multinomiale Regression*

Auch bei der **multinomialen (logistischen) Regression** kann die abhängige Variable mehr als zwei Ausprägungen annehmen. Dabei ist –

analog zur ordinalen Regression – auch die multinomiale Regression eine Abwandlung des Logit-Schätzansatzes. Die Ausprägungen der Y-Variablen bei der multinomialen Regression stehen allerdings nicht in einer hierarchischen Beziehung zueinander; sie sind vielmehr nominalskaliert.

Öffnen Sie in diesem Zusammenhang den Datensatz *REGRESSION_MULTINOMIAL.SAV*. Dort finden Sie die Variablen *alter* und *region*. Ihre Aufgabe soll es nun sein, einen Regressionszusammenhang zwischen der nominalen Regionsvariablen *region* (mit ihren vier Ausprägungen 1, 2, 3 und 4) als Y-Variable und der metrisch skalierten Altersvariablen als X-Variable herzustellen.

Für diesen inhaltlichen Zusammenhang schätzen Sie sinnvollerweise eine entsprechende multinomiale Regression. Da eine solche Regression iterativ geschätzt wird, sind verschiedene Kriterien anzugeben, zum Beispiel ein Kriterium im Sinne eines Abweichungswerts zwischen zwei aufeinanderfolgenden Schätzungen für das Ende der Schätzungen. Im Folgenden habe ich diesbezüglich die SPSS-Standardeinstellungen einfach übernommen.

Bitte wählen Sie im SPSS-Menü ANALYSIEREN|REGRESSION|MULTINOMIAL LOGISTISCH aus (siehe Abbildung 10.7) beziehungsweise geben Sie an Syntax ein (mit dem Hauptbefehl NOMREG):

```
NOMREG region (BASE=LAST ORDER=ASCENDING) WITH alter
    /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001)
SINGULAR(0.00000001)
    /MODEL
    /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
    /INTERCEPT=INCLUDE
    /PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

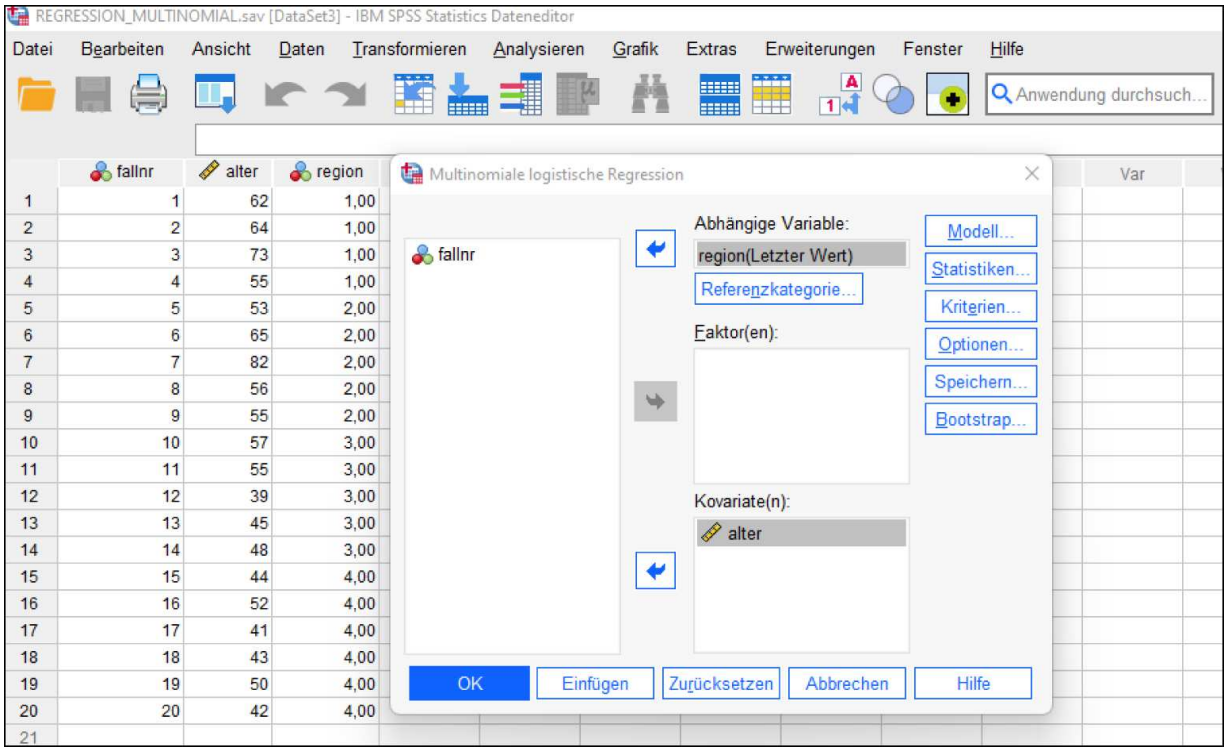



**Abbildung 10.7:** Anfordern einer multinomialen Regressionsschätzung zwischen *region* und *alter* in der Datendatei *REGRESSION_MULTINOMIAL.SAV*

Auf Basis dieser Setzungen ergibt sich unter anderem ein Nagelkerke-Bestimmtheitsmaß in Höhe von 0,629. Außerdem sind in Abbildung 10.8 die Parameterschätzungen für die Konstante sowie für den Altersparameter von SPSS angegeben worden, und zwar für die Regionen 1 bis 3; die Region 4 ist, wie ebenfalls aus Abbildung 10.8 ersichtlich wird, als Referenzregion ausgewählt worden. Die regionenspezifischen Parameterwerte sind daher in Relation zu Region 4 zu sehen. Aus der Spalte »Exp(B)« im unteren Teil von Abbildung 10.8 geht für die Regionen 1 bis 3 jeweils ein Wert größer 1 hervor. Inhaltlich meint dies, dass sich tendenziell in diesen drei Regionen ältere Menschen »tummeln« als in der Referenzregion 4 (und zwar in absteigender Relevanz von der erst- bis zur drittgenannten Region).

| Pseudo-R-Quadrat | |
|---|---|
| Cox und Snell | ,589 |
| Nagelkerke | ,629 |
| McFadden | ,323 |

**Parameterschätzer**

| region[a] | | B | Standard Fehler | Wald | Freiheitsgrade | Signifikanz | Exp(B) | 95% Konfidenzintervall für Exp(B) Untergrenze | Obergrenze |
|---|---|---|---|---|---|---|---|---|---|
| 1,00 | Konstanter Term | -29,272 | 15,028 | 3,794 | 1 | ,051 | | | |
| | alter | ,543 | ,279 | 3,793 | 1 | ,051 | 1,721 | ,997 | 2,974 |
| 2,00 | Konstanter Term | -28,074 | 14,847 | 3,575 | 1 | ,059 | | | |
| | alter | ,528 | ,276 | 3,643 | 1 | ,056 | 1,695 | ,986 | 2,914 |
| 3,00 | Konstanter Term | -5,677 | 5,566 | 1,040 | 1 | ,308 | | | |
| | alter | ,117 | ,118 | ,989 | 1 | ,320 | 1,124 | ,893 | 1,415 |

a. Die Referenzkategorie lautet: 4,00.

**Abbildung 10.8:** Ergebnis einer multinomialen Regressionsschätzung zwischen *bereich* und *alter* in der Datendatei *REGRESSION_MULTINOMIAL.SAV* (Ergebnisauszug)

# Kapitel 11

# Vom Gruppieren und Zusammenfassen von Daten: Diskriminanz-, Cluster- und Faktorenanalyse

## IN DIESEM KAPITEL

Diskriminanzanalyse

Clusteranalyse

Faktorenanalyse

Kapitel 11 behandelt Verfahren zum Gruppieren und Zusammenfassen von Daten. Zunächst wird auf die Diskriminanzanalyse eingegangen, die bestehende Gruppen möglichst trennscharf voneinander abzugrenzen versucht. Das Auffinden derartiger Gruppen ist Gegenstand der Clusteranalyse. Bei ihr werden Untersuchungseinheiten über verschiedene Variablen hinweg zu Gruppen zusammengeführt. Die Faktorenanalyse schließlich fasst verschiedene Erklärungsvariablen zu Variablengruppen (= Faktoren) zusammen. Ebenso wie die Clusteranalyse Zusammenfassungen in Form von Fallgruppen vornimmt, beinhaltet die Faktorenanalyse also ebenfalls Zusammenfassungen, allerdings nicht nach Fallgruppen, sondern nach Variablen.

## *Diskriminanzanalyse*

Ein Verfahren zur gruppenbezogenen Datenanalyse ist die **Diskriminanzanalyse**. Auf ihre Grundkonzeption und ihre Annahmen wird im Folgenden eingegangen, ehe an einem einfachen Beispiel die Durchführung einer solchen Analyse in SPSS gezeigt wird.

## *Idee und Annahmen*

Ziel der Diskriminanzanalyse ist die Erklärung von Gruppenunterschieden beziehungsweise die Ermittlung des Einflusses einzelner Variablen auf die Trennung in verschiedene Gruppen. Hierbei handelt es sich um *vorgegebene* Gruppen (im Unterschied zur unten in diesem Kapitel behandelten Clusteranalyse, bei der nach Gruppen gesucht wird). Als Gruppen kommt ganz Verschiedenes infrage: zum Beispiel Personen- oder Produktgruppen.

Der Name Diskriminanzanalyse kommt aus dem Lateinischen von discriminare und bedeutet »trennen« beziehungsweise »unterscheiden«. Es geht allein um die statistische Unterscheidung verschiedener Gruppen, ohne jegliche normative Konnotation. Wenn im Rahmen der Diskriminanzanalyse also diskriminiert wird, dann nur im Sinne des statistischen Trennens beziehungsweise Unterscheidens und keineswegs in einem negativ besetzten politischen Sinne!

Die Diskriminanzanalyse ist insofern – ebenso wie zum Beispiel die Regressionsanalyse (siehe Kapitel 9 und 10) – Bestandteil der *Dependenzanalyse*, als sie in eine abhängige Variable und mehrere metrisch skalierte unabhängige Variablen unterscheidet. Hierbei bildet die Gruppierungsvariable die abhängige Variable. Es kann sich bei der Gruppierungsvariablen um eine nominale oder allgemeiner um eine kategoriale Größe (zum Beispiel: verschiedene Produktarten oder Unternehmenstypen) handeln. Die metrisch skalierten unabhängigen Variablen sollen für die jeweiligen Gruppen charakteristische Merkmale sein.

Im Rahmen der Diskriminanzanalyse wird mathematisch als Linearkombination aus den unabhängigen Variablen eine

**Diskriminanzfunktion** gebildet, deren Werte darüber entscheiden, zu welcher Gruppe eine Untersuchungseinheit gehört. Der einfachste Fall besteht aus zwei Gruppen mit einer einzigen Diskriminanzfunktion. Ein solcher Fall wird nachfolgend betrachtet. Grundsätzlich sind aber auch Diskriminanzanalysen mit mehr als zwei Gruppen und dann auch mit mehr als einer Diskriminanzfunktion möglich. Dabei ist darauf zu achten, dass die Anzahl der Gruppen nicht größer als die Anzahl der unabhängigen Variablen ist.

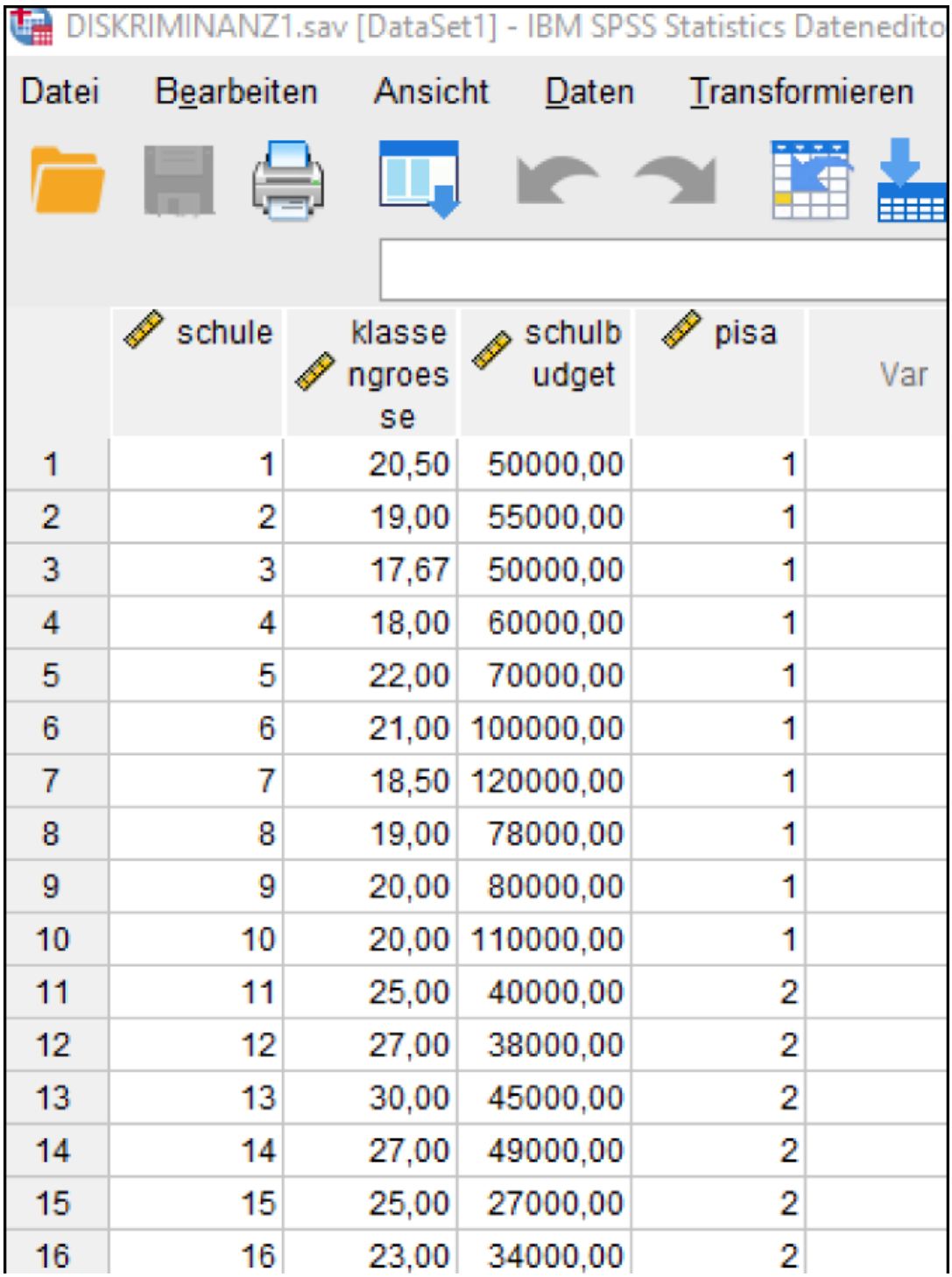
Es ist demnach das Hauptziel einer Diskriminanzanalyse, mithilfe eines *Diskriminanzkriteriums* die einzelnen Gruppen möglichst gut voneinander zu trennen. Anders ausgedrückt: Die Distanz zwischen den mittleren Diskriminanzwerten der Gruppen soll maximal, die Distanz in den Diskriminanzwerten innerhalb der Gruppen zum Gruppenmittelwert hingegen minimal sein.

Neben ihrem Ziel, Gruppen möglichst gut voneinander abzugrenzen, hat die Diskriminanzanalyse auch noch ein **Prognoseziel**. Dabei geht es darum, ein neues Untersuchungselement einer bestehenden Gruppe zuzuordnen. Haben Sie etwa als Gruppen »Impulsive Käufer« und »Rationale Käufer« als Käufertypen voneinander unterschieden, und Sie betrachten einen neu hinzukommenden Käufer, dann würden Sie ihn auf Basis der für ihn maßgeblichen Ausprägungen der unabhängigen Variablen beziehungsweise mittels der gebildeten Diskriminanzfunktion in eine der beiden Gruppen einsortieren.

## *Ein einfaches Beispiel in SPSS*

Das nachstehende Beispiel für die Diskriminanzanalyse bezieht sich auf den (fiktiven) Datensatz *DISKRIMINANZ1.SAV* (siehe Abbildung 11.1). Auf Basis von PISA-Ergebnissen (PISA = Programme for International Student Assessment) sind hier, wie erwähnt: fiktiv, 20 Schulen in überdurchschnittlich erfolgreiche Schulen (*pisa* = 1) und in unterdurchschnittlich erfolgreiche Schulen (*pisa* = 2) unterschieden worden. Es wird demnach von zwei Gruppen ausgegangen. Die für die beiden Gruppen charakteristischen Merkmale stellen zum einen die (durchschnittliche) Klassengröße einer Schule und zum anderen das Schulbudget der jeweiligen Schule dar. Es werden also zwei

unabhängige Variablen *klassengroesse* und *schulbudget* unterstellt. In Schule Nummer 1 zum Beispiel, die gemäß PISA-Ergebnis zu den überdurchschnittlich erfolgreichen Schulen zählt (*pisa* = 1), beträgt die durchschnittliche Klassengröße 20,50 Schüler und das betreffende Schulbudget 50.000,00 Geldeinheiten.




| | schule | klassengroesse | schulbudget | pisa | Var |
|---|---|---|---|---|---|
| 1 | 1 | 20,50 | 50000,00 | 1 | |
| 2 | 2 | 19,00 | 55000,00 | 1 | |
| 3 | 3 | 17,67 | 50000,00 | 1 | |
| 4 | 4 | 18,00 | 60000,00 | 1 | |
| 5 | 5 | 22,00 | 70000,00 | 1 | |
| 6 | 6 | 21,00 | 100000,00 | 1 | |
| 7 | 7 | 18,50 | 120000,00 | 1 | |
| 8 | 8 | 19,00 | 78000,00 | 1 | |
| 9 | 9 | 20,00 | 80000,00 | 1 | |
| 10 | 10 | 20,00 | 110000,00 | 1 | |
| 11 | 11 | 25,00 | 40000,00 | 2 | |
| 12 | 12 | 27,00 | 38000,00 | 2 | |
| 13 | 13 | 30,00 | 45000,00 | 2 | |
| 14 | 14 | 27,00 | 49000,00 | 2 | |
| 15 | 15 | 25,00 | 27000,00 | 2 | |
| 16 | 16 | 23,00 | 34000,00 | 2 | |

| | | | | | |
|---|---|---|---|---|---|
| 17 | 17 | 24,00 | 27000,00 | 2 | |
| 18 | 18 | 29,00 | 33000,00 | 2 | |
| 19 | 19 | 31,00 | 40000,00 | 2 | |
| 20 | 20 | 33,00 | 43000,00 | 2 | |

**Abbildung 11.1:** Datensatz DISKRIMINANZ1.SAV

## *Festlegung der unabhängigen und der Gruppierungsvariablen*

Auf der Grundlage der in Abbildung 11.1 dargestellten Datendatei soll jetzt eine Diskriminanzanalyse mit den beiden unabhängigen Variablen *klassengroesse* und *schulbudget* durchgeführt werden. Als Gruppierungsvariable fungiert *pisa*.

Grundsätzlich starten Sie eine Diskriminanzanalyse in SPSS im Menü mit ANALYSIEREN|KLASSIFIZIEREN|DISKRIMINANZANALYSE. Anschließend öffnet sich das in Abbildung 11.2 dargestellte Dialogfenster, in das Sie zunächst einmal die für Ihre Diskriminanzanalyse ausgewählte Gruppierungsvariable eintragen müssen – im vorliegenden Fall also: *pisa*. Für diese Variable muss darüber hinaus ihr Wertebereich angegeben werden: 1 (in der inhaltlichen Bedeutung von »überdurchschnittlich erfolgreiche Schule«) und 2 (in der inhaltlichen Bedeutung von »unterdurchschnittlich erfolgreiche Schule«). Hierdurch werden also die beiden voneinander unterschiedenen Gruppen definiert.

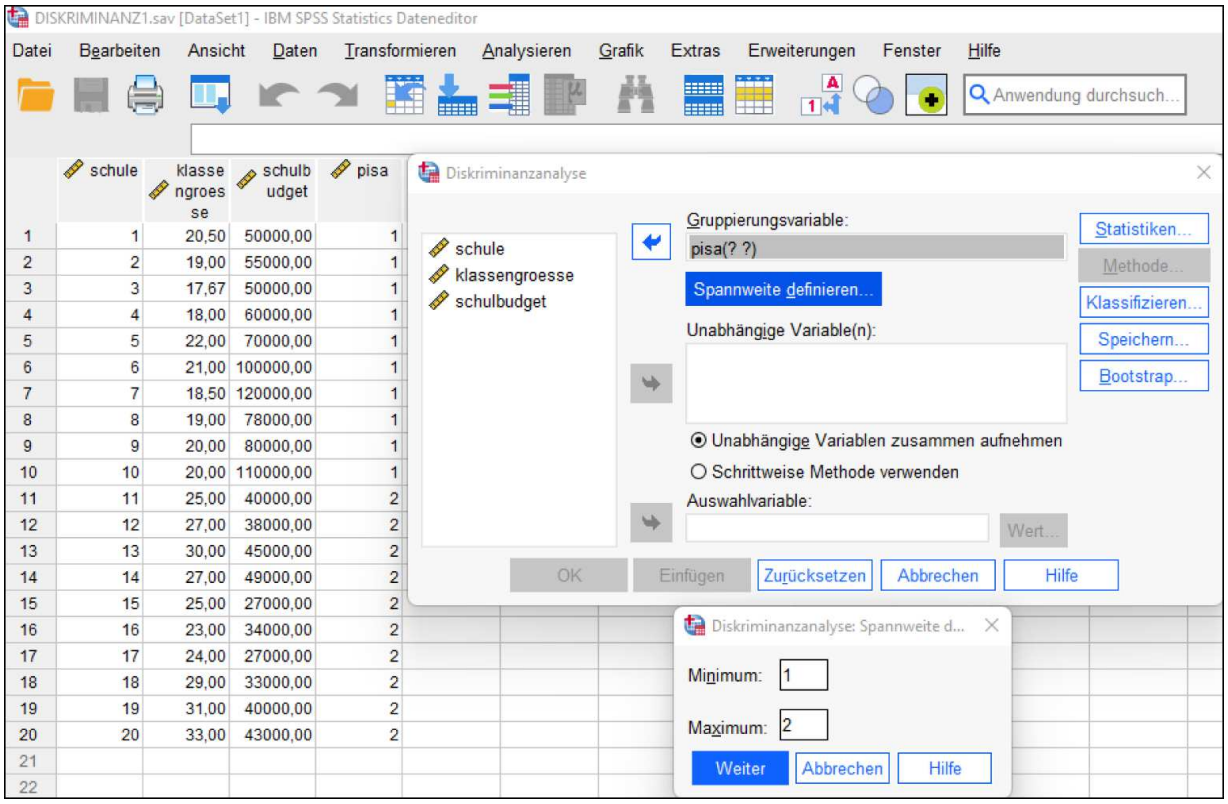



**Abbildung 11.2:** Auswahl der Gruppierungsvariablen *pisa* mit Werten 1 und 2 (DISKRIMINANZ1.SAV)

Der nächste Schritt beinhaltet die Angabe der für die Diskriminanzanalyse relevanten unabhängigen Variablen. In unserem Beispiel sind dies bekanntlich die Variablen klassengroesse und *schulbudget* (siehe Abbildung 11.3).

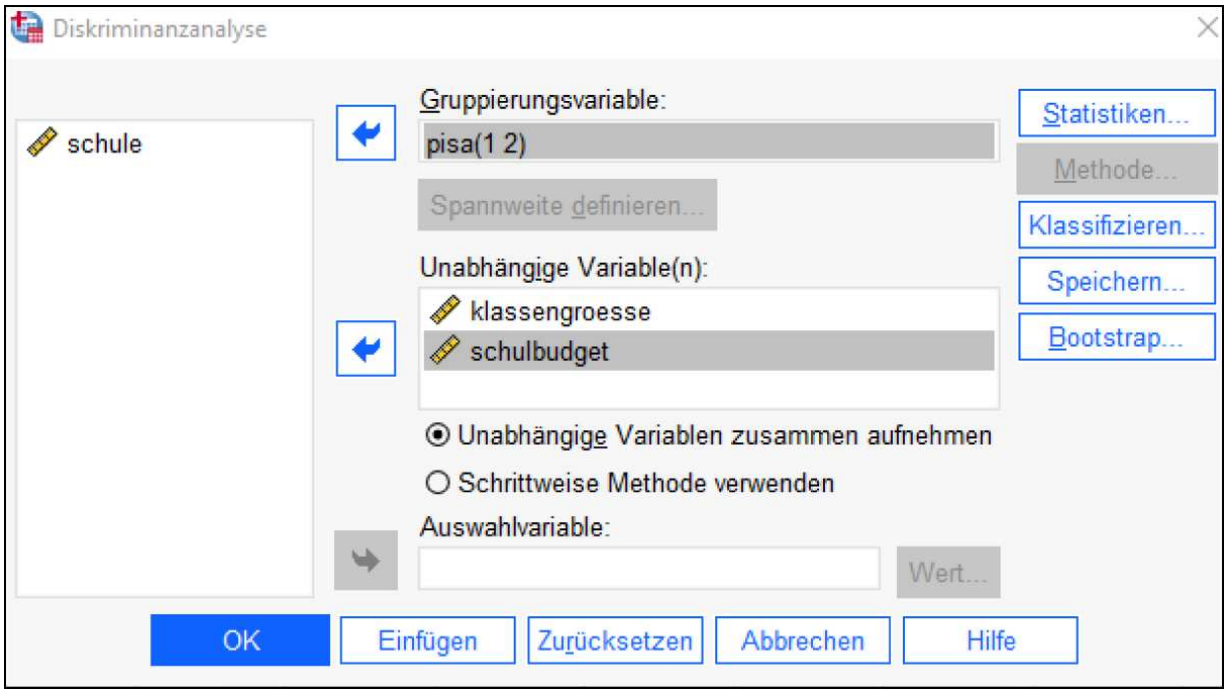



**Abbildung 11.3:** Auswahl der unabhängigen Diskriminanzvariablen: *klassengroesse* und *schulbudget*(DISKRIMINANZ1.SAV)

## *Spezifische Festlegungen*

Nachdem Sie sozusagen den Rahmen Ihrer Diskriminanzanalyse festgelegt haben, spezifizieren Sie Ihre Analyse dadurch, dass Sie angeben,

- ✔ welche deskriptiven Statistiken ausgegeben werden sollen,
- ✔ wie genau Ihre Daten gruppiert sind (gleiche oder unterschiedliche Gruppengröße und so weiter) und
- ✔ welche Prognosen zur Gruppenzugehörigkeit Sie in welcher Form erhalten möchten.

In der Subdialogbox DISKRIMINANZANALYSE: STATISTIK wählen Sie zunächst ein paar Statistiken aus. Um deskriptive gruppenspezifische Kennwerte zu erhalten, klicken Sie bitte auf MITTELWERT und auf UNIVARIATE ANOVA. Auch könnten Sie sich Korrelations- und Kovarianzergebnisse zwischen Ihren unabhängigen Variablen ausgeben lassen; aus Übersichtlichkeitsgründen habe ich davon abgesehen. Demgegenüber klicken Sie unter FUNKTIONSKOEFFIZIENT bitte auf

FISHER und auf NICHT STANDARDISIERT. Hierdurch werden Ihnen später die für gruppenbezogene Prognosen wichtigen Koeffizienten der betreffenden *Klassifizierungsfunktion* (bezüglich der Gruppen 1 und 2) und der (sogenannten kanonischen) *Diskriminanzfunktion* »ausgeworfen« (siehe hierzu Abbildung 11.4).

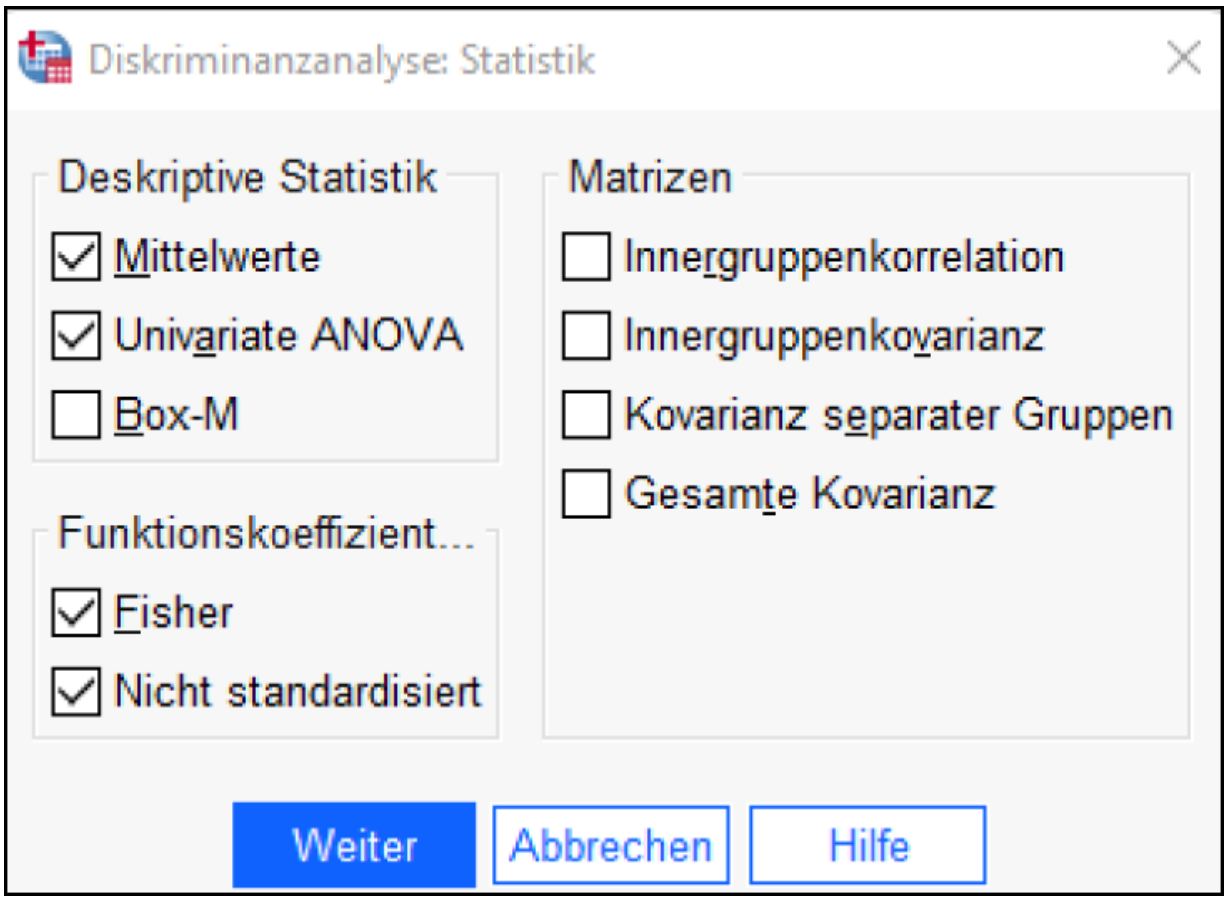



**Abbildung 11.4:** Auswahl von Diskriminanz-Statistiken (hier in Bezug auf DISKRIMINANZ1.SAV)

In der Subdialogbox DISKRIMINANZANALYSE: KLASSIFIZIEREN geben Sie ferner für unser Beispiel an, dass beide Gruppen die gleiche Gruppengröße haben, und zwar besteht jede der beiden Gruppen aus 10 Schulen. Hierdurch wird im Berechnungsprozedere von SPSS eine gleiche anfängliche Eintrittswahrscheinlichkeit – von jeweils 50 Prozent – für die Zuordnung einer Schule zu einer der beiden Gruppen unterstellt.

In der betreffenden Subdialogbox hätten Sie sich auch noch für die einzelnen Fälle diverse Befunde anzeigen lassen können – etwa FALLWEISE ERGEBNISSE. Aus Übersichtlichkeitsgründen sind aber in Abbildung 11.5 keine weiteren diesbezüglichen Auswahlen getroffen worden. Wie Sie gleich sehen werden, ist der SPSS-Output für unser kleines Beispiel zur Diskriminanzanalyse auch so schon recht umfangreich.

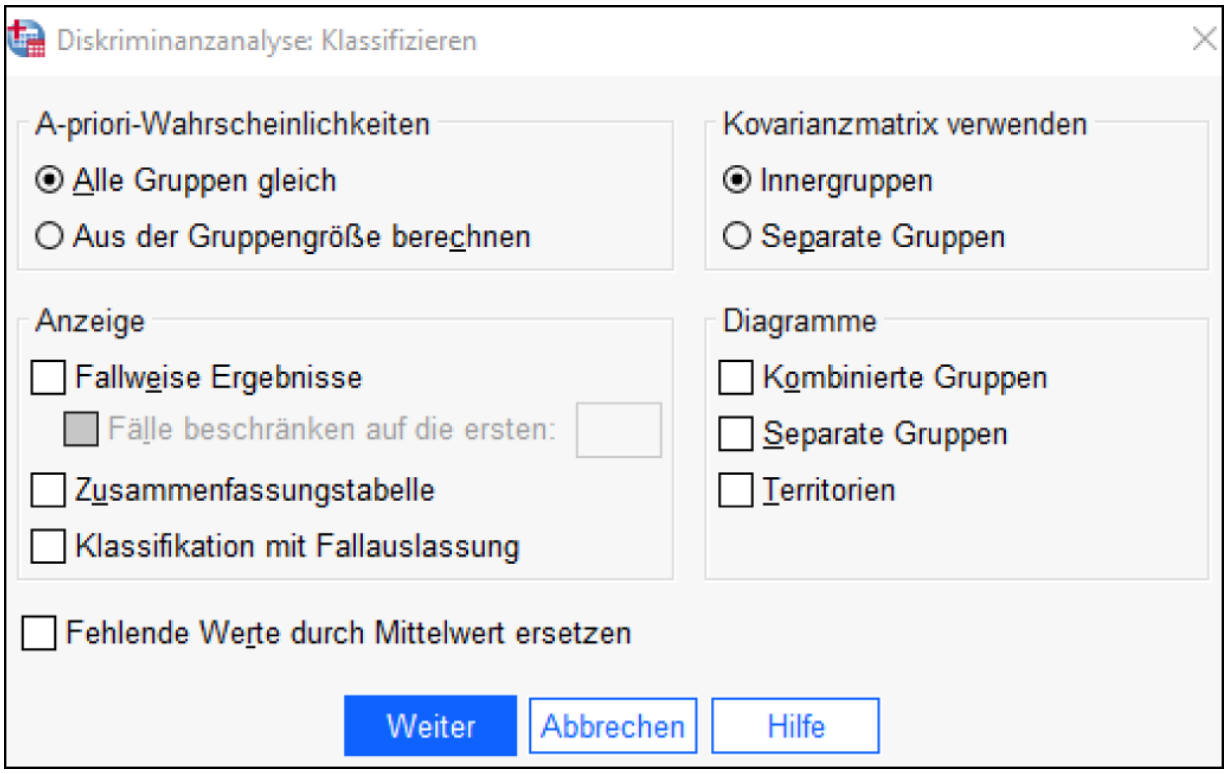



**Abbildung 11.5:** Auswahl der (gleichen) Gruppengröße im Rahmen der Diskriminanzanalyse (in Bezug auf DISKRIMINANZ1.SAV)

Auf der Dialogkarte DISKRIMINANZANALYSE: SPEICHERN geben Sie bitte noch an, dass Ihnen SPSS in die Datendatei *DISKRIMINANZ1.SAV* zwei neue Variablen einfügt, und zwar eine Variable, die die berechnete Gruppenzugehörigkeit direkt angibt, und eine andere Variable, die auf Basis der Diskriminanzfunktion diese Gruppenzugehörigkeit indirekt ausweist. Anschließend klicken Sie auf die OK-Taste, und es kann losgehen mit den SPSS-Berechnungsergebnissen zu unserem kleinen Beispiel für eine Diskriminanzanalyse (siehe hierzu Abbildung 11.6).

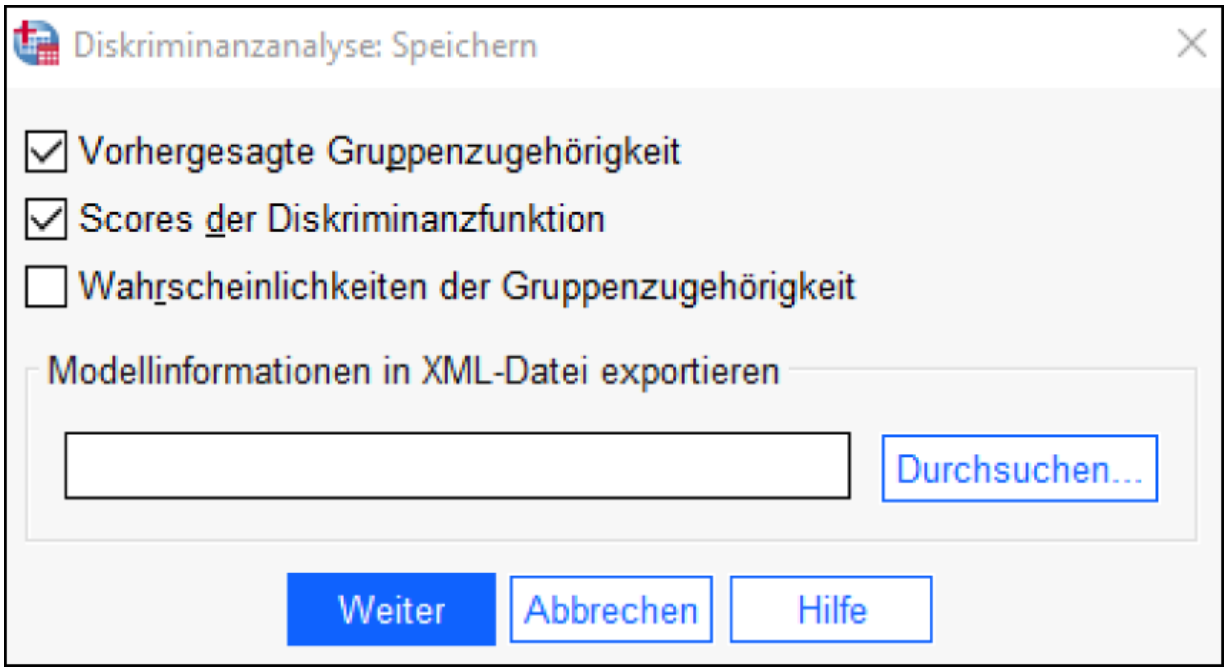
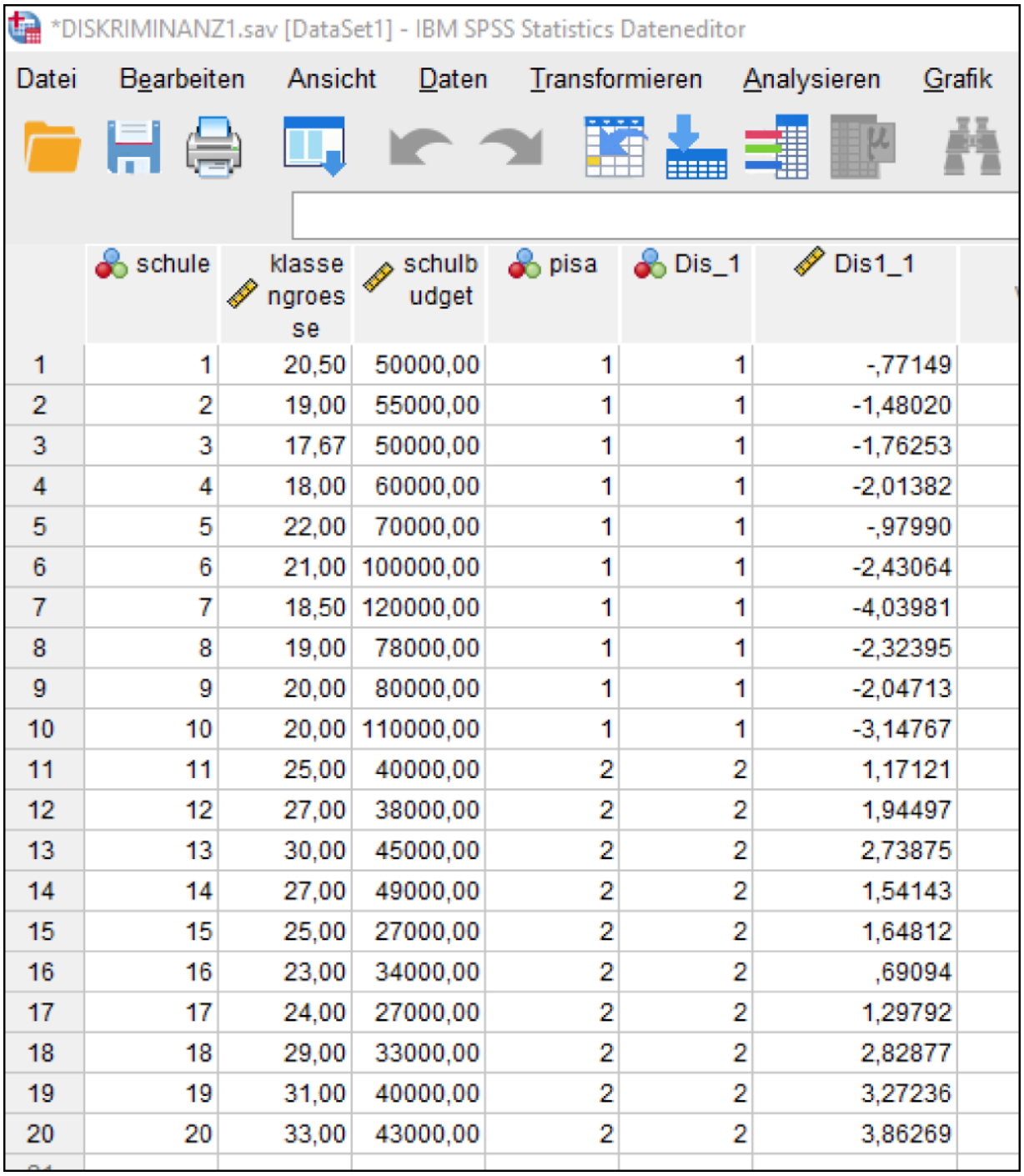



**Abbildung 11.6:** Auswahl zu speichernder Variablen im Rahmen der Diskriminanzanalyse (hier in Bezug auf DISKRIMINANZ1.SAV)

## *Output in der Datendatei*

Was da jetzt gespeichert wurde, kann aus Abbildung 11.7 abgelesen werden. SPSS hat in die Datendatei *DISKRIMINANZ1.SAV* die beiden Variablen *dis_1* und *dis1_1* eingefügt. Die Variable *dis_1* gibt direkt an, welcher Gruppe eine Schule jeweils zugewiesen worden ist. Wie der Vergleich der Ausprägungen der Gruppierungsvariablen *pisa* und der Zuordnungsvariablen *dis_1* zeigt, ergeben sich in unserem Beispiel keinerlei Abweichungen zwischen den Ausprägungen dieser beiden Variablen.




| | schule | klassengroesse | schulbudget | pisa | Dis_1 | Dis1_1 |
|---|---|---|---|---|---|---|
| 1 | 1 | 20,50 | 50000,00 | 1 | 1 | -,77149 |
| 2 | 2 | 19,00 | 55000,00 | 1 | 1 | -1,48020 |
| 3 | 3 | 17,67 | 50000,00 | 1 | 1 | -1,76253 |
| 4 | 4 | 18,00 | 60000,00 | 1 | 1 | -2,01382 |
| 5 | 5 | 22,00 | 70000,00 | 1 | 1 | -,97990 |
| 6 | 6 | 21,00 | 100000,00 | 1 | 1 | -2,43064 |
| 7 | 7 | 18,50 | 120000,00 | 1 | 1 | -4,03981 |
| 8 | 8 | 19,00 | 78000,00 | 1 | 1 | -2,32395 |
| 9 | 9 | 20,00 | 80000,00 | 1 | 1 | -2,04713 |
| 10 | 10 | 20,00 | 110000,00 | 1 | 1 | -3,14767 |
| 11 | 11 | 25,00 | 40000,00 | 2 | 2 | 1,17121 |
| 12 | 12 | 27,00 | 38000,00 | 2 | 2 | 1,94497 |
| 13 | 13 | 30,00 | 45000,00 | 2 | 2 | 2,73875 |
| 14 | 14 | 27,00 | 49000,00 | 2 | 2 | 1,54143 |
| 15 | 15 | 25,00 | 27000,00 | 2 | 2 | 1,64812 |
| 16 | 16 | 23,00 | 34000,00 | 2 | 2 | ,69094 |
| 17 | 17 | 24,00 | 27000,00 | 2 | 2 | 1,29792 |
| 18 | 18 | 29,00 | 33000,00 | 2 | 2 | 2,82877 |
| 19 | 19 | 31,00 | 40000,00 | 2 | 2 | 3,27236 |
| 20 | 20 | 33,00 | 43000,00 | 2 | 2 | 3,86269 |

**Abbildung 11.7:** Gespeicherte Variablen im Datensatz DISKRIMINANZ1.SAV

Dies ist darauf zurückzuführen, dass der den Berechnungen zugrunde liegende (fiktive) Datensatz von vorneherein klare Unterschiede bei der (durchschnittlichen) Klassengröße und beim Schulbudget zwischen den erfolgreichen und den nicht so erfolgreichen Schulen aufweist. Die

Schulen der Gruppe »Überdurchschnittliches PISA-Ergebnis« haben allesamt deutlich kleinere (durchschnittliche) Klassengrößen und markant höhere Schulbudgets als die Schulen der Gruppe »Unterdurchschnittliches PISA-Ergebnis«. Bei weniger klaren Unterschieden als im vorliegenden Beispiel kann es im Einzelfall zu Abweichungen zwischen der tatsächlichen und der von SPSS vorhergesagten Gruppenzugehörigkeit kommen.

Die Gruppenzugehörigkeit zu Gruppe 1 wird im vorliegenden Beispiel durch einen negativen Wert der Diskriminanzfunktion und die Gruppenzugehörigkeit zu Gruppe 2 durch einen positiven Wert der Diskriminanzfunktion bewirkt. Dies ist aus der letzten Spalte anhand der Ausprägungen der Variablen *dis1_1* ersichtlich.

Wenn Sie das Vorstehende zur Diskriminanzanalyse im Übrigen lieber in Syntaxform eingeben, müssen Sie schreiben:

```
DISCRIMINANT
    /GROUPS=pisa(1 2)
    /VARIABLES=klassengroesse schulbudget
    /ANALYSIS ALL
    /SAVE=CLASS SCORES
    /PRIORS EQUAL
    /STATISTICS=MEAN STDDEV UNIVF COEFF RAW
    /CLASSIFY=NONMISSING POOLED.
```

## *Viewer-Ergebnisse und ihre Interpretation*

Anhand der vorstehenden Syntax ergibt sich ein recht umfangreicher Viewer-Output im Rahmen Ihrer Diskriminanzanalyse. Zunächst werden – gemäß Ihrer Anforderung – einige deskriptive Statistiken ausgewiesen. Hierbei zeigt sich, dass die durchschnittliche Klassengröße in Gruppe 1 mit 19,57 Schülern deutlich niedriger als in Gruppe 2 mit 27,40 Schülern liegt und dass das durchschnittliche Schulbudget in Gruppe 1 mit 77.300,00 Geldeinheiten im Beispiel mehr als doppelt so hoch wie in Gruppe 2 (37.600,00 Geldeinheiten) ist.

Wie ein ergänzender »Gleichheitstest der Gruppenmittelwerte« darlegt, sind die Gruppenunterschiede bei beiden Variablen signifikant, was auch durch den Ausweis der sogenannten Eigenwerte und von Wilks Lambda angezeigt wird. Der ausgewiesene Eigenwert von 4,899 bedeutet, dass

die durch die Diskriminanzanalyse erklärte Streuung (zwischen den beiden Gruppen) circa das Fünffache der nicht erklärten Streuung beträgt. Darüber hinaus ist in Abbildung 11.8 eine »kanonische Korrelation« in Höhe von 0,911 angegeben. Dies bedeutet, dass 91,1 Prozent der Gesamtstreuung durch die Diskriminanzwerte erklärt werden. Hinzu kommt, dass Wilks Lambda mit einem Wert von 0,170 vergleichsweise gering ist, was ebenfalls ein Indikator für eine hohe Güte der durchgeführten Diskriminanzanalyse ist.

**Analyse der verarbeiteten Fälle.**

| Ungewichtete Fälle | | N | Prozent |
|---|---|---|---|
| Gültig | | 20 | 100,0 |
| Ausgeschlossen | Fehlende oder außerhalb des Bereichs liegende Gruppencodes | 0 | ,0 |
| | Wenigstens eine Diskriminanzvariable fehlt | 0 | ,0 |
| | Beide fehlenden oder außerhalb des Bereichs liegenden Gruppencodes und mindestens eine fehlende Diskriminanzvariable | 0 | ,0 |
| | Gesamt | 0 | ,0 |
| Gesamt | | 20 | 100,0 |

**Gruppenstatistiken**

| pisa | | Mittelwert | Std.-Abweichung | Gültige Werte (listenweise) Nicht gewichtet | Gewichtet |
|---|---|---|---|---|---|
| 1 | klassengroesse | 19,5670 | 1,37655 | 10 | 10,000 |
| | schulbudget | 77300,0000 | 25289,65533 | 10 | 10,000 |
| 2 | klassengroesse | 27,4000 | 3,27278 | 10 | 10,000 |
| | schulbudget | 37600,0000 | 7336,36301 | 10 | 10,000 |
| Gesamt | klassengroesse | 23,4835 | 4,70293 | 20 | 20,000 |
| | schulbudget | 57450,0000 | 27261,83953 | 20 | 20,000 |

**Gleichheitstest der Gruppenmittelwerte**

| | Wilks-Lambda | F | df1 | df2 | Sig. |
|---|---|---|---|---|---|
| klassengroesse | ,270 | 48,672 | 1 | 18 | <,001 |
| schulbudget | ,442 | 22,730 | 1 | 18 | <,001 |

**Eigenwerte**

| Funktion | Eigenwert | % der Varianz | Kumulierte % | Kanonische Korrelation |
|---|---|---|---|---|
| 1 | 4,899[a] | 100,0 | 100,0 | ,911 |

a. Die ersten 1 kanonischen Diskriminanzfunktionen werden in dieser Analyse verwendet.

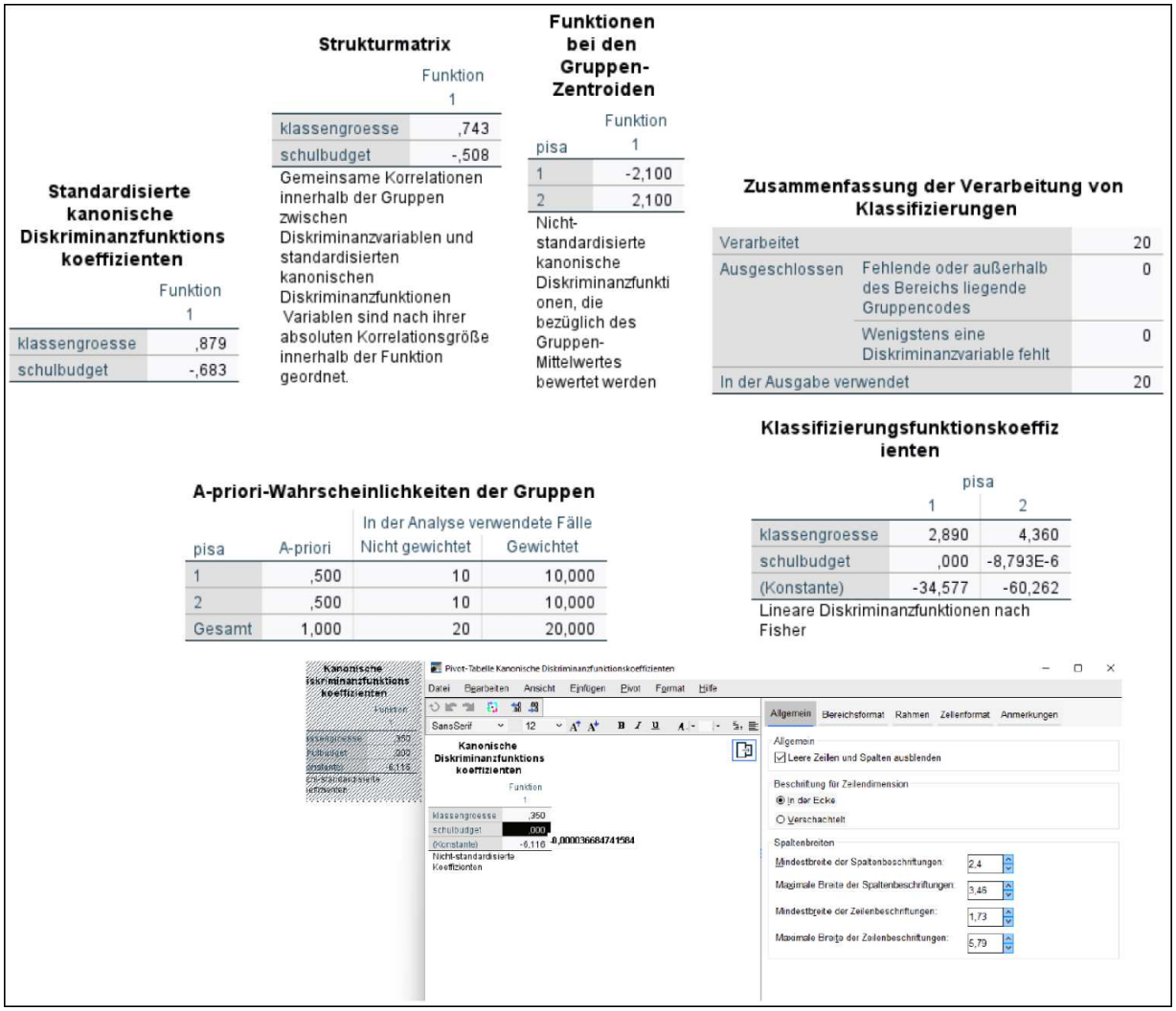
**Wilks-Lambda**

| Test der Funktion(en) | Wilks-Lambda | Chi-Quadrat | df | Sig. |
|---|---|---|---|---|
| 1 | ,170 | 30,170 | 2 | <,001 |

**Abbildung 11.8:** Grundlegende Beurteilungskennziffern für die mit DISKRIMINANZ1.SAV durchgeführte Diskriminanzanalyse

Die eigentlichen Schätzergebnisse der Diskriminanzanalyse sind in Abbildung 11.9 angegeben. Zunächst finden Sie dort die standardisierten Koeffizienten der geschätzten *Diskriminanzfunktion*. Wegen unterschiedlicher Dimensionen der unabhängigen Variablen ist eine solche Standardisierung sinnvoll. Prinzipiell werden derartige standardisierte Koeffizienten dadurch berechnet, dass die geschätzten Koeffizienten mit dem Verhältnis aus den Standardabweichungen der betreffenden unabhängigen und der abhängigen Variablen multipliziert werden. Aus den Absolutbeträgen der standardisierten Koeffizienten kann auf deren jeweilige Trennschärfe bei der Gruppenabgrenzung geschlossen werden. Im vorliegenden Beispielfall ist der Betrag bei

*klassengroesse* mit 0,879 etwas größer als der Betrag bei *schulbudget* mit 0,683. Dies bedeutet, dass die Aufteilung in zwei Gruppen hier in stärkerem Maße durch die Variable *klassengroesse* als durch die Variable *schulbudget* hervorgerufen wird.



**Standardisierte kanonische Diskriminanzfunktions koeffizienten**

| | Funktion 1 |
|---|---|
| klassengroesse | ,879 |
| schulbudget | -,683 |

**Strukturmatrix**

| | Funktion 1 |
|---|---|
| klassengroesse | ,743 |
| schulbudget | -,508 |

Gemeinsame Korrelationen innerhalb der Gruppen zwischen Diskriminanzvariablen und standardisierten kanonischen Diskriminanzfunktionen
Variablen sind nach ihrer absoluten Korrelationsgröße innerhalb der Funktion geordnet.

**Funktionen bei den Gruppen-Zentroiden**

| pisa | Funktion 1 |
|---|---|
| 1 | -2,100 |
| 2 | 2,100 |

Nicht-standardisierte kanonische Diskriminanzfunktionen, die bezüglich des Gruppen-Mittelwertes bewertet werden

**Zusammenfassung der Verarbeitung von Klassifizierungen**

| | | |
|---|---|---|
| Verarbeitet | | 20 |
| Ausgeschlossen | Fehlende oder außerhalb des Bereichs liegende Gruppencodes | 0 |
| | Wenigstens eine Diskriminanzvariable fehlt | 0 |
| In der Ausgabe verwendet | | 20 |

**A-priori-Wahrscheinlichkeiten der Gruppen**

| pisa | A-priori | In der Analyse verwendete Fälle Nicht gewichtet | Gewichtet |
|---|---|---|---|
| 1 | ,500 | 10 | 10,000 |
| 2 | ,500 | 10 | 10,000 |
| Gesamt | 1,000 | 20 | 20,000 |

**Klassifizierungsfunktionskoeffizienten**

| | pisa 1 | pisa 2 |
|---|---|---|
| klassengroesse | 2,890 | 4,360 |
| schulbudget | ,000 | -8,793E-6 |
| (Konstante) | -34,577 | -60,262 |

Lineare Diskriminanzfunktionen nach Fisher



**Abbildung 11.9:** Kernergebnisse für die mit DISKRIMINANZ1.SAV durchgeführte Diskriminanzanalyse

Die danach angegebene Strukturmatrix weist die Korrelation der unabhängigen Variablen mit den Diskriminanzwerten aus. Im Beispielfall haben sich diesbezüglich mittlere bis starke Korrelationen ergeben, was ebenfalls ein Indiz für eine recht gute Trennschärfe bezüglich der Gruppenabgrenzung im vorliegenden Beispielfall ist.

Anschließend finden Sie in Abbildung 11.9 die nichtstandardisierten Koeffizienten der Diskriminanzfunktion. Setzen Sie in die entsprechende Funktion die Werte der einzelnen Schulen für *klassengroesse* und *schulbudget* ein, erhalten Sie die in der obigen Abbildung 11.7 angegebenen Ausprägungen der Variablen *dis1_1*. Wie bereits beschrieben, führt ein negativer Wert von *dis1_1* hier zu einer Zuordnung zur Gruppe 1, ein positiver Wert hingegen zu einer Zuordnung zur Gruppe 2.

Die letzte Pivottabelle in Abbildung 11.9 schließlich gibt die Koeffizienten der Diskriminanzfunktion nach Fisher an. Mit ihrer Hilfe ist es im Sinne der Prognosefunktion der Diskriminanzanalyse möglich, ein neues Objekt einer der Gruppen zuzuordnen. Im vorliegenden Beispielfall berechnen Sie zum Beispiel für eine neu hinzukommende Schule mit einer durchschnittlichen Klassengröße von 24 Schülern und einem Schulbudget von 30.000 Geldeinheiten für die Gruppe 1 einen Klassifikationswert von 39,13 (= −34,577 + 2,890 · 24 + 0,000145 · 30.000) und für die Gruppe 2 einen solchen von 44,11 (= −60,262 + 4,360 · 24 − 0,000008793 · 30.000). Da in diesem Fall der Klassifikationswert für die Gruppe 2 höher als derjenige für die Gruppe 1 ist, wird diese neue Schule der Gruppe 2 zugewiesen.

# Clusteranalyse

Die im Folgenden vorgestellte **Clusteranalyse** kann im vorliegenden Rahmen nur ansatzweise behandelt werden. Auch hier dient wieder ein fiktives Beispiel der Veranschaulichung.

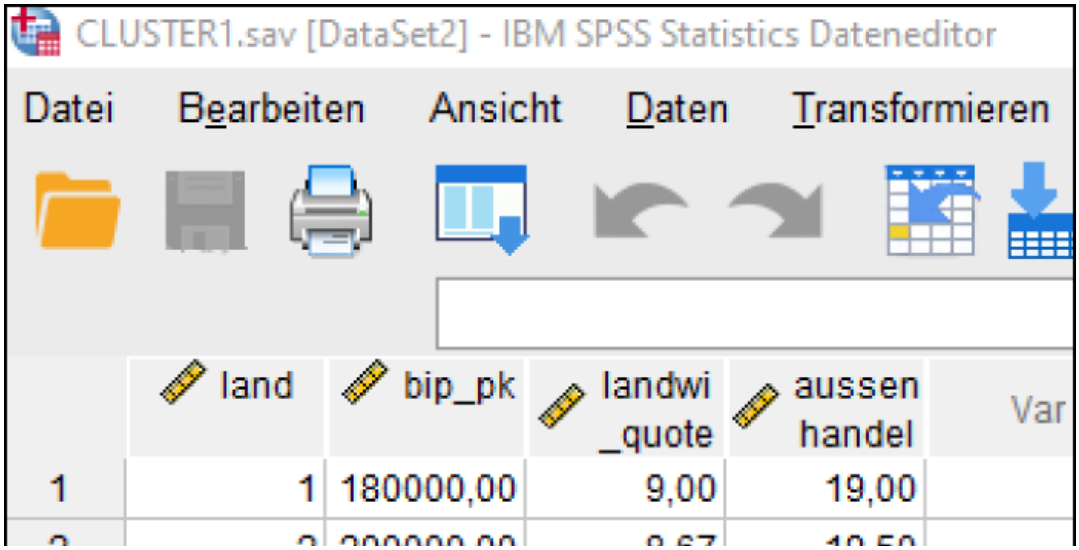
## Idee und Annahmen

Grundsätzlich ist zum einen eine ausreichend große Stichprobe sowie zum anderen die z-Transformation von Variablen wesentliche Voraussetzungen für eine Clusteranalyse. Letzteres bedeutet die bereits im vorhergehenden Abschnitt zur Diskriminanzanalyse erwähnte Standardisierung von Merkmalsausprägungen, um diese – wegen ihrer unterschiedlichen Variablendimensionen – miteinander vergleichbar zu machen.

Auf dieser Basis werden in einer Clusteranalyse für verschiedene Variablen Ähnlichkeiten bei den Fällen ermittelt. Es stehen dabei grundsätzlich verschiedene *Ähnlichkeitsmaße* für metrisch skalierte und für binäre Variablen zur Verfügung. Auch existieren mehrere Cluster-Algorithmen, wie im gleich folgenden Beispiel skizziert wird. Ein grafisches Hilfsmittel, das bei der Bildung von Gruppen, also von Clustern, zur Verfügung steht, sind sogenannte **Dendrogramme**. Hierfür wird auch gleich ein Beispiel gegeben.

## *Ein einfaches Beispiel in SPSS*

Das nachfolgende Beispiel zur Clusteranalyse bezieht sich auf die Datendatei *CLUSTER1.SAV* (siehe Abbildung 11.10). In dieser (fiktiven) Datendatei ist zum einen für insgesamt 20 Länder das Bruttoinlandsprodukt pro Kopf (grob ausgedrückt: deren Produktionsleistung) in Geldeinheiten in der Variablen *bip_pk* ausgewiesen. Zum anderen enthält die betreffende Datendatei für alle 20 Länder den Anteil der landwirtschaftlichen Produktionsleistung am gesamten Bruttoinlandsprodukt in Prozent in der Variablen *landwi_quote* (als indirekter Indikator für den Industrialisierungsgrad eines Landes: Je geringer der entsprechende Anteil ist, desto höher ist der Industrialisierungsgrad und umgekehrt) sowie ebenfalls in Prozent in der Variablen *aussenhandel* den Anteil der Summe aus Exporten und Importen am gesamten Bruttoinlandsprodukt als Indikator für den sogenannten (wirtschaftlichen) Offenheitsgrad eines Landes.



| | | | | | |
|---|---|---|---|---|---|
| 2 | 2 | 200000,00 | 8,67 | 19,50 | |
| 3 | 3 | 300000,00 | 7,00 | 22,00 | |
| 4 | 4 | 175000,00 | 9,20 | 18,80 | |
| 5 | 5 | 200000,00 | 8,50 | 20,00 | |
| 6 | 6 | 250000,00 | 7,95 | 21,67 | |
| 7 | 7 | 300000,00 | 6,00 | 22,00 | |
| 8 | 8 | 400000,00 | 2,50 | 22,50 | |
| 9 | 9 | 150000,00 | 9,20 | 17,50 | |
| 10 | 10 | 140000,00 | 9,30 | 17,00 | |
| 11 | 11 | 9000,00 | 19,50 | 9,95 | |
| 12 | 12 | 8000,00 | 20,00 | 8,75 | |
| 13 | 13 | 7000,00 | 22,50 | 8,25 | |
| 14 | 14 | 6000,00 | 23,00 | 8,00 | |
| 15 | 15 | 5000,00 | 43,00 | 7,50 | |
| 16 | 16 | 4000,00 | 45,00 | 5,00 | |
| 17 | 17 | 3000,00 | 60,00 | 4,50 | |
| 18 | 18 | 8000,00 | 33,33 | 7,90 | |
| 19 | 19 | 2000,00 | 75,00 | 4,00 | |
| 20 | 20 | 1500,00 | 80,00 | 3,00 | |
| 21 | | | | | |

**Abbildung 11.10:** Datendatei CLUSTER1.SAV

Mittels Clusteranalyse sollen über diese 20 Länder hinweg einander ähnliche Gruppen (**Cluster**) gebildet werden. Hier führen viele Wege nach Rom; das heißt: Es gibt mehrere Clusterverfahren, die in unserem Beispiel dieselben (Cluster-)Ergebnisse produzieren.

## *Varianten der Clusteranalyse*

Im Einzelnen wird nachfolgend auf drei Clusterverfahren Bezug genommen (ohne aber aus Platzgründen detailliert auf die methodischen Unterschiede eingehen zu können; für weiterführende diesbezügliche

Erörterungen wird deshalb auf Spezialliteratur verwiesen, wie zum Beispiel Schendera, Christian F. G. (2009): Clusteranalyse mit SPSS. Mit Faktorenanalyse, München: Oldenbourg Verlag):

- ✔ die zweistufige Clusteranalyse,
- ✔ die K-Mittelwerte-Clusteranalyse und
- ✔ die hierarchische Clusteranalyse.

Bei der zweistufigen Clusteranalyse wird die Clusterbildung in zwei Schritten durchgeführt: Im ersten Schritt erfolgt eine vorläufige Verdichtung auf einige »Pre-Cluster«, die dann im zweiten Schritt zu den endgültigen Clustern zusammengeführt werden. Im Rahmen der K-Mittelwerte-Clusteranalyse wird die Anzahl der zu bildenden Cluster vorgegeben (zum Beispiel: drei Cluster mit entsprechend K = 3). Bei der hierarchischen Clusteranalyse schließlich werden, ausgehend vom gesamten Datensatz als ein einzelnes Cluster, durch »Ähnlichkeits-Vergleiche« in den folgenden Analyseschritten einander ähnliche Gruppen anhand bestimmter »Ähnlichkeits-Kriterien« schrittweise in Clustern abgebildet.

## *Zweistufige Clusteranalyse*

Die **zweistufige Clusteranalyse** rufen Sie im SPSS-Menü über ANALYSIEREN|KLASSIFIZIEREN|TWO-STEP-CLUSTERANALYSE auf.

In der sich dann öffnenden Dialogbox TWO-STEP-CLUSTERANALYSE wählen Sie die Variablen für Ihre Clusteranalyse aus. Da keine der von Ihnen verwendeten Variablen eine kategoriale Größe ist, lassen Sie entsprechend in der folgenden Abbildung das Feld KATEGORIALE VARIABLEN leer. Ihre Variablen sind allesamt stetige Geld- beziehungsweise Prozentgrößen, sodass Sie die drei hier relevanten Variablen *bip_pk, landwi_quote* und *aussenhandel* in das Feld STETIGE VARIABLEN eingeben.

Bezüglich der Anzahl der maximalen Cluster lassen Sie einfach die Voreinstellung von zehn Clustern stehen. Dass das eigentlich eine zu

große Anzahl ist, wird allein schon daraus ersichtlich, dass in unserem Beispiel insgesamt ja nur 20 Fälle (Länder) vorliegen. Da würden zehn Cluster im Durchschnitt zwei Fälle beinhalten, was ja nicht gerade von großer Aussagekraft zeugt.

Aus Abbildung 11.11 erkennen Sie ferner, dass SPSS angesichts Ihrer Eingabe von drei stetigen Variablen in der Rubrik ANZAHL STETIGER VARIABLEN die zu standardisierenden (stetigen) Variablen folgerichtig auf »3« gesetzt hat. Als **Clusterkriterium** für die – auf Basis der *Maximum-Likelihood-Schätzmethode* ermittelten – Cluster wird im Übrigen an dieser Stelle das *Schwarz'sche Bayes-Kriterium* (BIC = Bayesian Information Criterion) ausgewählt, was hier aber im Grunde genommen im Detail nicht interessieren muss. Alternativ könnten Sie auf das *AIC-Kriterium* Bezug nehmen (AIC = Akaike's Information Criterion).

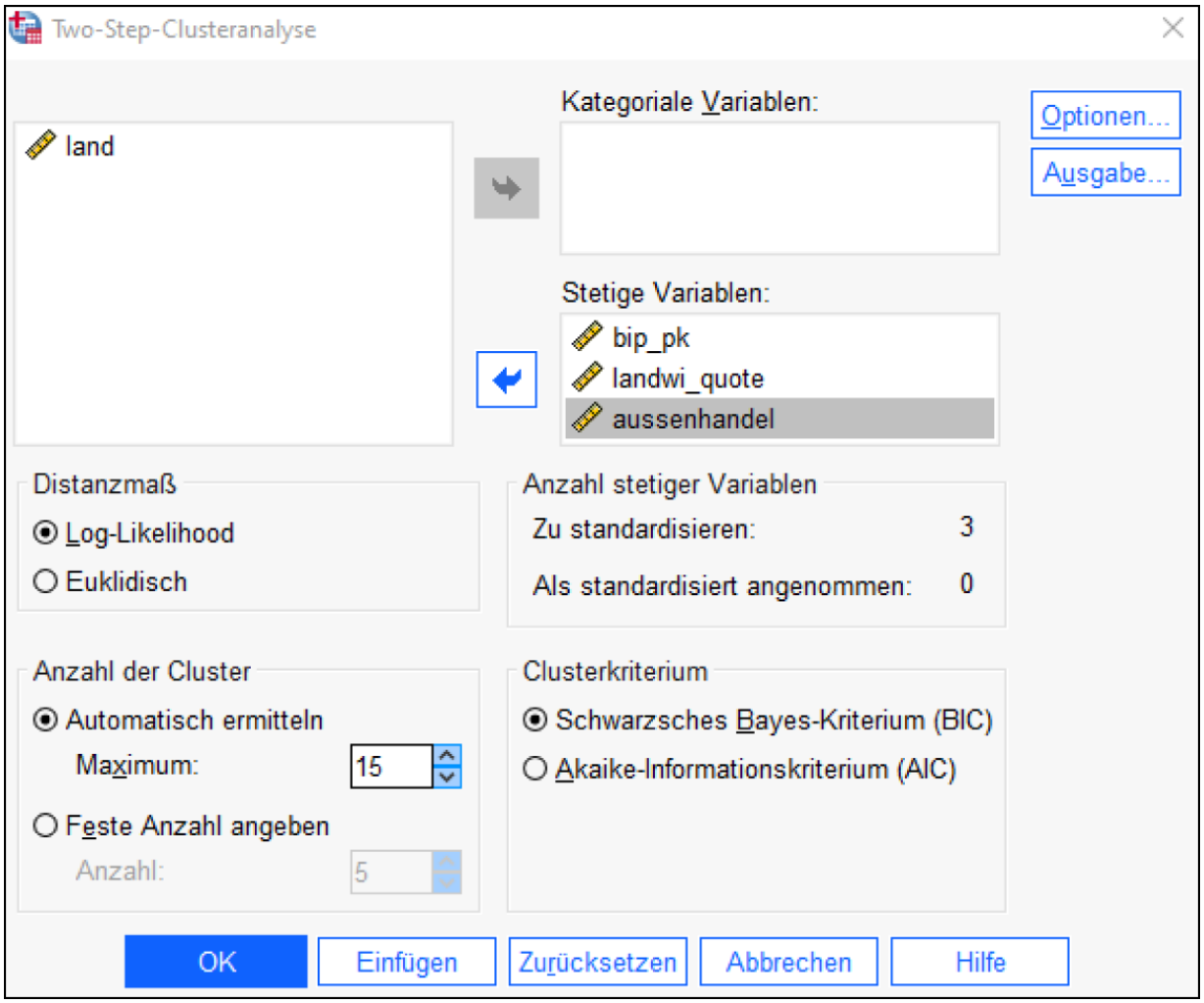



**Abbildung 11.11:** Variablenauswahl bei der zweistufigen Clusteranalyse (CLUSTER1.SAV)

Über die vorstehende Einstellung hinaus nehmen Sie bitte noch eine weitere Einstellung bei TWO-STEP-CLUSTERANALYSE: AUSGABE vor, und zwar setzen Sie bei PIVOT-TABELLEN und bei DIAGRAMME UND TABELLEN IM MODELLVIEWER jeweils ein Häkchen (siehe Abbildung 11.12). Auf diese Art und Weise werden die Ergebnisse der hier vorgenommenen Clusterbildung sowohl tabellarisch als auch visuell veranschaulicht, was gleich noch gezeigt wird.

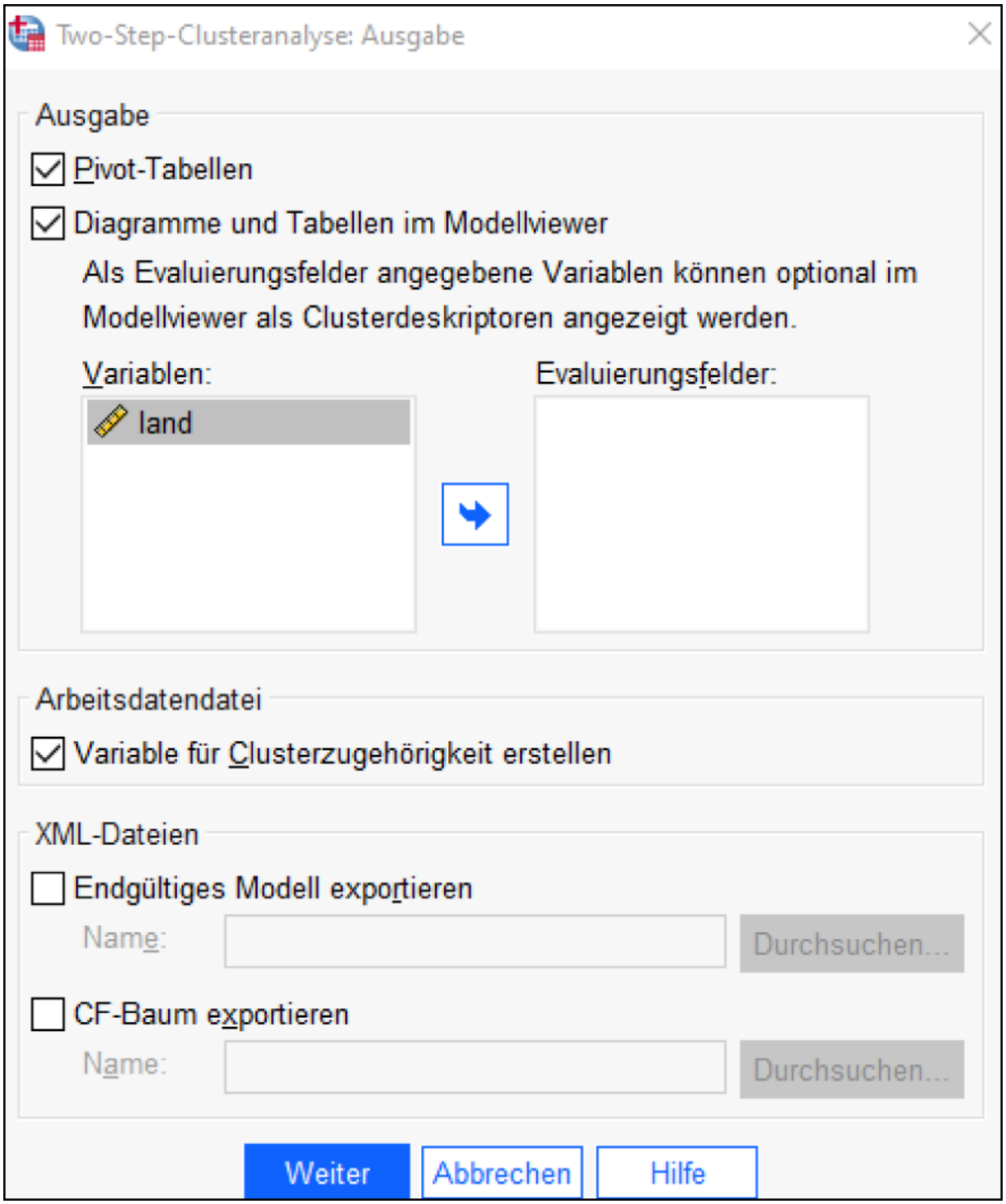



**Abbildung 11.12:** Outputsetzungen bei der zweistufigen Clusteranalyse (CLUSTER1.SAV)

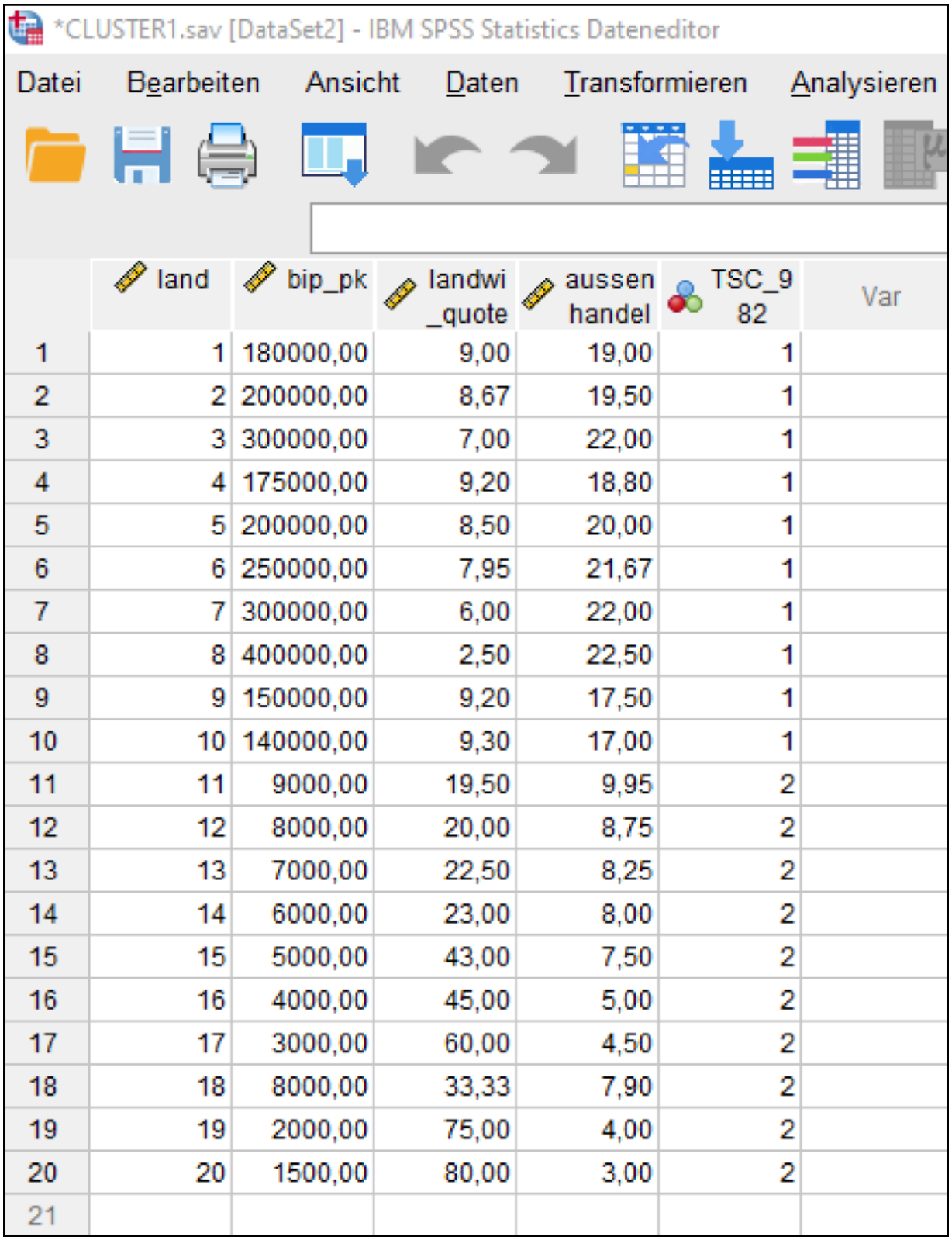
Als letzte Einstellung setzen Sie unter ARBEITSDATENDATEI ein Häkchen bei VARIABLE FÜR CLUSTERZUGEHÖRIGKEIT ERSTELLEN. Dadurch wird eine neue Variable in Ihren Datensatz

*CLUSTER1.SAV* eingefügt, die angibt, zu welchem der erstellten Cluster der jeweilige Fall gehört.

Die zugehörige SPSS-Syntax sieht so aus:

```
TWOSTEP CLUSTER
    /CONTINUOUS VARIABLES=bip_pk landwi_quote aussenhandel
    /DISTANCE LIKELIHOOD
    /NUMCLUSTERS AUTO 15 BIC
    /HANDLENOISE 0
    /MEMALLOCATE 64
    /CRITERIA INITHRESHOLD(0) MXBRANCH(8) MXLEVEL(3)
    /VIEWMODEL DISPLAY=YES
    /PRINT IC COUNT SUMMARY
    /SAVE VARIABLE=tsc_982.
```

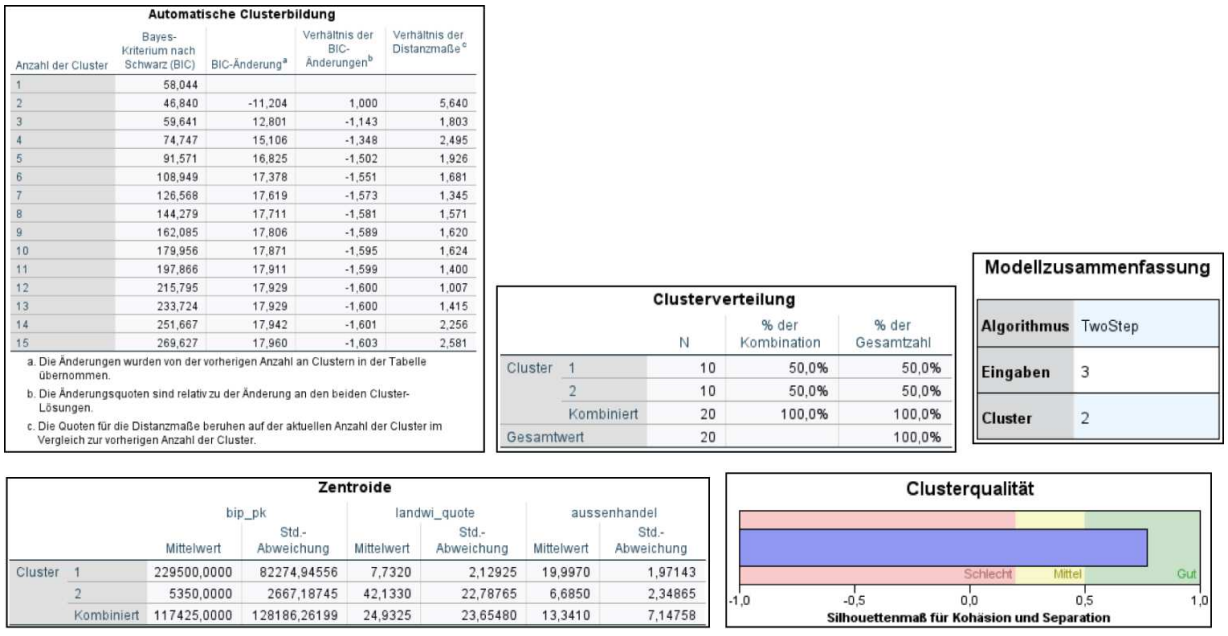
Mithilfe der neuen Variable *tsc_982* (bei Ihnen kann im Übrigen eine andere Indizierung als »982« zutage treten, wenn Sie menügesteuert vorgehen, aber das spielt keine Rolle) kann aus der zugrunde liegenden Datendatei direkt abgelesen werden, zu welchem der gebildeten Cluster der jeweilige Fall gehört. Aus Abbildung 11.13 geht in diesem Zusammenhang hervor, dass hier die Länder 1 bis 10 dem Cluster 1 und die Länder 11 bis 20 dem Cluster 2 zugeordnet worden sind. Insgesamt sind also zwei Cluster gebildet worden.




| | land | bip_pk | landwi_quote | aussenhandel | TSC_982 | Var |
|---|---|---|---|---|---|---|
| 1 | 1 | 180000,00 | 9,00 | 19,00 | 1 | |
| 2 | 2 | 200000,00 | 8,67 | 19,50 | 1 | |
| 3 | 3 | 300000,00 | 7,00 | 22,00 | 1 | |
| 4 | 4 | 175000,00 | 9,20 | 18,80 | 1 | |
| 5 | 5 | 200000,00 | 8,50 | 20,00 | 1 | |
| 6 | 6 | 250000,00 | 7,95 | 21,67 | 1 | |
| 7 | 7 | 300000,00 | 6,00 | 22,00 | 1 | |
| 8 | 8 | 400000,00 | 2,50 | 22,50 | 1 | |
| 9 | 9 | 150000,00 | 9,20 | 17,50 | 1 | |
| 10 | 10 | 140000,00 | 9,30 | 17,00 | 1 | |
| 11 | 11 | 9000,00 | 19,50 | 9,95 | 2 | |
| 12 | 12 | 8000,00 | 20,00 | 8,75 | 2 | |
| 13 | 13 | 7000,00 | 22,50 | 8,25 | 2 | |
| 14 | 14 | 6000,00 | 23,00 | 8,00 | 2 | |
| 15 | 15 | 5000,00 | 43,00 | 7,50 | 2 | |
| 16 | 16 | 4000,00 | 45,00 | 5,00 | 2 | |
| 17 | 17 | 3000,00 | 60,00 | 4,50 | 2 | |
| 18 | 18 | 8000,00 | 33,33 | 7,90 | 2 | |
| 19 | 19 | 2000,00 | 75,00 | 4,00 | 2 | |
| 20 | 20 | 1500,00 | 80,00 | 3,00 | 2 | |
| 21 | | | | | | |

**Abbildung 11.13:** Um die Variable *tsc_982* ergänzte Datendatei CLUSTER1.SAV bei der zweistufigen Clusteranalyse

Abbildung 11.14 illustriert den Abbildung 11.13 vorgeschalteten Prozess der Clusterbildung, der am Ende zu zwei Clustern geführt hat. Kennziffern für diese beiden Cluster sind unter der Überschrift »Zentroide« in Abbildung 11.14 angegeben, wobei mit »Zentroiden« hier die clusterspezifischen Ähnlichkeiten gemeint sind. Für die Länder in Cluster 1 ergeben sich höhere arithmetische Mittelwerte beim Bruttoinlandsprodukt pro Kopf (229.500,00 Geldeinheiten gegenüber 5.350,00 Geldeinheiten) und beim Offenheitsgrad (20,00 Prozent versus 6,69 Prozent) sowie ein niedriger durchschnittlicher Anteil der landwirtschaftlichen Produktionsleistung am Bruttoinlandsprodukt (7,73 Prozent gegenüber 42,13 Prozent) im Vergleich zu den Ländern aus Cluster 2. Man kann demnach die Länder in Cluster 1 als hochindustrialisierte, wohlhabende Länder mit einer vergleichsweise großen außenwirtschaftlichen Verflechtung bezeichnen. Demgegenüber stellen die Länder in Cluster 2 eher stark landwirtschaftlich geprägte, wenig wohlhabende Länder mit einem relativ geringen außenwirtschaftlichen Offenheitsgrad dar.



**Automatische Clusterbildung**

| Anzahl der Cluster | Bayes-Kriterium nach Schwarz (BIC) | BIC-Änderung[a] | Verhältnis der BIC-Änderungen[b] | Verhältnis der Distanzmaße[c] |
|---|---|---|---|---|
| 1 | 58,044 | | | |
| 2 | 46,840 | -11,204 | 1,000 | 5,640 |
| 3 | 59,641 | 12,801 | -1,143 | 1,803 |
| 4 | 74,747 | 15,106 | -1,348 | 2,495 |
| 5 | 91,571 | 16,825 | -1,502 | 1,926 |
| 6 | 108,949 | 17,378 | -1,551 | 1,681 |
| 7 | 126,568 | 17,619 | -1,573 | 1,345 |
| 8 | 144,279 | 17,711 | -1,581 | 1,571 |
| 9 | 162,085 | 17,806 | -1,589 | 1,620 |
| 10 | 179,956 | 17,871 | -1,595 | 1,624 |
| 11 | 197,866 | 17,911 | -1,599 | 1,400 |
| 12 | 215,795 | 17,929 | -1,600 | 1,007 |
| 13 | 233,724 | 17,929 | -1,600 | 1,415 |
| 14 | 251,667 | 17,942 | -1,601 | 2,256 |
| 15 | 269,627 | 17,960 | -1,603 | 2,581 |

a. Die Änderungen wurden von der vorherigen Anzahl an Clustern in der Tabelle übernommen.
b. Die Änderungsquoten sind relativ zu der Änderung an den beiden Cluster-Lösungen.
c. Die Quoten für die Distanzmaße beruhen auf der aktuellen Anzahl der Cluster im Vergleich zur vorherigen Anzahl der Cluster.

**Clusterverteilung**

| | | N | % der Kombination | % der Gesamtzahl |
|---|---|---|---|---|
| Cluster | 1 | 10 | 50,0% | 50,0% |
| | 2 | 10 | 50,0% | 50,0% |
| | Kombiniert | 20 | 100,0% | 100,0% |
| Gesamtwert | | 20 | | 100,0% |

**Modellzusammenfassung**

| | |
|---|---|
| Algorithmus | TwoStep |
| Eingaben | 3 |
| Cluster | 2 |

**Zentroide**

| | | bip_pk | | landwi_quote | | aussenhandel | |
|---|---|---|---|---|---|---|---|
| | | Mittelwert | Std.-Abweichung | Mittelwert | Std.-Abweichung | Mittelwert | Std.-Abweichung |
| Cluster | 1 | 229500,0000 | 82274,94556 | 7,7320 | 2,12925 | 19,9970 | 1,97143 |
| | 2 | 5350,0000 | 2667,18745 | 42,1330 | 22,78765 | 6,6850 | 2,34865 |
| | Kombiniert | 117425,0000 | 128186,26199 | 24,9325 | 23,65480 | 13,3410 | 7,14758 |

**Clusterqualität**



**Abbildung 11.14:** Output bei der zweistufigen Clusteranalyse (CLUSTER1.SAV)

Unten rechts in Abbildung 11.14 liefert SPSS noch eine grafische Beurteilung der Qualität der gebildeten (beiden) Cluster. Wie ersichtlich

wird die Qualität der hier vorgenommenen Clusterbildung als »gut« eingestuft.

## *K-Mittelwerte-Clusteranalyse*

Eine methodische Alternative zur vorstehenden zweistufigen Clusteranalyse stellt die K-Mittelwerte-Clusteranalyse dar. Diese wählen Sie im SPSS-Menü – wieder für *CLUSTER1.SAV* – über ANALYSIEREN|KLASSIFIZIEREN|K-MEANS-CLUSTER aus. Es öffnet sich dann ein Dialogfenster, in das Sie unter VARIABLEN die Variablen *bip_pk, aussenhandel* und *landwi_quote* eintragen (und, wenn Sie mögen, unter FALLBESCHRIFTUNG die Variable *land*, was aber nicht unbedingt erforderlich ist). Außerdem geben Sie bitte unter ANZAHL DER CLUSTER die Zahl 2 ein. Letzteres ist ein Wesensmerkmal der K-Mittelwerte-Clusteranalyse; das heißt, hier müssen Sie die Anzahl der Cluster vorgeben. Da ein Blick auf *CLUSTER1.SAV* deutliche Unterschiede zwischen den ersten zehn und den letzten zehn Ländern offenbart, ist in Abbildung 11.15 die Entscheidung zugunsten von zwei Clustern gefallen. Die K-Mittelwerte-Clusteranalyse ist demnach dann zu empfehlen, wenn vergleichsweise klare Vorstellungen über die Anzahl der zu bildenden Cluster im Vorhinein existieren.

Bei ITERIEREN brauchen Sie an dieser Stelle nichts zu ändern; Sie können ruhig die Voreinstellung von zehn Iterationen übernehmen, da – ohne Ihnen zu viel zu verraten – SPSS diese maximale Zahl an Näherungsberechnungen an eine endgültige Lösung unseres Clusterproblems bei Weitem nicht benötigt. Des Weiteren setzen Sie bitte bei SPEICHERN zwei Häkchen, und zwar bei CLUSTERZUGEHÖRIGKEIT und bei DISTANZ VOM CLUSTERZENTRUM (siehe Abbildung 11.16). Hierdurch werden in Ihre Datendatei *CLUSTER1.SAV* zwei neue Variablen eingefügt, die für jeden Fall (das heißt für jedes Land) zum einen Auskunft geben, zu welchem der beiden Cluster der jeweilige Fall gehört, und die zum anderen den jeweiligen Abstand vom »Mittel-/Gravitationspunkt« des betreffenden Clusters angeben.

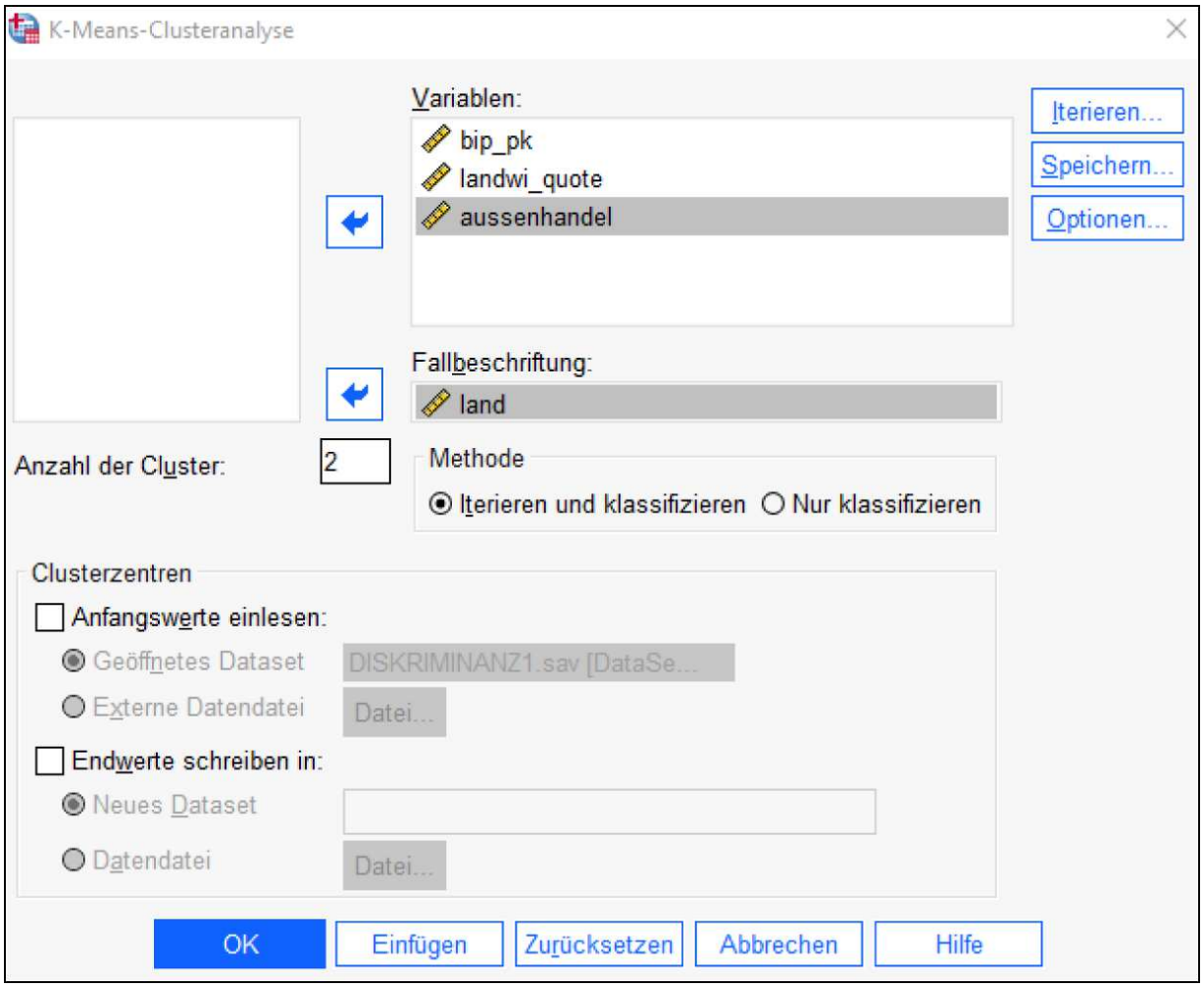



**Abbildung 11.15:** Variablenauswahl mit der Vorgabe von zwei Clustern bei der K-Mittelwerte-Clusteranalyse (CLUSTER1.SAV)

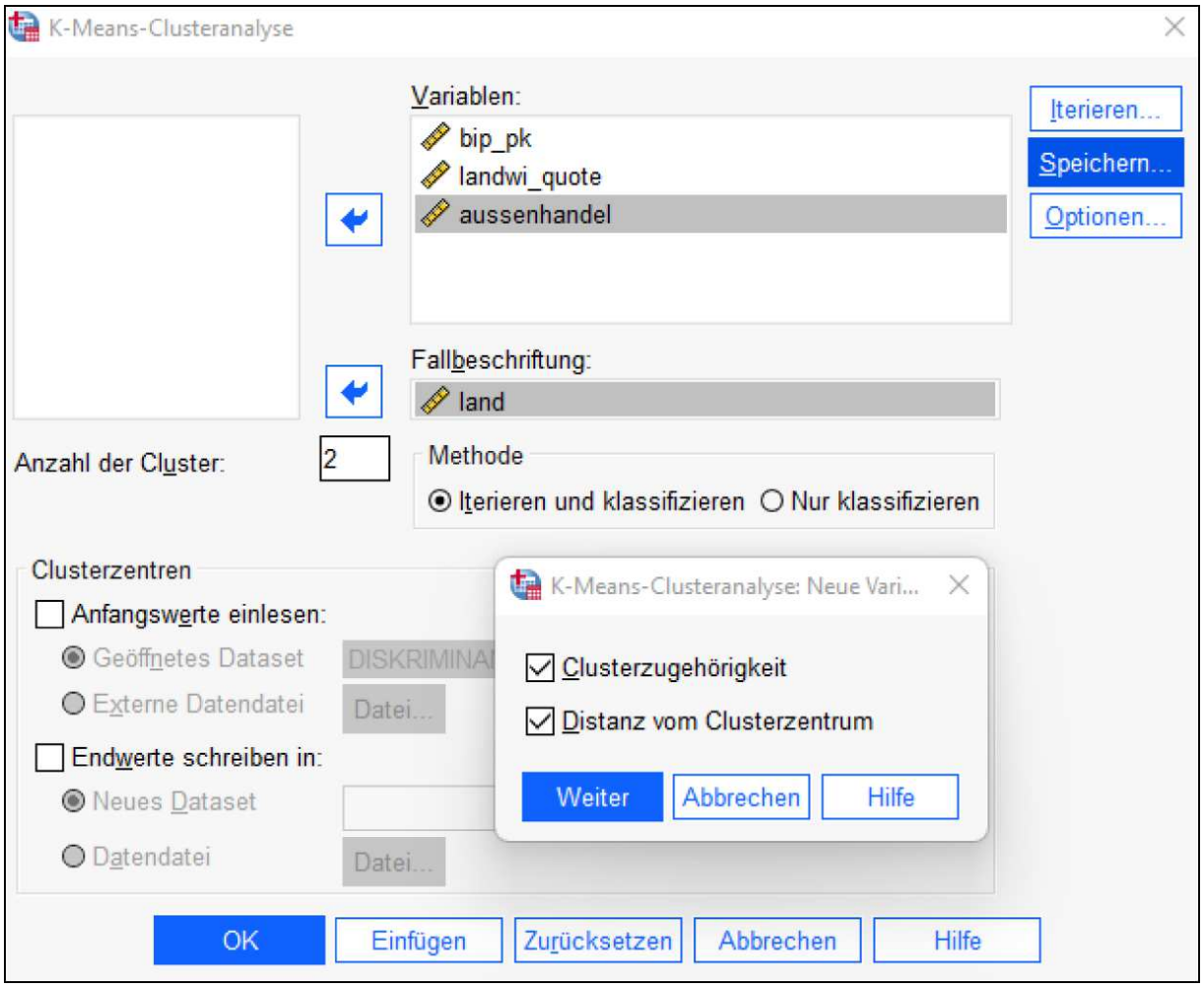



**Abbildung 11.16:** Abzuspeichernde Variablen bei der K-Mittelwerte-Clusteranalyse (CLUSTER1.SAV)

4Auch nehmen Sie bitte noch ein paar Einstellungen unter OPTIONEN vor. Setzen Sie doch hier einfach überall ein Häkchen. Auf diese Art und Weise gibt Ihnen SPSS ein paar zusätzliche statistische Kennziffern aus (siehe hierzu Abbildung 11.17).

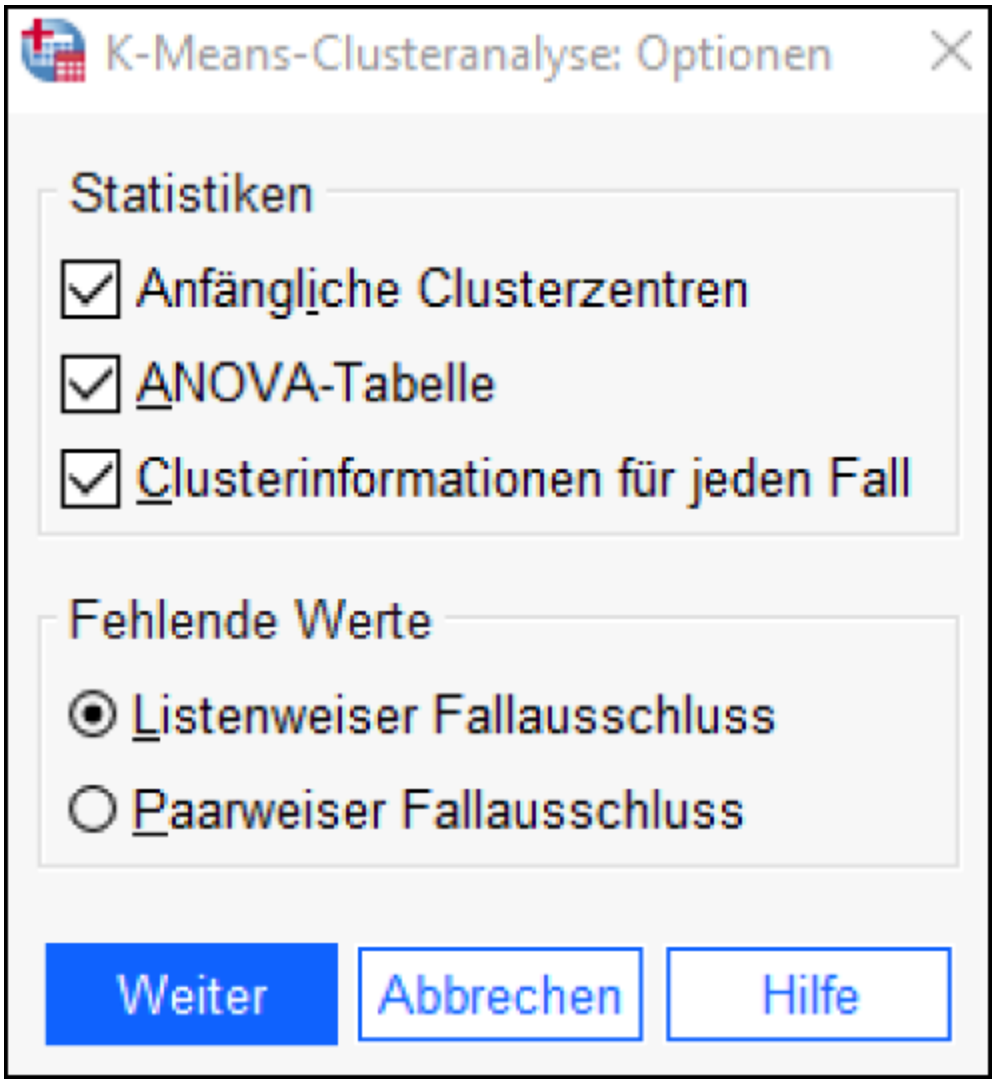
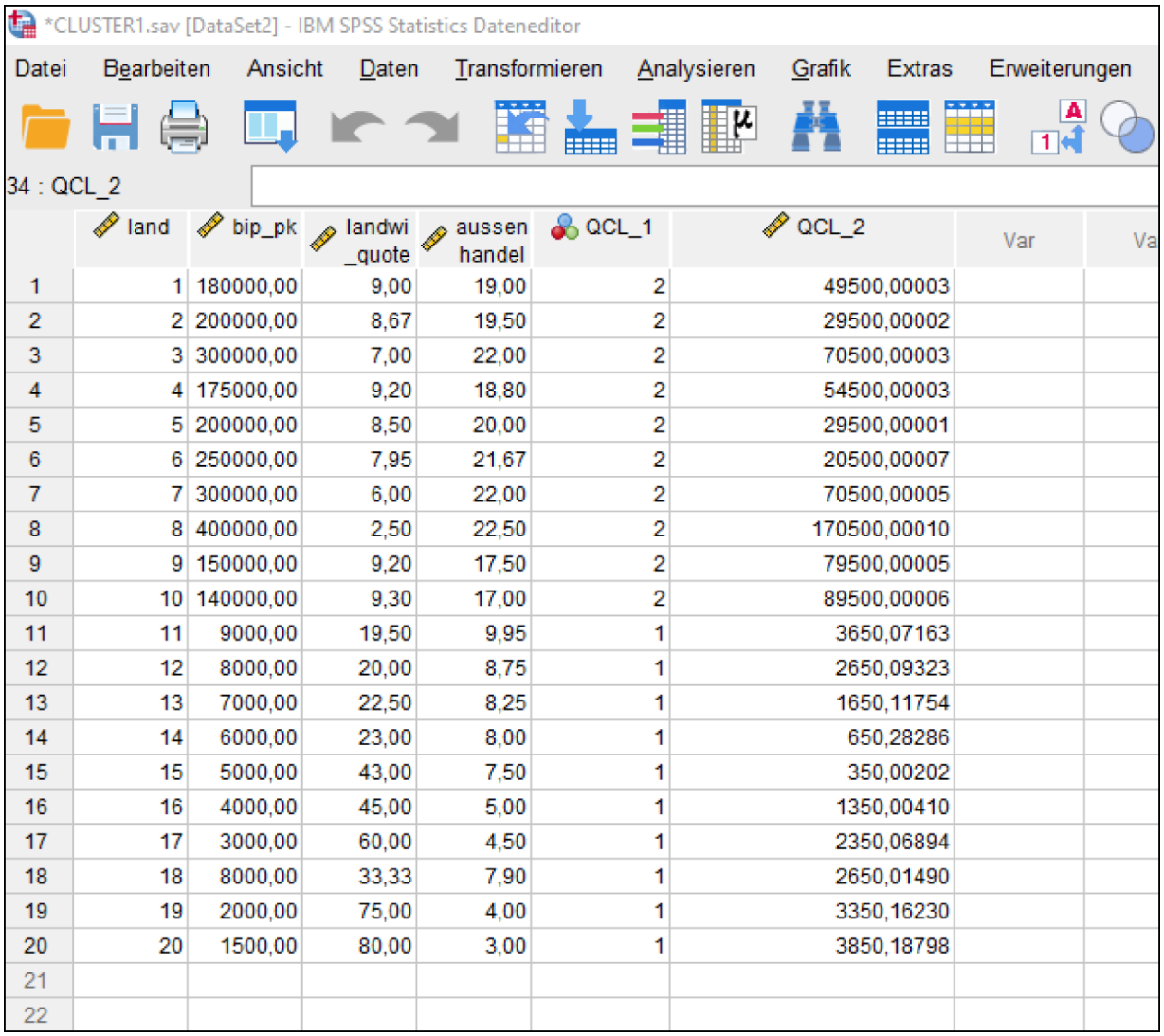



**Abbildung 11.17:** Auswahl statistischer Kennziffern bei der K-Mittelwerte-Clusteranalyse (CLUSTER1.SAV)

Wenn Sie jetzt die OK-Taste anklicken, können Sie zunächst im Dateneditor erkennen, dass SPSS die beiden neuen Variablen *qcl_1* und *qcl_2* in Ihre Datendatei eingefügt hat (siehe Abbildung 11.18). Dabei gibt die Variable *qcl_1* die Clusterzugehörigkeiten der einzelnen Länder an (hier: Länder 1 bis 10 als Elemente von Cluster 2 und Länder 11 bis 20 als Elemente von Cluster 1).

Anhand der Ausprägungen der Variablen *qcl_2* können Sie die Abstände der einzelnen Länder vom jeweiligen Cluster-»Gravitationszentrum« erkennen, das letztlich aus einer Linearkombination aus den drei Variablen *bip_pk, aussenhandel* und *landwi_quote* hervorgeht. Den

niedrigsten Distanzwert weist in Cluster 2 Land Nummer 6 auf, und zwar mit Werten von 250.000 Geldeinheiten beim Bruttoinlandsprodukt pro Kopf (arithmetischer Mittelwert von Cluster 2 hier: 229.500 Geldeinheiten), 7,95 Prozent beim Anteil der landwirtschaftlichen Produktionsleistung am Bruttoinlandsprodukt (arithmetischer Mittelwert von Cluster 2 hier: 7,73 Prozent) und 21,67 Prozent beim Offenheitsgrad (arithmetischer Mittelwert von Cluster 2 hier: 20,00 Prozent). In Cluster 1 hat Land Nummer 15 den geringsten gruppenspezifischen Distanzwert inne – mit Werten von 5.000 Geldeinheiten beim Bruttoinlandsprodukt pro Kopf (arithmetischer Mittelwert von Cluster 1 hier: 5.350 Geldeinheiten), 43,00 Prozent beim Anteil der landwirtschaftlichen Produktionsleistung am Bruttoinlandsprodukt (arithmetischer Mittelwert von Cluster 1 hier: 42,13 Prozent) und 7,50 Prozent beim Offenheitsgrad (arithmetischer Mittelwert von Cluster 1 hier: 6,69 Prozent; zu den arithmetischen Mittelwerten in den beiden Clustern siehe im Übrigen Abbildung 11.19). Insofern können in unserem (fiktiven) Beispiel Land 6 als repräsentativstes Land für Cluster 2 und Land 15 als repräsentativstes Land für Cluster 1 angesehen werden.




| | land | bip_pk | landwi_quote | aussenhandel | QCL_1 | QCL_2 | Var | Va |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 180000,00 | 9,00 | 19,00 | 2 | 49500,00003 | | |
| 2 | 2 | 200000,00 | 8,67 | 19,50 | 2 | 29500,00002 | | |
| 3 | 3 | 300000,00 | 7,00 | 22,00 | 2 | 70500,00003 | | |
| 4 | 4 | 175000,00 | 9,20 | 18,80 | 2 | 54500,00003 | | |
| 5 | 5 | 200000,00 | 8,50 | 20,00 | 2 | 29500,00001 | | |
| 6 | 6 | 250000,00 | 7,95 | 21,67 | 2 | 20500,00007 | | |
| 7 | 7 | 300000,00 | 6,00 | 22,00 | 2 | 70500,00005 | | |
| 8 | 8 | 400000,00 | 2,50 | 22,50 | 2 | 170500,00010 | | |
| 9 | 9 | 150000,00 | 9,20 | 17,50 | 2 | 79500,00005 | | |
| 10 | 10 | 140000,00 | 9,30 | 17,00 | 2 | 89500,00006 | | |
| 11 | 11 | 9000,00 | 19,50 | 9,95 | 1 | 3650,07163 | | |
| 12 | 12 | 8000,00 | 20,00 | 8,75 | 1 | 2650,09323 | | |
| 13 | 13 | 7000,00 | 22,50 | 8,25 | 1 | 1650,11754 | | |
| 14 | 14 | 6000,00 | 23,00 | 8,00 | 1 | 650,28286 | | |
| 15 | 15 | 5000,00 | 43,00 | 7,50 | 1 | 350,00202 | | |
| 16 | 16 | 4000,00 | 45,00 | 5,00 | 1 | 1350,00410 | | |
| 17 | 17 | 3000,00 | 60,00 | 4,50 | 1 | 2350,06894 | | |
| 18 | 18 | 8000,00 | 33,33 | 7,90 | 1 | 2650,01490 | | |
| 19 | 19 | 2000,00 | 75,00 | 4,00 | 1 | 3350,16230 | | |
| 20 | 20 | 1500,00 | 80,00 | 3,00 | 1 | 3850,18798 | | |
| 21 | | | | | | | | |
| 22 | | | | | | | | |

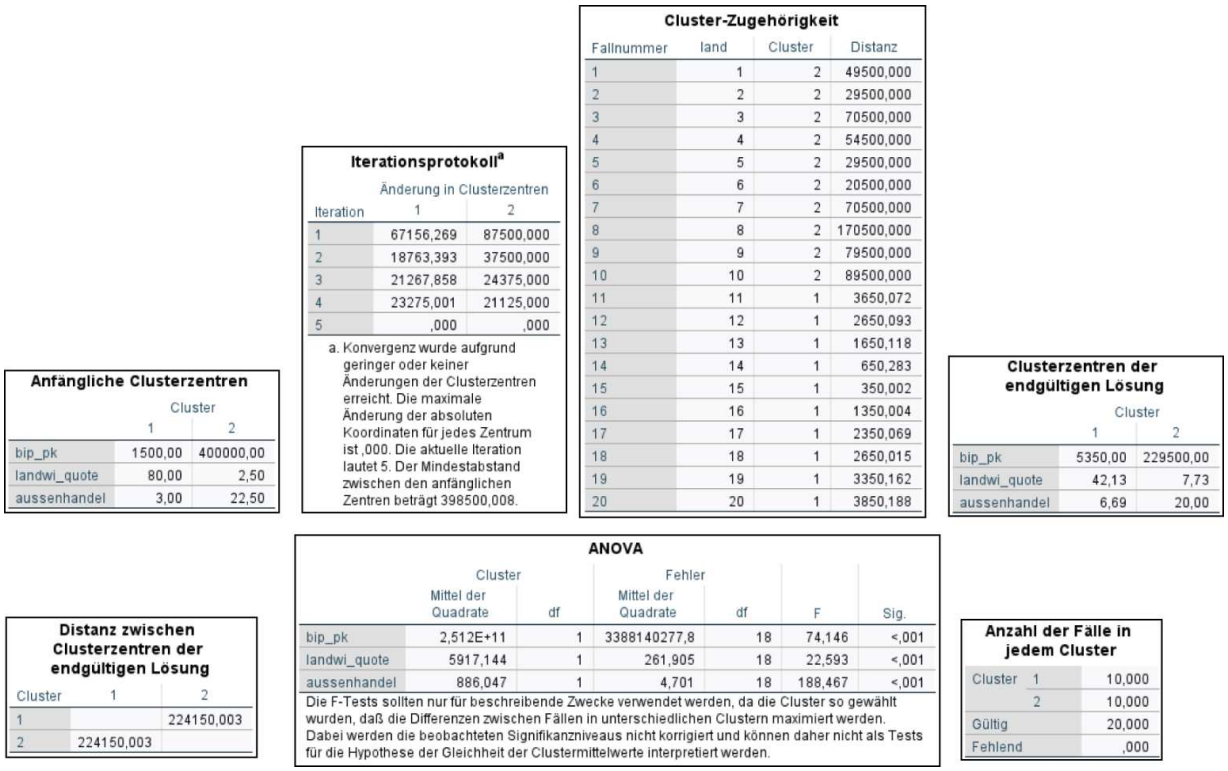
**Abbildung 11.18:** Um die Variablen *qcl_1* und *qcl_2* ergänzte Datendatei CLUSTER1.SAV bei der K-Mittelwerte-Clusteranalyse

Sie hätten das Vorstehende (sowie den folgenden Viewer-Output) auch mit folgendem Syntaxprogramm zustande gebracht:

```
QUICK CLUSTER bip_pk aussenhandel landwi_quote
    /MISSING=LISTWISE
    /CRITERIA=CLUSTER(2) MXITER(10) CONVERGE(0)
    /METHOD=KMEANS(NOUPDATE)
    /SAVE CLUSTER DISTANCE
    /PRINT ID(land) INITIAL ANOVA CLUSTER DISTAN.
```

Wenn Sie dieses Syntaxprogramm starten beziehungsweise wenn Sie die K-Mittelwerte-Clusteranalyse über das SPSS-Menü zur Ausführung bringen, erhalten Sie im Viewer den eben bereits zum Teil beschriebenen

Output in Abbildung 11.19. Die einzelnen Befunde zeigen unter anderem eine sehr große Distanz – das heißt eine große Trennschärfe – zwischen den beiden Clustern. Die Länder 1 bis 10 werden Cluster 2 und die Länder 11 bis 20 Cluster 1 zugeordnet (siehe »Cluster-Zugehörigkeit«), und zwar nach – wie das Iterationsprotokoll ausweist – fünf Iterationen. Die in Abbildung 11.19 ebenfalls angegebenen »Clusterzentren der endgültigen Lösung« zeigen Ihnen die arithmetischen Mittelwerte der drei Variablen *bip_pk*, *aussenhandel* und *landwi_quote* in den beiden (endgültigen) Clustern, die – wie ersichtlich – deutlich zwischen beiden Clustern voneinander abweichen. Unter »ANOVA« schließlich finden sich in Abbildung 11.19 Hinweise auf die Signifikanz der Clusterbildung. Da in der Spalte »Sig.« jeweils »< 0,001« bei den einzelnen Variablen steht, kann diesbezüglich auf eine hohe Signifikanz geschlossen werden.



**Anfängliche Clusterzentren**

| | Cluster 1 | Cluster 2 |
|---|---|---|
| bip_pk | 1500,00 | 400000,00 |
| landwi_quote | 80,00 | 2,50 |
| aussenhandel | 3,00 | 22,50 |

**Iterationsprotokoll[a]**

| Iteration | Änderung in Clusterzentren 1 | 2 |
|---|---|---|
| 1 | 67156,269 | 87500,000 |
| 2 | 18763,393 | 37500,000 |
| 3 | 21267,858 | 24375,000 |
| 4 | 23275,001 | 21125,000 |
| 5 | ,000 | ,000 |

a. Konvergenz wurde aufgrund geringer oder keiner Änderungen der Clusterzentren erreicht. Die maximale Änderung der absoluten Koordinaten für jedes Zentrum ist ,000. Die aktuelle Iteration lautet 5. Der Mindestabstand zwischen den anfänglichen Zentren beträgt 398500,008.

**Cluster-Zugehörigkeit**

| Fallnummer | land | Cluster | Distanz |
|---|---|---|---|
| 1 | 1 | 2 | 49500,000 |
| 2 | 2 | 2 | 29500,000 |
| 3 | 3 | 2 | 70500,000 |
| 4 | 4 | 2 | 54500,000 |
| 5 | 5 | 2 | 29500,000 |
| 6 | 6 | 2 | 20500,000 |
| 7 | 7 | 2 | 70500,000 |
| 8 | 8 | 2 | 170500,000 |
| 9 | 9 | 2 | 79500,000 |
| 10 | 10 | 2 | 89500,000 |
| 11 | 11 | 1 | 3650,072 |
| 12 | 12 | 1 | 2650,093 |
| 13 | 13 | 1 | 1650,118 |
| 14 | 14 | 1 | 650,283 |
| 15 | 15 | 1 | 350,002 |
| 16 | 16 | 1 | 1350,004 |
| 17 | 17 | 1 | 2350,069 |
| 18 | 18 | 1 | 2650,015 |
| 19 | 19 | 1 | 3350,162 |
| 20 | 20 | 1 | 3850,188 |

**Clusterzentren der endgültigen Lösung**

| | Cluster 1 | Cluster 2 |
|---|---|---|
| bip_pk | 5350,00 | 229500,00 |
| landwi_quote | 42,13 | 7,73 |
| aussenhandel | 6,69 | 20,00 |

**Distanz zwischen Clusterzentren der endgültigen Lösung**

| Cluster | 1 | 2 |
|---|---|---|
| 1 | | 224150,003 |
| 2 | 224150,003 | |

**ANOVA**

| | Cluster Mittel der Quadrate | df | Fehler Mittel der Quadrate | df | F | Sig. |
|---|---|---|---|---|---|---|
| bip_pk | 2,512E+11 | 1 | 3388140277,8 | 18 | 74,146 | <,001 |
| landwi_quote | 5917,144 | 1 | 261,905 | 18 | 22,593 | <,001 |
| aussenhandel | 886,047 | 1 | 4,701 | 18 | 188,467 | <,001 |

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.

**Anzahl der Fälle in jedem Cluster**

| | | |
|---|---|---|
| Cluster | 1 | 10,000 |
| | 2 | 10,000 |
| Gültig | | 20,000 |
| Fehlend | | ,000 |

**Abbildung 11.19:** Output bei der K-Mittelwerte-Clusteranalyse (CLUSTER1.SAV)

## *Hierarchische Clusteranalyse*

Eine dritte Variante der Clusteranalyse stellt die sogenannte **hierarchische Clusteranalyse** dar. In SPSS gelangen Sie zu dieser Variante menügesteuert über ANALYSIEREN|KLASSIFIZIEREN|HIERARCHISCHE CLUSTER. Sie erhalten hierdurch das in Abbildung 11.20 angegebene Dialogfeld, in dem Sie die drei für die Clusterung relevanten Variablen *bip_pk, aussenhandel* und *landwi_quote* eingeben.

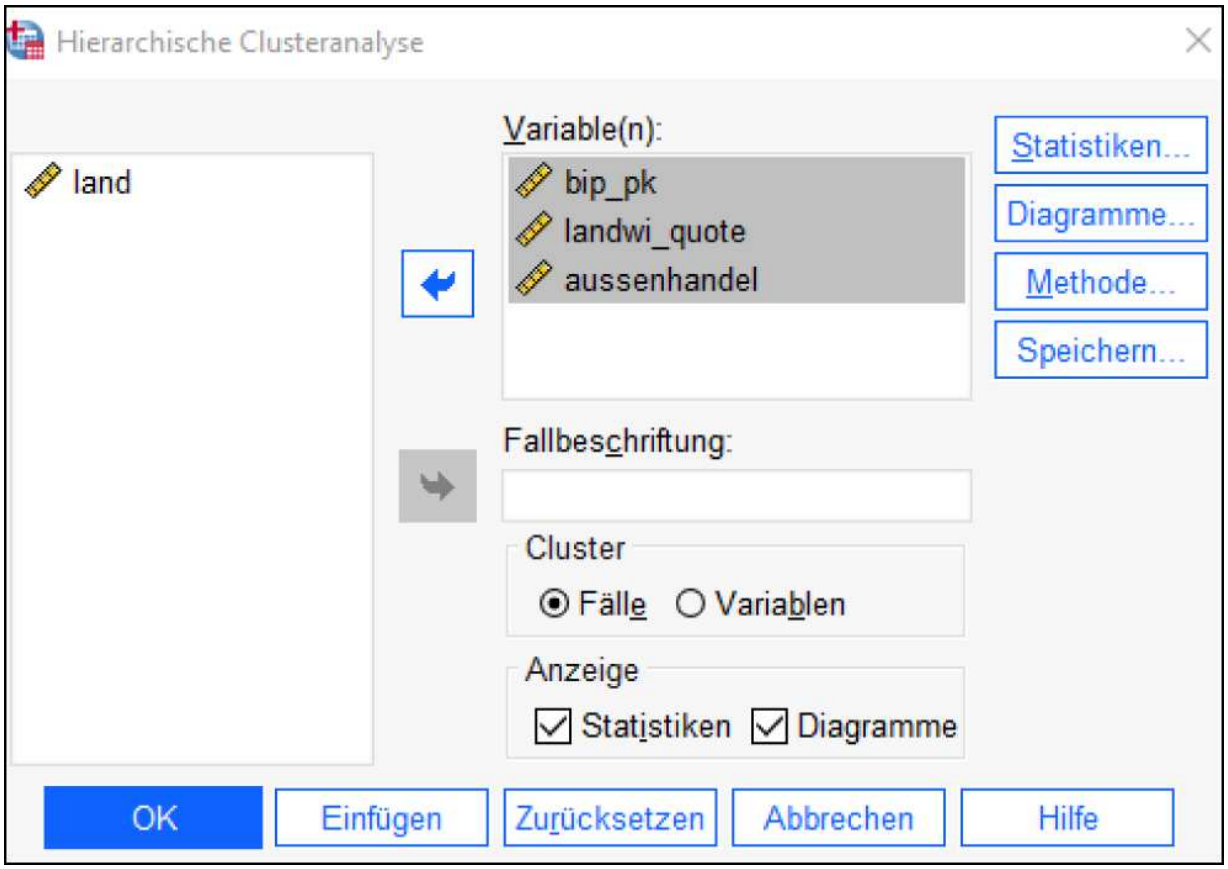



**Abbildung 11.20:** Variablenauswahl bei der hierarchischen Clusteranalyse (CLUSTER1.SAV)

Des Weiteren wählen Sie bitte unter HIERARCHISCHE CLUSTERANALYSE: DIAGRAMME die Option DENDROGRAMM aus (siehe Abbildung 11.21). Mit einem **Dendrogramm** wird der Prozess der Clusterbildung bis hin zu den finalen Clustern visualisiert.

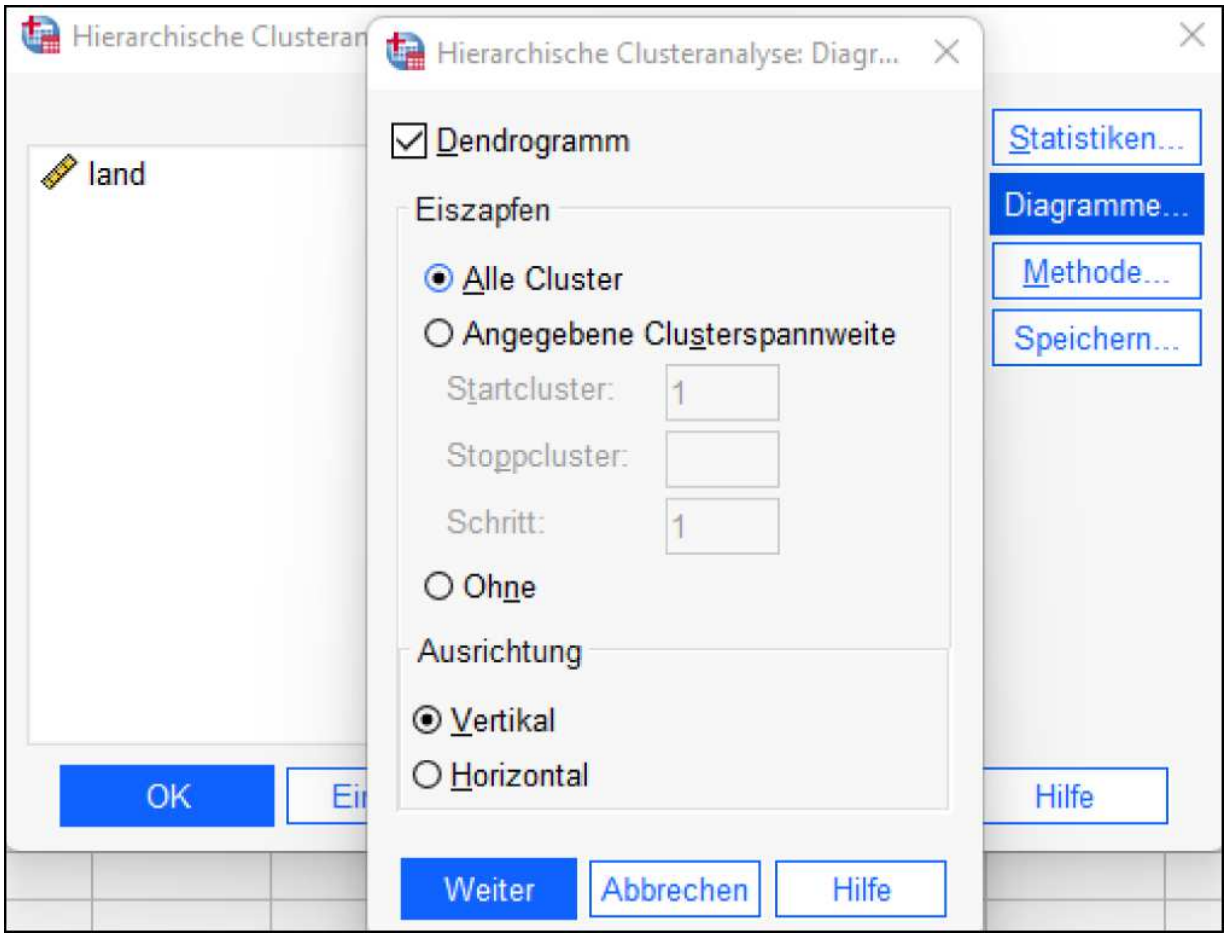



**Abbildung 11.21:** Auswahl eines Dendrogramms bei der hierarchischen Clusteranalyse (CLUSTER1.SAV)

In SPSS-Syntaxform sieht das Ganze so aus:

```
CLUSTER bip_pk landwi_quote aussenhandel
    /METHOD BAVERAGE
    /MEASURE=SEUCLID
    /PRINT SCHEDULE
    /PLOT DENDROGRAM VICICLE.
```

Auch mittels der hierarchischen Clusteranalyse werden im vorliegenden Beispiel am Ende zwei Cluster gebildet – mit den Ländern 1 bis 10 in dem einen Cluster und den Ländern 11 bis 20 in dem anderen Cluster, wie aus Abbildung 11.22 und Abbildung 11.23 ersichtlich ist.

Anhand des Dendrogramms in Abbildung 11.23 erkennen Sie anschaulich den Prozess der Clusterbildung bis zu zwei Clustern am Ende des betreffenden Berechnungsprozesses. Sie können das

Dendogramm so interpretieren, dass am Anfang der Clusterbildung – ganz links am Bildrand der Abbildung 11.23 – fünf vorläufige Cluster gebildet worden sind, und zwar das erste vorläufige Cluster mit den Ländern 1, 2, 4, 5, 9 und 10, das zweite vorläufige Cluster mit den Ländern 3 und 7, das dritte vorläufige Cluster nur mit Land 6, das vierte vorläufige Cluster nur mit Land 8 sowie das fünfte vorläufige Cluster mit den Ländern 11 bis 20. Anschließend sind die ersten drei vorläufigen Cluster zu einem Cluster fusioniert worden, ehe danach auch noch Land 8 in dieses Cluster einbezogen worden ist. Dadurch hat sich das Endergebnis mit den beiden endgültigen Clustern (das eine Cluster mit den Ländern 1 bis 10; das andere Cluster mit den Ländern 11 bis 20) ganz rechts am Bildrand von Abbildung 11.23 ergeben.

**Zusammenfassung der Fallverarbeitung**[a,b]

| Fälle | | | | | |
|---|---|---|---|---|---|
| Gültig | | Fehlend | | Gesamt | |
| N | Prozent | N | Prozent | N | Prozent |
| 20 | 100,0 | 0 | ,0 | 20 | 100,0 |

a. Quadriertes euklidisches Distanzmaß verwendet
b. Linkage zwischen den Gruppen

**Linkage zwischen den Gruppen**

**Zuordnungsübersicht**

| Schritt | Zusammengeführte Cluster Cluster 1 | Cluster 2 | Koeffizienten | Erstes Vorkommen des Clusters Cluster 1 | Cluster 2 | Nächster Schritt |
|---|---|---|---|---|---|---|
| 1 | 2 | 5 | ,279 | 0 | 0 | 14 |
| 2 | 3 | 7 | 1,000 | 0 | 0 | 16 |
| 3 | 12 | 18 | 178,411 | 0 | 0 | 7 |
| 4 | 19 | 20 | 250026,000 | 0 | 0 | 8 |
| 5 | 13 | 14 | 1000000,313 | 0 | 0 | 9 |
| 6 | 15 | 16 | 1000010,250 | 0 | 0 | 10 |
| 7 | 11 | 12 | 1000098,581 | 0 | 3 | 9 |
| 8 | 17 | 19 | 1625313,750 | 0 | 4 | 10 |
| 9 | 11 | 13 | 3833378,023 | 7 | 5 | 11 |
| 10 | 15 | 17 | 6084179,792 | 6 | 8 | 11 |
| 11 | 11 | 15 | 22931633,794 | 9 | 10 | 19 |
| 12 | 1 | 4 | 25000000,080 | 0 | 0 | 14 |
| 13 | 9 | 10 | 100000000,26 | 0 | 0 | 15 |
| 14 | 1 | 2 | 512500001,08 | 12 | 1 | 15 |
| 15 | 1 | 9 | 2068750004,8 | 14 | 13 | 17 |
| 16 | 3 | 6 | 2500000002,5 | 2 | 0 | 17 |
| 17 | 1 | 3 | 12993055572 | 15 | 16 | 18 |
| 18 | 1 | 8 | 39069444490 | 17 | 0 | 19 |
| 19 | 1 | 11 | 56341876841 | 18 | 11 | 0 |

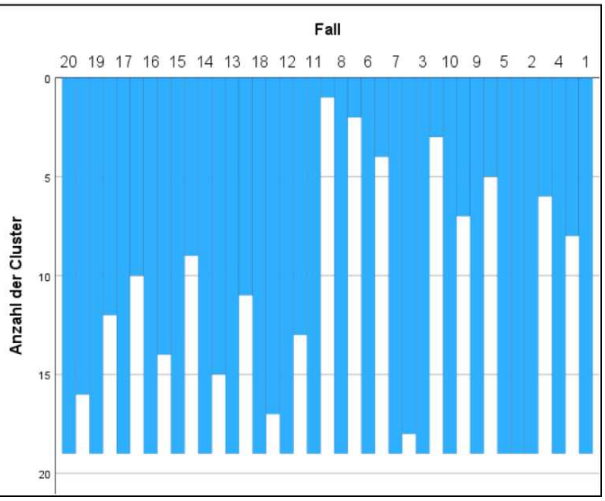



**Abbildung 11.22:** Output der hierarchischen Clusteranalyse (CLUSTER1.SAV) (1)

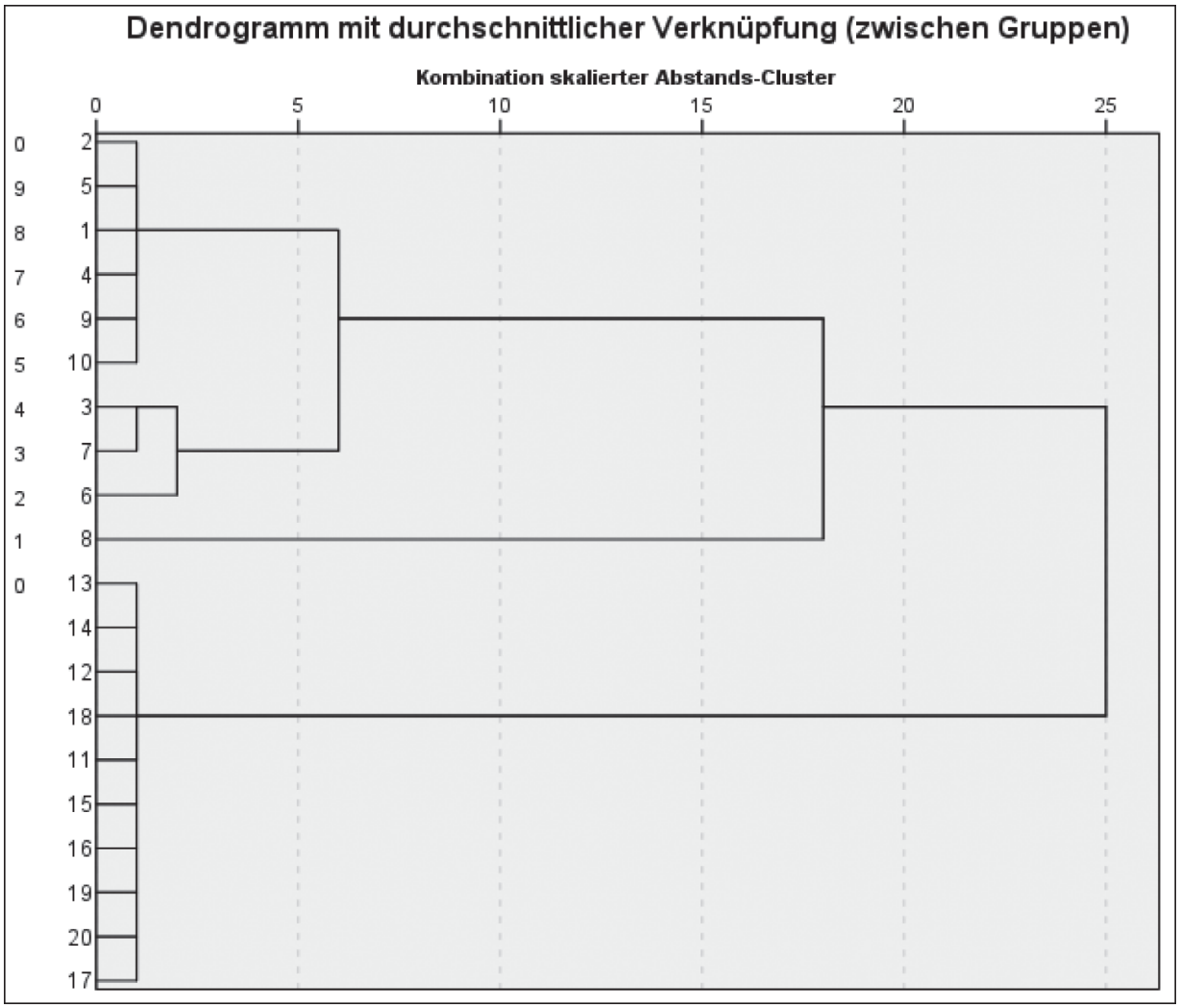



**Abbildung 11.23:** Output der hierarchischen Clusteranalyse (CLUSTER1.SAV): Dendrogramm (2)

# *Faktorenanalyse*

Bei der **Faktorenanalyse** werden metrisch skalierte Variablen zu Gruppen beziehungsweise zu Faktoren zusammengefasst. Hierdurch unterscheidet sich die Faktorenanalyse etwa von der Clusteranalyse, bei der – wie im vorstehenden Abschnitt gezeigt worden ist – Fälle zu Gruppen zusammengefasst werden. Die nachfolgende Darstellung der Faktorenanalyse ist stark an das instruktive Paper von Müller, Wolfgang (2015): Marketing Analytics: Faktorenanalyse, Dortmund, Wintersemester 2015/2016 (mimeo) angelehnt.

## *Idee und Annahmen*

Die Zusammenfassung von Variablen zu Faktoren im Rahmen der Faktorenanalyse erfolgt aus Gründen der Datenstrukturierung beziehungsweise -reduktion. Letztlich dient die Faktorenanalyse also der Dimensionsreduktion.

Die Voraussetzungen der Faktorenanalyse sind streng genommen:

- ✔ genügend große Stichproben (mindestens zehn Fälle pro Variable als Faustregel),
- ✔ eine ausreichende Anzahl an Variablen (als Faustregel: mindestens vier Variablen pro Faktor) und
- ✔ metrisch skalierte Variablen.

In der Praxis werden aber zum Beispiel auch ordinalskalierte Variablen verwendet.

## *Ein Beispiel in SPSS*

Das hier gewählte (fiktive) Beispiel zur Faktorenanalyse gründet sich auf die in Abbildung 11.24 dargestellte Datendatei *FAKTOREN1.SAV*, und zwar sollen die Variablen *werbebudget, werbeart, design, eigenkapital, liquidität* und *umsatz* zu Faktoren zusammengefasst werden. Bei den ersten drei Variablen soll es sich in diesem fiktiven Beispiel um Marketing-Kennziffern und bei den drei letztgenannten Variablen um Finanzierungs-Kennziffern von insgesamt 30 Betrieben handeln. Aus Illustrationszwecken wird hier ein Beispiel mit einer geringen Stichprobengröße und einer vergleichsweise geringen Variablenanzahl gewählt, das – streng genommen – die im vorstehenden Unterabschnitt genannten ersten beiden Voraussetzungen nicht erfüllt.

## *Einstellungen im Rahmen der Faktorenanalyse*

Um mit der Faktorenanalyse beginnen zu können, müssen Sie im SPSS-Menü zunächst ANALYSIEREN|DIMENSIONSREDUKTION|FAKTORENANALYSE anklicken. In dem sich dann öffnenden Dialogfenster müssen Sie die Variablen angeben, für die eine Faktorenanalyse durchgeführt werden

soll. In unserem Beispiel sind also *werbebudget, werbeart, design, eigenkapital, liquidität* und *umsatz* auszuwählen (siehe Abbildung 11.25).

FAKTOREN1.sav [DataSet4] - IBM SPSS Statistics Dateneditor

Datei Bearbeiten Ansicht Daten Transformieren Analysieren Grafik Extras

| | betrieb | werbebudget | werbeart | design | eigenkapital | liquidität | umsatz | Var |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 10,00 | 5 | 4 | 75,00 | 80,00 | 40,00 | |
| 2 | 2 | 20,00 | 6 | 5 | 39,00 | 44,00 | 25,00 | |
| 3 | 3 | 15,00 | 4 | 6 | 42,00 | 40,00 | 30,00 | |
| 4 | 4 | 14,00 | 6 | 6 | 50,00 | 50,00 | 35,00 | |
| 5 | 5 | 10,00 | 6 | 6 | 70,00 | 75,00 | 45,00 | |
| 6 | 6 | 15,00 | 5 | 4 | 40,00 | 49,00 | 27,00 | |
| 7 | 7 | 15,00 | 3 | 4 | 37,00 | 43,00 | 25,00 | |
| 8 | 8 | 17,00 | 5 | 6 | 35,00 | 43,00 | 23,00 | |
| 9 | 9 | 16,00 | 6 | 5 | 37,00 | 40,00 | 24,00 | |
| 10 | 10 | 15,00 | 5 | 5 | 44,00 | 45,00 | 33,00 | |
| 11 | 11 | 20,00 | 2 | 4 | 29,00 | 32,00 | 25,00 | |
| 12 | 12 | 22,00 | 4 | 3 | 30,00 | 35,00 | 25,00 | |
| 13 | 13 | 24,00 | 3 | 4 | 62,00 | 65,00 | 38,00 | |
| 14 | 14 | 22,00 | 4 | 3 | 40,00 | 52,00 | 26,00 | |
| 15 | 15 | 20,00 | 4 | 4 | 80,00 | 75,00 | 45,00 | |
| 16 | 16 | 31,00 | 4 | 3 | 70,00 | 68,00 | 38,00 | |
| 17 | 17 | 29,00 | 3 | 4 | 20,00 | 33,00 | 19,00 | |
| 18 | 18 | 32,00 | 4 | 3 | 48,00 | 45,00 | 32,00 | |
| 19 | 19 | 30,00 | 3 | 3 | 50,00 | 50,00 | 34,00 | |
| 20 | 20 | 40,00 | 3 | 5 | 20,00 | 30,00 | 20,00 | |
| 21 | 21 | 35,00 | 2 | 2 | 15,00 | 20,00 | 15,00 | |
| 22 | 22 | 40,00 | 3 | 3 | 16,00 | 21,00 | 16,00 | |
| 23 | 23 | 30,00 | 2 | 2 | 17,00 | 25,00 | 16,00 | |
| 24 | 24 | 40,00 | 2 | 2 | 15,00 | 15,00 | 14,00 | |
| 25 | 25 | 40,00 | 2 | 2 | 18,00 | 15,00 | 13,00 | |
| 26 | 26 | 50,00 | 1 | 2 | 33,00 | 30,00 | 23,00 | |
| 27 | 27 | 60,00 | 2 | 1 | 24,00 | 25,00 | 20,00 | |
| 28 | 28 | 50,00 | 1 | 1 | 29,00 | 30,00 | 25,00 | |
| 29 | 29 | 50,00 | 1 | 1 | 13,00 | 20,00 | 14,00 | |
| 30 | 30 | 50,00 | 1 | 1 | 15,00 | 15,00 | 12,00 | |
| 31 | | | | | | | | |

**Abbildung 11.24:** Datendatei FAKTOREN1.SAV

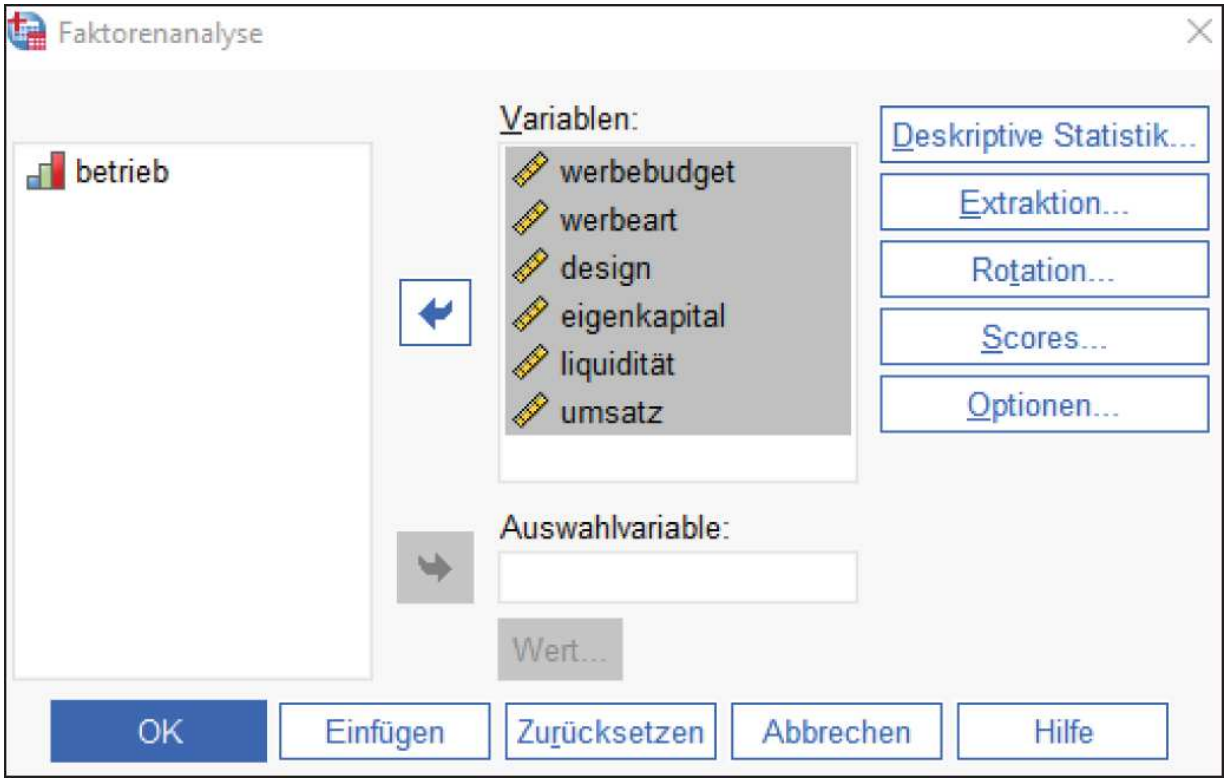



**Abbildung 11.25:** Variablenauswahl bei der Faktorenanalyse (FAKTOREN1.SAV)

Unter FAKTORENANALYSE: DESKRIPTIVE STATISTIK können Sie nun verschiedene statistische Kennziffern auswählen. Neben einfachen (deskriptiven) Statistiken und den Korrelationen zwischen den ausgewählten Variablen (inklusive der betreffenden Signifikanzangaben) wählen Sie bitte auch den sogenannten **KMO-** beziehungsweise *Bartlett-Test* und die **Anti-Image-Tabelle** aus (siehe Abbildung 11.26).

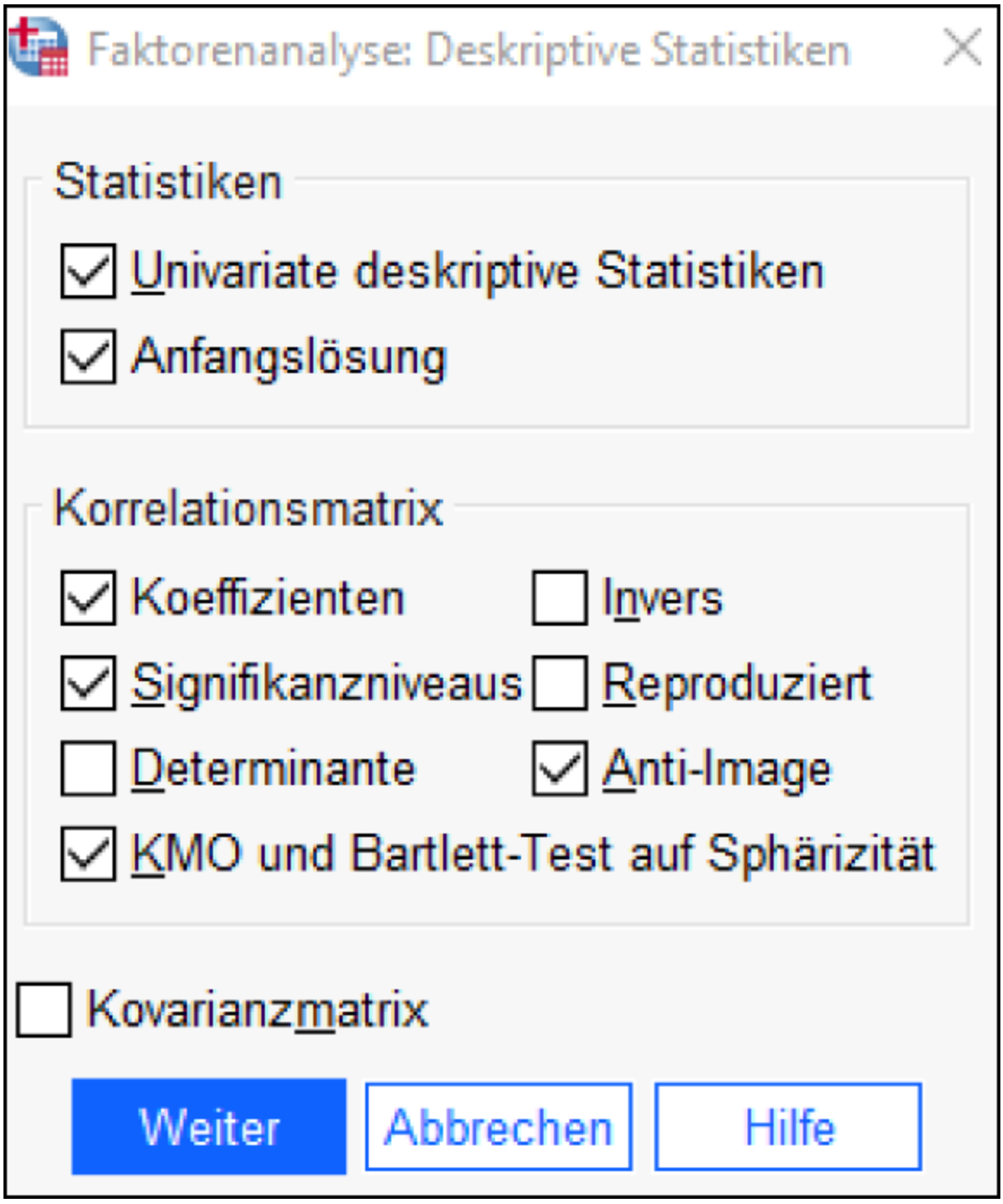



**Abbildung 11.26:** Statistiken bei der Faktorenanalyse (FAKTOREN1.SAV)

Bei dem KMO- beziehungsweise Bartlett-Test wird gecheckt, ob die ausgewählten Variablen in der Grundgesamtheit unkorreliert sind. Mithilfe der Anti-Image-Tabelle wird untersucht, ob sämtliche ausgewählten Variablen in die Faktorenbildung einbezogen werden oder ob bestimmte Variablen aufgrund fehlender korrelativer

Zusammenhänge von der Faktorenbildung ausgeschlossen werden. Sowohl der KMO/Bartlett-Test als auch die Anti-Image-Tabelle liefern Hinweise auf die Güte der Variablenkorrelationen, die der Faktorenanalyse zugrunde liegen.

Im nächsten Schritt wird unter FAKTORENANALYSE: EXTRAKTION die **Hauptkomponentenmethode** ausgewählt, da dieses – in SPSS voreingestellte – Verfahren in der Praxis am häufigsten eingesetzt wird und als besonders leistungsfähig gilt. Außerdem wird ein sogenannter **Screeplot** angefordert (Abbildung 11.27). Mit einem Screeplot kann grafisch beurteilt werden, wie viele Faktoren letztlich gebildet werden sollen.

In dem Dialogfeld der Abbildung 11.27 ist zudem angegeben, dass ein Faktor bei einem **Eigenwert** größer 1 gebildet werden soll. Der Eigenwert einer Variablen gibt dabei an, wie hoch der Anteil der gesamten Varianz aller Variablen ist, der durch diesen Faktor erklärt wird. Je höher ein Eigenwert ist, desto höher ist also der Erklärungsbeitrag eines Faktors. Dadurch, dass die Faktorenanalyse von standardisierten Variablen mit einer Varianz dieser standardisierten Variablen von jeweils 1 ausgeht, leitet sich die »Faktorenbildungs-Regel« eines Eigenwerts größer 1 ab: Ein Faktor soll also mehr von der Gesamtvarianz erklären als eine einzelne Variable. (Alternativ hätten Sie auch eine feste Faktorenzahl vorgeben können.)

Unter FAKTORENANALYSE: ROTATION wird ferner hier die VARIMAX-Rotationsform ausgewählt, die sich in der Praxis bewährt hat (Abbildung 11.28). Im Rahmen einer Faktorenanalyse wird nämlich mit mathematischen Methoden typischerweise eine sogenannte **Rotation** der Faktorenstruktur durchgeführt, um die Interpretation der gewonnenen Faktoren zu erleichtern.

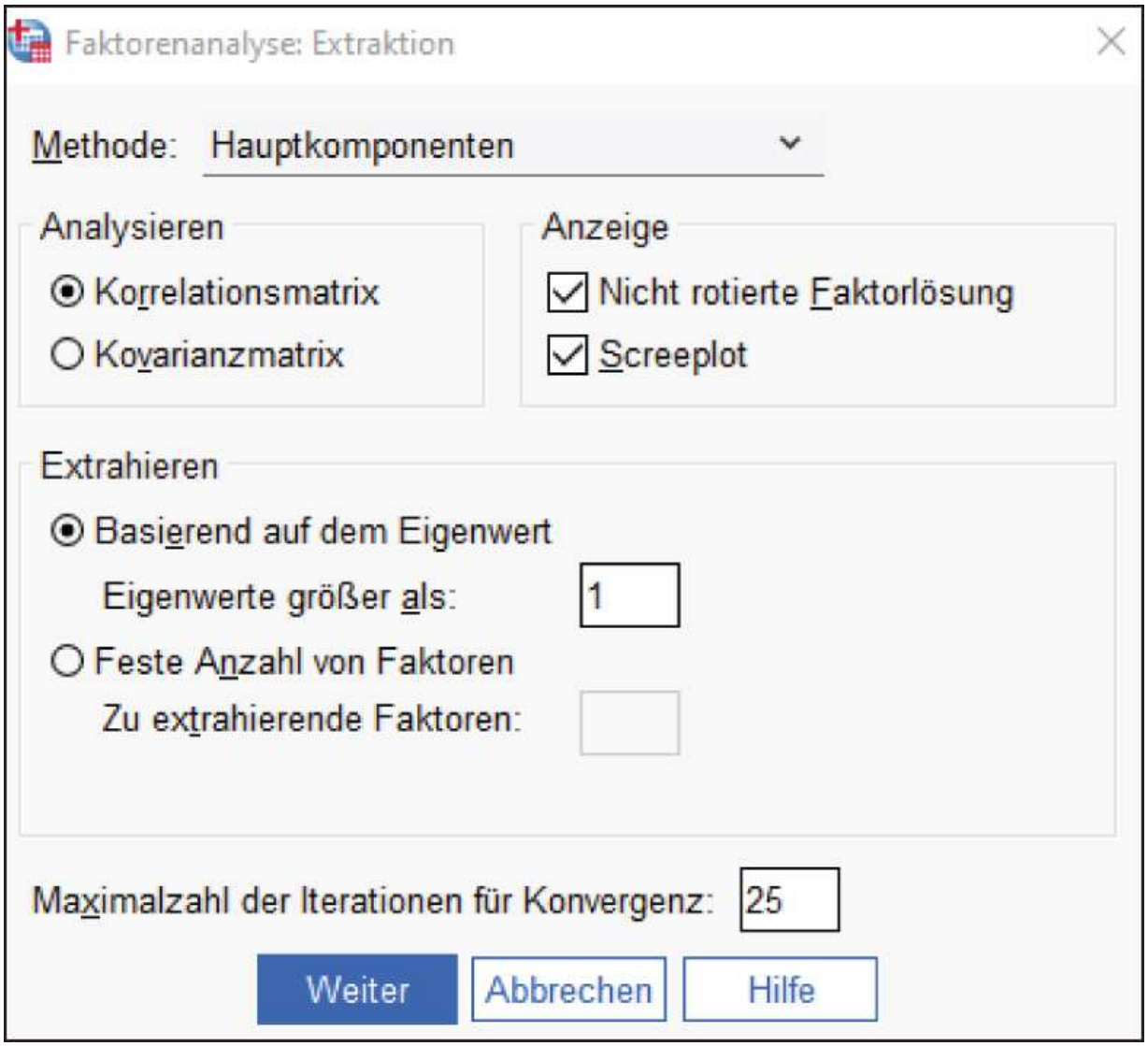



**Abbildung 11.27:** Wahl der konkreten Faktorenanalyse und »Faktoren-Entscheidungsregel« (FAKTOREN1.SAV)

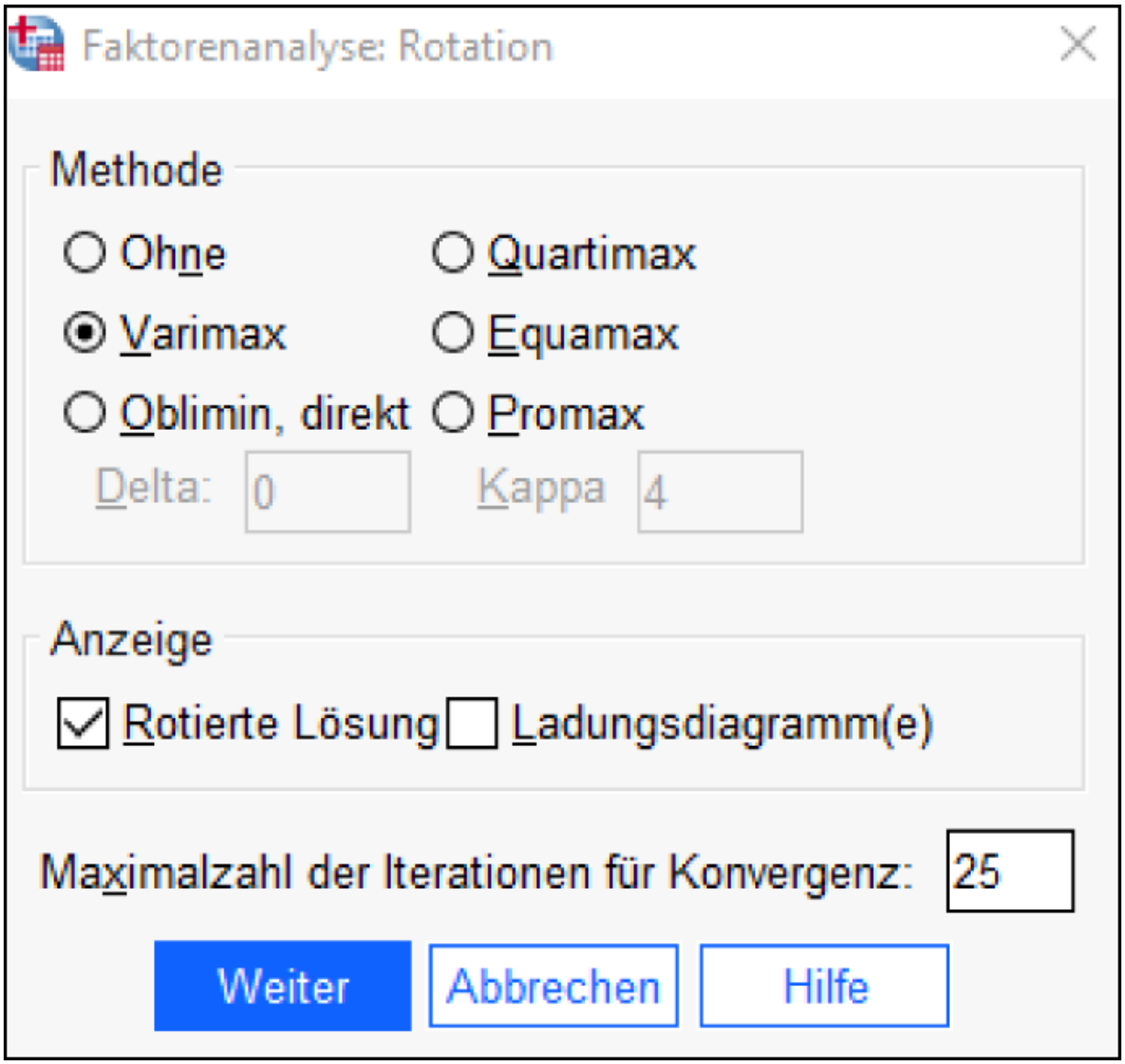



**Abbildung 11.28:** Auswahl eines Rotationsverfahrens (VARIMAX) bei der Faktorenanalyse (FAKTOREN1.SAV)

Bei FAKTORENANALYSE: FACTORSCORES werden Faktorwerte für die einzelnen Fälle (hier: Länder) angefordert, die SPSS als eigene Variablen in den Datensatz schreibt (siehe hierzu Abbildung 11.29), wie später noch gezeigt wird. Auf der betreffenden Basis können dann die einzelnen Fälle der Datendatei hinsichtlich der Stärke der einzelnen Faktoren beurteilt werden.

Unter FAKTORENANALYSE: OPTIONEN schließlich lassen Sie sich die Koeffizienten der einzelnen Faktoren in der Sortierung nach ihrer

Größe anzeigen (siehe Abbildung 11.30). Das ist aber letztlich keine allzu wichtige Einstellung; sie dient nur der Übersichtlichkeit der Ergebnisdarstellung.

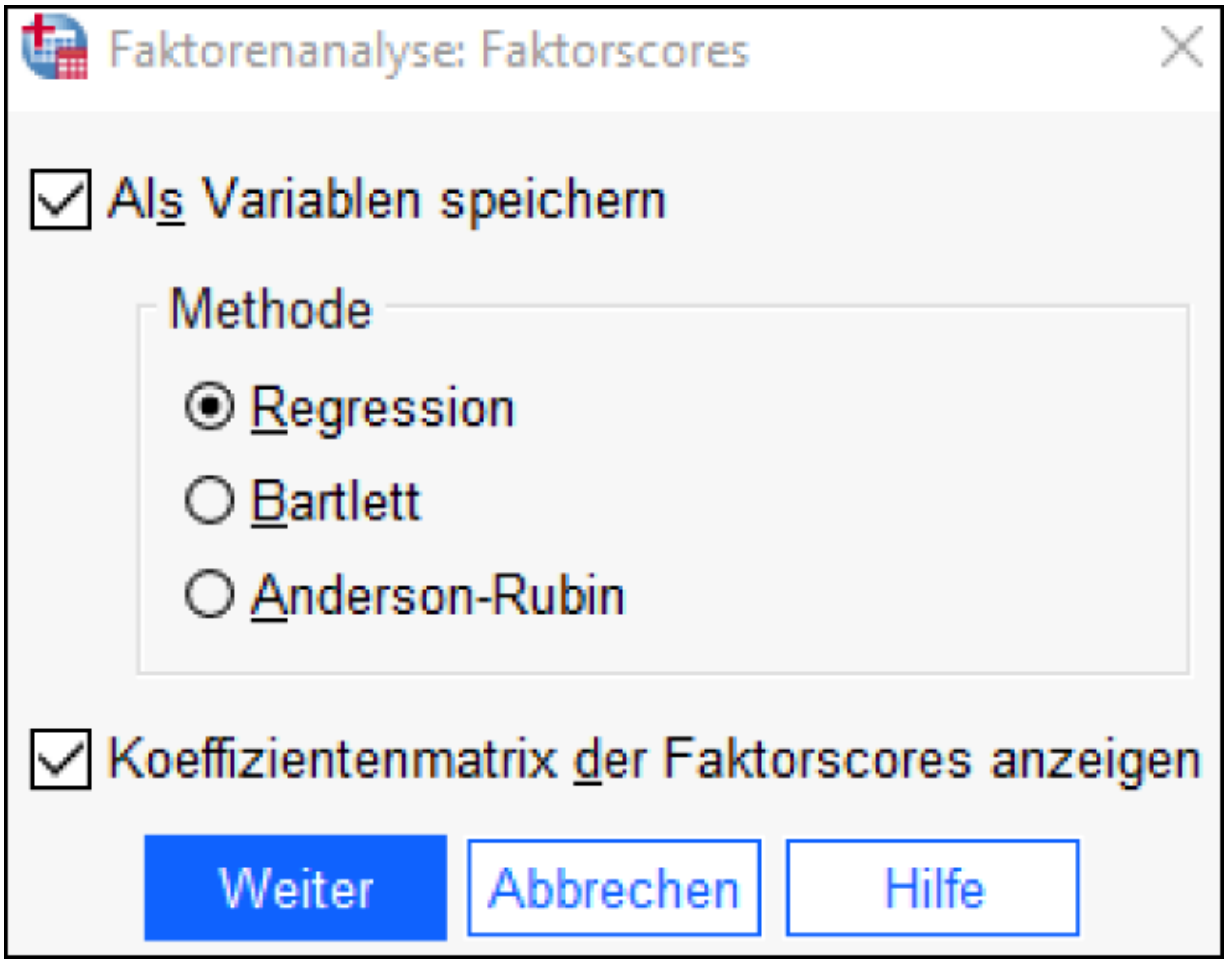



**Abbildung 11.29:** Faktorwerte für die einzelnen Fälle bei der Faktorenanalyse (FAKTOREN1.SAV)

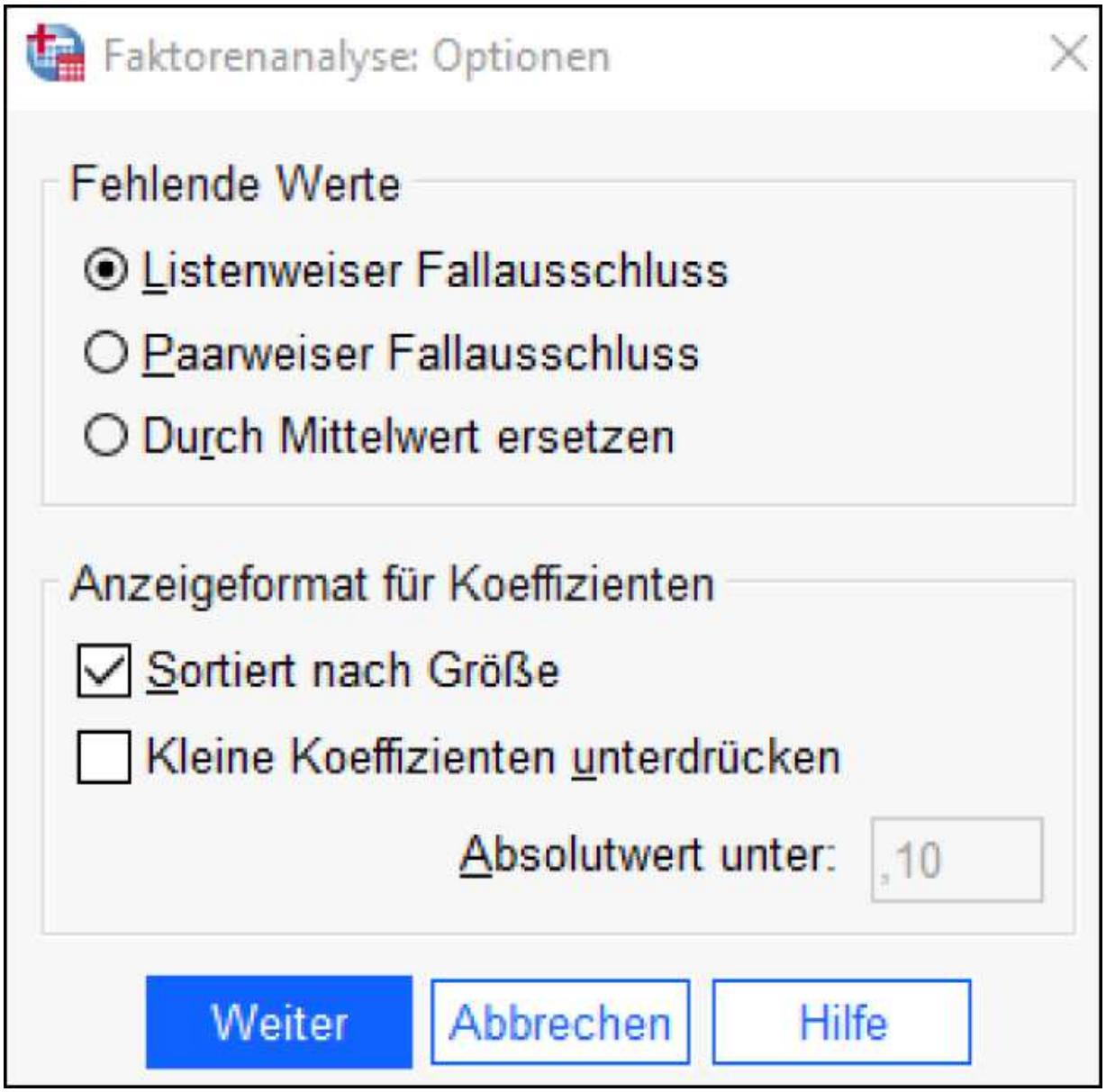



**Abbildung 11.30:** Faktorenanalyse: Optionen (hier in Bezug auf FAKTOREN1.SAV)

Alle vorgenannten Setzungen korrespondieren mit folgender Syntax:

```
FACTOR
    /VARIABLES werbebudget werbeart design eigenkapital liquidität umsatz
    /MISSING LISTWISE
    /ANALYSIS werbebudget werbeart design eigenkapital liquidität umsatz
    /PRINT UNIVARIATE INITIAL CORRELATION SIG KMO AIC EXTRACTION ROTATION
FSCORE
    /FORMAT SORT
    /PLOT EIGEN
    /CRITERIA MINEIGEN(1) ITERATE(25)
```

```
/EXTRACTION PC
/CRITERIA ITERATE(25)
/ROTATION VARIMAX
/SAVE REG(ALL)
/METHOD=CORRELATION.
```

## *Allgemeiner Faktorenoutput*

Abbildung 11.31 und Abbildung 11.32 geben den umfangreichen Output Ihrer Faktorenanalyse im SPSS-Viewer an, wobei zunächst deskriptive Statistiken ausgewiesen werden – in Form der arithmetischen Mittelwerte und Standardabweichungen der einzelnen Variablen.

Die SPSS-Hauptkomponentenmethode hat im vorliegenden Beispielfall zwei Faktoren gebildet, was aufgrund der diesbezüglichen Eigenwerte größer 1 deutlich wird. Allerdings ist der Eigenwert des zweiten gebildeten Faktors nur ganz knapp größer 1, was auch aus dem Screeplot in Abbildung 11.32 ersichtlich ist.

Aus der rotierten Komponentenmatrix sind die Elemente der beiden Faktoren ableitbar. Allgemein gilt in diesem Zusammenhang, dass die Korrelationswerte in der rotierten Komponentenmatrix in Absolutbeträgen größer 0,5 sein sollten, damit eine Variable einem Faktor zugehörig ist. Legt man dieses Kriterium an unsere Ergebnisse an, kann man die Variablen *eigenkapital, umsatz* und *liquidität* dem ersten Faktor und die Variablen *design, werbeart* und *werbebudget* offenkundig dem zweiten Faktor zuordnen. Aufgrund der inhaltlichen Variablenbedeutungen könnte man daher den ersten Faktor als »Finanz-Faktor« und den zweiten Faktor als »Marketing-Faktor« bezeichnen.

In Abbildung 11.31 sind im Übrigen auch »*Kommunalitäten*« ausgewiesen. Hierunter wird der von den einzelnen Faktoren erklärte Anteilswert an der Gesamtstreuung der einzelnen Variablen verstanden. Dabei gilt diese faktorenbezogene »Erklärung« als umso besser, desto näher der Wert für die Kommunalität am Wert 1 liegt. Im vorliegenden Beispielfall sind aus der Spalte »Extraktion« (das heißt: nach der endgültigen Faktorenbildung) diesbezüglich Werte zwischen 0,877 (*werbeart*) und 0,989 (*eigenkapital*) ersichtlich. Diese Werte liegen also recht nahe am Zielwert von 1, sodass hier auf eine recht gute

Erklärungskraft des gebildeten Zweifaktorenmodells geschlossen werden kann.

**Deskriptive Statistiken**

| | Mittelwert | Std.-Abweichung | Analyse N |
|---|---|---|---|
| werbebudget | 28,7333 | 13,98012 | 30 |
| werbeart | 3,40 | 1,610 | 30 |
| design | 3,47 | 1,592 | 30 |
| eigenkapital | 37,1000 | 19,32008 | 30 |
| liquidität | 40,3333 | 18,46587 | 30 |
| umsatz | 25,9000 | 9,42246 | 30 |

**Anti-Image-Matrizen**

| | | werbebudget | werbeart | design | eigenkapital | liquidität | umsatz |
|---|---|---|---|---|---|---|---|
| Anti-Image-Kovarianz | werbebudget | ,217 | ,055 | ,100 | -,008 | ,029 | -,007 |
| | werbeart | ,055 | ,228 | -,112 | -,010 | -,010 | ,021 |
| | design | ,100 | -,112 | ,208 | ,019 | ,006 | -,036 |
| | eigenkapital | -,008 | -,010 | ,019 | ,027 | -,023 | -,025 |
| | liquidität | ,029 | -,010 | ,006 | -,023 | ,044 | -,001 |
| | umsatz | -,007 | ,021 | -,036 | -,025 | -,001 | ,047 |
| Anti-Image-Korrelation | werbebudget | ,868[a] | ,247 | ,471 | -,101 | ,300 | -,065 |
| | werbeart | ,247 | ,865[a] | -,514 | -,131 | -,105 | ,200 |
| | design | ,471 | -,514 | ,762[a] | ,253 | ,060 | -,367 |
| | eigenkapital | -,101 | -,131 | ,253 | ,735[a] | -,663 | -,703 |
| | liquidität | ,300 | -,105 | ,060 | -,663 | ,848[a] | -,023 |
| | umsatz | -,065 | ,200 | -,367 | -,703 | -,023 | ,811[a] |

a. Maß der Stichprobeneignung

**Korrelationsmatrix**

| | | werbebudget | werbeart | design | eigenkapital | liquidität | umsatz |
|---|---|---|---|---|---|---|---|
| Korrelation | werbebudget | 1,000 | -,818 | -,828 | -,599 | -,675 | -,616 |
| | werbeart | -,818 | 1,000 | ,826 | ,602 | ,661 | ,607 |
| | design | -,828 | ,826 | 1,000 | ,487 | ,551 | ,544 |
| | eigenkapital | -,599 | ,602 | ,487 | 1,000 | ,970 | ,971 |
| | liquidität | -,675 | ,661 | ,551 | ,970 | 1,000 | ,948 |
| | umsatz | -,616 | ,607 | ,544 | ,971 | ,948 | 1,000 |
| Sig. (1-seitig) | werbebudget | | <,001 | <,001 | <,001 | <,001 | <,001 |
| | werbeart | ,000 | | ,000 | ,000 | ,000 | ,000 |
| | design | ,000 | ,000 | | ,003 | ,001 | ,001 |
| | eigenkapital | ,000 | ,000 | ,003 | | ,000 | ,000 |
| | liquidität | ,000 | ,000 | ,001 | ,000 | | ,000 |
| | umsatz | ,000 | ,000 | ,001 | ,000 | ,000 | |

**Kommunalitäten**

| | Anfänglich | Extraktion |
|---|---|---|
| werbebudget | 1,000 | ,879 |
| werbeart | 1,000 | ,877 |
| design | 1,000 | ,907 |
| eigenkapital | 1,000 | ,989 |
| liquidität | 1,000 | ,973 |
| umsatz | 1,000 | ,968 |

Extraktionsmethode: Hauptkomponentenanalyse.

**KMO- und Bartlett-Test**

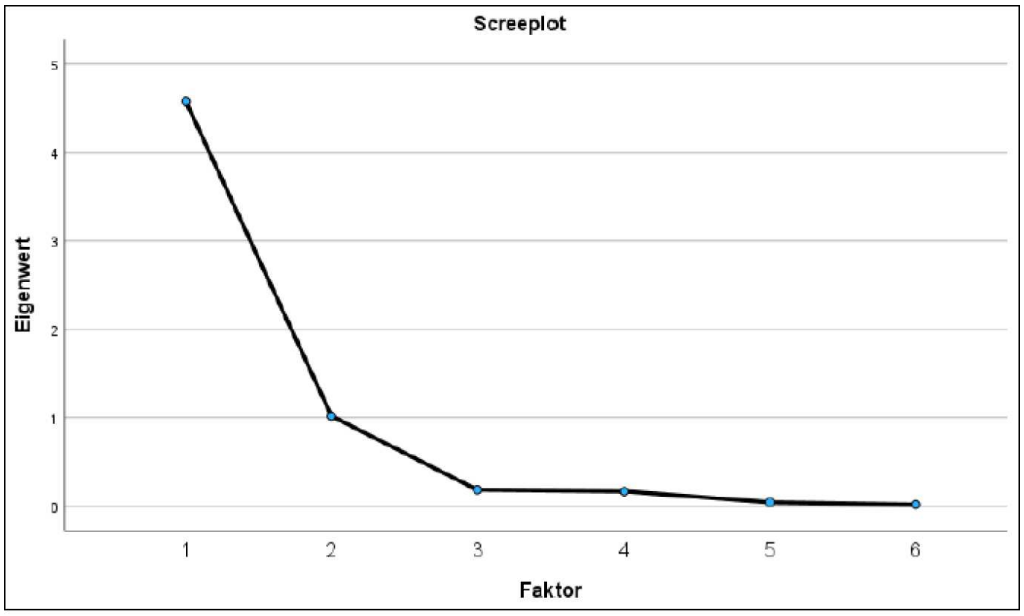
| | | |
|---|---|---|
| Maß der Stichprobeneignung nach Kaiser-Meyer-Olkin. | | ,812 |
| Bartlett-Test auf Sphärizität | Ungefähres Chi-Quadrat | 240,878 |
| | df | 15 |
| | Signifikanz nach Bartlett | <,001 |

**Erklärte Gesamtvarianz**

| | Anfängliche Eigenwerte | | | Summen von quadrierten Faktorladungen für Extraktion | | | Rotierte Summe der quadrierten Ladungen | | |
|---|---|---|---|---|---|---|---|---|---|
| Komponente | Gesamt | % der Varianz | Kumulierte % | Gesamt | % der Varianz | Kumulierte % | Gesamt | % der Varianz | Kumulierte % |
| 1 | 4,576 | 76,261 | 76,261 | 4,576 | 76,261 | 76,261 | 2,924 | 48,736 | 48,736 |
| 2 | 1,017 | 16,953 | 93,214 | 1,017 | 16,953 | 93,214 | 2,669 | 44,478 | 93,214 |
| 3 | ,183 | 3,049 | 96,263 | | | | | | |
| 4 | ,165 | 2,752 | 99,015 | | | | | | |
| 5 | ,042 | ,702 | 99,717 | | | | | | |
| 6 | ,017 | ,283 | 100,000 | | | | | | |

Extraktionsmethode: Hauptkomponentenanalyse.

**Abbildung 11.31:** Output der Faktorenanalyse (FAKTOREN1.SAV) (1)




**Komponentenmatrix[a]**

| | Komponente 1 | Komponente 2 |
|---|---|---|
| liquidität | ,925 | ,343 |
| umsatz | ,903 | ,392 |
| eigenkapital | ,893 | ,438 |

**Komponententransformationsmatrix**

| Komponente | 1 | 2 |
|---|---|---|
| 1 | ,732 | ,681 |
| 2 | ,681 | -,732 |

Extraktionsmethode: Hauptkomponentenanalyse.
Rotationsmethode: Varimax mit Kaiser-Normalisierung.

| | | |
|---|---|---|
| werbebudget | -,860 | ,374 |
| werbeart | ,856 | -,381 |
| design | ,798 | -,519 |

Extraktionsmethode: Hauptkomponentenanalyse.

a. 2 Komponenten extrahiert

**Rotierte Komponentenmatrix[a]**

| | Komponente 1 | 2 |
|---|---|---|
| eigenkapital | ,952 | ,288 |
| umsatz | ,927 | ,328 |
| liquidität | ,911 | ,379 |
| design | ,230 | ,924 |
| werbeart | ,367 | ,862 |
| werbebudget | -,375 | -,859 |

Extraktionsmethode: Hauptkomponentenanalyse. Rotationsmethode: Varimax mit Kaiser-Normalisierung.

a. Die Rotation ist in 3 Iterationen konvergiert.

**Koeffizientenmatrix der Komponentenwerte**

| | Komponente 1 | 2 |
|---|---|---|
| werbebudget | ,113 | -,397 |
| werbeart | -,118 | ,402 |
| design | -,220 | ,493 |
| eigenkapital | ,436 | -,182 |
| liquidität | ,378 | -,109 |
| umsatz | ,407 | -,147 |

Extraktionsmethode: Hauptkomponentenanalyse. Rotationsmethode: Varimax mit Kaiser-Normalisierung. Komponentenscores.

**Kovarianzmatrix des Komponentenwerts**

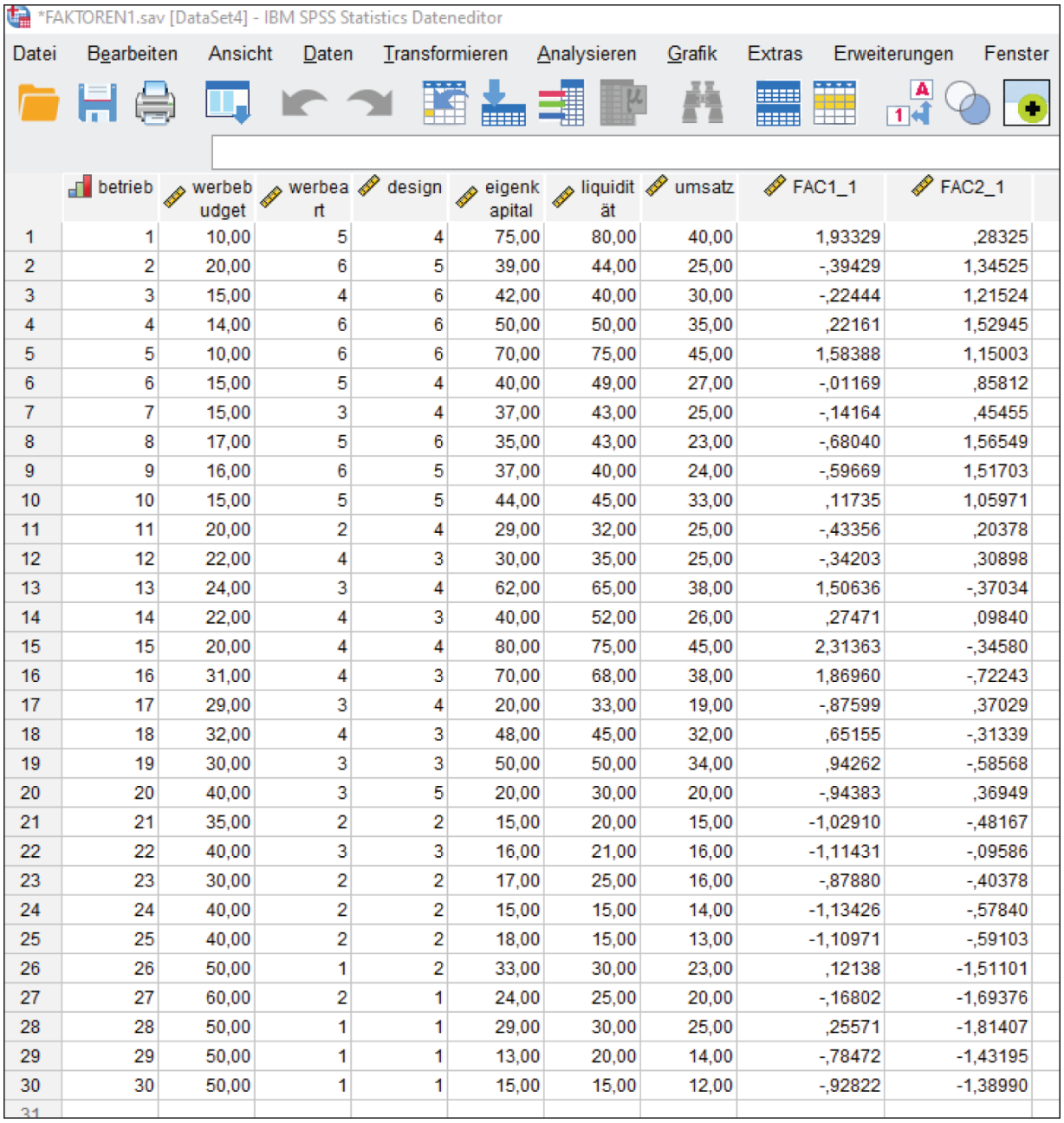
| Komponente | 1 | 2 |
|---|---|---|
| 1 | 1,000 | ,000 |
| 2 | ,000 | 1,000 |

Extraktionsmethode: Hauptkomponentenanalyse. Rotationsmethode: Varimax mit Kaiser-Normalisierung. Komponentenscores.

**Abbildung 11.32:** Output der Faktorenanalyse (FAKTOREN1.SAV) (2)

## *Fallbezogener Output*

Gleichzeitig mit diesem Output hat SPSS, wie gewünscht, die Faktorwerte für die beiden Faktoren in die Datendatei geschrieben und dort den 30 Betrieben zugeordnet. Die Faktorwerte des ersten Faktors (»Finanz-Faktor«) sind in der Variablen *fac1_1* und diejenigen des zweiten Faktors (»Marketing-Faktor«) in der Variablen *fac2_1* enthalten (siehe Abbildung 11.33).




| | betrieb | werbebudget | werbeart | design | eigenkapital | liquidität | umsatz | FAC1_1 | FAC2_1 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 10,00 | 5 | 4 | 75,00 | 80,00 | 40,00 | 1,93329 | ,28325 |
| 2 | 2 | 20,00 | 6 | 5 | 39,00 | 44,00 | 25,00 | -,39429 | 1,34525 |
| 3 | 3 | 15,00 | 4 | 6 | 42,00 | 40,00 | 30,00 | -,22444 | 1,21524 |
| 4 | 4 | 14,00 | 6 | 6 | 50,00 | 50,00 | 35,00 | ,22161 | 1,52945 |
| 5 | 5 | 10,00 | 6 | 6 | 70,00 | 75,00 | 45,00 | 1,58388 | 1,15003 |
| 6 | 6 | 15,00 | 5 | 4 | 40,00 | 49,00 | 27,00 | -,01169 | ,85812 |
| 7 | 7 | 15,00 | 3 | 4 | 37,00 | 43,00 | 25,00 | -,14164 | ,45455 |
| 8 | 8 | 17,00 | 5 | 6 | 35,00 | 43,00 | 23,00 | -,68040 | 1,56549 |
| 9 | 9 | 16,00 | 6 | 5 | 37,00 | 40,00 | 24,00 | -,59669 | 1,51703 |
| 10 | 10 | 15,00 | 5 | 5 | 44,00 | 45,00 | 33,00 | ,11735 | 1,05971 |
| 11 | 11 | 20,00 | 2 | 4 | 29,00 | 32,00 | 25,00 | -,43356 | ,20378 |
| 12 | 12 | 22,00 | 4 | 3 | 30,00 | 35,00 | 25,00 | -,34203 | ,30898 |
| 13 | 13 | 24,00 | 3 | 4 | 62,00 | 65,00 | 38,00 | 1,50636 | -,37034 |
| 14 | 14 | 22,00 | 4 | 3 | 40,00 | 52,00 | 26,00 | ,27471 | ,09840 |
| 15 | 15 | 20,00 | 4 | 4 | 80,00 | 75,00 | 45,00 | 2,31363 | -,34580 |
| 16 | 16 | 31,00 | 4 | 3 | 70,00 | 68,00 | 38,00 | 1,86960 | -,72243 |
| 17 | 17 | 29,00 | 3 | 4 | 20,00 | 33,00 | 19,00 | -,87599 | ,37029 |
| 18 | 18 | 32,00 | 4 | 3 | 48,00 | 45,00 | 32,00 | ,65155 | -,31339 |
| 19 | 19 | 30,00 | 3 | 3 | 50,00 | 50,00 | 34,00 | ,94262 | -,58568 |
| 20 | 20 | 40,00 | 3 | 5 | 20,00 | 30,00 | 20,00 | -,94383 | ,36949 |
| 21 | 21 | 35,00 | 2 | 2 | 15,00 | 20,00 | 15,00 | -1,02910 | -,48167 |
| 22 | 22 | 40,00 | 3 | 3 | 16,00 | 21,00 | 16,00 | -1,11431 | -,09586 |
| 23 | 23 | 30,00 | 2 | 2 | 17,00 | 25,00 | 16,00 | -,87880 | -,40378 |
| 24 | 24 | 40,00 | 2 | 2 | 15,00 | 15,00 | 14,00 | -1,13426 | -,57840 |
| 25 | 25 | 40,00 | 2 | 2 | 18,00 | 15,00 | 13,00 | -1,10971 | -,59103 |
| 26 | 26 | 50,00 | 1 | 2 | 33,00 | 30,00 | 23,00 | ,12138 | -1,51101 |
| 27 | 27 | 60,00 | 2 | 1 | 24,00 | 25,00 | 20,00 | -,16802 | -1,69376 |
| 28 | 28 | 50,00 | 1 | 1 | 29,00 | 30,00 | 25,00 | ,25571 | -1,81407 |
| 29 | 29 | 50,00 | 1 | 1 | 13,00 | 20,00 | 14,00 | -,78472 | -1,43195 |
| 30 | 30 | 50,00 | 1 | 1 | 15,00 | 15,00 | 12,00 | -,92822 | -1,38990 |
| 31 | | | | | | | | | |

**Abbildung 11.33:** Um zwei Variablen erweitere Datendatei FAKTOREN1.SAV im Rahmen der Faktorenanalyse

Die neue Datendatei ist im Übrigen von mir unter *FAKTOREN1A.SAV* abgespeichert worden. Hierbei sind in der Variablenansicht die Variable *betrieb* als »Betrieb« etikettiert und – was wichtiger ist – die Merkmalsausprägungen dieser Variablen von 1, 2, 3, …, 30 in B1, B2, B3, … , B30 verändert worden. Auch sind die beiden Faktoren-Variablen *fac1_1* und *fac1_2* als Finanz-Faktor und Marketing-Faktor

beschriftet worden (siehe hierzu Abbildung 11.34). Diese Formatierungen dienen dazu, die Darstellungen in Tabellen- und Grafikform anschaulicher zu gestalten.

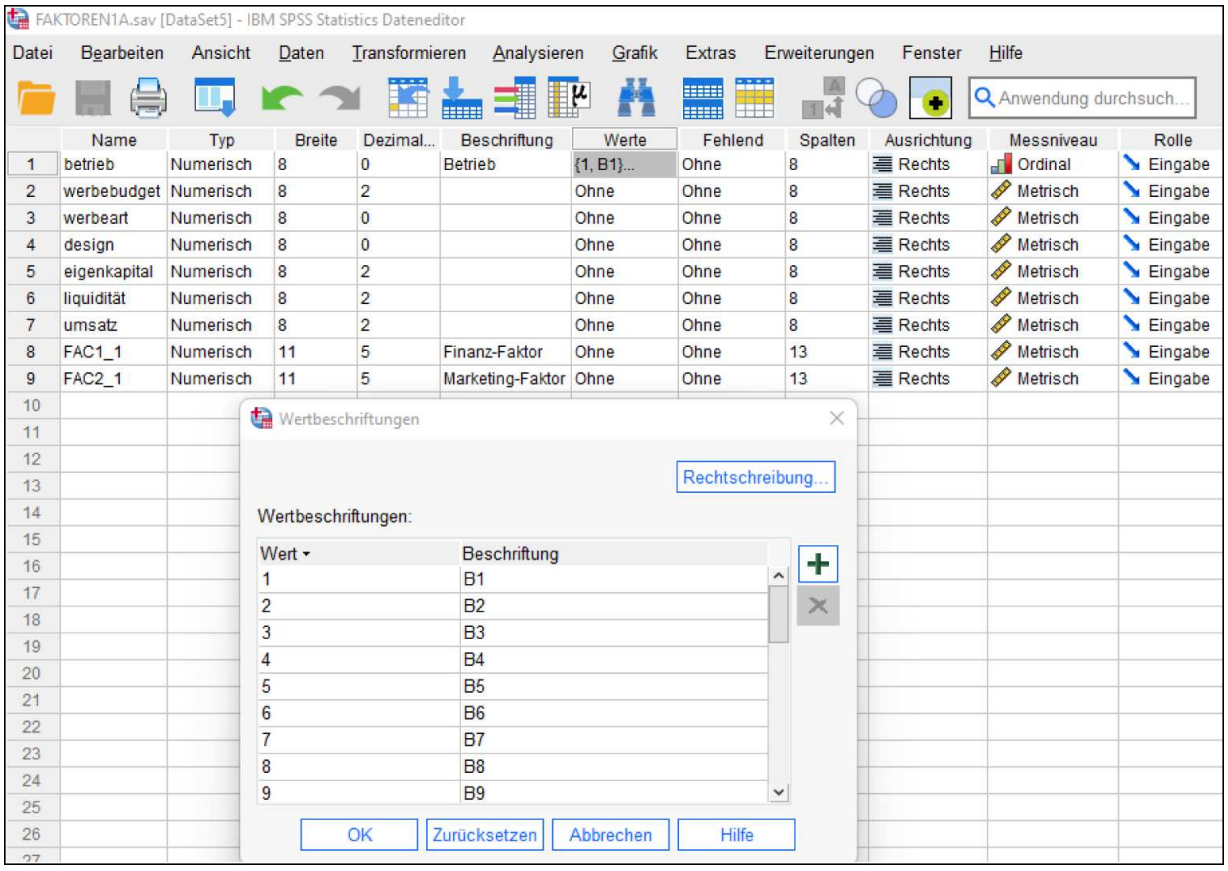



**Abbildung 11.34:** Variablenveränderungen in FAKTOREN1A.SAV

Mit den in Abbildung 11.33 und Abbildung 11.34 dargelegten »Neuerungen« kann jetzt in SPSS über ANALYSIEREN|TABELLEN|BENUTZERDEFINIERTE TABELLEN eine **Tabelle** mit den 30 Betrieben in den Zeilen und den beiden Faktoren in den beiden Spalten erstellt werden. Sie müssen hierzu die Variable *betrieb* in die Rubrik ZEILEN und die Faktoren-Variablen *fac1_1* und *fac1_2* in die Rubrik SPALTEN mit der Maus ziehen (siehe hierzu Abbildung 11.35).

Das Ganze hätten Sie auch via SPSS-Syntax bewerkstelligen können:

```
CTABLES
    /VLABELS VARIABLES=betrieb fac1_1 fac2_1 DISPLAY=LABEL
```

```
/TABLE betrieb [C] BY fac1_1 [S][MEAN] + fac2_1 [S][MEAN]
/CATEGORIES VARIABLES=betrieb ORDER=A KEY=VALUE EMPTY=INCLUDE
/CRITERIA CILEVEL=95.
```

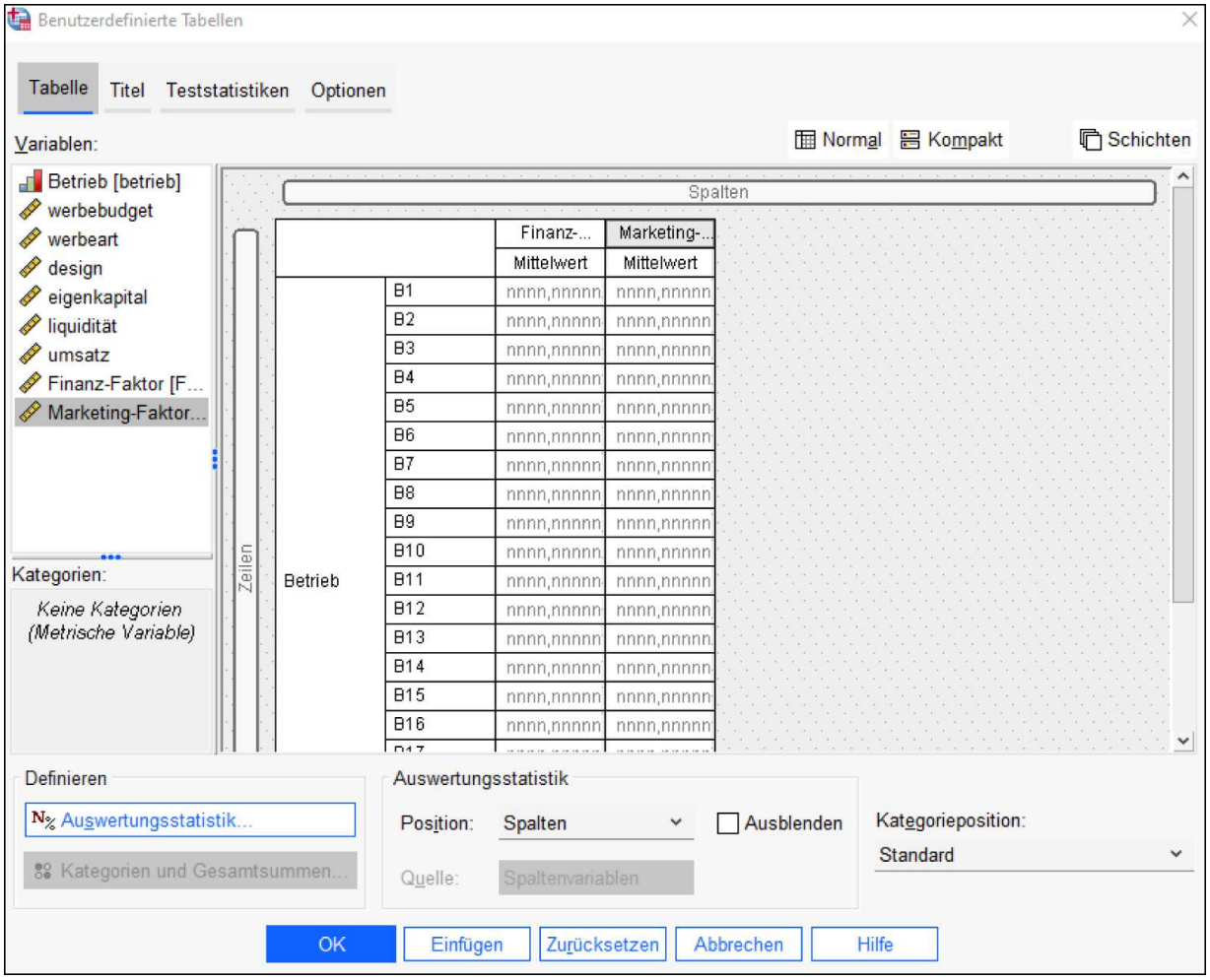



**Abbildung 11.35:** Anfordern und Erstellen einer Tabelle in SPSS (FAKTOREN1A.SAV)

Wie erkennbar, sind für die Zellenwerte der Tabelle die fallbezogenen arithmetischen Mittelwerte der jeweiligen Faktorwerte ausgewählt worden. SPSS zeigt Ihnen mit diesen Vorgaben dann die entsprechende Tabelle im Viewer an (siehe Abbildung 11.36).

Um die vorstehende Tabelle grafisch umzusetzen, klicken Sie bitte als Nächstes auf GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM. Dies tun Sie, um ein **Streudiagramm** zwischen den beiden Faktoren zu kreieren. Für die Fallbezeichnungen wählen Sie

hierbei via BETRIEB letztlich die Betriebsbezeichnungen B1, B2, B3, …, B30 aus (siehe Abbildung 11.37).

Mit den vorstehenden Setzungen erhalten Sie in SPSS das in Abbildung 11.38 dargestellte Streudiagramm zwischen den beiden Faktoren.

Im nächsten Schritt klicken Sie doppelt auf das Bild von Abbildung 11.38, und es öffnet sich der DIAGRAMMEDITOR. In dem **Diagrammeditor** nehmen Sie jetzt bitte einige grafische Anpassungen vor: Sie fügen jeweils beim Wert 0 eine Linie ein. In Abbildung 11.39 ist dieser Vorgang für die Y-Achse dargestellt. Sie müssen lediglich mit der rechten Maustaste auf die Y-Achse klicken und schon öffnet sich nach erneutem Klick auf EIGENSCHAFTSFENSTER ein Fenster EIGENSCHAFTEN, in dem Sie unter METRISCH ein Häkchen bei LINIE AM URSPRUNG ANZEIGEN setzen. Für die X-Achse gehen Sie ganz genauso vor.

| | | Finanz-Faktor Mittelwert | Marketing-Faktor Mittelwert |
|---|---|---|---|
| Betrieb | B1 | 1,93329 | ,28325 |
| | B2 | -,39429 | 1,34525 |
| | B3 | -,22444 | 1,21524 |
| | B4 | ,22161 | 1,52945 |
| | B5 | 1,58388 | 1,15003 |
| | B6 | -,01169 | ,85812 |
| | B7 | -,14164 | ,45455 |
| | B8 | -,68040 | 1,56549 |
| | B9 | -,59669 | 1,51703 |
| | B10 | ,11735 | 1,05971 |
| | B11 | -,43356 | ,20378 |
| | B12 | -,34203 | ,30898 |
| | B13 | 1,50636 | -,37034 |
| | B14 | ,27471 | ,09840 |
| | B15 | 2,31363 | -,34580 |
| | B16 | 1,86960 | -,72243 |

| | | |
|---|---|---|
| B17 | -,87599 | ,37029 |
| B18 | ,65155 | -,31339 |
| B19 | ,94262 | -,58568 |
| B20 | -,94383 | ,36949 |
| B21 | -1,02910 | -,48167 |
| B22 | -1,11431 | -,09586 |
| B23 | -,87880 | -,40378 |
| B24 | -1,13426 | -,57840 |
| B25 | -1,10971 | -,59103 |
| B26 | ,12138 | -1,51101 |
| B27 | -,16802 | -1,69376 |
| B28 | ,25571 | -1,81407 |
| B29 | -,78472 | -1,43195 |
| B30 | -,92822 | -1,38990 |

**Abbildung 11.36:** Fallbezogene arithmetische Faktoren-Mittelwerte als Output einer Faktorenanalyse (FAKTOREN1A.SAV)

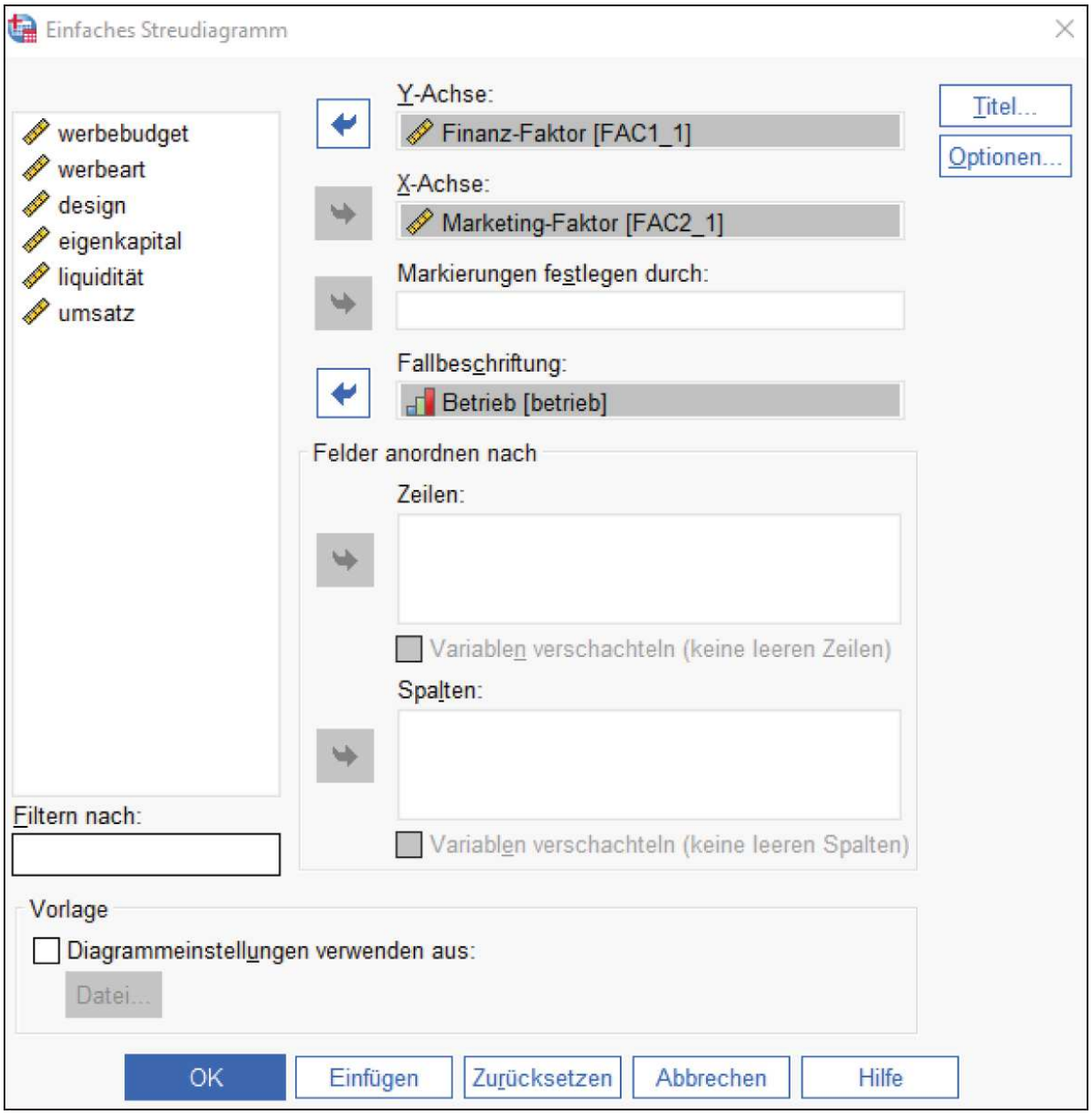



**Abbildung 11.37:** Anforderung eines Streudiagramms zwischen den beiden Faktoren einer Faktorenanalyse (FAKTOREN1A.SAV)

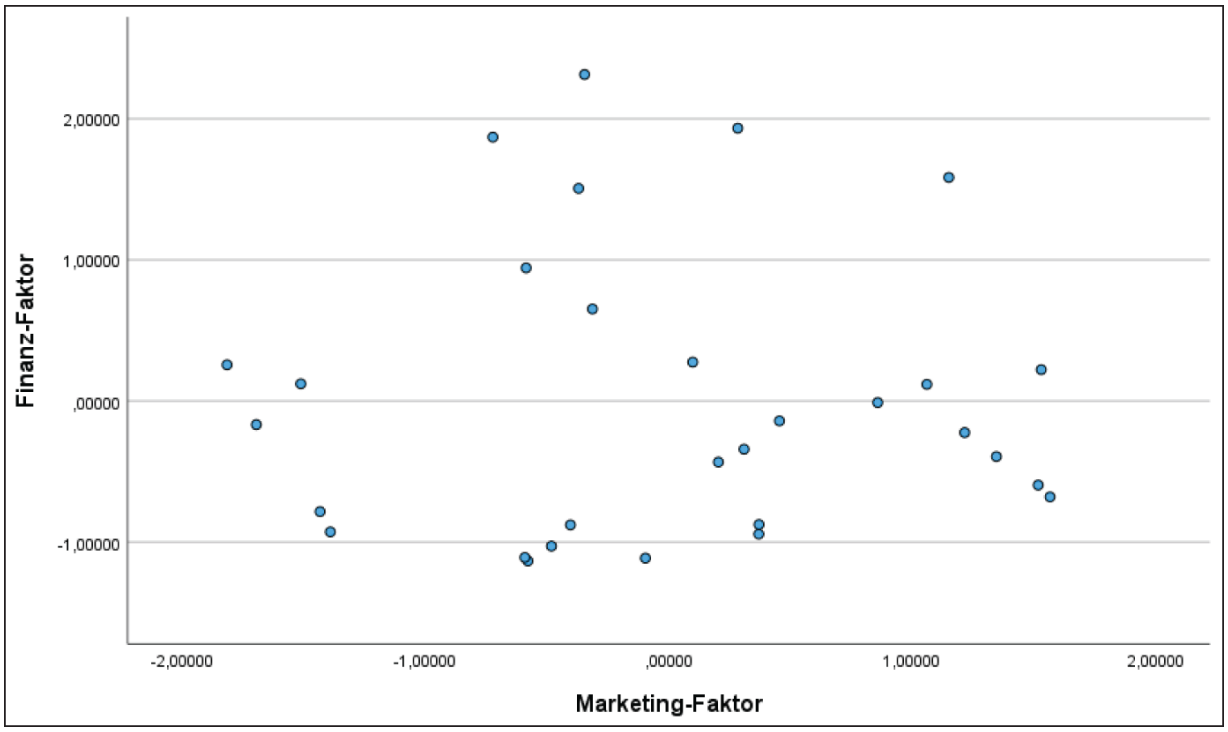



**Abbildung 11.38:** Unbearbeitetes Streudiagramm zwischen den beiden Faktoren (FAKTOREN1A.SAV)

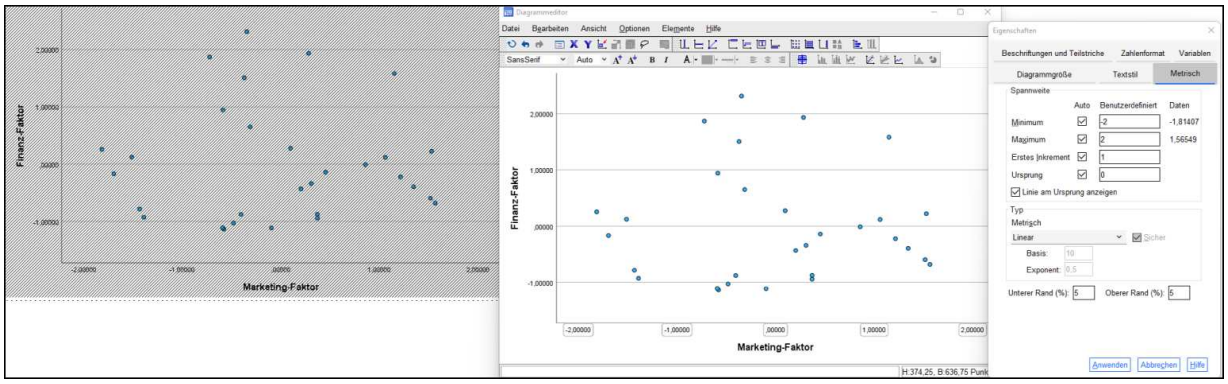



**Abbildung 11.39:** Bearbeitetes Streudiagramm zwischen den beiden Faktoren: Linien einfügen (FAKTOREN1A.SAV)

Nun fügen Sie noch die Datenbeschriftungen B1, B2, B3, …, B30 in das Diagramm ein. Dies tun Sie, indem Sie die Datenpunkte markieren und dann via ELEMENTE|DATENBESCHRIFTUNGEN EINBLENDEN vorgehen (Abbildung 11.40).

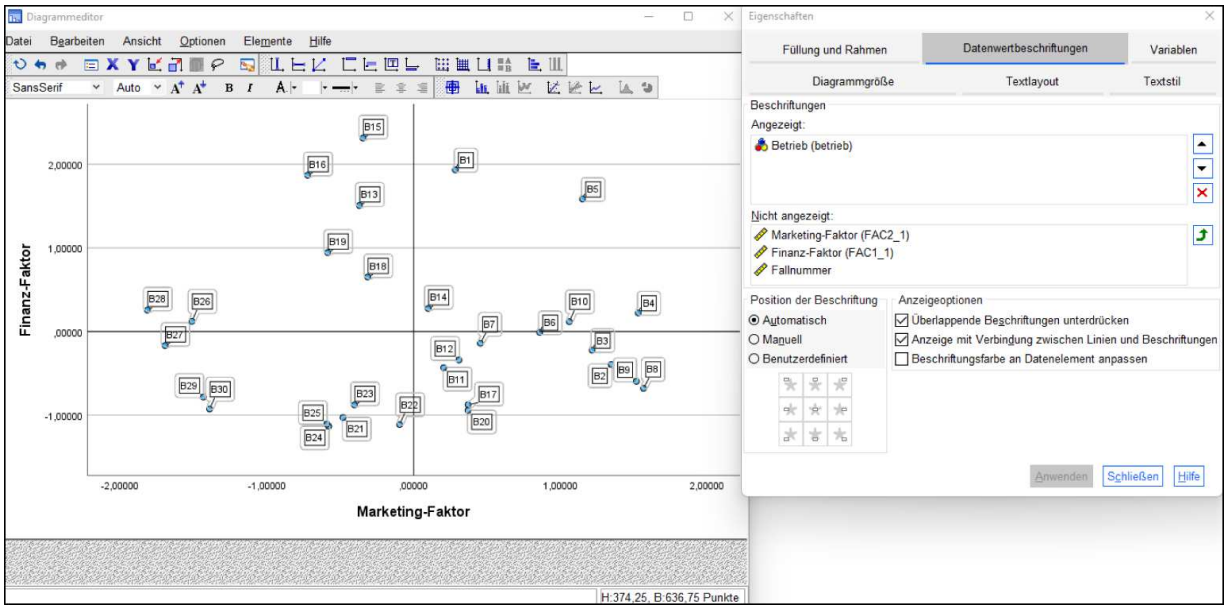



**Abbildung 11.40:** Bearbeitetes Streudiagramm zwischen den beiden Faktoren: Datenbeschriftungen einfügen (FAKTOREN1A.SAV)

Jetzt können Sie das Datenfenster über SCHLIESSEN beenden, und Sie erhalten im Viewer den in Abbildung 11.41 angegebenen Output. Die Betriebe B1, B4, B5, B10 und B14 sind sowohl beim Marketing- als auch beim Finanz-Faktor überdurchschnittlich; demgegenüber sind die Betriebe B21, B22, B23, B24, B25, B27, B29 und B30 bei beiden Faktoren unterdurchschnittlich. Beim Marketing-Faktor überdurchschnittlich, beim Finanz-Faktor hingegen unterdurchschnittlich sind die Betriebe B2, B3, B6, B7, B8, B9, B11, B12, B17 und B20. Schließlich sind die Betriebe B13, B15, B16, B18, B19, B26 und B28 beim Marketing-Faktor unterdurchschnittlich und beim Finanz-Faktor überdurchschnittlich. In diesem Sinne bietet die Faktorenanalyse hier also eine inhaltliche Interpretation über die 30 Betriebe hinweg an. Und außerdem sieht das Ganze jetzt auch ganz nett aus – oder?

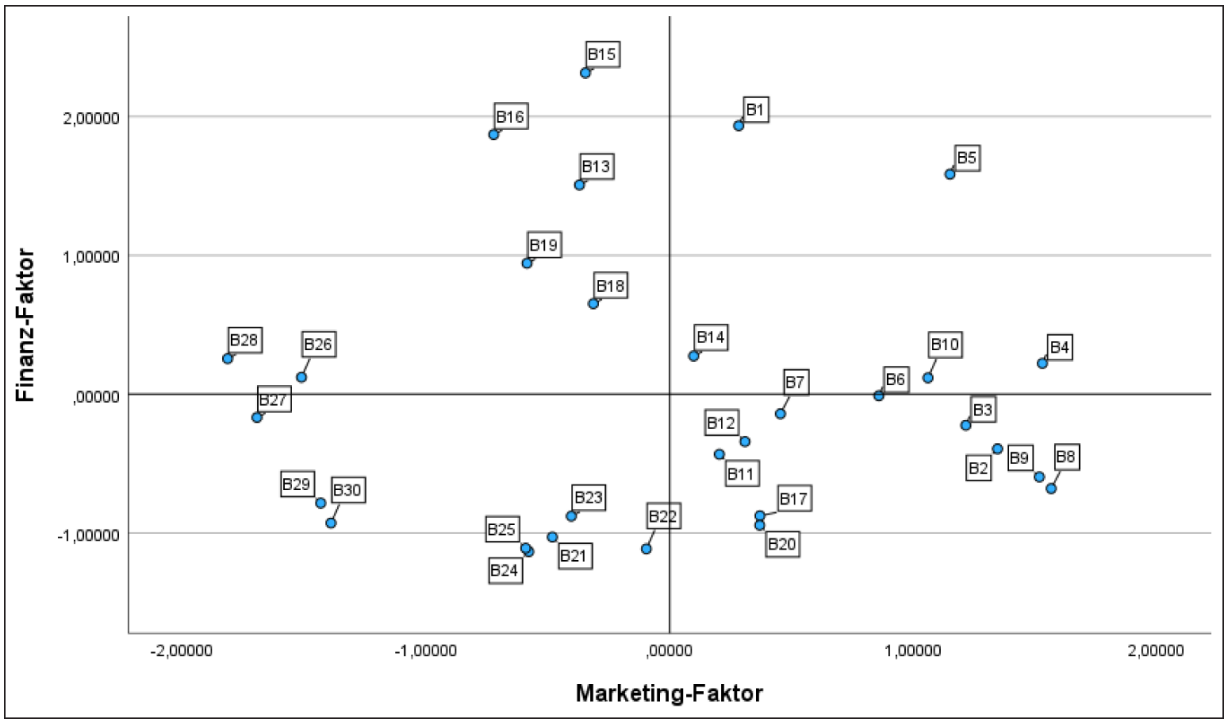



**Abbildung 11.41:** Beschriftetes und in den Nullpunkt verschobenes Streudiagramm zwischen zwei Faktoren im Rahmen der Faktorenanalyse (FAKTOREN1A.SAV)

# Kapitel 12

# Zeit spielt eine Rolle – und überhaupt: Überleben erwünscht!

## IN DIESEM KAPITEL

Zeitreihenanalyse

Sterbetafelanalyse

Kaplan/Meier-Methode

Cox-Regression

Kapitel 12 beschäftigt sich mit Zeitreihen. Zunächst wird auf die Möglichkeiten eingegangen, eine Zeitreihe in ihre einzelnen Komponenten zu zerlegen (vor allem in Bezug auf Trend- und saisonale Komponente). Danach werden im Panel-Längsschnitt Überlebensanalysen von identischen Untersuchungseinheiten (also zum Beispiel von Personen) in Bezug auf den Eintritt eines bestimmten Ereignisses (wie zum Beispiel »Tod«) thematisiert. Im Einzelnen werden die Sterbetafelanalyse, die Kaplan/Meier-Methode und die Cox-Regression angesprochen.

## *Zeitreihenanalyse*

Eine **Zeitreihe** stellt die zeitliche Auflistung der Ausprägungen einer Variablen dar. Bereits in Kapitel 9 sind im Rahmen der Regressionsschätzungen Zeitdaten als unabhängige Größe präsentiert worden. Vielfach beziehen sich Zeitreihen auf unterjährige Daten (Monate, Quartale und dergleichen) mit einer Saisonkomponente.

Entsprechend werden im Folgenden geglättete, das heißt saisonbereinigte, (unterjährige) Zeitreihen beziehungsweise die Möglichkeiten derartiger Glättungen – auf einem recht einfachen methodischen Niveau – behandelt.

# Zeitreihenkomponenten und ihre Verknüpfung

Im Großen und Ganzen kann eine (unterjährige) Zeitreihe in drei Komponenten zerlegt werden:

- in eine **glatte Komponente,** womit im Wesentlichen der zeitliche *Trend* einer Zeitreihe gemeint ist,
- in eine **Saisonkomponente,** die periodisch wiederkehrende Einflüsse auf eine Zeitreihe beschreibt (etwa als Folge klimatischer Bedingungen), und
- in eine **Zufallskomponente**, in der alle durch die glatte und die Saisonkomponente nicht erfassten Einflüsse auf die Zeitreihe erfasst werden.

Ganz grundsätzlich können diese drei Komponenten additiv oder multiplikativ miteinander verknüpft sein.

## Additive Komponentenverknüpfung

Ein Beispiel für die **additive Zeitreihenverknüpfung** finden Sie in Abbildung 12.1. Dort sind für zehn Jahre – also etwa von 2013 bis 2022 – 40 Quartalswerte dargestellt. Mit t = 0 ist dabei das erste Quartal 2013, mit t = 1 das zweite Quartal 2013, mit t = 2 das dritte Quartal 2013, mit t = 3 das vierte Quartal 2013, mit t = 4 das erste Quartal 2014 gemeint und so weiter. Die durchgezogene Gerade in Abbildung 12.1 repräsentiert einen linearen Anstiegstrend der dort behandelten Variablen Y.

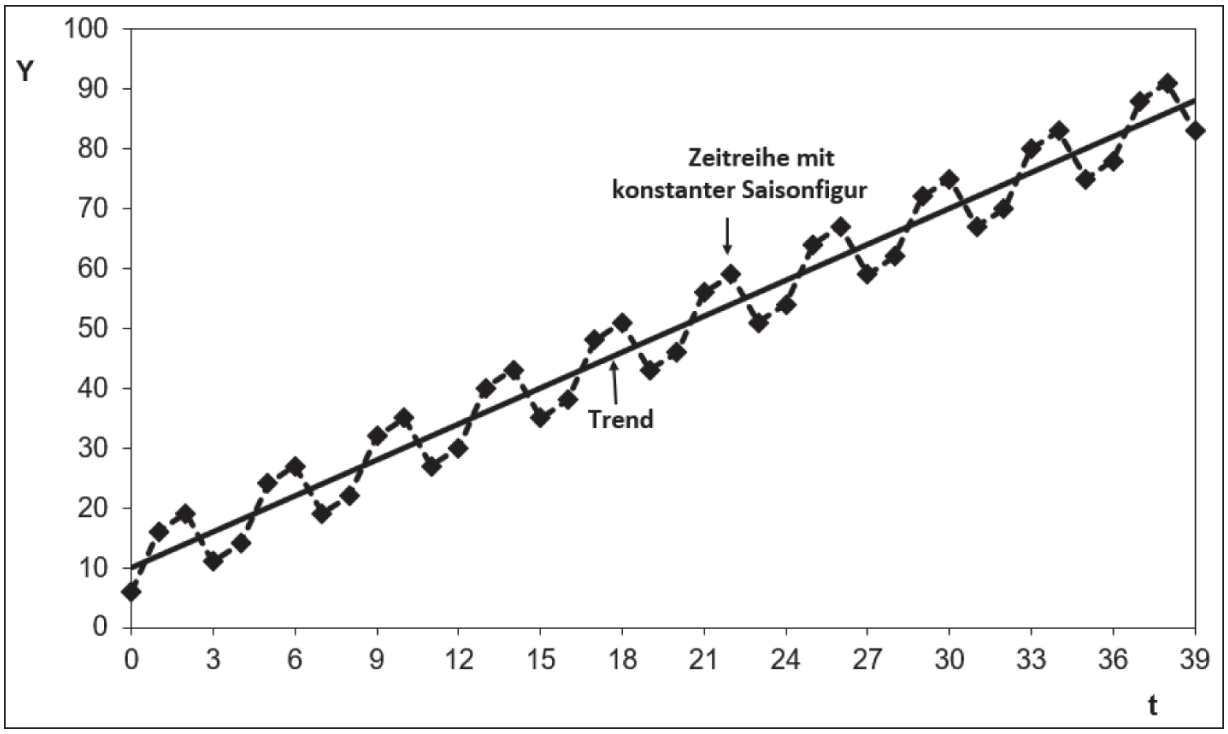



**Abbildung 12.1:** Additive Zeitreihenverknüpfung: Konstante Saisonfigur (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley-VCH, S. 213)

Die Werte dieser Variablen schwanken konstant um den linearen Anstiegstrend herum. Dabei ist der Wert des zweiten Quartals eines Jahrs immer 10 Einheiten höher als der Wert von Y im ersten Quartal desselben Jahrs. Ferner ist der Wert des dritten Quartals eines Jahrs durchgängig um 3 Einheiten höher als im zweiten Quartal desselben Jahrs, und im vierten Quartal eines Jahrs ist der Wert von Y um 8 Einheiten niedriger als im dritten Quartal desselben Jahrs. Im ersten Quartal eines Jahrs schließlich liegt der Wert von Y stets um 3 Einheiten oberhalb des Werts im vierten Quartal des Vorjahrs. Man spricht deshalb von einer **konstanten Saisonfigur**.

## *Multiplikative Komponentenverknüpfung*

Bei der **multiplikativen Zeitreihenverknüpfung** der einzelnen Zeitreihenkomponenten zeigt sich eine **variable Saisonfigur** derart, dass die Abweichungen der Werte einer Variablen Y vom Trend nach oben und nach unten über die Zeit hinweg immer größer werden. Dies ist in Abbildung 12.2 dargestellt. Dort bleiben zwar die Relationen zwischen

den einzelnen Quartalen über die Zeit hinweg konstant, die Niveauunterschiede erhöhen sich aber offenkundig.

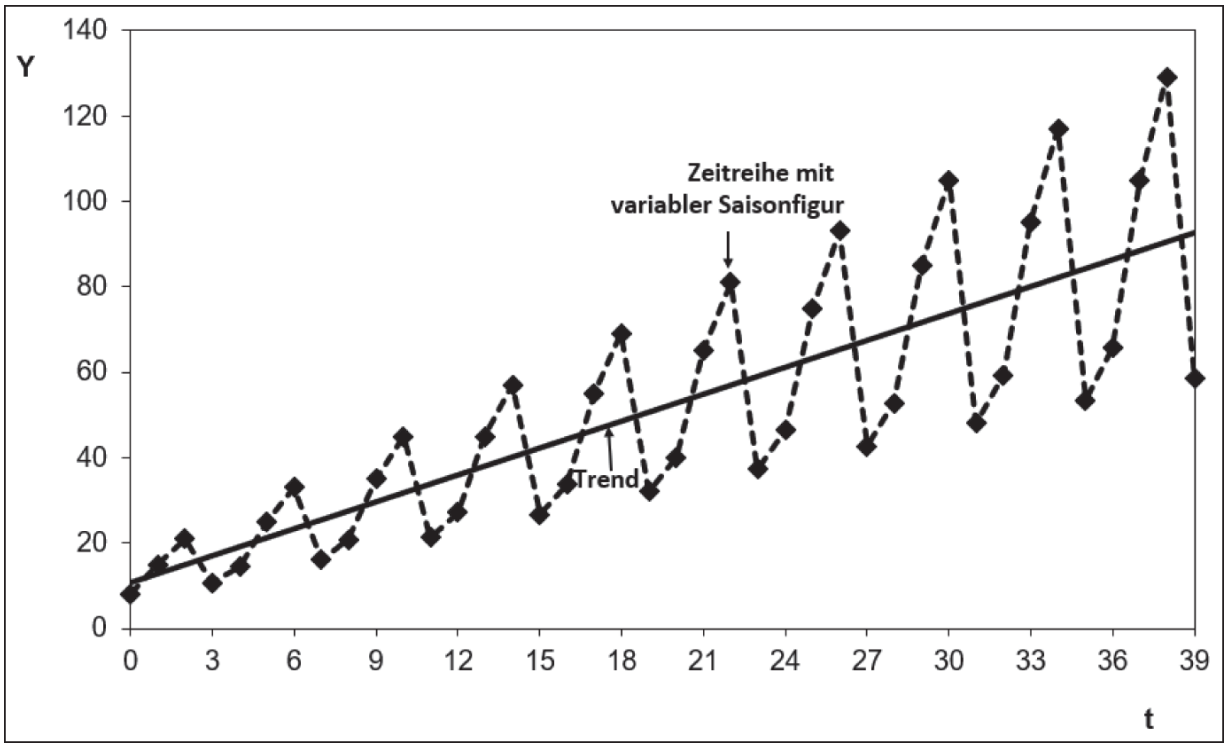



**Abbildung 12.2:** Multiplikative Zeitreihenverknüpfung: Variable Saisonfigur (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley-VCH, S. 214)

# *Trendermittlung*

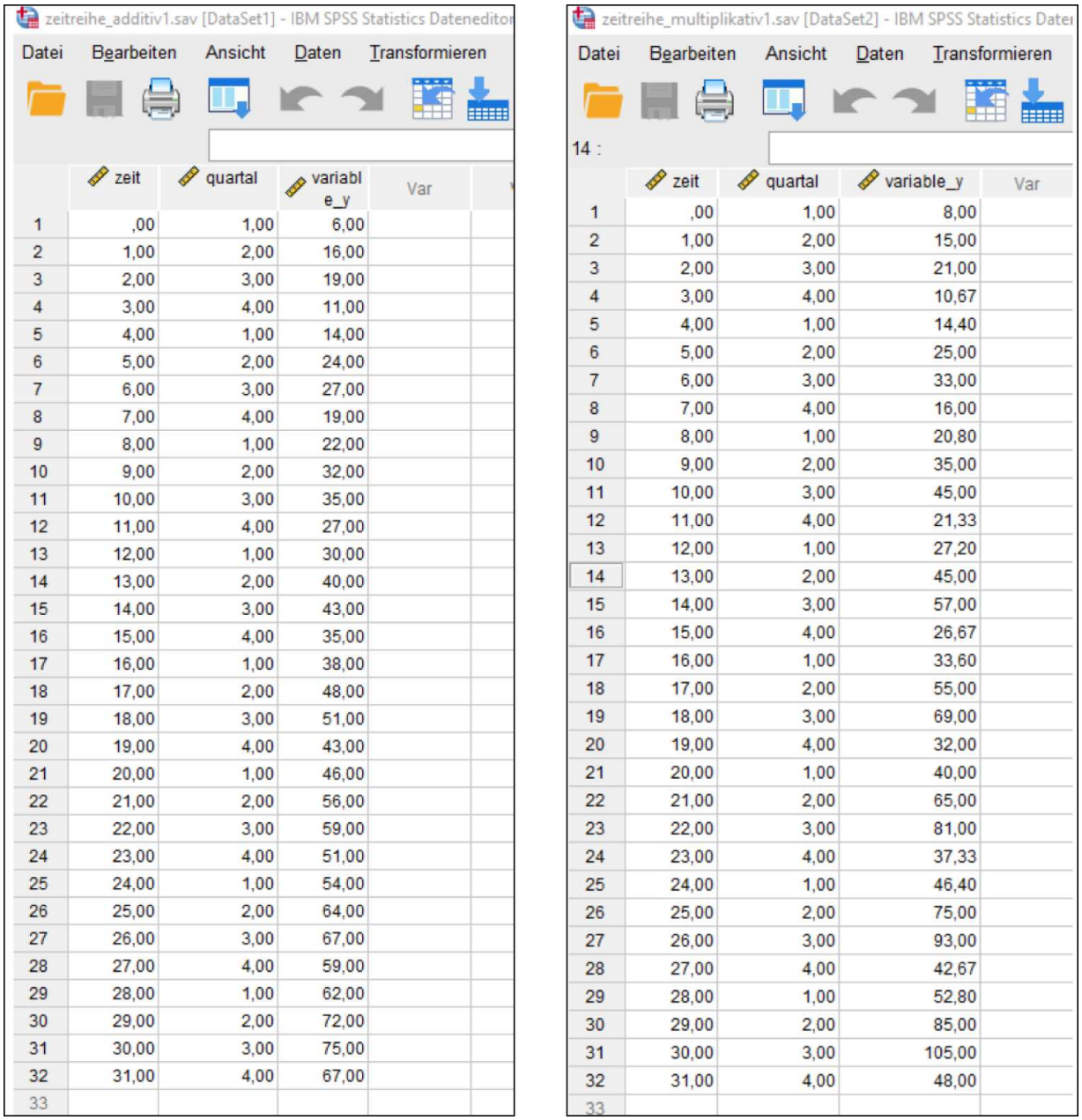
In Abbildung 12.1 und Abbildung 12.2 ist jeweils ein linearer (Anstiegs-)**Trend** dargestellt. Einen Trend kann man prinzipiell auf zweierlei Arten ermitteln, und zwar

- ✔ mittels der bereits aus Kapitel 9 bekannten Kleinst-Quadrate-Methode und
- ✔ mithilfe der Methode gleitender Durchschnitte.

## *Kleinst-Quadrate-Methode*

Bei der **Kleinst-Quadrate-Methode** wird mit dem Instrumentarium der (linearen) Regressionsanalyse ein Zusammenhang zwischen der

Variablen *y* und einer Zeitvariablen geschätzt (wie bereits in Kapitel 9 skizziert). Das entsprechende Vorgehen wird nachfolgend für zwei Datensätze präsentiert: für *ZEITREIHE_ADDITIV1.SAV* und für *ZEITREIHE_MULTIPLIKATIV1.SAV*. Die beiden in Abbildung 12.3 dargestellten Datensätze bestehen aus insgesamt 32 Werten für die Zeitvariable. Beispielsweise kann es sich hierbei um Quartalswerte für die (früheren) Jahre 2010 bis 2017 handeln.



zeitreihe_additiv1.sav [DataSet1] - IBM SPSS Statistics Dateneditor

Datei Bearbeiten Ansicht Daten Transformieren

| | zeit | quartal | variable_y | Var |
|---|---|---|---|---|
| 1 | ,00 | 1,00 | 6,00 | |
| 2 | 1,00 | 2,00 | 16,00 | |
| 3 | 2,00 | 3,00 | 19,00 | |
| 4 | 3,00 | 4,00 | 11,00 | |
| 5 | 4,00 | 1,00 | 14,00 | |
| 6 | 5,00 | 2,00 | 24,00 | |
| 7 | 6,00 | 3,00 | 27,00 | |
| 8 | 7,00 | 4,00 | 19,00 | |
| 9 | 8,00 | 1,00 | 22,00 | |
| 10 | 9,00 | 2,00 | 32,00 | |
| 11 | 10,00 | 3,00 | 35,00 | |
| 12 | 11,00 | 4,00 | 27,00 | |
| 13 | 12,00 | 1,00 | 30,00 | |
| 14 | 13,00 | 2,00 | 40,00 | |
| 15 | 14,00 | 3,00 | 43,00 | |
| 16 | 15,00 | 4,00 | 35,00 | |
| 17 | 16,00 | 1,00 | 38,00 | |
| 18 | 17,00 | 2,00 | 48,00 | |
| 19 | 18,00 | 3,00 | 51,00 | |
| 20 | 19,00 | 4,00 | 43,00 | |
| 21 | 20,00 | 1,00 | 46,00 | |
| 22 | 21,00 | 2,00 | 56,00 | |
| 23 | 22,00 | 3,00 | 59,00 | |
| 24 | 23,00 | 4,00 | 51,00 | |
| 25 | 24,00 | 1,00 | 54,00 | |
| 26 | 25,00 | 2,00 | 64,00 | |
| 27 | 26,00 | 3,00 | 67,00 | |
| 28 | 27,00 | 4,00 | 59,00 | |
| 29 | 28,00 | 1,00 | 62,00 | |
| 30 | 29,00 | 2,00 | 72,00 | |
| 31 | 30,00 | 3,00 | 75,00 | |
| 32 | 31,00 | 4,00 | 67,00 | |
| 33 | | | | |

zeitreihe_multiplikativ1.sav [DataSet2] - IBM SPSS Statistics Datei

Datei Bearbeiten Ansicht Daten Transformieren

14 :

| | zeit | quartal | variable_y | Var |
|---|---|---|---|---|
| 1 | ,00 | 1,00 | 8,00 | |
| 2 | 1,00 | 2,00 | 15,00 | |
| 3 | 2,00 | 3,00 | 21,00 | |
| 4 | 3,00 | 4,00 | 10,67 | |
| 5 | 4,00 | 1,00 | 14,40 | |
| 6 | 5,00 | 2,00 | 25,00 | |
| 7 | 6,00 | 3,00 | 33,00 | |
| 8 | 7,00 | 4,00 | 16,00 | |
| 9 | 8,00 | 1,00 | 20,80 | |
| 10 | 9,00 | 2,00 | 35,00 | |
| 11 | 10,00 | 3,00 | 45,00 | |
| 12 | 11,00 | 4,00 | 21,33 | |
| 13 | 12,00 | 1,00 | 27,20 | |
| 14 | 13,00 | 2,00 | 45,00 | |
| 15 | 14,00 | 3,00 | 57,00 | |
| 16 | 15,00 | 4,00 | 26,67 | |
| 17 | 16,00 | 1,00 | 33,60 | |
| 18 | 17,00 | 2,00 | 55,00 | |
| 19 | 18,00 | 3,00 | 69,00 | |
| 20 | 19,00 | 4,00 | 32,00 | |
| 21 | 20,00 | 1,00 | 40,00 | |
| 22 | 21,00 | 2,00 | 65,00 | |
| 23 | 22,00 | 3,00 | 81,00 | |
| 24 | 23,00 | 4,00 | 37,33 | |
| 25 | 24,00 | 1,00 | 46,40 | |
| 26 | 25,00 | 2,00 | 75,00 | |
| 27 | 26,00 | 3,00 | 93,00 | |
| 28 | 27,00 | 4,00 | 42,67 | |
| 29 | 28,00 | 1,00 | 52,80 | |
| 30 | 29,00 | 2,00 | 85,00 | |
| 31 | 30,00 | 3,00 | 105,00 | |
| 32 | 31,00 | 4,00 | 48,00 | |
| 33 | | | | |

**Abbildung 12.3:** Datendateien *ZEITREIHE_ADDITIV1.SAV* (links) und *ZEITREIHE_MULTIPLIKATIV1.SAV* (rechts)

Im jeweiligen Datensatz wird nun jeweils für die Variable *variable_y* eine lineare Regression mit *zeit* als unabhängiger Variable geschätzt. Dies bewerkstelligen Sie im SPSS-Menü über ANALYSIEREN|REGRESSION|LINEAR. Alternativ können Sie als SPSS-Syntax schreiben:

```
REGRESSION
    /MISSING LISTWISE
    /STATISTICS COEFF OUTS R ANOVA
    /CRITERIA=PIN(.05) POUT(.10)
    /NOORIGIN
    /DEPENDENT variable_y
    /METHOD=ENTER zeit.
```

Für die beiden Datensätze erhalten Sie in SPSS den nachfolgenden Output:

für *ZEITREIHE_ADDITIV1.SAV*:

> *variable_y* = 10,045 + 1,9570 · *zeit* (korrigiertes Bestimmtheitsmaß: 0,941),

für *ZEITREIHE_MULTIPLIKATIV1.SAV*:

> *variable_y* = 10,761 + 2,092 · *zeit* (korrigiertes Bestimmtheitsmaß: 0,599),

jeweils mit *zeit* = 0, 1, 2, …, 31.

In beiden Beispielen ergibt sich also ein linearer Anstiegstrend.

## *Gleitende Durchschnitte*

Alternativ zur Anwendung der Kleinst-Quadrate-Methode können Sie den Trend einer Zeitreihe über die Methode der **gleitenden Durchschnitte** bestimmen. Bei diesem Verfahren wird für eine bestimmte Anzahl jeweils aufeinanderfolgender Werte einer Variablen Y ein Durchschnittswert berechnet.

Bei einer ungeradzahligen Anzahl aufeinanderfolgender Werte von Y wird jeweils durch die Anzahl der aufeinanderfolgenden Werte dividiert, also zum Beispiel bei jeweils drei aufeinanderfolgenden Werten von Y

werden zunächst die ersten drei Werte von Y addiert und durch 3 dividiert, dann werden der zweite bis vierte Wert von Y addiert und durch 3 dividiert und so weiter.

Demgegenüber werden bei einer geradzahligen Anzahl aufeinanderfolgender Werte von Y der erste und der letzte Wert jeweils halbiert. Für einen gleitenden 4er-Durchschnitt zum Beispiel müssen Sie insgesamt fünf Werte zugrunde legen. Bei dann jeweils fünf aufeinanderfolgenden Werten von Y werden also der erste und der fünfte Wert von Y jeweils halbiert; anschließend wird die Summe der fünf Werte von Y jeweils durch 4 dividiert.

Für die Datendatei *ZEITREIHE_ADDITIV1.SAV* wird das Verfahren der gleitenden Durchschnitte am Beispiel eines gleitenden 4er-Durchschnitts nachfolgend demonstriert. In diesem Fall werden zunächst die ersten fünf Werte von *variable_y* folgendermaßen addiert und dann durch 4 dividiert: $\frac{\frac{6}{2}+16+19+11+\frac{14}{2}}{4}=14.$ Dieser Wert von 14 wird dem dritten Quartal des Jahrs 2010 zugewiesen. Entsprechend wird der gleitende 4er-Durchschnitt für das vierte Quartal des Jahrs 2010 berechnet: $\frac{\frac{16}{2}+19+11+14+\frac{24}{2}}{4}=16.$ Offenkundig gehen bei der Berechnung von gleitenden Durchschnitten am Anfang der Zeitreihe – ebenso wie am Ende der Zeitreihe – Werte »verloren«. Bei einem gleitenden 4er-Durchschnitt gehen für die beiden ersten und für die beiden letzten Perioden (hier: Quartale) Werte »verloren«.

Die betreffenden Berechnungen in SPSS werden über den Menüpunkt TRANSFORMIEREN|ZEITREIHEN ERSTELLEN in die Wege geleitet. In dem sich öffnenden Dialogfenster ist die Reihenfolge der Eingaben wichtig. Zunächst wählen Sie bitte FUNKTION aus. Im vorliegenden Fall geht es ja um gleitende Durchschnitte. Deshalb wählen Sie bitte ZENTRIERTER GLEITENDER DURCHSCHNITT aus und bestätigen diese Auswahl für die vorgegebene Variable *variab_1* durch Anklicken der ÄNDERN-Taste. Unter dieser Angabe befindet sich das Eingabefeld SPANNE. Hierin wird die Art der gleitenden Durchschnittsbildung festgelegt. Da ein 4er-Durchschnitt berechnet

werden soll, muss hier die Zahl »4« stehen. Hiermit haben Sie also angegeben, dass Sie sich einen gleitenden 4er-Durchschnitt berechnen lassen möchten, und Sie müssen natürlich noch die Variable auswählen, für die diese (gleitenden) Durchschnitte ausgerechnet werden sollen: Im vorliegenden Fall ist dies die Variable *variable_y*. SPSS legt die gleitenden 4er-Durchschnitte für diese Variable in einer neuen Variablen ab, nämlich – wie aus Abbildung 12.4 ersichtlich ist – in der Variablen *variab_1*.

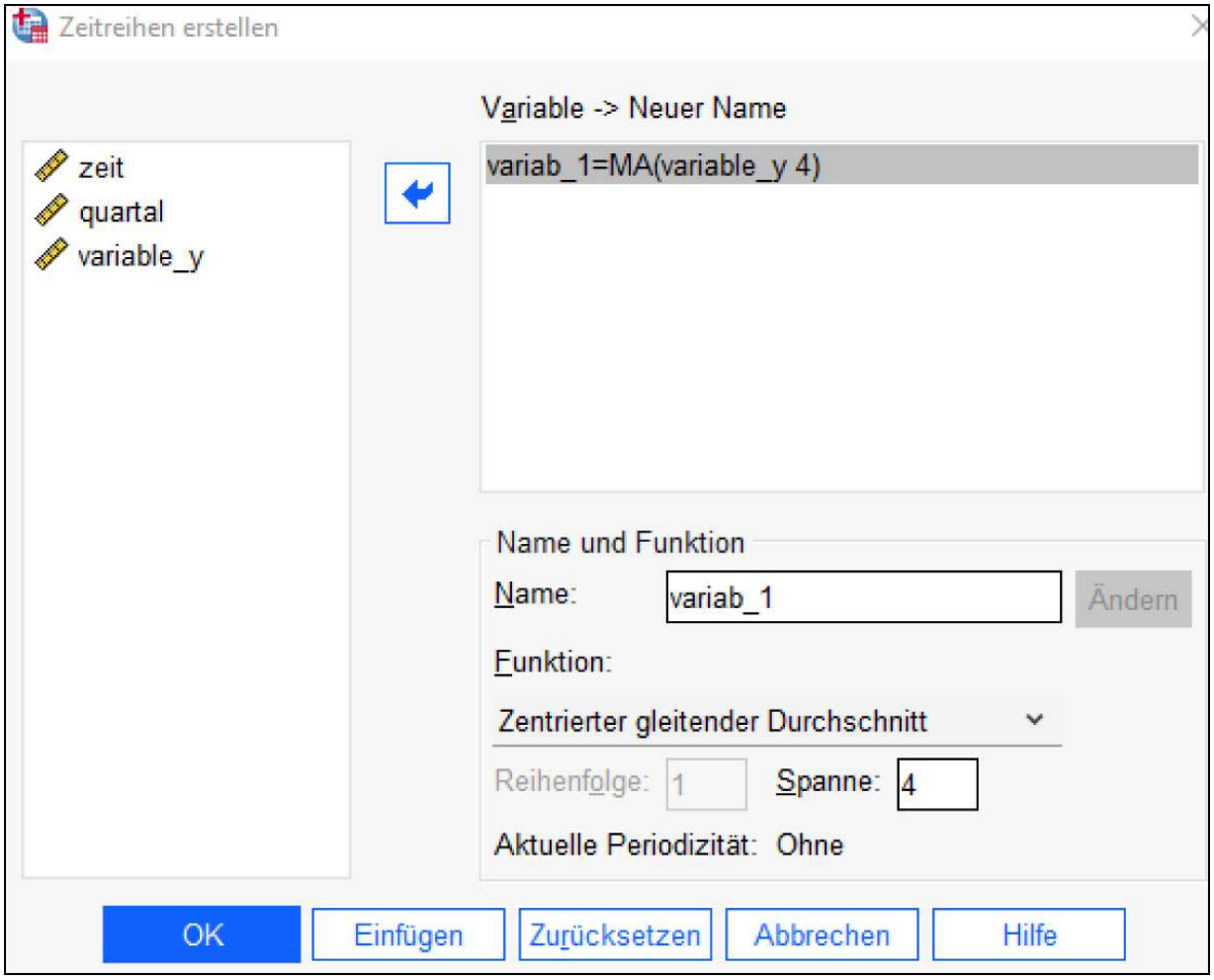
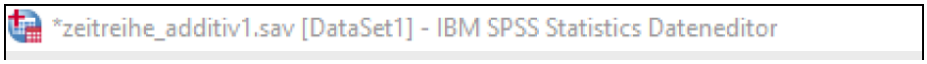



**Abbildung 12.4:** Festlegung der Berechnung von gleitenden 4er-Durchschnitten für *variable_y* in *ZEITREIHE_ADDITIV1.SAV*

Mit diesen Vorgaben gibt SPSS in der neuen Variablen *variab_1* die gleitenden 4er-Durchschnitte aus (siehe Abbildung 12.5).



26 :

| | zeit | quartal | variable_y | variab_1 | Var | |
|---|---|---|---|---|---|---|
| 1 | ,00 | 1,00 | 6,00 | . | | |
| 2 | 1,00 | 2,00 | 16,00 | . | | |
| 3 | 2,00 | 3,00 | 19,00 | 14,00 | | |
| 4 | 3,00 | 4,00 | 11,00 | 16,00 | | |
| 5 | 4,00 | 1,00 | 14,00 | 18,00 | | |
| 6 | 5,00 | 2,00 | 24,00 | 20,00 | | |
| 7 | 6,00 | 3,00 | 27,00 | 22,00 | | |
| 8 | 7,00 | 4,00 | 19,00 | 24,00 | | |
| 9 | 8,00 | 1,00 | 22,00 | 26,00 | | |
| 10 | 9,00 | 2,00 | 32,00 | 28,00 | | |
| 11 | 10,00 | 3,00 | 35,00 | 30,00 | | |
| 12 | 11,00 | 4,00 | 27,00 | 32,00 | | |
| 13 | 12,00 | 1,00 | 30,00 | 34,00 | | |
| 14 | 13,00 | 2,00 | 40,00 | 36,00 | | |
| 15 | 14,00 | 3,00 | 43,00 | 38,00 | | |
| 16 | 15,00 | 4,00 | 35,00 | 40,00 | | |
| 17 | 16,00 | 1,00 | 38,00 | 42,00 | | |
| 18 | 17,00 | 2,00 | 48,00 | 44,00 | | |
| 19 | 18,00 | 3,00 | 51,00 | 46,00 | | |
| 20 | 19,00 | 4,00 | 43,00 | 48,00 | | |
| 21 | 20,00 | 1,00 | 46,00 | 50,00 | | |
| 22 | 21,00 | 2,00 | 56,00 | 52,00 | | |
| 23 | 22,00 | 3,00 | 59,00 | 54,00 | | |
| 24 | 23,00 | 4,00 | 51,00 | 56,00 | | |
| 25 | 24,00 | 1,00 | 54,00 | 58,00 | | |
| 26 | 25,00 | 2,00 | 64,00 | 60,00 | | |
| 27 | 26,00 | 3,00 | 67,00 | 62,00 | | |
| 28 | 27,00 | 4,00 | 59,00 | 64,00 | | |
| 29 | 28,00 | 1,00 | 62,00 | 66,00 | | |
| 30 | 29,00 | 2,00 | 72,00 | 68,00 | | |
| 31 | 30,00 | 3,00 | 75,00 | . | | |
| 32 | 31,00 | 4,00 | 67,00 | . | | |
| 33 | | | | | | |
| 34 | | | | | | |

**Abbildung 12.5:** Berechnung von gleitenden 4er-Durchschnitten für *variable_y* in der neuen Variablen *variab_1* (*ZEITREIHE_ADDITIV1.SAV*)

Die Glättung der Zeitreihe durch die Bildung gleitender 4er-Durchschnitte ist sehr schön aus Abbildung 12.6 zu erkennen, die in SPSS als Scatterplot zwischen *variab_1* und *zeit* über GRAFIK|STREU-/PUNKTDIAGRAMM|DEFINIEREN:EINFACHES STREUDIAGRAMM erzeugt worden ist; die entsprechende Syntax ist:

```
GRAPH
    /SCATTERPLOT(BIVAR)=zeit WITH variab_1
    /MISSING=LISTWISE.
```

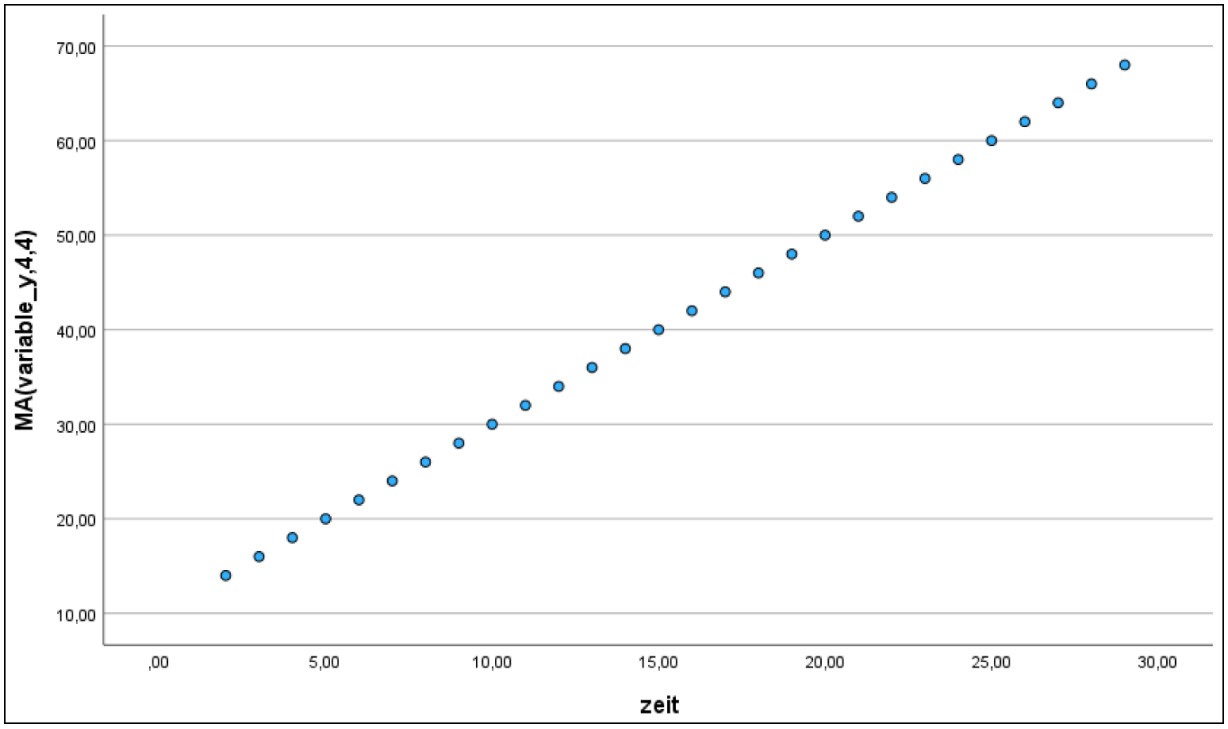



**Abbildung 12.6:** Glättung der Zeitreihe von *variable_y* in *ZEITREIHE_ADDITIV1.SAV* mittels gleitender 4er-Durchschnitte

# *Saisonkomponente*

Neben dem Trend können für eine Zeitreihe auch noch die saisonalen Einflüsse berechnet werden. Genauer ausgedrückt ist es hier das Ziel, eine **Saisonfigur** zu berechnen und über die betreffenden Saisonfigur-Werte dann die Ursprungszeitreihe um diese Saisoneinflüsse zu

bereinigen, sodass die einzelnen Zeitpunkte der Zeitreihe aussagekräftiger miteinander zu vergleichen sind. In den einfachsten Fällen unterscheidet man eine konstante Saisonfigur (siehe Abbildung 12.1) von einer variablen Saisonfigur (siehe Abbildung 12.2). Je nachdem, welche dieser beiden Saisonfiguren unterstellt wird, führt man die Berechnungen etwas unterschiedlich durch. Methodisch sind die betreffenden Berechnungen dem **Phasendurchschnittsverfahren** zugeordnet.

## *Phasendurchschnittsverfahren mit konstanter Saisonfigur*

Das **Phasendurchschnittsverfahren mit konstanter Saisonfigur** wird bei *additiver* Verknüpfung der Zeitreihenelemente angewandt.

Dieses Verfahren wird nachfolgend für die Datendatei *ZEITREIHE_ADDITIV1.SAV* beziehungsweise für die Variable *variable_y* in dieser Datendatei illustriert. Für die dortigen Quartale werden aus den Differenzen zwischen Ursprungs- und Trendwerten arithmetische Mittelwerte berechnet, die – nach einer Normierung auf eine mittlere Abweichungssumme von 0 über alle Quartale hinweg – als *Saisonziffern* für die Quartale interpretiert werden. Indem diese quartalsbezogenen Saisonziffern von den jeweiligen Ursprungswerten abgezogen werden, erhält man die *saisonbereinigte Zeitreihe*.

Um die entsprechenden Berechnungen in SPSS überhaupt durchführen zu können, sind ein paar Vorarbeiten notwendig. Zunächst müssen Sie SPSS mitteilen, dass sich Ihre Zeitdaten auf Quartale beziehen. Dies tun Sie dadurch, dass Sie eine sogenannte Datumskomponente definieren, und zwar über DATEN|DATUM UND UHRZEIT DEFINIEREN.

Wie Abbildung 12.7 zu entnehmen ist, wird in der Dialogbox DATUM DEFINIEREN links angegeben, dass es sich bei der Zeitreihe in der Datendatei *ZEITREIHE_ADDITIV1.SAV* um »Jahre, Quartale« handelt. Des Weiteren geben Sie auf der rechten Seite des Dialogfelds DATUM DEFINIEREN für den ersten Fall an, dass dieser sich auf das Jahr 2010 und das erste Quartal dieses Jahrs bezieht.

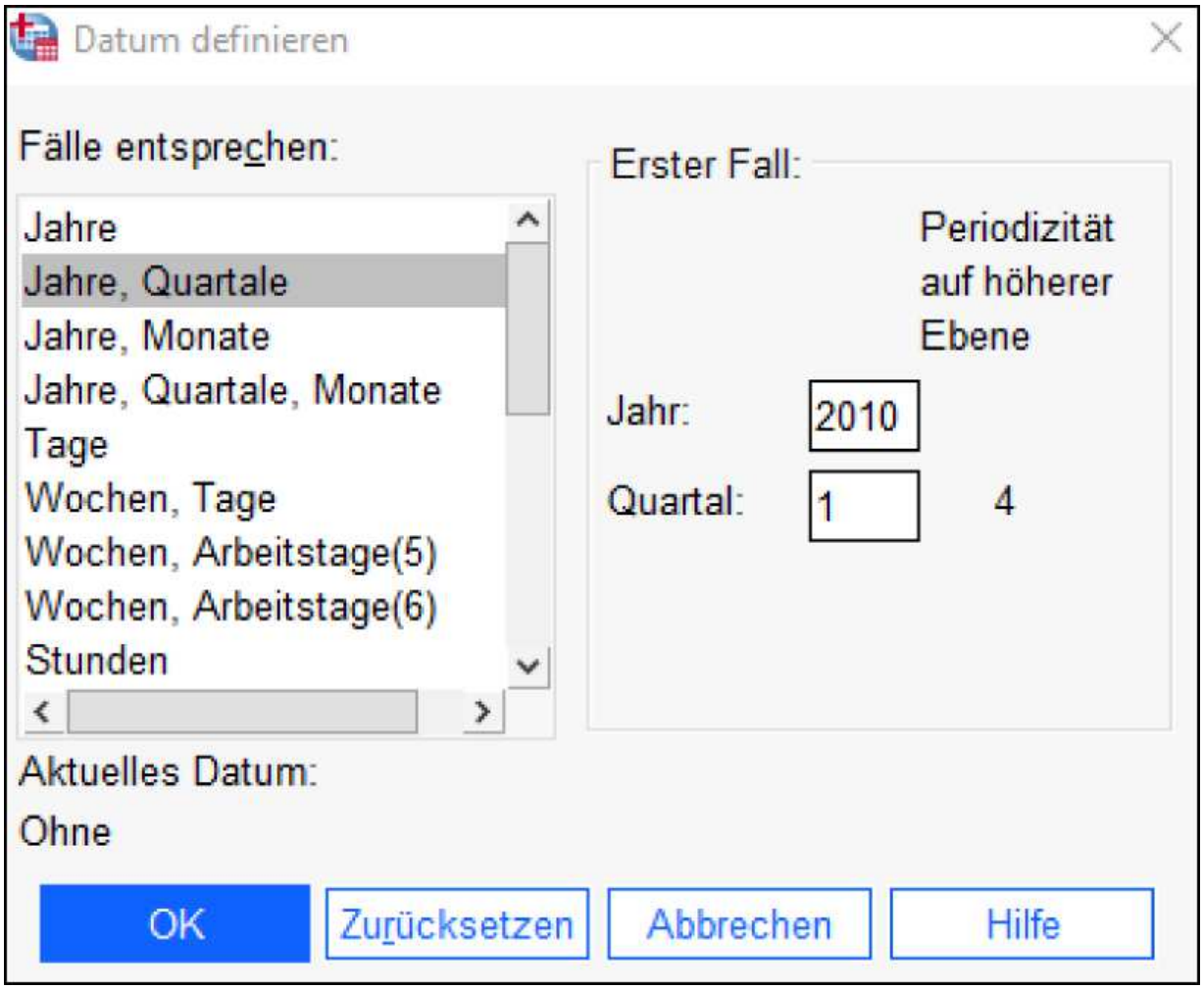
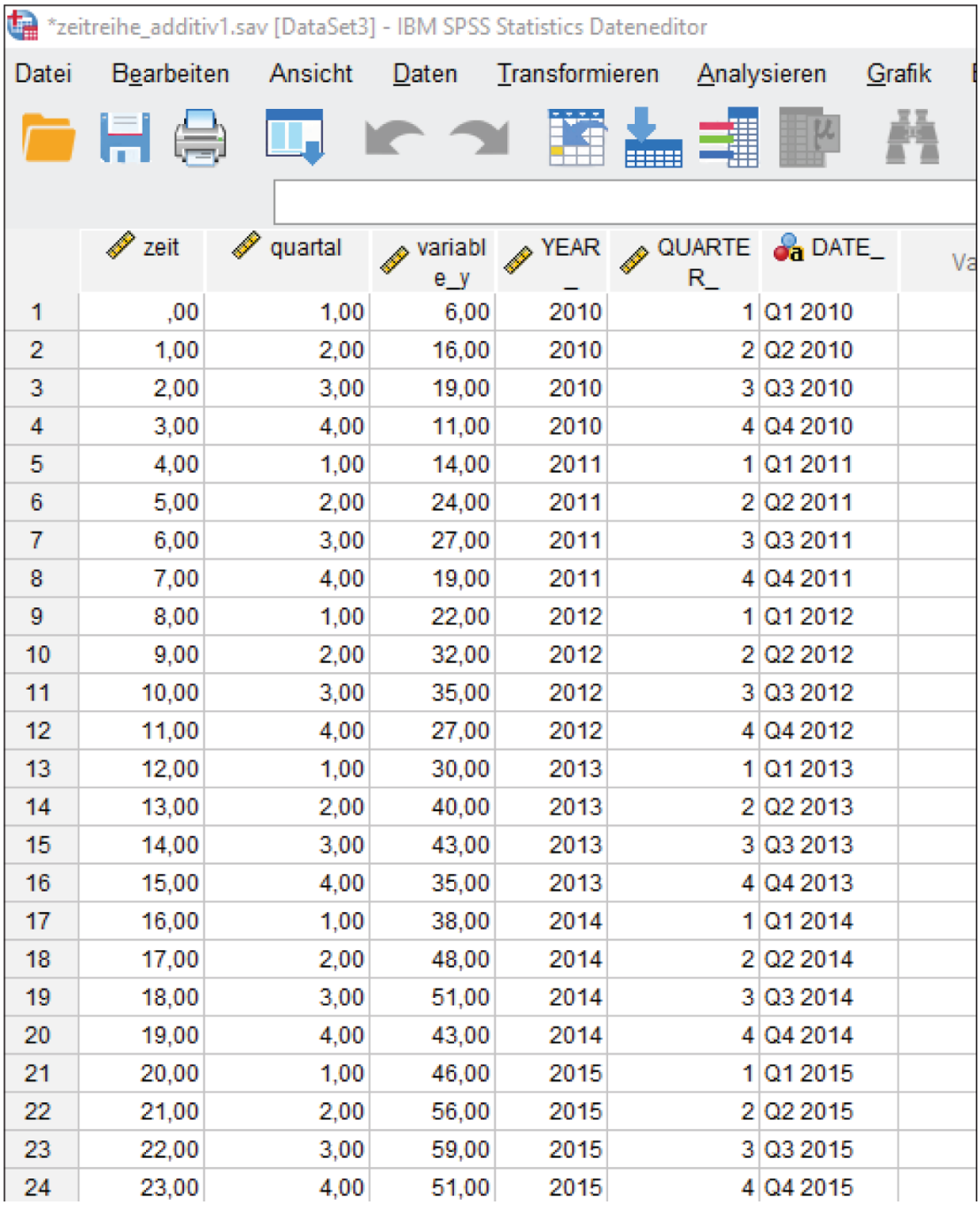



**Abbildung 12.7:** Definition der relevanten Datumskomponente für die Datendatei *ZEITREIHE_ADDITIV1.SAV*

SPSS fügt auf der Grundlage dieser Angaben drei neue Variablen direkt in Ihren Datensatz ein: *year_*, *quarter_* und *date_*. Die Inhalte dieser Variablen spiegeln Ihre gemäß Abbildung 12.7 getroffenen Datumsangaben wider. In *year_* beziehungsweise *quarter_* finden Sie eine Angabe darüber, um welches Jahr beziehungsweise um welches Quartal eines Jahrs es sich handelt, und in *date_* sind zusätzlich diese beiden Informationen zusammen angegeben (siehe Abbildung 12.8).

So, nach diesen vielleicht etwas lästigen, aber leider notwendigen Vorarbeiten können Sie nun endlich mit der Berechnung Ihrer Saisonziffern beginnen! Hierzu wählen Sie bitte im SPSS-Menü ANALYSIEREN|VORHERSAGE|SAISONALE ZERLEGUNG aus. In dem betreffenden Dialogfenster tragen Sie die zu analysierende Variable

unter VARIABLE(N) ein (hier: variable_y) und klicken bei MODELLTYP auf ADDITIV, da Sie ja für den Zeitreihen-Modelltyp einer additiven Komponentenverknüpfung Saisonziffern berechnen möchten. SPSS weist Sie nach Ihren mit OK beendeten Eingaben darauf hin, dass vier weitere Variablen direkt in Ihre Datendatei eingefügt werden. Bitte klicken Sie dann an dieser Stelle nochmals auf OK (siehe Abbildung 12.9).




| | zeit | quartal | variable_y | YEAR_ | QUARTER_ | DATE_ |
|---|---|---|---|---|---|---|
| 1 | ,00 | 1,00 | 6,00 | 2010 | 1 | Q1 2010 |
| 2 | 1,00 | 2,00 | 16,00 | 2010 | 2 | Q2 2010 |
| 3 | 2,00 | 3,00 | 19,00 | 2010 | 3 | Q3 2010 |
| 4 | 3,00 | 4,00 | 11,00 | 2010 | 4 | Q4 2010 |
| 5 | 4,00 | 1,00 | 14,00 | 2011 | 1 | Q1 2011 |
| 6 | 5,00 | 2,00 | 24,00 | 2011 | 2 | Q2 2011 |
| 7 | 6,00 | 3,00 | 27,00 | 2011 | 3 | Q3 2011 |
| 8 | 7,00 | 4,00 | 19,00 | 2011 | 4 | Q4 2011 |
| 9 | 8,00 | 1,00 | 22,00 | 2012 | 1 | Q1 2012 |
| 10 | 9,00 | 2,00 | 32,00 | 2012 | 2 | Q2 2012 |
| 11 | 10,00 | 3,00 | 35,00 | 2012 | 3 | Q3 2012 |
| 12 | 11,00 | 4,00 | 27,00 | 2012 | 4 | Q4 2012 |
| 13 | 12,00 | 1,00 | 30,00 | 2013 | 1 | Q1 2013 |
| 14 | 13,00 | 2,00 | 40,00 | 2013 | 2 | Q2 2013 |
| 15 | 14,00 | 3,00 | 43,00 | 2013 | 3 | Q3 2013 |
| 16 | 15,00 | 4,00 | 35,00 | 2013 | 4 | Q4 2013 |
| 17 | 16,00 | 1,00 | 38,00 | 2014 | 1 | Q1 2014 |
| 18 | 17,00 | 2,00 | 48,00 | 2014 | 2 | Q2 2014 |
| 19 | 18,00 | 3,00 | 51,00 | 2014 | 3 | Q3 2014 |
| 20 | 19,00 | 4,00 | 43,00 | 2014 | 4 | Q4 2014 |
| 21 | 20,00 | 1,00 | 46,00 | 2015 | 1 | Q1 2015 |
| 22 | 21,00 | 2,00 | 56,00 | 2015 | 2 | Q2 2015 |
| 23 | 22,00 | 3,00 | 59,00 | 2015 | 3 | Q3 2015 |
| 24 | 23,00 | 4,00 | 51,00 | 2015 | 4 | Q4 2015 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 25 | 24,00 | 1,00 | 54,00 | 2016 | 1 | Q1 2016 | |
| 26 | 25,00 | 2,00 | 64,00 | 2016 | 2 | Q2 2016 | |
| 27 | 26,00 | 3,00 | 67,00 | 2016 | 3 | Q3 2016 | |
| 28 | 27,00 | 4,00 | 59,00 | 2016 | 4 | Q4 2016 | |
| 29 | 28,00 | 1,00 | 62,00 | 2017 | 1 | Q1 2017 | |
| 30 | 29,00 | 2,00 | 72,00 | 2017 | 2 | Q2 2017 | |
| 31 | 30,00 | 3,00 | 75,00 | 2017 | 3 | Q3 2017 | |
| 32 | 31,00 | 4,00 | 67,00 | 2017 | 4 | Q4 2017 | |
| 33 | | | | | | | |

**Abbildung 12.8:** Von SPSS in die Datendatei *ZEITREIHE_ADDITIV1.SAV* eingefügte Datumsvariablen *year_*, *quarter_* und *date_*

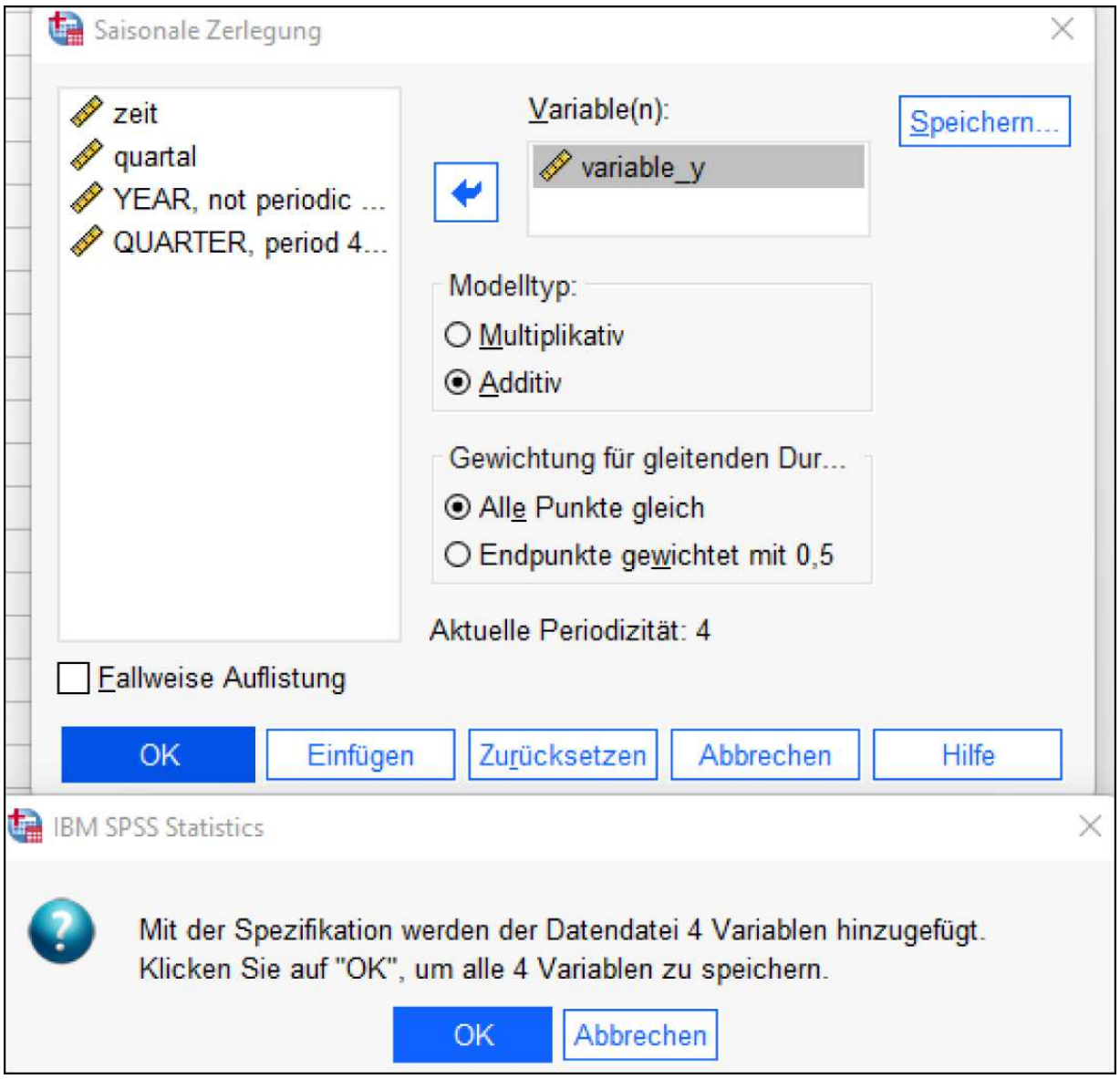


**Abbildung 12.9:** Anfordern einer additiven Zeitreihen-Komponentenzerlegung (*ZEITREIHE_ADDITIV1.SAV*)

Anstelle des Menübezugs können Sie auch folgende SPSS-Befehlsfolge eingeben:

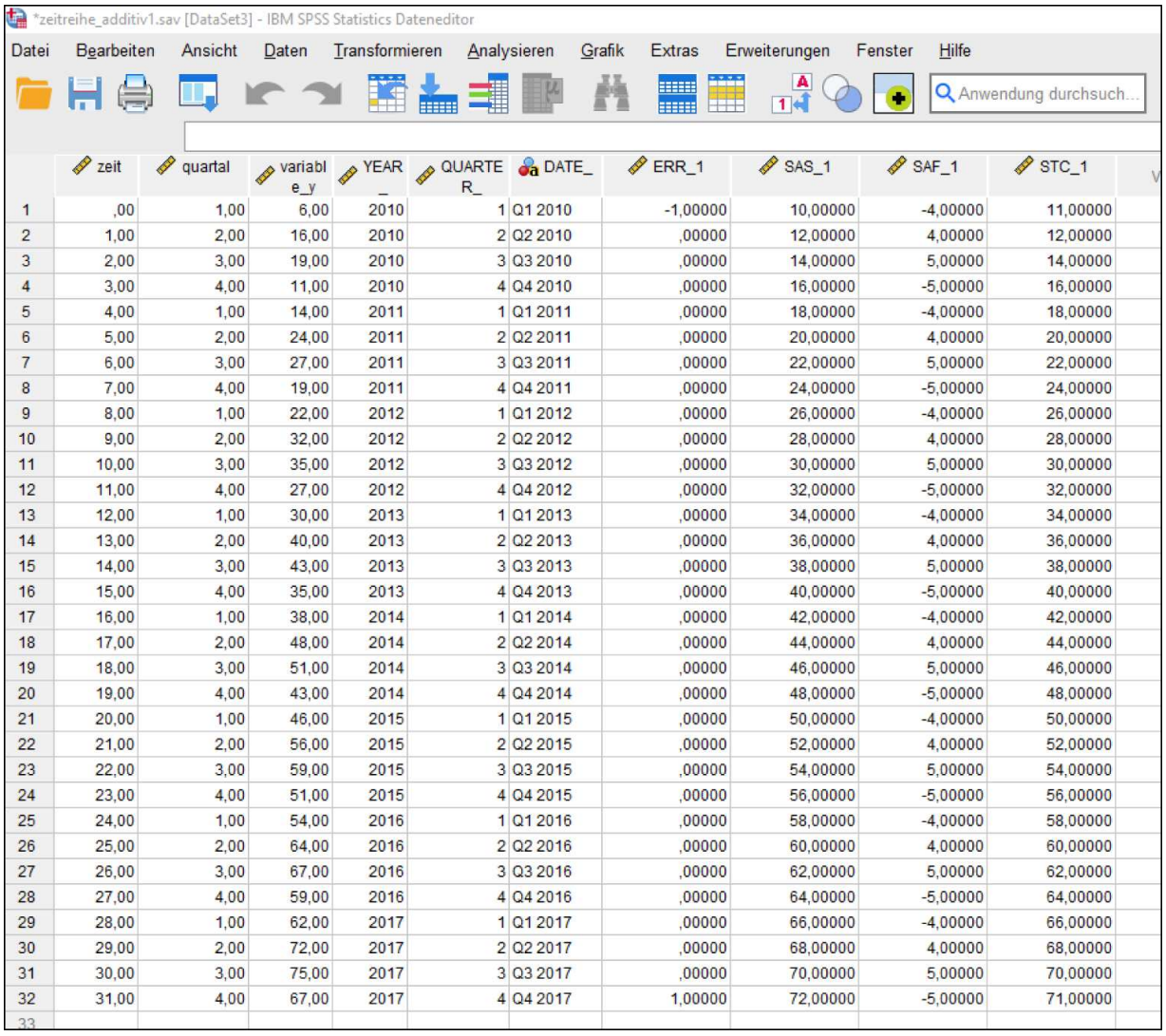
```
TSET PRINT=BRIEF NEWVAR=ALL MXNEWVAR=4.
SEASON
/VARIABLES=variable_y
/MODEL=ADDITIVE
/MA=EQUAL.
```

Von den vier hinzugefügten Variablen beinhaltet *saf_1* die jeweiligen Saisonziffern, die von den Ursprungswerten abgezogen werden und solcherart die *saisonbereinigte Zeitreihe* repräsentieren (siehe Abbildung 12.10).

Die **Saisonziffern** lauten also im Beispielfall: erstes Quartal: −4; zweites Quartal: +4; drittes Quartal: +5; viertes Quartal: −5. SPSS stellt diese Werte nicht nur im erweiterten Datensatz, sondern auch übersichtlicher im Viewer dar (siehe Abbildung 12.11).

## *Phasendurchschnittsverfahren mit variabler Saisonfigur*

Bei *multiplikativer* Verknüpfung der Zeitreihenelemente kommt das **Phasendurchschnittsverfahren mit variabler Saisonfigur** zur Anwendung.



*zeitreihe_additiv1.sav [DataSet3] - IBM SPSS Statistics Dateneditor

Datei Bearbeiten Ansicht Daten Transformieren Analysieren Grafik Extras Erweiterungen Fenster Hilfe

Anwendung durchsuch...

| | zeit | quartal | variable_y | YEAR_ | QUARTER_ | DATE_ | ERR_1 | SAS_1 | SAF_1 | STC_1 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ,00 | 1,00 | 6,00 | 2010 | 1 | Q1 2010 | -1,00000 | 10,00000 | -4,00000 | 11,00000 |
| 2 | 1,00 | 2,00 | 16,00 | 2010 | 2 | Q2 2010 | ,00000 | 12,00000 | 4,00000 | 12,00000 |
| 3 | 2,00 | 3,00 | 19,00 | 2010 | 3 | Q3 2010 | ,00000 | 14,00000 | 5,00000 | 14,00000 |
| 4 | 3,00 | 4,00 | 11,00 | 2010 | 4 | Q4 2010 | ,00000 | 16,00000 | -5,00000 | 16,00000 |
| 5 | 4,00 | 1,00 | 14,00 | 2011 | 1 | Q1 2011 | ,00000 | 18,00000 | -4,00000 | 18,00000 |
| 6 | 5,00 | 2,00 | 24,00 | 2011 | 2 | Q2 2011 | ,00000 | 20,00000 | 4,00000 | 20,00000 |
| 7 | 6,00 | 3,00 | 27,00 | 2011 | 3 | Q3 2011 | ,00000 | 22,00000 | 5,00000 | 22,00000 |
| 8 | 7,00 | 4,00 | 19,00 | 2011 | 4 | Q4 2011 | ,00000 | 24,00000 | -5,00000 | 24,00000 |
| 9 | 8,00 | 1,00 | 22,00 | 2012 | 1 | Q1 2012 | ,00000 | 26,00000 | -4,00000 | 26,00000 |
| 10 | 9,00 | 2,00 | 32,00 | 2012 | 2 | Q2 2012 | ,00000 | 28,00000 | 4,00000 | 28,00000 |
| 11 | 10,00 | 3,00 | 35,00 | 2012 | 3 | Q3 2012 | ,00000 | 30,00000 | 5,00000 | 30,00000 |
| 12 | 11,00 | 4,00 | 27,00 | 2012 | 4 | Q4 2012 | ,00000 | 32,00000 | -5,00000 | 32,00000 |
| 13 | 12,00 | 1,00 | 30,00 | 2013 | 1 | Q1 2013 | ,00000 | 34,00000 | -4,00000 | 34,00000 |
| 14 | 13,00 | 2,00 | 40,00 | 2013 | 2 | Q2 2013 | ,00000 | 36,00000 | 4,00000 | 36,00000 |
| 15 | 14,00 | 3,00 | 43,00 | 2013 | 3 | Q3 2013 | ,00000 | 38,00000 | 5,00000 | 38,00000 |
| 16 | 15,00 | 4,00 | 35,00 | 2013 | 4 | Q4 2013 | ,00000 | 40,00000 | -5,00000 | 40,00000 |
| 17 | 16,00 | 1,00 | 38,00 | 2014 | 1 | Q1 2014 | ,00000 | 42,00000 | -4,00000 | 42,00000 |
| 18 | 17,00 | 2,00 | 48,00 | 2014 | 2 | Q2 2014 | ,00000 | 44,00000 | 4,00000 | 44,00000 |
| 19 | 18,00 | 3,00 | 51,00 | 2014 | 3 | Q3 2014 | ,00000 | 46,00000 | 5,00000 | 46,00000 |
| 20 | 19,00 | 4,00 | 43,00 | 2014 | 4 | Q4 2014 | ,00000 | 48,00000 | -5,00000 | 48,00000 |
| 21 | 20,00 | 1,00 | 46,00 | 2015 | 1 | Q1 2015 | ,00000 | 50,00000 | -4,00000 | 50,00000 |
| 22 | 21,00 | 2,00 | 56,00 | 2015 | 2 | Q2 2015 | ,00000 | 52,00000 | 4,00000 | 52,00000 |
| 23 | 22,00 | 3,00 | 59,00 | 2015 | 3 | Q3 2015 | ,00000 | 54,00000 | 5,00000 | 54,00000 |
| 24 | 23,00 | 4,00 | 51,00 | 2015 | 4 | Q4 2015 | ,00000 | 56,00000 | -5,00000 | 56,00000 |
| 25 | 24,00 | 1,00 | 54,00 | 2016 | 1 | Q1 2016 | ,00000 | 58,00000 | -4,00000 | 58,00000 |
| 26 | 25,00 | 2,00 | 64,00 | 2016 | 2 | Q2 2016 | ,00000 | 60,00000 | 4,00000 | 60,00000 |
| 27 | 26,00 | 3,00 | 67,00 | 2016 | 3 | Q3 2016 | ,00000 | 62,00000 | 5,00000 | 62,00000 |
| 28 | 27,00 | 4,00 | 59,00 | 2016 | 4 | Q4 2016 | ,00000 | 64,00000 | -5,00000 | 64,00000 |
| 29 | 28,00 | 1,00 | 62,00 | 2017 | 1 | Q1 2017 | ,00000 | 66,00000 | -4,00000 | 66,00000 |
| 30 | 29,00 | 2,00 | 72,00 | 2017 | 2 | Q2 2017 | ,00000 | 68,00000 | 4,00000 | 68,00000 |
| 31 | 30,00 | 3,00 | 75,00 | 2017 | 3 | Q3 2017 | ,00000 | 70,00000 | 5,00000 | 70,00000 |
| 32 | 31,00 | 4,00 | 67,00 | 2017 | 4 | Q4 2017 | 1,00000 | 72,00000 | -5,00000 | 71,00000 |
| 33 | | | | | | | | | | |

**Abbildung 12.10:** Um weitere vier Variablen erweiterte Datendatei *ZEITREIHE_ADDITIV1.SAV* bei additiver Zeitreihen-Komponentenverknüpfung

**Saisonale Faktoren**

Bezeichnung der Zeitreihe: variable_y

| Periode | Saisonaler Faktor |
|---|---|
| 1 | -4,00000 |
| 2 | 4,00000 |
| 3 | 5,00000 |
| 4 | -5,00000 |

**Abbildung 12.11:** Saisonziffern bei konstanter Saisonfigur in Bezug auf *variable_y* (*ZEITREIHE_ADDITIV1.SAV*)

Nachstehend wird diese Methode auf die Datendatei *ZEITREIHE_MULTIPLIKATIV1.SAV* beziehungsweise die dortige Variable *variable_y* bezogen. Im Rahmen dieser Methode werden aus den Relationen zwischen Ursprungs- und Trendwerten arithmetische Mittelwerte berechnet, die – nach einer Normierung auf eine mittlere Abweichungssumme von 0 über alle Quartale hinweg – als *Saisonziffern* für die Quartale interpretiert werden. Anschließend können Sie auch bei dieser Spielart des Phasendurchschnittsverfahrens eine saisonbereinigte Zeitreihe erzeugen, und zwar indem Sie hier die Ursprungswerte durch die berechneten quartalsbezogenen Saisonziffern dividieren.

Um die Saisonziffern in SPSS berechnen zu können, sind auch hier ein paar Vorarbeiten notwendig. Es sind dies grundsätzlich die gleichen Vorarbeiten wie beim Phasendurchschnittsverfahren mit konstanter Saisonfigur. Das heißt: Auch hier müssen Sie SPSS mitteilen, dass sich Ihre Zeitdaten auf Quartale beziehen. Die Datumskomponente definieren Sie auch hier über DATEN|DATUM UND UHRZEIT DEFINIEREN. Gehen Sie bitte genauso vor, wie für die konstante Saisonfigur in Abbildung 12.7 dargelegt. SPSS fügt dann auch hier mit *year_*, *quarter_*

und *date_*drei neue Variablen direkt in Ihren Datensatz ein (hier: in *ZEITREIHE_MULTIPLIKATIV1.SAV* ). In *year_* beziehungsweise *quarter_* befinden sich wieder Angaben, um welches Jahr beziehungsweise um welches Quartal eines Jahrs es sich handelt, und in *date_* sind diese Informationen jeweils zusammen angegeben (siehe hierzu gegebenenfalls nochmals Abbildung 12.8 zum grundsätzlichen Datensatzaufbau).

Nach diesen Vorarbeiten können Sie auch hier mit der Berechnung der Saisonziffern loslegen. Dazu wählen Sie erneut im SPSS-Menü ANALYSIEREN|VORHERSAGE|SAISONALE ZERLEGUNG aus. In dem sich öffnenden Dialogfenster tragen Sie die zu analysierende Variable unter VARIABLE(N) ein (hier: *variable_y*) und klicken zudem bei MODELLTYP auf MULTIPLIKATIV; schließlich möchten Sie ja jetzt für den Zeitreihen-Modelltyp einer multiplikativen Komponentenverknüpfung Saisonziffern berechnen. SPSS fügt auch hier wieder vier weitere Variablen direkt in Ihre Datendatei ein (siehe Abbildung 12.12).

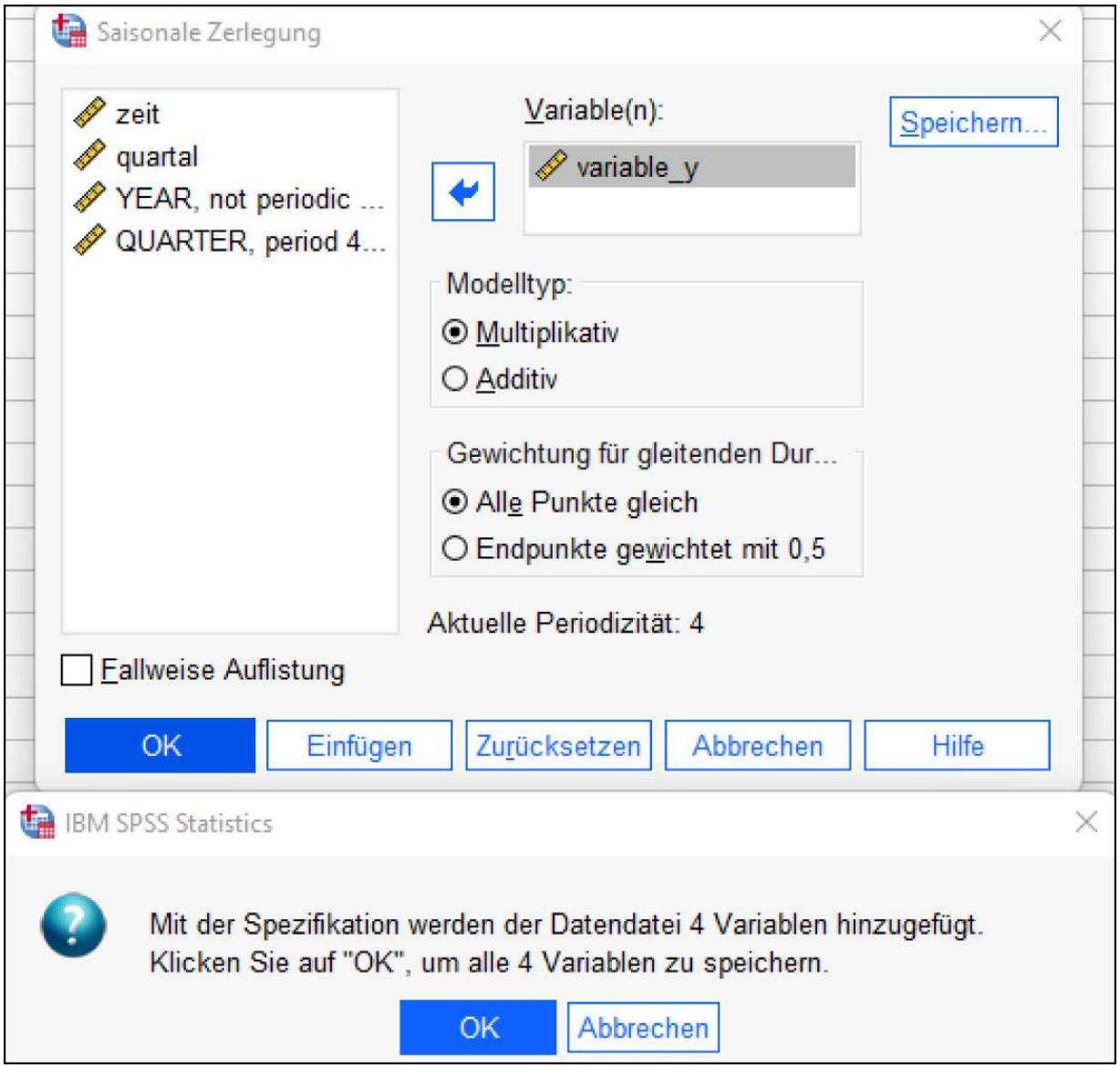


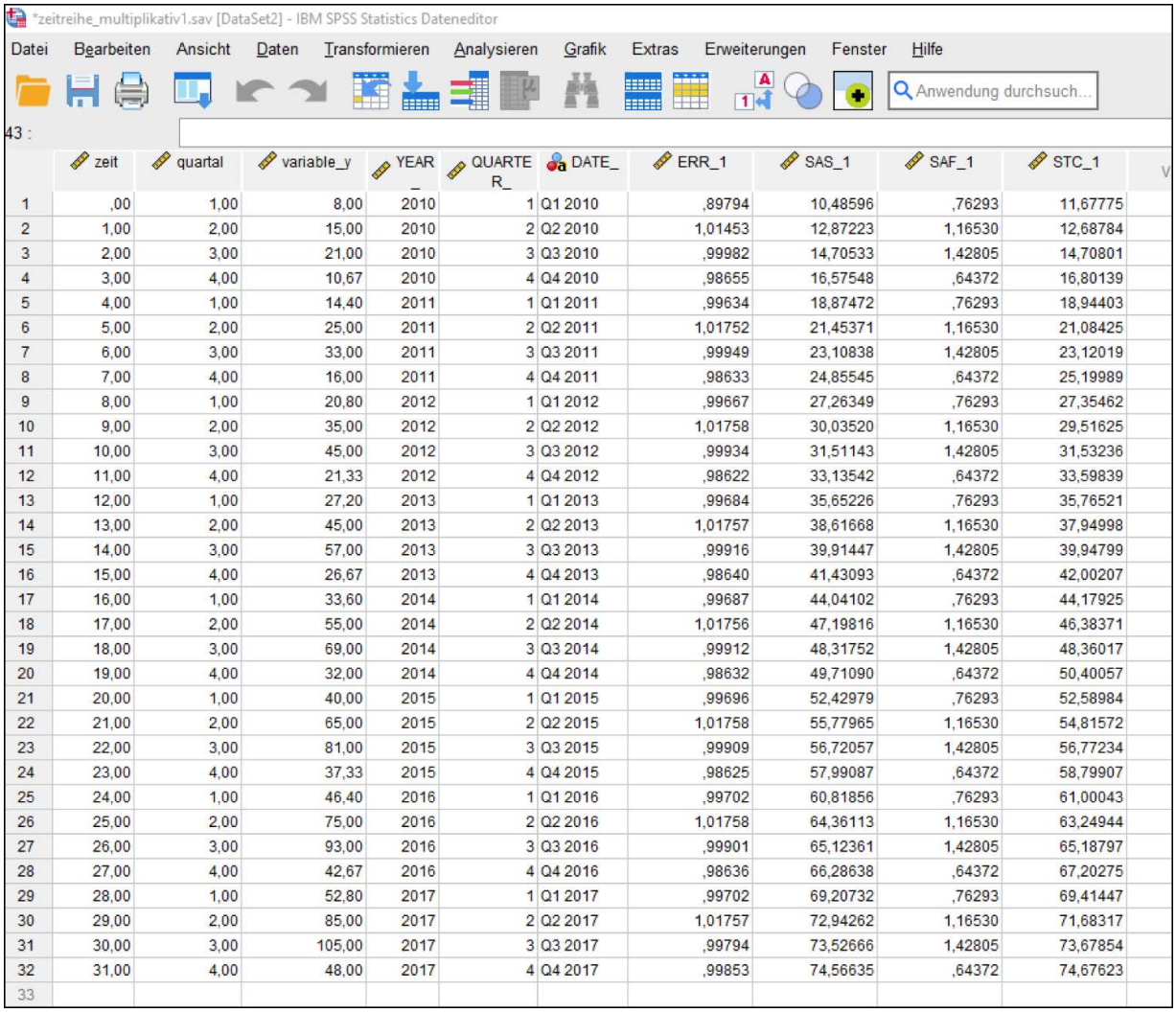

**Abbildung 12.12:** Anfordern einer multiplikativen Zeitreihen-Komponentenzerlegung (*ZEITREIHE_MULTIPLIKATIV1.SAV*)

Die zugehörige SPSS-Syntax lautet:

```
TSET PRINT=BRIEF NEWVAR=ALL MXNEWVAR=4.
SEASON
    /VARIABLES=variable_y
    /MODEL=MULTIPLICATIVE
    /MA=EQUAL.
```

Auch hier enthält von den vier hinzugefügten Variablen *saf_1* die jeweiligen Saisonziffern (siehe Abbildung 12.13). Die saisonbereinigte

Zeitreihe erstellen Sie dabei dadurch, dass Sie die Ursprungswerte durch diese Saisonziffern teilen.


*zeitreihe_multiplikativ1.sav [DataSet2] - IBM SPSS Statistics Dateneditor

Datei Bearbeiten Ansicht Daten Transformieren Analysieren Grafik Extras Erweiterungen Fenster Hilfe

Anwendung durchsuch...

43 :

| | zeit | quartal | variable_y | YEAR_ | QUARTER_ | DATE_ | ERR_1 | SAS_1 | SAF_1 | STC_1 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ,00 | 1,00 | 8,00 | 2010 | 1 | Q1 2010 | ,89794 | 10,48596 | ,76293 | 11,67775 |
| 2 | 1,00 | 2,00 | 15,00 | 2010 | 2 | Q2 2010 | 1,01453 | 12,87223 | 1,16530 | 12,68784 |
| 3 | 2,00 | 3,00 | 21,00 | 2010 | 3 | Q3 2010 | ,99982 | 14,70533 | 1,42805 | 14,70801 |
| 4 | 3,00 | 4,00 | 10,67 | 2010 | 4 | Q4 2010 | ,98655 | 16,57548 | ,64372 | 16,80139 |
| 5 | 4,00 | 1,00 | 14,40 | 2011 | 1 | Q1 2011 | ,99634 | 18,87472 | ,76293 | 18,94403 |
| 6 | 5,00 | 2,00 | 25,00 | 2011 | 2 | Q2 2011 | 1,01752 | 21,45371 | 1,16530 | 21,08425 |
| 7 | 6,00 | 3,00 | 33,00 | 2011 | 3 | Q3 2011 | ,99949 | 23,10838 | 1,42805 | 23,12019 |
| 8 | 7,00 | 4,00 | 16,00 | 2011 | 4 | Q4 2011 | ,98633 | 24,85545 | ,64372 | 25,19989 |
| 9 | 8,00 | 1,00 | 20,80 | 2012 | 1 | Q1 2012 | ,99667 | 27,26349 | ,76293 | 27,35462 |
| 10 | 9,00 | 2,00 | 35,00 | 2012 | 2 | Q2 2012 | 1,01758 | 30,03520 | 1,16530 | 29,51625 |
| 11 | 10,00 | 3,00 | 45,00 | 2012 | 3 | Q3 2012 | ,99934 | 31,51143 | 1,42805 | 31,53236 |
| 12 | 11,00 | 4,00 | 21,33 | 2012 | 4 | Q4 2012 | ,98622 | 33,13542 | ,64372 | 33,59839 |
| 13 | 12,00 | 1,00 | 27,20 | 2013 | 1 | Q1 2013 | ,99684 | 35,65226 | ,76293 | 35,76521 |
| 14 | 13,00 | 2,00 | 45,00 | 2013 | 2 | Q2 2013 | 1,01757 | 38,61668 | 1,16530 | 37,94998 |
| 15 | 14,00 | 3,00 | 57,00 | 2013 | 3 | Q3 2013 | ,99916 | 39,91447 | 1,42805 | 39,94799 |
| 16 | 15,00 | 4,00 | 26,67 | 2013 | 4 | Q4 2013 | ,98640 | 41,43093 | ,64372 | 42,00207 |
| 17 | 16,00 | 1,00 | 33,60 | 2014 | 1 | Q1 2014 | ,99687 | 44,04102 | ,76293 | 44,17925 |
| 18 | 17,00 | 2,00 | 55,00 | 2014 | 2 | Q2 2014 | 1,01756 | 47,19816 | 1,16530 | 46,38371 |
| 19 | 18,00 | 3,00 | 69,00 | 2014 | 3 | Q3 2014 | ,99912 | 48,31752 | 1,42805 | 48,36017 |
| 20 | 19,00 | 4,00 | 32,00 | 2014 | 4 | Q4 2014 | ,98632 | 49,71090 | ,64372 | 50,40057 |
| 21 | 20,00 | 1,00 | 40,00 | 2015 | 1 | Q1 2015 | ,99696 | 52,42979 | ,76293 | 52,58984 |
| 22 | 21,00 | 2,00 | 65,00 | 2015 | 2 | Q2 2015 | 1,01758 | 55,77965 | 1,16530 | 54,81572 |
| 23 | 22,00 | 3,00 | 81,00 | 2015 | 3 | Q3 2015 | ,99909 | 56,72057 | 1,42805 | 56,77234 |
| 24 | 23,00 | 4,00 | 37,33 | 2015 | 4 | Q4 2015 | ,98625 | 57,99087 | ,64372 | 58,79907 |
| 25 | 24,00 | 1,00 | 46,40 | 2016 | 1 | Q1 2016 | ,99702 | 60,81856 | ,76293 | 61,00043 |
| 26 | 25,00 | 2,00 | 75,00 | 2016 | 2 | Q2 2016 | 1,01758 | 64,36113 | 1,16530 | 63,24944 |
| 27 | 26,00 | 3,00 | 93,00 | 2016 | 3 | Q3 2016 | ,99901 | 65,12361 | 1,42805 | 65,18797 |
| 28 | 27,00 | 4,00 | 42,67 | 2016 | 4 | Q4 2016 | ,98636 | 66,28638 | ,64372 | 67,20275 |
| 29 | 28,00 | 1,00 | 52,80 | 2017 | 1 | Q1 2017 | ,99702 | 69,20732 | ,76293 | 69,41447 |
| 30 | 29,00 | 2,00 | 85,00 | 2017 | 2 | Q2 2017 | 1,01757 | 72,94262 | 1,16530 | 71,68317 |
| 31 | 30,00 | 3,00 | 105,00 | 2017 | 3 | Q3 2017 | ,99794 | 73,52666 | 1,42805 | 73,67854 |
| 32 | 31,00 | 4,00 | 48,00 | 2017 | 4 | Q4 2017 | ,99853 | 74,56635 | ,64372 | 74,67623 |
| 33 | | | | | | | | | | |

**Abbildung 12.13:** Um weitere vier Variablen erweiterte Datendatei *ZEITREIHE_MULTIPLIKATIV1.SAV* bei multiplikativer Zeitreihen-Komponentenverknüpfung

Die **Saisonziffern** werden auch hier wieder übersichtlich im Viewer ausgewiesen – dort als Prozentwerte. Im Beispielfall ergeben sich folgende Saisonziffern (als Prozentwerte): erstes Quartal: 76,3 Prozent; zweites Quartal: 116,5 Prozent; drittes Quartal: 142,8 Prozent; viertes Quartal: 64,4 Prozent (siehe Abbildung 12.14).

**Saisonale Faktoren**

Bezeichnung der Zeitreihe: variable_y

| Periode | Saisonaler Faktor (%) |
|---|---|
| 1 | 76,3 |
| 2 | 116,5 |
| 3 | 142,8 |
| 4 | 64,4 |

**Abbildung 12.14:** Saisonziffern bei variabler Saisonfigur in Bezug auf *variable_y* in Prozent (*ZEITREIHE_MULTIPLIKATIV1.SAV* )

# *Sterbetafelanalyse*

Die **Sterbetafelanalyse** ist eine Statistikart im Bereich der Überlebensanalysen. Ausgehend von einem Startzeitpunkt, ist es der Inhalt von Überlebensanalysen, dass für verschiedene Ereignisse eine **Überlebensfunktion** ermittelt wird.

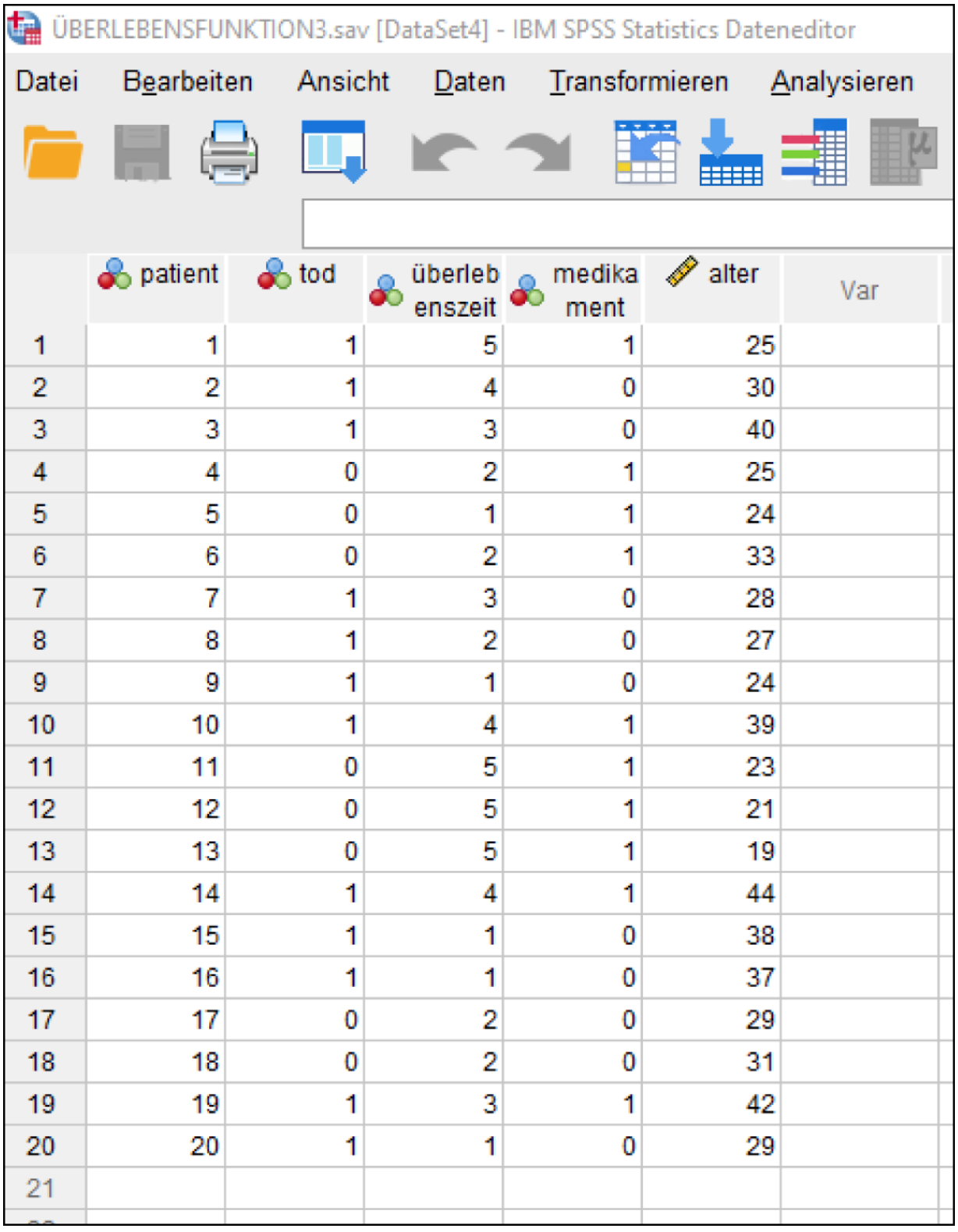
Man bildet dabei im Zusammenhang mit den Ereigniszeitpunkten eine *Überlebens*- und eine *Hazardrate*. Die Überlebensrate gibt an, wie viele Untersuchungseinheiten aus einer früheren Periode bis zum Eintritt eines bestimmten Ereignisses überlebt haben, während die Hazardrate genau umgekehrt die Anzahl der Untersuchungseinheiten aus einer früheren Periode angibt, die bis zum Eintritt eines bestimmten Ereignisses weggefallen sind.

Dabei ist der Begriff des »Überlebens« inhaltlich nicht notwendigerweise mit »Leben und Tod« in Verbindung zu bringen. In

Beschäftigungsanalysen kann es zum Beispiel bei der Überlebensfunktion darum gehen, wie viele Personen bis zu einem bestimmten Jahr in einem Beschäftigungsverhältnis »überlebt« haben. In diesem Fall würde die Hazardrate angeben, wie viele Personen bis zu dem betreffenden Jahr aus einem Beschäftigungsverhältnis ausgeschieden sind.

Die Überlebensanalysen zugrunde liegenden Daten sind typischerweise zensiert. Dabei wird hier unter **Zensierung** verstanden, dass bis zu einem bestimmten Zeitpunkt für bestimmte Untersuchungseinheiten noch kein Ereignis eingetreten ist. Eine Zensierung kann aber auch bedeuten, dass Untersuchungseinheiten die Überlebensstudie verlassen haben. Da die Zensierung sich hier auf das Ende der Studie bezieht, spricht man auch von *rechtszensierten Daten*.

Im Folgenden soll zunächst eine Sterbetafelanalyse auf Basis der Datendatei *ÜBERLEBENSFUNKTION3.SAV* durchgeführt werden. In dieser Datei sind – wie aus Abbildung 12.15 zu erkennen ist – Angaben vorhanden, ob eine Person im Untersuchungszeitraum verstorben ist (*tod* = 1) oder nicht (*tod* = 0), wie viele Jahre sie überlebt hat (in der Variablen *überlebenszeit*), ob sie ein Medikament verabreicht bekommen hat (*medikament* = 1) oder nicht (*medikament* = 0) und wie alt sie ist (in der Variablen *alter*). Aus Vereinfachungsgründen wird hier unterstellt, dass die verstorbenen Personen alle an der betreffenden Krankheit, für die das Medikament entwickelt worden ist, verstorben sind und dass dieses Medikament gegebenenfalls in der Startperiode der Untersuchung verabreicht worden ist.




| | patient | tod | überlebenszeit | medikament | alter | Var |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 5 | 1 | 25 | |
| 2 | 2 | 1 | 4 | 0 | 30 | |
| 3 | 3 | 1 | 3 | 0 | 40 | |
| 4 | 4 | 0 | 2 | 1 | 25 | |
| 5 | 5 | 0 | 1 | 1 | 24 | |
| 6 | 6 | 0 | 2 | 1 | 33 | |
| 7 | 7 | 1 | 3 | 0 | 28 | |
| 8 | 8 | 1 | 2 | 0 | 27 | |
| 9 | 9 | 1 | 1 | 0 | 24 | |
| 10 | 10 | 1 | 4 | 1 | 39 | |
| 11 | 11 | 0 | 5 | 1 | 23 | |
| 12 | 12 | 0 | 5 | 1 | 21 | |
| 13 | 13 | 0 | 5 | 1 | 19 | |
| 14 | 14 | 1 | 4 | 1 | 44 | |
| 15 | 15 | 1 | 1 | 0 | 38 | |
| 16 | 16 | 1 | 1 | 0 | 37 | |
| 17 | 17 | 0 | 2 | 0 | 29 | |
| 18 | 18 | 0 | 2 | 0 | 31 | |
| 19 | 19 | 1 | 3 | 1 | 42 | |
| 20 | 20 | 1 | 1 | 0 | 29 | |
| 21 | | | | | | |

**Abbildung 12.15:** Datendatei *ÜBERLEBENSFUNKTION3.SAV*

Die Sterbetafelanalyse rufen Sie in SPSS über ANALYSIEREN| ÜBERLEBEN|STERBETAFELN auf. Dabei müssen Sie auf der sich öffnenden Dialogkarte bei ZEIT in unserem Fall *überlebenszeit*, bei STATUS *tod(1)* und bei FAKTOR *medikament (0 1)* eingeben. Dann klicken Sie einfach nur noch auf die OK-Taste, und los geht's mit der Sterbetafelanalyse (siehe hierzu Abbildung 12.16)!

Die zugehörige Syntax lautet:

```
SURVIVAL TABLE=überlebenszeit BY medikament(0 1)
    /INTERVAL=THRU 5 BY 1
    /STATUS=tod(1)
    /PRINT=TABLE.
```

Mit diesen Angaben »liefert« SPSS die in Abbildung 12.17 wiedergegebene Sterbetafelanalyse. Im oberen Abbildungsteil sind für zwei Gruppen die Überlebens- und Hazardraten angegeben, und zwar zum einen für die Gruppe, die das hier fragliche Medikament nicht eingenommen hat (*medikament* = 0), und zum anderen für die Gruppe, die dieses Medikament eingenommen hat (*medikament* = 1). Aus dem unteren Abbildungsteil ist die – am Median gemessene – mittlere Überlebenszeit dieser beiden Personengruppen dargelegt. Während diese mittlere Überlebenszeit für die Personen ohne Medikamenteneinnahme 2,83 Jahre beträgt, beläuft sie sich für die Personen mit Medikamenteneinnahme auf immerhin 5,00 Jahre.

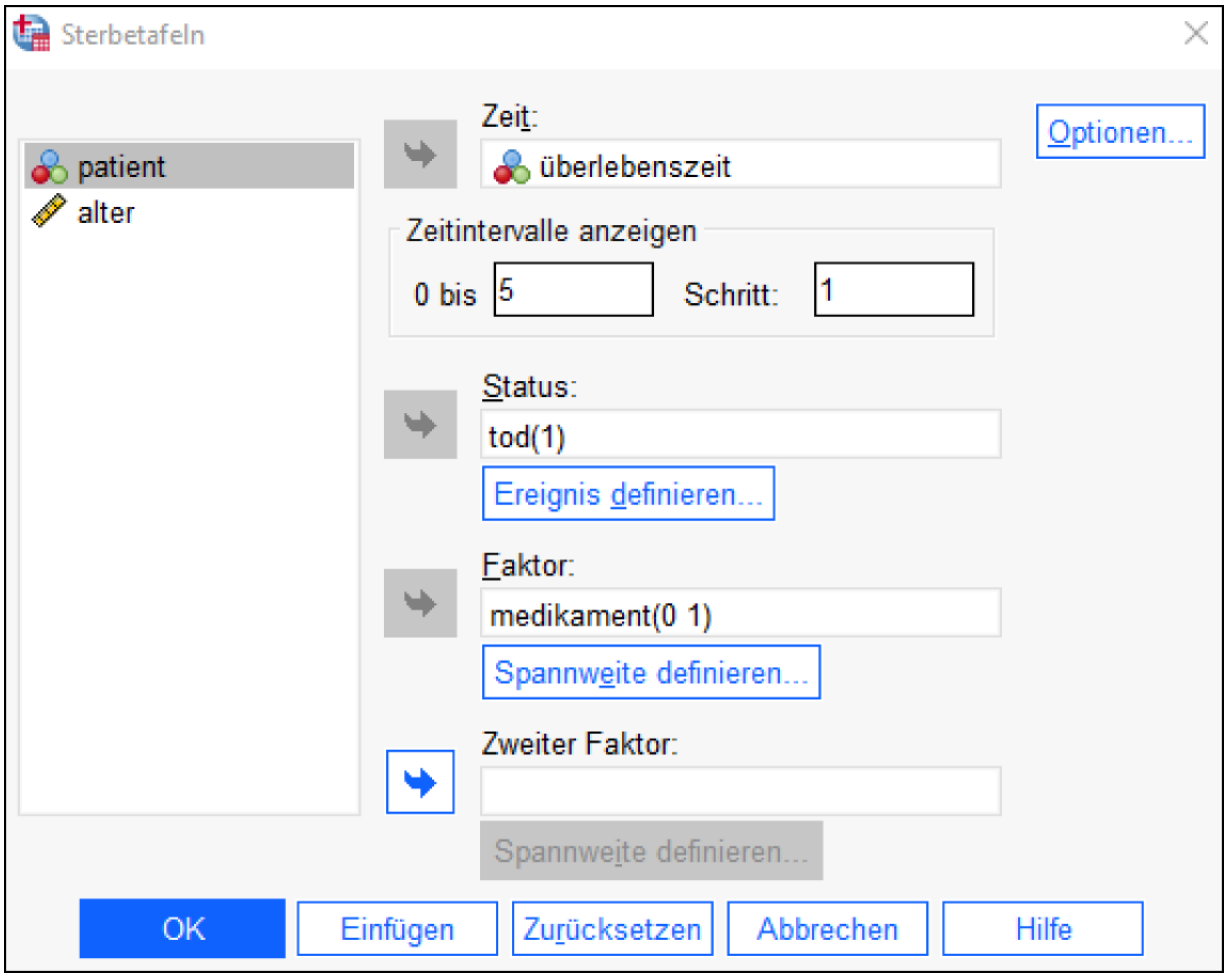



**Abbildung 12.16:** Aufruf einer Sterbetafelanalyse (*ÜBERLEBENSFUNKTION3.SAV*)

**Sterbetafel**

| Kontrollgrößen erster Ordnung | | Anfangszeit des Intervalls | Anzahl der zur Anfangszeit Überlebenden | Anzahl der Ausgeschiedenen | Anzahl der dem Risiko ausgesetzten | Anzahl terminaler Ereignisse | Anteil der Terminierenden | Anteil der Überlebenden | Kumulierter Anteil der Überlebenden am Intervallende | Standardfehler des kumulierten Anteils der Überlebenden am Intervallende | Wahrscheinlichkeitsdichte | Standardfehler der Wahrscheinlichkeitsdichte | Hazard-Rate | Standardfehler der Hazard-Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| medikament | 0 | 0 | 10 | 0 | 10,000 | 0 | ,00 | 1,00 | 1,00 | ,00 | ,000 | ,000 | ,00 | ,00 |
| | | 1 | 10 | 0 | 10,000 | 4 | ,40 | ,60 | ,60 | ,15 | ,400 | ,155 | ,50 | ,24 |
| | | 2 | 6 | 2 | 5,000 | 1 | ,20 | ,80 | ,48 | ,16 | ,120 | ,112 | ,22 | ,22 |
| | | 3 | 3 | 0 | 3,000 | 2 | ,67 | ,33 | ,16 | ,14 | ,320 | ,170 | 1,00 | ,61 |
| | | 4 | 1 | 0 | 1,000 | 1 | 1,00 | ,00 | ,00 | ,00 | ,160 | ,142 | 2,00 | ,00 |
| | 1 | 0 | 10 | 0 | 10,000 | 0 | ,00 | 1,00 | 1,00 | ,00 | ,000 | ,000 | ,00 | ,00 |
| | | 1 | 10 | 1 | 9,500 | 0 | ,00 | 1,00 | 1,00 | ,00 | ,000 | ,000 | ,00 | ,00 |
| | | 2 | 9 | 2 | 8,000 | 0 | ,00 | 1,00 | 1,00 | ,00 | ,000 | ,000 | ,00 | ,00 |
| | | 3 | 7 | 0 | 7,000 | 1 | ,14 | ,86 | ,86 | ,13 | ,143 | ,132 | ,15 | ,15 |
| | | 4 | 6 | 0 | 6,000 | 2 | ,33 | ,67 | ,57 | ,19 | ,286 | ,171 | ,40 | ,28 |
| | | 5 | 4 | 3 | 2,500 | 1 | ,40 | ,60 | ,34 | ,21 | ,000 | ,000 | ,00 | ,00 |

**Median der Überlebenszeit**

| Kontrollgrößen erster Ordnung | | Median der Zeit |
|---|---|---|
| medikament | 0 | 2,83 |
| | 1 | 5,00 |

**Abbildung 12.17:** Ergebnis einer Sterbetafelanalyse (*ÜBERLEBENSFUNKTION3.SAV*)

# *Kaplan/Meier-Methode*

Wie die Sterbetafelanalyse ist auch die **Kaplan/Meier-Methode** dem Bereich der Überlebensanalysen zuzuordnen. Dabei hat man es auch im Zusammenhang mit der Kaplan/Meier-Methode mit (rechts-)zensierten Beobachtungen zu tun.

Auf der Basis entsprechender Beobachtungsdaten wird im Rahmen der Kaplan/Meier-Methode und ohne Annahmen über die zeitliche Entwicklung der zugrunde liegenden Wahrscheinlichkeitsverteilung eine **Überlebensfunktion** mit einer *Überlebens*- und einer *Hazardrate* geschätzt. Konkret wird für eine Untersuchungseinheit die Wahrscheinlichkeit geschätzt, dass ein bestimmtes Ereignis innerhalb eines Zeitintervalls *nicht* eintritt. Es handelt sich um eine nichtparametrische Schätzung einer Überlebensfunktion (mit rechtszensierten Daten). Zudem ist die Kaplan/Meier-Methode nur für kategoriale Variablen (mit vergleichsweise wenigen Ausprägungen) sinnvoll.

Die Kaplan/Meier-Schätzungen starten Sie im SPSS-Menü über ANALYSIEREN|ÜBERLEBEN|KAPLAN-MEIER, wobei Sie bitte – erneut für die Datendatei *ÜBERLEBENSFUNKTION3.SAV* – im sich öffnenden Dialogfenster bei ZEIT *überlebenszeit*, bei STATUS *tod(1)* und bei FAKTOR *medikament* eingeben (siehe Abbildung 12.18).

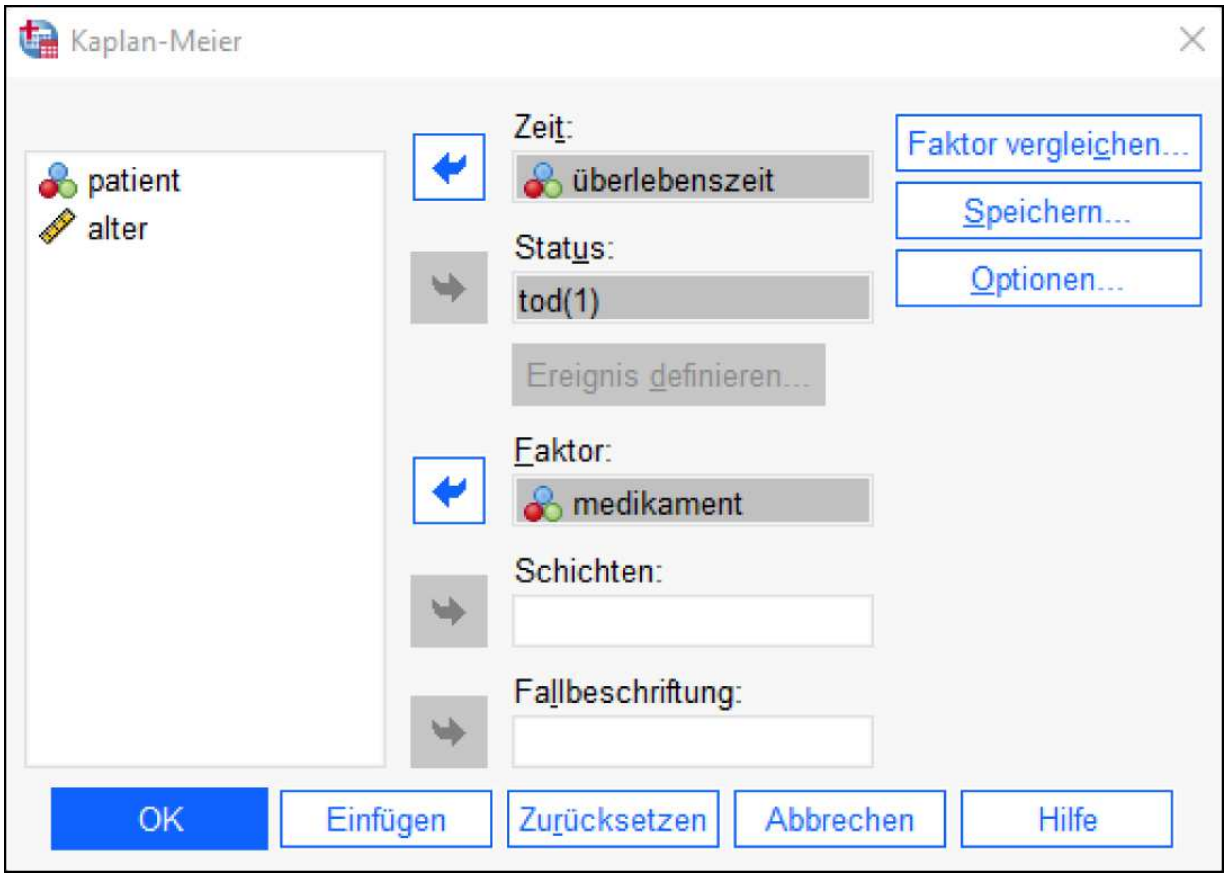



**Abbildung 12.18:** Aufruf der Kaplan/Meier-Methode (*ÜBERLEBENSFUNKTION3.SAV*) (1)

Zusätzlich müssen Sie noch unter OPTIONEN genauer festlegen (mit einem Fachbegriff: spezifizieren), was genau im Zusammenhang mit der Kaplan/Meier-Methode an SPSS-Output ausgegeben werden soll. In Abbildung 12.19 sind diesbezüglich ausgewählt worden: tabellierte Überlebensangaben, das arithmetische Mittel und der Median der Überlebenszeit sowie die Überlebensfunktion als Diagramm. (Auf die Darstellung der Hazard-Funktion wird nachfolgend verzichtet, da diese im Grunde genommen ja nur das Gegenstück zur Überlebensfunktion ist.)

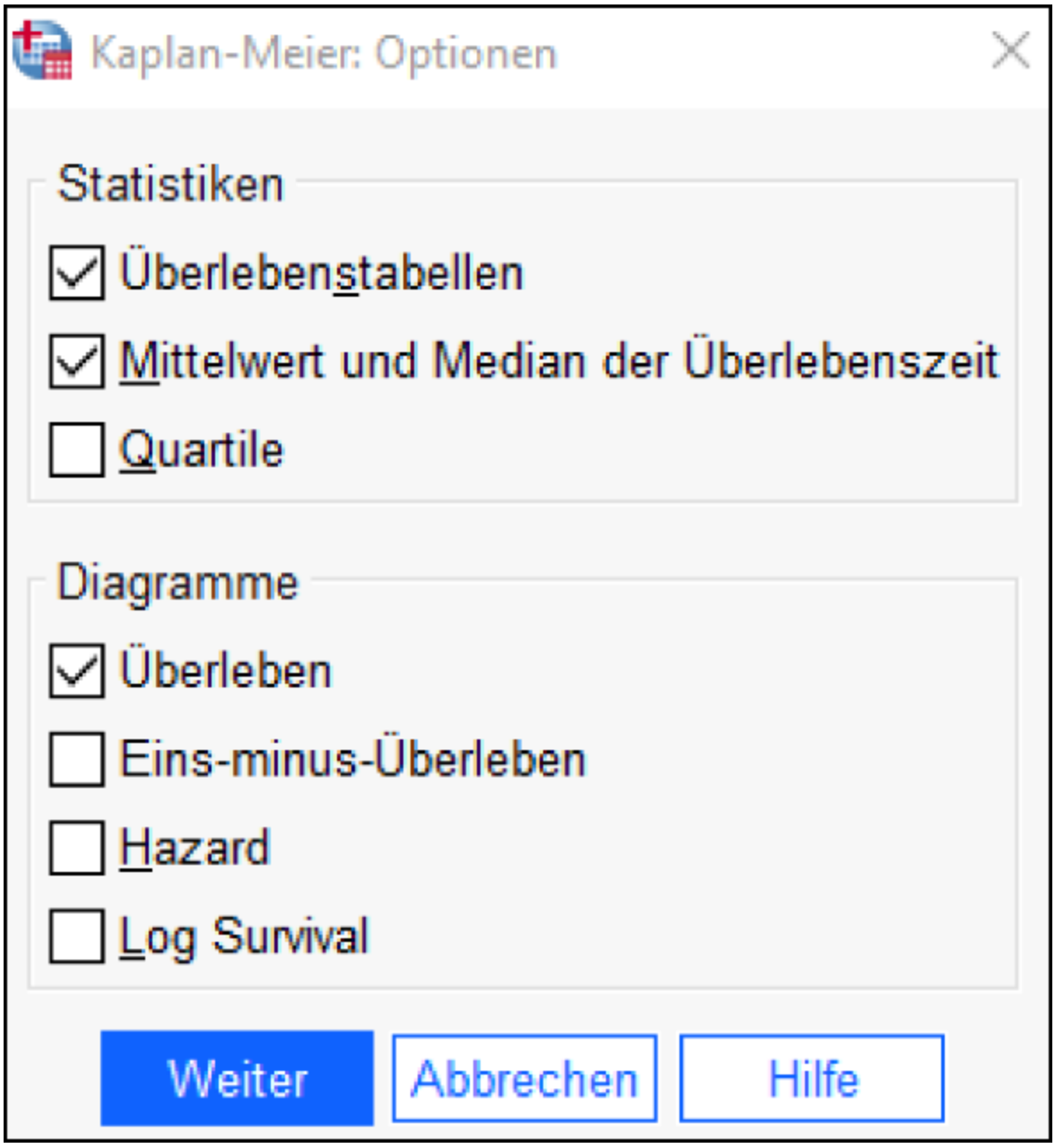



**Abbildung 12.19:** Aufruf der Kaplan/Meier-Methode (*ÜBERLEBENSFUNKTION3.SAV*) (2)

Die entsprechende Syntax lautet:

```
KM überlebenszeit BY medikament
    /STATUS=tod(1)
    /PRINT TABLE MEAN
    /PLOT SURVIVAL.
```

In Abbildung 12.20 und Abbildung 12.21 finden Sie die Berechnungsergebnisse für die Kaplan/Meier-Methode im vorliegenden

Beispielfall. Abbildung 12.20 beinhaltet im Wesentlichen die Überlebenstabelle, während Abbildung 12.21 die Überlebensunterschiede zwischen den Personen ohne und denjenigen mit Medikamenteneinnahme eindrucksvoll veranschaulicht. Aus dem Diagramm mit der Überlebensfunktion ist zum Beispiel ersichtlich, dass zu allen Zeitpunkten die Überlebensfunktion der Personen mit Medikamenteneinnahme oberhalb der Funktion der Personen ohne Medikamenteneinnahme liegt, was auf die Wirksamkeit des entsprechenden Medikaments in unserem einfachen Beispiel schließen lässt. Die Striche auf den »Stufen« der jeweiligen Überlebensfunktion zeigen im Übrigen die Zensierungen an.

**Zusammenfassung der Fallverarbeitung**

| medikament | Gesamtzahl | Anzahl von Ereignissen | Zensiert | |
|---|---|---|---|---|
| | | | N | Prozent |
| 0 | 10 | 8 | 2 | 20,0% |
| 1 | 10 | 4 | 6 | 60,0% |
| Gesamt | 20 | 12 | 8 | 40,0% |

**Überlebenstabelle**

| medikament | | Zeit | Status | Kumulierter Anteil Überlebender zum Zeitpunkt | | Anzahl der kumulativen Ereignisse | Anzahl der verbliebenen Fälle |
|---|---|---|---|---|---|---|---|
| | | | | Schätzer | Standardfehler | | |
| 0 | 1 | 1,000 | 1 | . | . | 1 | 9 |
| | 2 | 1,000 | 1 | . | . | 2 | 8 |
| | 3 | 1,000 | 1 | . | . | 3 | 7 |
| | 4 | 1,000 | 1 | ,600 | ,155 | 4 | 6 |
| | 5 | 2,000 | 1 | ,500 | ,158 | 5 | 5 |
| | 6 | 2,000 | 0 | . | . | 5 | 4 |
| | 7 | 2,000 | 0 | . | . | 5 | 3 |
| | 8 | 3,000 | 1 | . | . | 6 | 2 |
| | 9 | 3,000 | 1 | ,167 | ,146 | 7 | 1 |
| | 10 | 4,000 | 1 | ,000 | ,000 | 8 | 0 |
| 1 | 1 | 1,000 | 0 | . | . | 0 | 9 |
| | 2 | 2,000 | 0 | . | . | 0 | 8 |
| | 3 | 2,000 | 0 | . | . | 0 | 7 |
| | 4 | 3,000 | 1 | ,857 | ,132 | 1 | 6 |
| | 5 | 4,000 | 1 | . | . | 2 | 5 |
| | 6 | 4,000 | 1 | ,571 | ,187 | 3 | 4 |
| | 7 | 5,000 | 1 | ,429 | ,187 | 4 | 3 |
| | 8 | 5,000 | 0 | . | . | 4 | 2 |
| | 9 | 5,000 | 0 | . | . | 4 | 1 |
| | 10 | 5,000 | 0 | . | . | 4 | 0 |

**Abbildung 12.20:** Ergebnisse einer Kaplan/Meier-Analyse (*ÜBERLEBENSFUNKTION3.SAV*) (1)

**Mittelwerte und Mediane für die Überlebenszeit**

| | Mittelwert[a] | | | | Median | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 95%-Konfidenzintervall | | | | 95%-Konfidenzintervall | |
| medikament | Schätzer | Standardfehler | Untergrenze | Obergrenze | Schätzer | Standardfehler | Untergrenze | Obergrenze |
| 0 | 2,267 | ,400 | 1,482 | 3,051 | 2,000 | ,730 | ,570 | 3,430 |
| 1 | 4,429 | ,318 | 3,805 | 5,052 | 5,000 | 1,309 | 2,434 | 7,566 |
| Gesamt | 3,368 | ,362 | 2,658 | 4,078 | 4,000 | ,545 | 2,931 | 5,069 |

a. Die Schätzung ist auf die längste Überlebenszeit begrenzt, wenn sie zensiert ist.

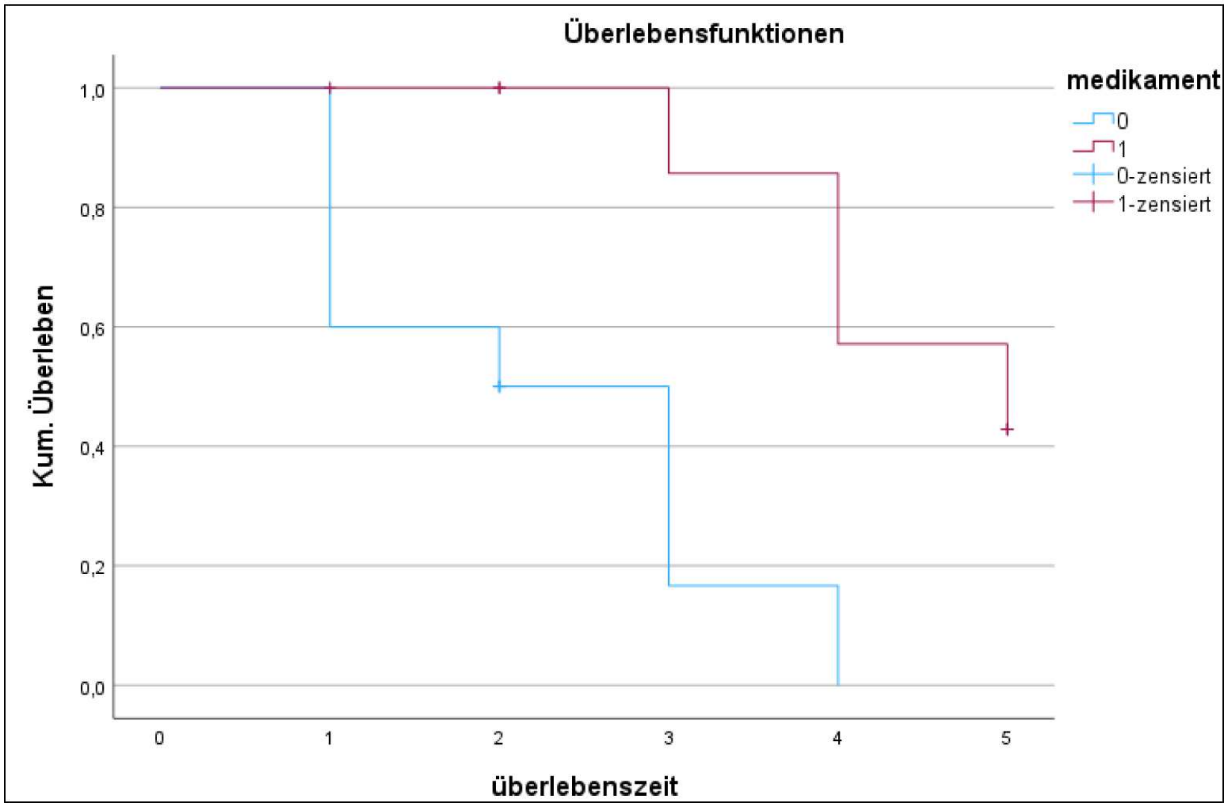



**Abbildung 12.21:** Ergebnisse einer Kaplan/Meier-Analyse (*ÜBERLEBENSFUNKTION3.SAV*) (2)

# *Cox-Regression*

Die **Regressionsanalyse nach Cox** bezieht sich – das ist ein Unterschied zur Kaplan/Meier-Methode – auf *mehrere* unabhängige Variablen.

Sie steuern die Cox-Regression im SPSS-Menü folgendermaßen an: ANALYSIEREN|ÜBERLEBEN|COX-REGRESSION. Erneut in Bezug auf die Datendatei *ÜBERLEBENSFUNKTION3.SAV* geben Sie bitte in dem sich öffnenden Dialogfenster bei ZEIT *überlebenszeit*, bei STATUS

*tod(1)* sowie bei KOVARIATEN *medikament* und *alter* ein (siehe Abbildung 12.22).

Für die Variable *medikament* müssen Sie zusätzlich noch angeben, dass es sich bei ihr um eine kategoriale Variable handelt. Dies ist in Abbildung 12.23 dargestellt und funktioniert über KATEGORIAL (siehe hierzu Abbildung 12.22). Als REFERENZKATEGORIE wurde hier im Übrigen LETZTE ausgewählt. (Sie hätten aber auch ERSTE auswählen können.)

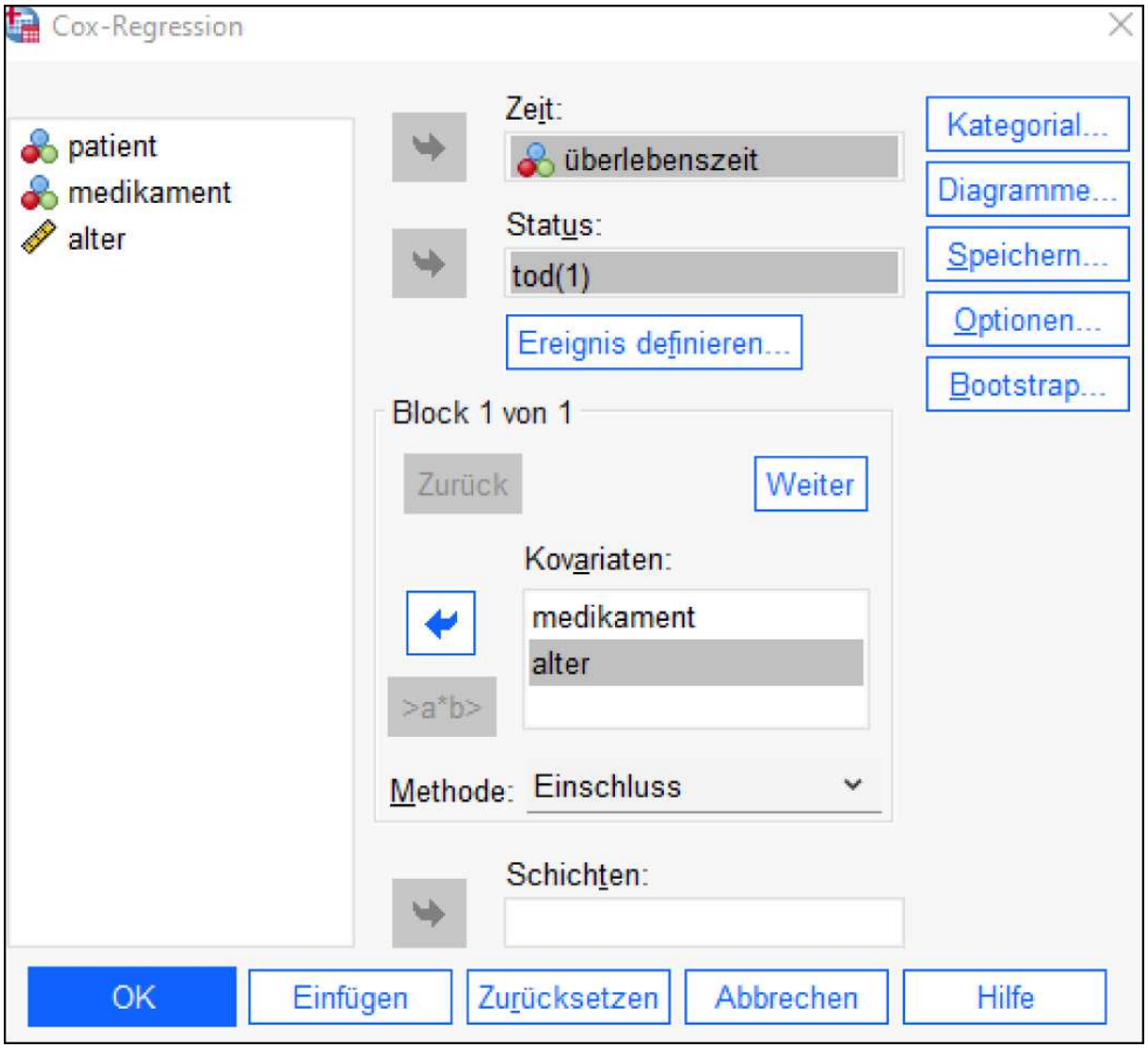



**Abbildung 12.22:** Aufruf einer Cox-Regression (*ÜBERLEBENSFUNKTION3.SAV*) (1)

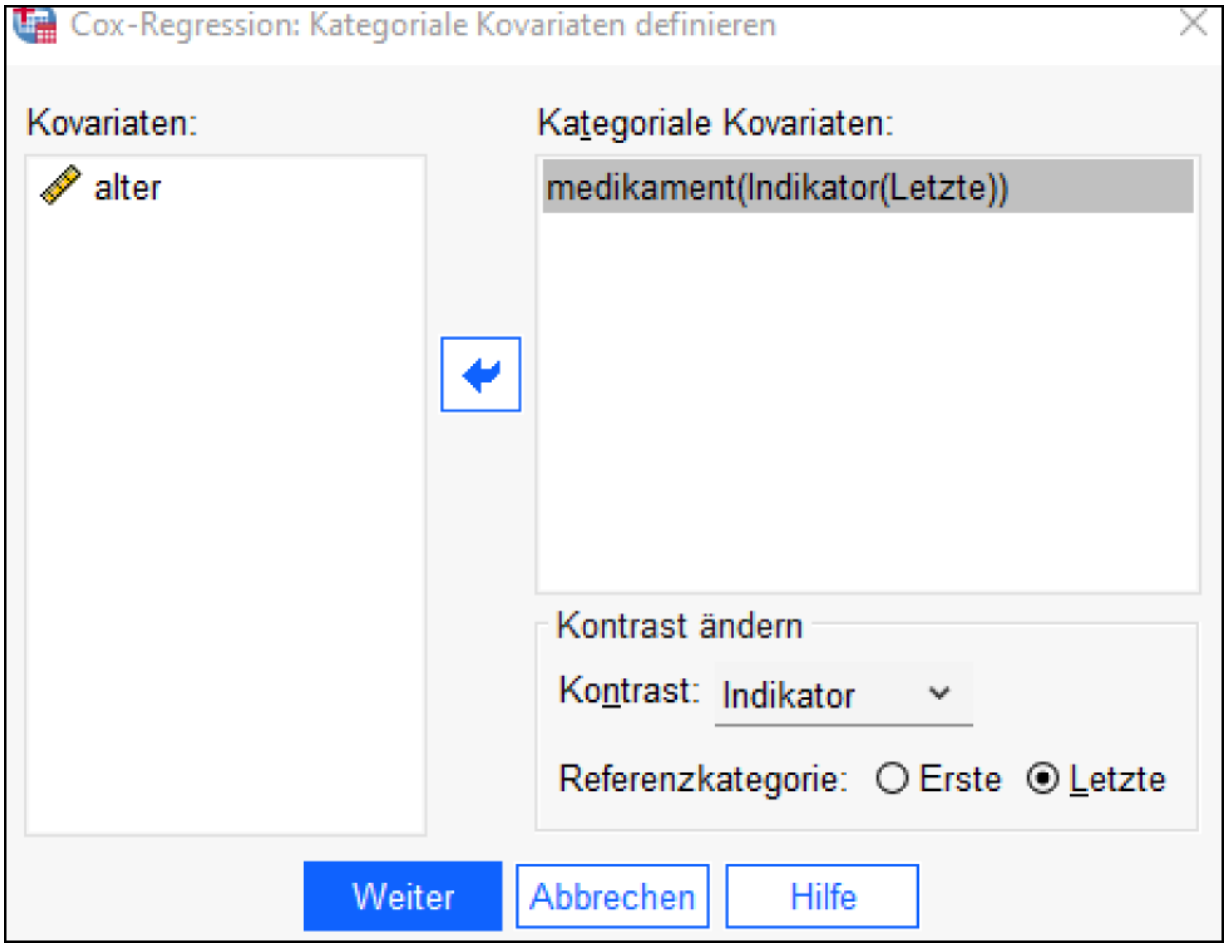



**Abbildung 12.23:** Aufruf einer Cox-Regression (*ÜBERLEBENSFUNKTION3.SAV*) (2)

Die zugehörige SPSS-Syntax lautet wie folgt:

```
COXREG überlebenszeit
    /STATUS=tod(1)
    /CONTRAST (medikament)=Indicator
    /METHOD=ENTER medikament alter
    /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

In unserem Beispielfall ergeben sich die nachfolgenden Ergebnisse (siehe Abbildung 12.24). Unten rechts stehen meines Erachtens die interessantesten Ergebnisse, nämlich die Koeffizientenwerte der (logistischen) Cox-Regression. Aus der letzten Spalte ist entsprechend abzulesen, dass im vorliegenden Beispiel die Überlebenswahrscheinlichkeit mit jedem zusätzlichen Lebensjahr um 9,8 Prozent (= 109,8 Prozent – 100 Prozent) steigt und dass sich die Überlebenswahrscheinlichkeit durch die Einnahme des betreffenden

Medikaments um sagenhafte 647,4 Prozent (= 747,4 Prozent – 100 Prozent) erhöht (gegenüber der Nichteinnahme dieses Medikaments).

**Auswertung der Fallverarbeitung**

| | | N | Prozent |
|---|---|---|---|
| Für Analyse verfügbare Fälle | Ereignis[a] | 12 | 60,0% |
| | Zensiert | 8 | 40,0% |
| | Insgesamt | 20 | 100,0% |
| Nicht verwendete Fälle | Fälle mit fehlenden Werten | 0 | 0,0% |
| | Fälle mit negativer Zeit | 0 | 0,0% |
| | Zensierte Fälle vor dem frühesten Ereignis in einer Schicht | 0 | 0,0% |
| | Insgesamt | 0 | 0,0% |
| Insgesamt | | 20 | 100,0% |

a. Abhängige Variable: überlebenszeit

**Codierungen für kategoriale Variablen[a]**

| | | Häufigkeit | (1)[c] |
|---|---|---|---|
| medikament[b] | 0 | 10 | 1 |
| | 1 | 10 | 0 |

a. Kategorie-Variable: medikament

b. Kodierung für Indikatorparameter

c. Die Variable (0,1) wurde umkodiert. Die Koeffizienten stimmen daher nicht mit den Koeffizienten für die Indikatorkodierung (0,1) überein.

**Block 0: Anfangsblock**

**Omnibus-Tests der Modellkoeffizienten**

| -2 Log-Likelihood |
|---|
| 57,646 |

**Block 1: Methode = Einschluß**

**Omnibus-Tests der Modellkoeffizienten[a]**

| -2 Log-Likelihood | Gesamt (Wert) | | | Änderung aus vorangegangenem Schritt | | | Änderung aus vorangegangenem Block | | |
|---|---|---|---|---|---|---|---|---|---|
| | Chi-Quadrat | df | Signifikanz | Chi-Quadrat | df | Signifikanz | Chi-Quadrat | df | Signifikanz |
| 46,043 | 10,581 | 2 | ,005 | 11,602 | 2 | ,003 | 11,602 | 2 | ,003 |

a. Beginnen mit Block-Nr. 1. Methode = Einschluß

**Variablen in der Gleichung**

| | B | SE | Wald | df | Signifikanz | Exp(B) |
|---|---|---|---|---|---|---|
| medikament | 2,011 | ,763 | 6,951 | 1 | ,008 | 7,474 |
| alter | ,094 | ,050 | 3,519 | 1 | ,061 | 1,098 |

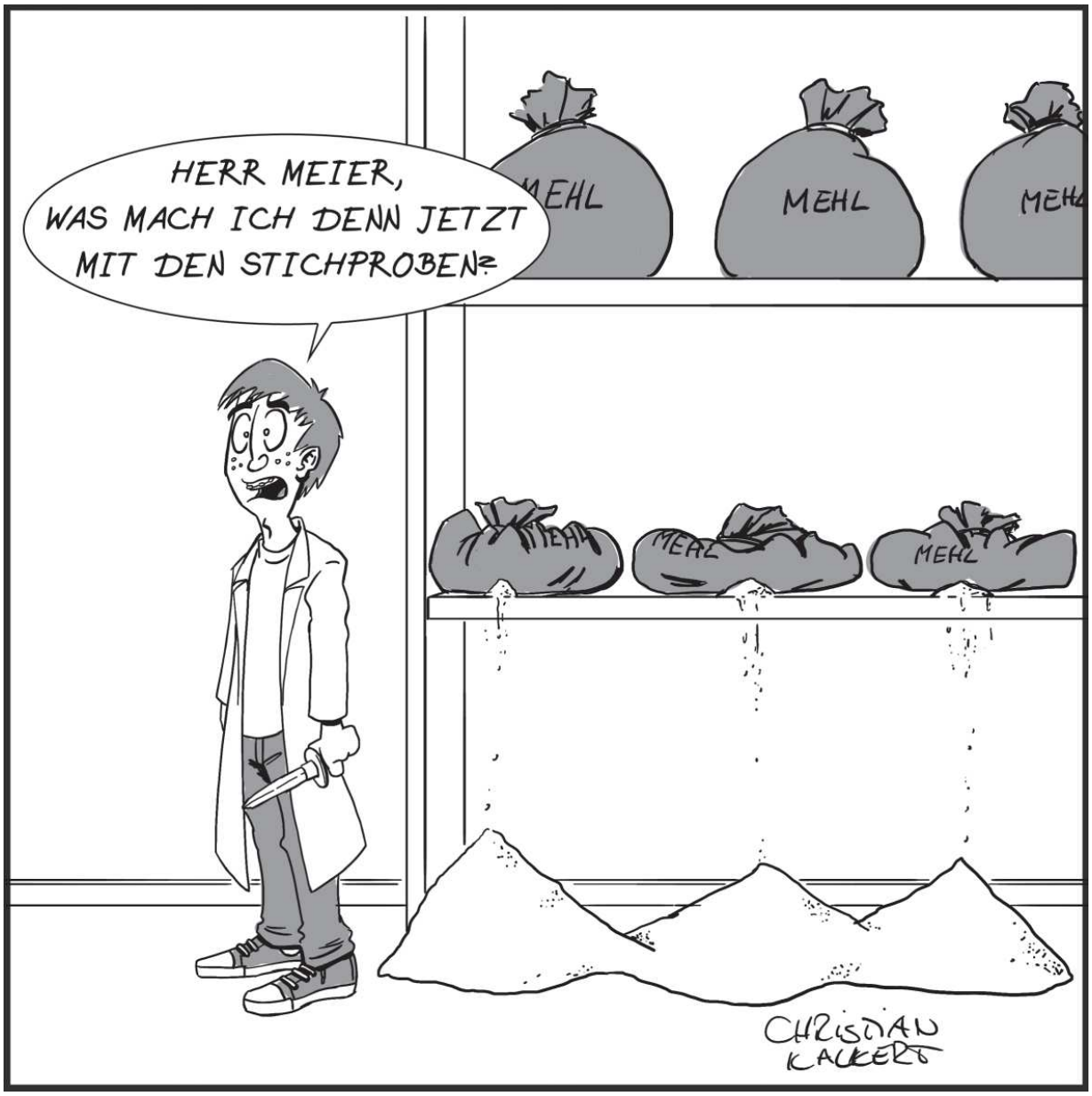
**Kovariaten-Mittelwerte**

| | Mittelwert |
|---|---|
| medikament | ,500 |
| alter | 30,400 |

**Abbildung 12.24:** Ergebnisse einer Cox-Regression (*ÜBERLEBENSFUNKTION3.SAV*)

# Teil IV

# Induktive SPSS-Statistiken



# IN DIESEM TEIL …

- ✔ Mittelwert- und Varianztests
- ✔ Verteilungstests
- ✔ Korrelationstests

# Kapitel 13

# Tests I: Welcher Mittelwert ist am größten?

**IN DIESEM KAPITEL**

Allgemeines zu statistischen Tests

Mittelwerttests für eine Variable

Mittelwerttests für zwei Gruppen

Mittelwerttests für mehr als zwei Gruppen

In Kapitel 13 werden nach einer kurzen, allgemein gehaltenen Einführung in die statistische Testtheorie verschiedene Mittelwerttests erklärt. Diese werden gemeinhin in Ein-, Zweistichprobentests und Tests für mehr als zwei Stichproben unterschieden. Eine für meinen Geschmack bessere Bezeichnung ist es allerdings, wenn anstelle von Stichproben- von Variablentests (beziehungsweise von Gruppentests) die Rede ist, wenn also der Wortbestandteil »Stichprobe« jeweils durch »Variable« (beziehungsweise auch durch »Gruppe«) ersetzt wird. (Über die in diesem Kapitel präsentierten Mittelwerttests hinaus könnten Sie im Übrigen in SPSS grundsätzlich auch zum Beispiel einen Einstichprobentest bezüglich der Streuung, das heißt hinsichtlich der Varianz, durchführen.)

Die im Folgenden behandelten Mittelwerttests beziehen sich alternativ auf das arithmetische Mittel oder den Median. Dabei spiegelt der *arithmetische Mittelwert* die Merkmalssumme dividiert durch die Fallzahl wider (zum Beispiel: Für die fünf Ausprägungen 1, 2, 4, 7 und 9 beträgt das arithmetische Mittel $\frac{1+2+4+7+9}{5}=4{,}6$).

Der *Median* kennzeichnet die genau in der Mitte einer (geordneten) Variablenverteilung liegende Merkmalsausprägung (zum Beispiel: Bei fünf Ausprägungen 1, 2, 4, 7 und 9 ist der Wert 4 der in der Mitte liegende Wert, also der Median).

## *Allgemeines*

**Statistische Tests** werden genutzt, um mit statistischen, besser: mit wahrscheinlichkeitstheoretischen »Bordmitteln«, bestimmte Fragestellungen zu beantworten. Hierbei werden zwei Hypothesen formuliert: die **Nullhypothese** und die konträre **Alternativhypothese.**

## *Null- und Alternativhypothese*

Um die Nullhypothese entweder zu verwerfen oder nicht zu verwerfen, benötigt man zum einen eine aus den beobachteten Daten konstruierte Kennziffer, die sogenannte **Teststatistik**, und zum anderen den sogenannten **kritischen Prüfbereich**. Für diesen kritischen Prüfbereich müssen eine

Irrtumswahrscheinlichkeit und eine Verteilungsannahme vorgegeben werden. Die Irrtumswahrscheinlichkeit bezieht sich dabei auf den Fehler, eine Nullhypothese abzulehnen, obwohl sie richtig ist. Mit diesen »Bordmitteln« folgt aus dem Vergleich von Test- und kritischer Prüfgröße als Ergebnis entweder die Ablehnung oder die Nichtablehnung der Nullhypothese.

Da Sie sich hier auf dem Gebiet der Wahrscheinlichkeitstheorie bewegen, können Sie, streng genommen, eine Nullhypothese nie annehmen im Sinne von »akzeptieren«. Sie können nämlich aufgrund Ihrer Testergebnisse nicht mit letzter Sicherheit sagen, ob die Nullhypothese tatsächlich zutrifft. Es bleibt immer ein gewisses Restrisiko im Sinne von »Zufall«, dass das Testergebnis fehlerhaft ist. Diesem Restrisiko wird durch die Vorgabe der genannten Irrtumswahrscheinlichkeit Rechnung getragen.

# Zweiseitiger Test

Bei den einzelnen Tests wird grundsätzlich zwischen zwei- und einseitigen statistischen Tests unterschieden. Dabei wird bei einem **zweiseitigen statistischen Test** die Nullhypothese $H_0$ so formuliert, dass für die interessierende Größe $\Delta$ (zum Beispiel – wie hier – für einen Mittelwert) ein konkreter Wert $\Delta_0$ angegeben wird, was dann mit der gegenteiligen Alternativhypothese $H_A$ kontrastiert wird:

$$H_0 : \Delta = \Delta_0 \text{ versus } H_A : \Delta \neq \Delta_0 .$$

Nochmals deutlich (und weniger technisch) formuliert: Die Abkürzung $\Delta$ ist hier in diesem Kapitel eine Abkürzung für Mittelwert (arithmetischer Mittelwert oder Median).

Fällt der Wert $\Delta_0$ in den **Nichtablehnungsbereich** des Tests, wird die Nullhypothese nicht verworfen; andernfalls wird sie verworfen. Im oberen Teil von Abbildung 13.1 ist – auf Basis der Standardnormalverteilung – bei einer vorgegebenen *Irrtumswahrscheinlichkeit* von $\alpha = 0{,}05$ der **Ablehnungsbereich** durch die Flächen am unteren (linken) und am oberen (rechten) Rand der Dichtefunktion (hier – wie erwähnt – bei der Standardnormalverteilung) dargestellt. Unterschreitet der beobachtete Wert, also die oben bereits erwähnte *Teststatistik*, hier als hinsichtlich Erwartungswert und Standardabweichung standardisierte Z-Variable dargestellt, den unteren kritischen Wert – das heißt hier: $z_{1-\frac{\alpha}{2}}$ – beziehungsweise überschreitet er den oberen kritischen Wert – das heißt hier: $+z_{1-\frac{\alpha}{2}}$, wird die Nullhypothese abgelehnt, weil dann der beobachtete z-Wert in den Ablehnungsbereich fällt.

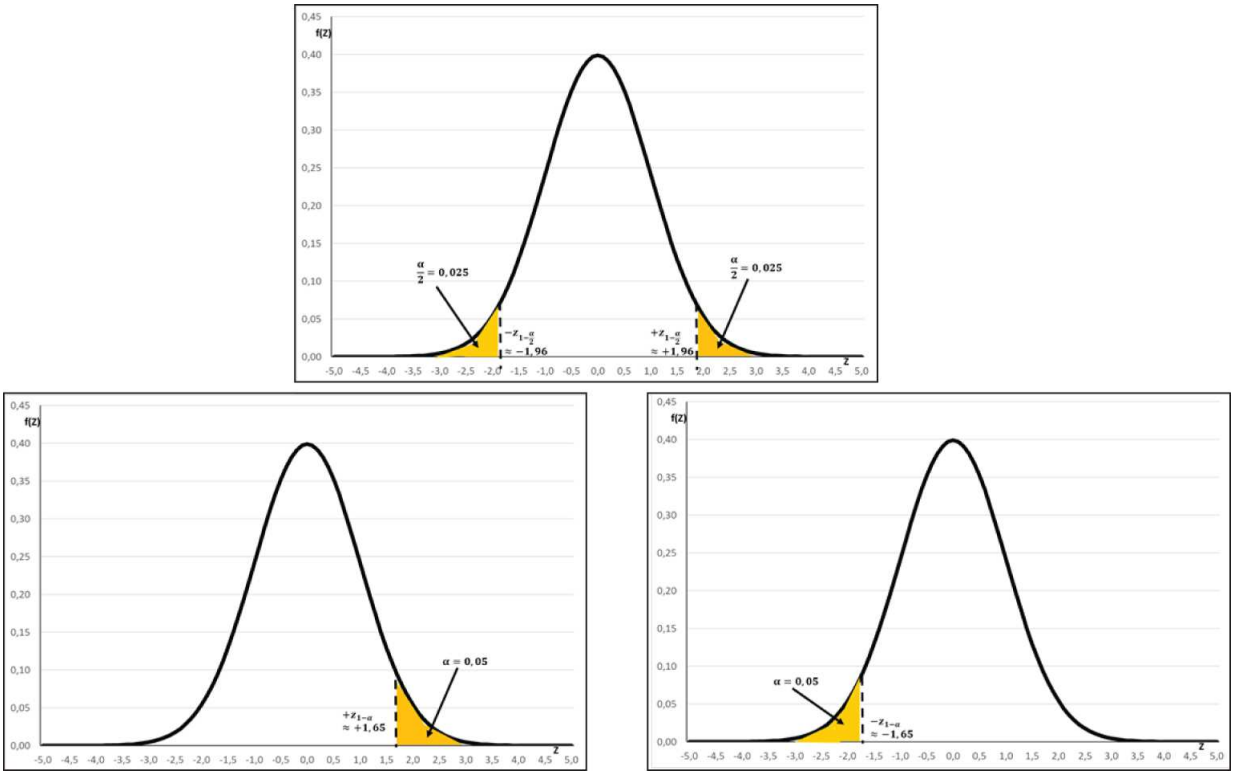



**Abbildung 13.1:** Beispiele für einen zweiseitigen Hypothesentest (oben) sowie für einen rechtsseitigen Hypothesentest (unten links) und einen linksseitigen Hypothesentest (unten rechts) (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 306–308)

# *Einseitiger Test*

Bei einem **einseitigen statistischen Test** lautet die Nullhypothese entweder $H_0$: $\Delta \leq \Delta_0$ (mit der Alternativhypothese $H_A$: $\Delta > \Delta_0$) oder $H_0$: $\Delta \geq \Delta_0$ (mit der Alternativhypothese $H_A$: $\Delta < \Delta_0$). Im erstgenannten Fall wird die Nullhypothese abgelehnt, wenn die beobachtete Größe größer als der kritische Wert ist (siehe unten links in Abbildung 13.1; sogenannter *rechtsseitiger statistischer Test*), im zweitgenannten Fall, wenn die beobachtete Größe kleiner als der kritische Wert ist (siehe unten rechts in Abbildung 13.1; sogenannter linksseitiger statistischer *Test*).

# *Schema für einen statistischen Test*

Grundsätzlich gibt es für statistische Tests ein **Grundschema**:

1. Formulierung einer Forschungshypothese als Alternativhypothese und Aufstellen der entgegengesetzten Nullhypothese;
2. Treffen bestimmter Annahmen: Verteilungsannahmen, Vorgabe einer Irrtumswahrscheinlichkeit (Fehler erster Art);
3. Bildung einer Teststatistik aus den beobachteten Werten der Zufallsvariablen;
4. Ermittlung des beziehungsweise der kritischen Werte;
5. Vergleich von Teststatistik und kritischem Wert beziehungsweise kritischen Werten;

6. Urteil, ob Nullhypothese zu verwerfen oder nicht zu verwerfen ist.

**Parametrische Tests** erfordern eine Verteilungsannahme (in der Regel die Normalverteilungsannahme), **nichtparametrische Tests** hingegen nicht (das heißt üblicherweise, die Normalverteilungsannahme ist nicht erforderlich). Nichtparametrische Tests gelten allerdings in der Regel als weniger effizient im Vergleich zu parametrischen Tests und sind zudem durch ein relativ kleines Methodenspektrum gekennzeichnet.

Bezüglich der für viele Fragestellungen »angesagten« Analyse von Mittelwertunterschieden können unter anderem die in Tabelle 13.1 dargestellten Tests durchgeführt werden.

| *Mindestskalenniveau der Daten* | *2 Gruppen: Unabhängige Stichproben* | *2 Gruppen: Abhängige Stichproben* | *Mehr als 2 Gruppen: Unabhängige Stichproben* | *Mehr als 2 Gruppen: Abhängige Stichproben* |
|---|---|---|---|---|
| Metrische Daten | Mittelwertdifferenzentest (t-Test) für unabhängige Stichproben beziehungsweise Welch-Test | Mittelwertdifferenzentest (t-Test) für abhängige Stichproben | Einfaktorielle ANOVA ohne Messwiederholungen | Einfaktorielle ANOVA mit Messwiederholungen |
| Ordinale Daten | Mann/Whitney-U-Test | Wilcoxon-Test beziehungsweise Vorzeichentest | Kruskal/Wallis-Test | Friedman-Test |

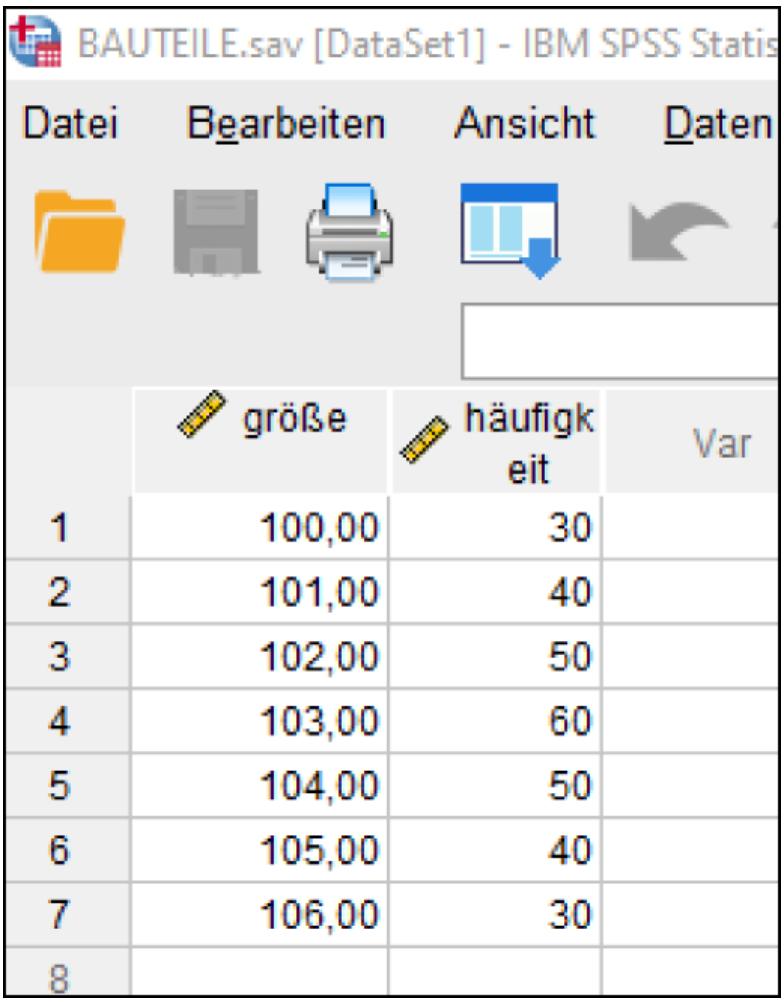
**Tabelle 13.1**: Tests zur Analyse von Unterschieden

Bei den t-Tests ist – wie aus Tabelle 13.1 hervorgeht – zwischen verbundenen (abhängigen) und unverbundenen (unabhängigen) Stichproben zu unterscheiden, wobei bezüglich der betreffenden Differenzvariablen üblicherweise die Normalverteilungsannahme getroffen wird, die ihrerseits vom SPSS-Anwender überprüft werden sollte. Allerdings kann man bei genügend großem (Gesamt-)Stichprobenumfang (üblicherweise reichen mindestens 30 Fälle hierfür) eine Normalverteilung sozusagen guten Gewissens unterstellen und damit den betreffenden t-Test durchführen.

Die nachfolgenden Testdarstellungen sind inspiriert von den Darlegungen auf der Website der Universität Zürich http://www.methodenberatung.uzh.ch/de.html (Aufruf am 22.09.2022).

# *Eine Variable: t-Test*

In diesem Abschnitt geht es jetzt aber erst einmal um einen **t-Test** für eine Variable. Aufbauend auf dem in Abbildung 13.2 angegebenen Datensatz *BAUTEILE.SAV* soll der gefundene arithmetische Mittelwert für die Größe eines bestimmten Bauteils (siehe Variable *größe*) auf statistische Aussagekraft (beziehungsweise mit einem Fachbegriff: auf statistische *Signifikanz*) hin überprüft werden.




| | größe | häufigkeit | Var |
|---|---|---|---|
| 1 | 100,00 | 30 | |
| 2 | 101,00 | 40 | |
| 3 | 102,00 | 50 | |
| 4 | 103,00 | 60 | |
| 5 | 104,00 | 50 | |
| 6 | 105,00 | 40 | |
| 7 | 106,00 | 30 | |
| 8 | | | |

**Abbildung 13.2:** Datendatei *BAUTEILE.SAV*

Da der Datensatz *BAUTEILE.SAV* offenkundig als Häufigkeitstabelle abgespeichert ist, müssen Sie diesen Datensatz über die Gewichtungsvariable *häufigkeit* noch in das erforderliche SPSS-Verarbeitungsformat bringen. Dies tun Sie entweder über das SPSS-Menü DATEN|FÄLLE GEWICHTEN oder über den SPSS-Befehl WEIGHT BY *häufigkeit*.

Anschließend berechnen Sie via ANALYSIEREN|DESKRIPTIVE STATISTIKEN|DESKRIPTIVE STATISTIK oder mittels des SPSS-Befehls DESCRIPTIVE den arithmetischen Mittelwert von *größe*. Es ergibt sich im vorliegenden Beispiel über die 300 Fälle hinweg ein arithmetischer Mittelwert von 103,00 Millimetern, und zwar bei einer Standardabweichung von 1,79184 Millimetern.

Diese Werte werden genutzt, um in SPSS die Teststatistik t zu berechnen (mit $\bar{X}$ als arithmetischem Mittelwert, $X_0$ als gemäß der Nullhypothese vorgegebenem Wert, $S_X$ als Standardabweichung von X und n als Stichprobengröße):

$$t = \frac{\bar{X} - X_0}{\frac{S_X}{\sqrt{n-1}}}.$$

## *Beispiel 1: Zweiseitiger Test*

Zunächst soll auf der Datenbasis von *BAUTEILE.SAV* ein zweiseitiger t-Test durchgeführt werden. Konkret soll geprüft werden, ob sich der oben genannte arithmetische Mittelwert von 103,00 Millimetern signifikant von einem früheren arithmetischen Mittelwert von 104,00 Millimetern unterscheidet. Hierzu wählen Sie im SPSS-Menü einfach ANALYSIEREN|MITTELWERTE UND PROPORTIONEN VERGLEICHEN|T-TEST BEI EINER STICHPROBE aus und geben in das sich dann öffnende Dialogfeld bei TESTVARIABLE(N) die hier relevante Variable *größe* ein; zusätzlich fügen Sie noch bei TESTWERT Ihren früheren arithmetischen Mittelwert von »104« ein (siehe Abbildung 13.3). Damit lautet Ihre Nullhypothese, dass die mittlere Größe eines Bauteils gleich 104 Millimetern ist; entsprechend formuliert Ihre Alternativhypothese, dass die durchschnittliche Größe eines Bauteils ungleich 104 Millimetern ist.

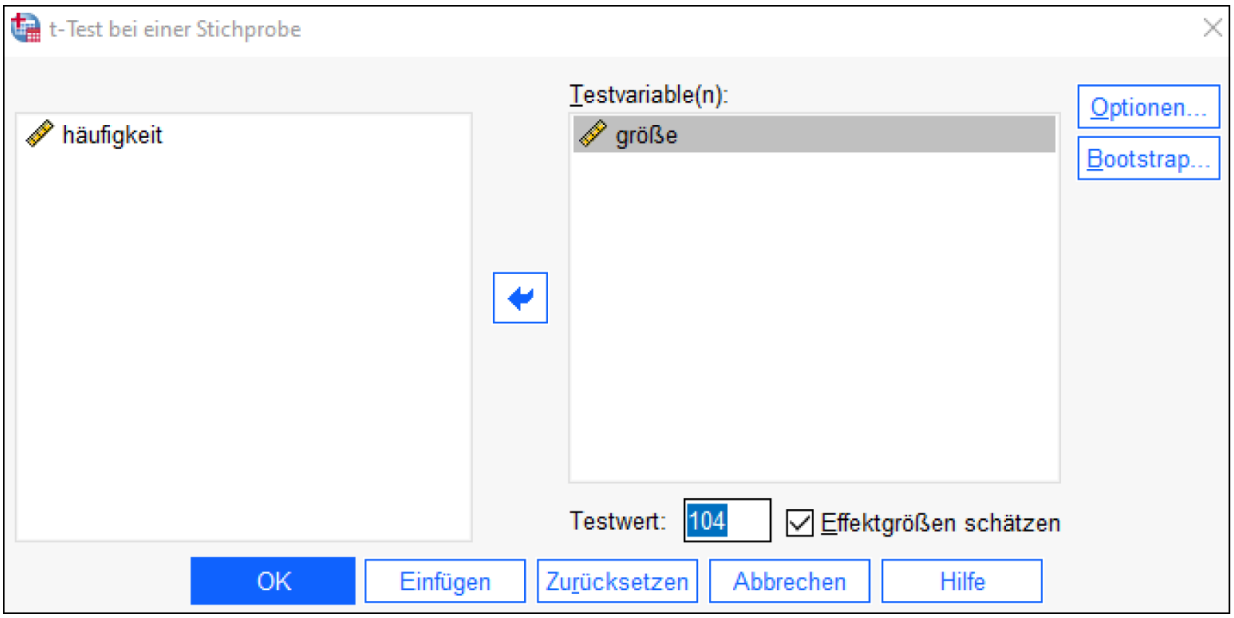



**Abbildung 13.3:** Anfordern eines t-Mittelwerttests für die Variable *größe* (*BAUTEILE.SAV*)

Die zugehörige SPSS-Syntax ist:

```
T-TEST
    /TESTVAL=104
    /MISSING=ANALYSIS
    /VARIABLES=größe
    /CRITERIA=CI(.95).
```

SPSS liefert Ihnen das in Abbildung 13.4 dargelegte Ergebnis. Aus der Spalte »Signifikanz: Zweiseitiges p« können Sie unmittelbar ablesen, dass die aktuelle mittlere Bauteile-Größe sich signifikant von der früheren unterscheidet. Die Irrtumswahrscheinlichkeit (beziehungsweise in alternativer Formulierung: die Überschreitungswahrscheinlichkeit) geht gegen 0, sodass die Nullhypothese gleich großer Bauteile hier also verworfen wird.

Es ist kein Zufall, dass Sie der vorstehende Test möglicherweise an die Signifikanzbeurteilungen der Regressionsparameter in Kapitel 9 erinnert. Dort lautet der Testwert für die einzelnen Regressionsparameter »0«. Es wird dort also geprüft, ob sich die Regressionsparameter signifikant von null unterscheiden.

### *Beispiel 2: Einseitiger Test (rechtsseitig)*

Die eben behandelte Fragestellung soll nunmehr etwas abgewandelt werden. Stellen Sie sich bitte vor, für die Bauteile sei eine Mindestgröße von 104 Millimetern vorgegeben. Dann lautet Ihre Alternativhypothese, dass *größe* kleiner als 104 ist, während die Nullhypothese ist, dass *größe* größer gleich 104 ist. Es handelt sich um einen einseitigen t-Test, genauer ausgedrückt: um einen rechtsseitigen t-Test.

**Test bei einer Stichprobe**

Testwert = 104

| | T | df | Signifikanz Einseitiges p | Signifikanz Zweiseitiges p | Mittlere Differenz | 95% Konfidenzintervall der Differenz Unterer Wert | 95% Konfidenzintervall der Differenz Oberer Wert |
|---|---|---|---|---|---|---|---|
| größe | -9,666 | 299 | <,001 | <,001 | -1,00000 | -1,2036 | -,7964 |

**Abbildung 13.4:** Ergebnis eines t-Mittelwerttests für *größe* (*BAUTEILE.SAV*) (Output-Auszug)

Sie können in diesem Beispiel Abbildung 13.4 zur Signifikanzbeurteilung verwenden. In unserem Fall steht in der Spalte »Signifikanz: Einseitiges p« als Angabe »< 0,001«, sodass Sie die genannte Nullhypothese ablehnen. Die ermittelte durchschnittliche Bauteilegröße ist demnach signifikant niedriger als die geforderte Mindestgröße.

# *Zwei Gruppen: Unabhängigkeit*

Eben wurde für eine Variable ein statistischer (t-)Test durchgeführt. Es gibt darüber hinaus weitere Testmöglichkeiten, zum Beispiel zwei Variablen beziehungsweise zwei Gruppen miteinander zu vergleichen. In diesem Sinne können zwei Gruppen statistischer Untersuchungseinheiten, also zum Beispiel zwei Personengruppen, hinsichtlich ihrer Mittelwerte miteinander verglichen werden. Sind die beiden miteinander verglichenen Gruppen unabhängig voneinander (zum Beispiel nicht jeweils dieselben Personen oder nicht miteinander verheiratete Personen), können Sie einen **unabhängigen Zweistichproben-Test** durchführen. Hierzu stehen alternativ ein t-Test und der Mann/Whitney-U-Test zur Auswahl (siehe gegebenenfalls nochmals Tabelle 13.1). Auf beide Tests wird im Folgenden eingegangen.

# *t-Test bei unabhängigen Stichproben*

Unabhängige Stichprobenbefunde können Sie über ANALYSIEREN|MITTELWERTE UND PROPORTIONEN VERGLEICHEN|T-TEST FÜR UNABHÄNGIGE STICHPROBEN nach Festlegung der Testvariablen und der Gruppenvariablen (mittels GRUPPEN DEFINIEREN) einem **t-Test** unterziehen. Der t-Test für unabhängige Stichproben bezieht sich also auf den Vergleich von Durchschnitten für zwei Gruppen.

Die **Voraussetzungen** dieses t-Tests sind:

- ✔ metrische Skalierung der für den Test relevanten Variablen
- ✔ eine weitere Variable zur Bildung von zwei Gruppen
- ✔ Normalverteilung der untersuchten Merkmale (aber Anwendung des zentralen Grenzwertsatzes bei einem Stichprobenumfang größer gleich 30 Fälle möglich!)

- ✔ Varianzhomogenität in der Grundgesamtheit (hier Bezugnahme auf den sogenannten Levene-Test)
- ✔ Unabhängigkeit der Messwerte untereinander

Grundsätzlich wird bei diesem t-Test folgende Teststatistik gebildet (mit: $\bar{X}_1 - \bar{X}_2$ als Differenz der arithmetischen Mittelwerte der Gruppen 1 und 2, $n_1$: Fallzahl der Gruppe 1, $n_2$: Fallzahl der Gruppe 2, df: Anzahl der Freiheitsgrade; wegen der Abgrenzung zur deskriptiven Statistik in Teil II des Buchs werden in Teil IV bei der induktiven Statistik die Fallzahlen nicht mehr mit einem großen, sondern mit einem kleinen n dargelegt):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\text{Standardfehler Mittelwertdifferenz}}, df = n_1 + n_2 - 2.$$

Hierbei bestimmt sich der Standardfehler der Mittelwertdifferenz wie folgt:

- ✔ bei bekannten Grundgesamtheitsvarianzen (mit: $\sigma_1^2$ als Grundgesamtheitsvarianz der Gruppe 1, $\sigma_2^2$ als Grundgesamtheitsvarianz der Gruppe 2):

$$\hat{\sigma}_{X_1 - X_2} = \sqrt[2]{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}};$$

- ✔ bei unbekannten Grundgesamtheitsvarianzen (mit: $S_1^2$ als Stichprobenvarianz der Gruppe 1, $S_2^2$ als Stichprobenvarianz der Gruppe 2):

$$\hat{\sigma}_{X_1 - X_2} = \sqrt[2]{\frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}} \cdot \sqrt[2]{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

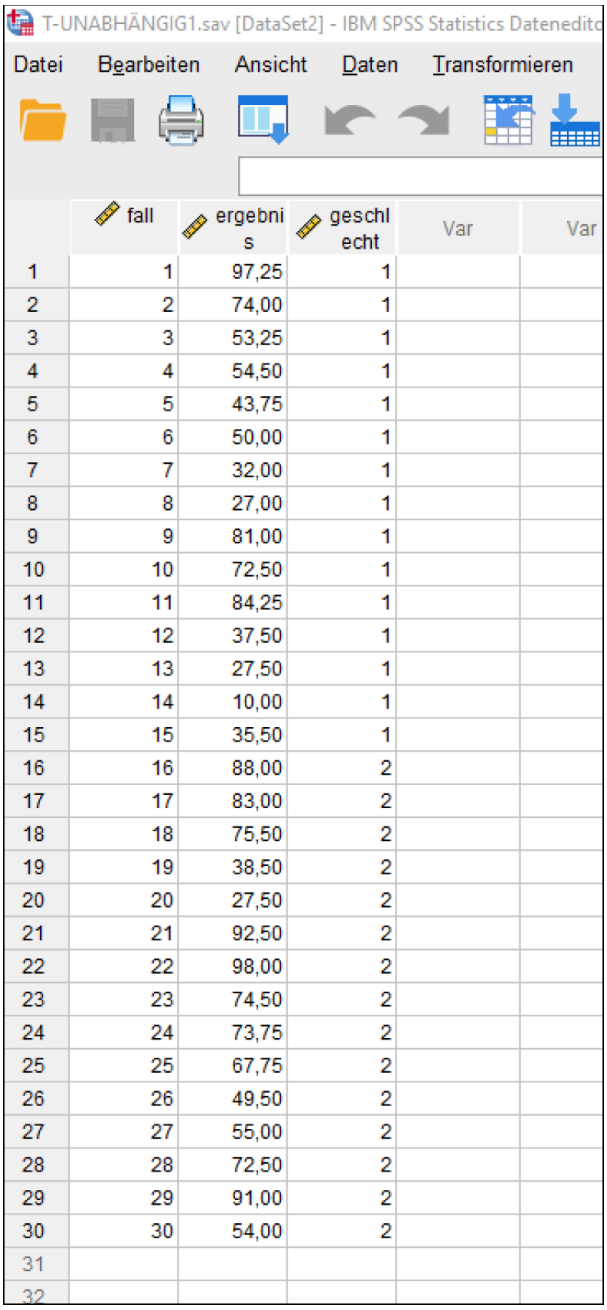
Wenn $|t| > |t_{krit}|$ ist (also wenn der Betrag des t-Testwerts größer als der Betrag des kritischen t-Werts ist), wird die Nullhypothese $H_0$ verworfen.

Der skizzierte t-Test wird nachfolgend auf Basis der Datendatei *T-UNABHÄNGIG1.SAV* exemplarisch erläutert. In dieser Datei befinden sich in der Variablen *ergebnis* die Wurfweiten (mit einem Ball) von Frauen und Männern. Nach Frauen und Männern wird in der Variablen *geschlecht* unterschieden, und zwar mit 1 = Frauen und 2 = Männer (siehe Abbildung 13.5).

Die Frage lautet nun, ob sich die durchschnittlichen Wurfweiten zwischen Frauen und Männern statistisch-signifikant voneinander unterscheiden. Hierzu rufen Sie, wie bekannt, ANALYSIEREN|MITTELWERTE UND PROPORTIONEN VERGLEICHEN|T-TEST BEI UNABHÄNGIGEN STICHPROBEN auf und tragen als Testvariable *ergebnis* sowie als Gruppenvariable *geschlecht* mit den beiden Ausprägungen 1 und 2 ein (siehe Abbildung 13.6) und klicken danach auf OK.

Die SPSS-Syntax sieht an dieser Stelle so aus:

```
T-TEST GROUPS=geschlecht(1 2)
    /MISSING=ANALYSIS
    /VARIABLES=ergebnis
    /CRITERIA=CI(.95).
```




| | fall | ergebnis | geschlecht | Var | Var |
|---|---|---|---|---|---|
| 1 | 1 | 97,25 | 1 | | |
| 2 | 2 | 74,00 | 1 | | |
| 3 | 3 | 53,25 | 1 | | |
| 4 | 4 | 54,50 | 1 | | |
| 5 | 5 | 43,75 | 1 | | |
| 6 | 6 | 50,00 | 1 | | |
| 7 | 7 | 32,00 | 1 | | |
| 8 | 8 | 27,00 | 1 | | |
| 9 | 9 | 81,00 | 1 | | |
| 10 | 10 | 72,50 | 1 | | |
| 11 | 11 | 84,25 | 1 | | |
| 12 | 12 | 37,50 | 1 | | |
| 13 | 13 | 27,50 | 1 | | |
| 14 | 14 | 10,00 | 1 | | |
| 15 | 15 | 35,50 | 1 | | |
| 16 | 16 | 88,00 | 2 | | |
| 17 | 17 | 83,00 | 2 | | |
| 18 | 18 | 75,50 | 2 | | |
| 19 | 19 | 38,50 | 2 | | |
| 20 | 20 | 27,50 | 2 | | |
| 21 | 21 | 92,50 | 2 | | |
| 22 | 22 | 98,00 | 2 | | |
| 23 | 23 | 74,50 | 2 | | |
| 24 | 24 | 73,75 | 2 | | |
| 25 | 25 | 67,75 | 2 | | |
| 26 | 26 | 49,50 | 2 | | |
| 27 | 27 | 55,00 | 2 | | |
| 28 | 28 | 72,50 | 2 | | |
| 29 | 29 | 91,00 | 2 | | |
| 30 | 30 | 54,00 | 2 | | |
| 31 | | | | | |
| 32 | | | | | |

**Abbildung 13.5:** Datendatei *T-UNABHÄNGIG1.SAV*

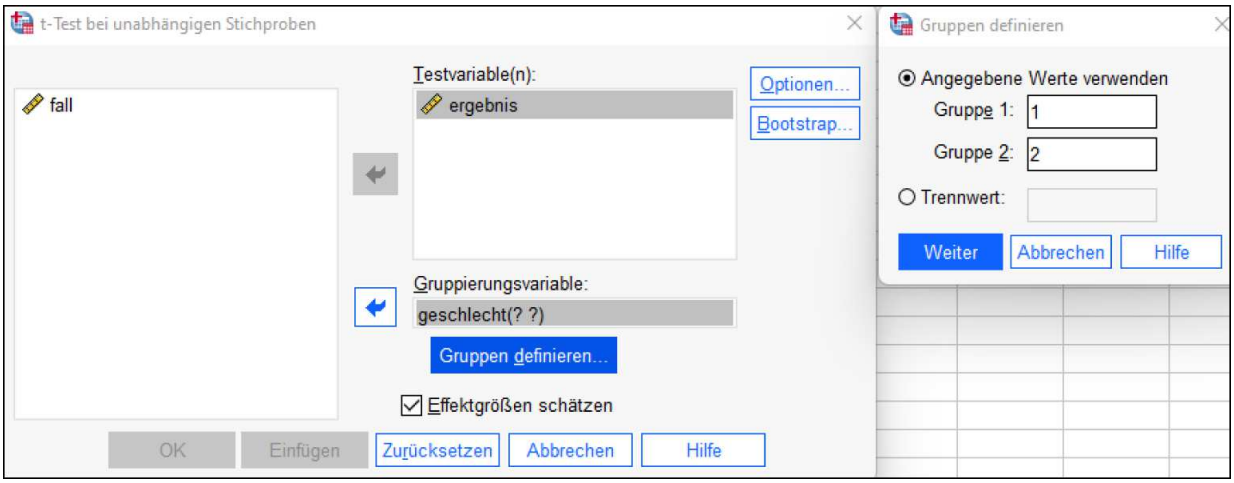



**Abbildung 13.6:** Aufruf des t-Tests bei unabhängigen Stichproben mit *ergebnis* als Testvariable und mit *geschlecht* als Gruppierungsvariable (*T-UNABHÄNGIG1.SAV*)

Wenn Sie diese Syntax »laufen« lassen, erhalten Sie Abbildung 13.7 als Ergebnis. Zunächst erkennen Sie, dass die Wurfweite der Männer mit 69,40 Metern im Durchschnitt größer als diejenige der Frauen mit 52,00 Metern ist. Wenn Sie danach auf die eigentlichen Testergebnisse blicken, müssen Sie zunächst einmal die erste Zeile »Varianzen sind gleich« betrachten, da die Varianzhomogenität zwischen den Gruppen eine wichtige Voraussetzung für die Durchführung des hier diskutierten t-Tests ist.

SPSS führt an dieser Stelle den sogenannten **Levene-Test** aus. Beim Levene-Test lautet die Nullhypothese: »Die Varianzen sind gleich.« Wird diese Nullhypothese nicht verworfen, können Sie sich in der ersten Zeile der betreffenden Pivottabelle auf die Spalten mit der Überschrift »T-Test für die Mittelwertgleichheit« stürzen. In unserem Fall ergibt sich eine Irrtumswahrscheinlichkeit im Sinne einer Überschreitungswahrscheinlichkeit von 41,4 Prozent, die deutlich darauf hinweist, dass die genannte Nullhypothese nicht zu verwerfen ist. Wäre dies nicht der Fall gewesen, hätten Sie Ihre t-Test-Interpretation auf Basis der zweiten Zeile der betreffenden Pivottabelle durchführen müssen (von SPSS dann dort automatisch mit dem sogenannten **Welch-Test** bewerkstelligt).

Schauen Sie nun in den hinteren Teil der ersten Ergebniszeile, erhalten Sie als Signifikanzangabe für den hier relevanten Mittelwertunterschied eine Irrtums-/Überschreitungswahrscheinlichkeit von 4,7 Prozent, die »gerade so« unterhalb der verbreiteten Irrtumswahrscheinlichkeit von 5 Prozent liegt. Sie können diesen Befund in Abbildung 13.7 so interpretieren, dass ein (einigermaßen) signifikanter Mittelwertunterschied in den Wurfweiten von Frauen und Männern im Beispielfall vorliegt (in dem Sinne, dass in unserem Beispiel Männer im Durchschnitt weiter als Frauen werfen).

**Gruppenstatistiken**

| | geschlecht | N | Mittelwert | Std.-Abweichung | Standardfehler des Mittelwertes |
|---|---|---|---|---|---|
| ergebnis | 1 | 15 | 52,0000 | 25,04924 | 6,46769 |
| | 2 | 15 | 69,4000 | 20,69243 | 5,34276 |

**Test bei unabhängigen Stichproben**

| | | Levene-Test der Varianzgleichheit | | t-Test für die Mittelwertgleichheit | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Signifikanz | | Mittlere Differenz | Differenz für Standardfehler | 95% Konfidenzintervall der Differenz | |
| | | F | Sig. | T | df | Einseitiges p | Zweiseitiges p | | | Unterer Wert | Oberer Wert |
| ergebnis | Varianzen sind gleich | ,687 | ,414 | -2,074 | 28 | ,024 | ,047 | -17,40000 | 8,38904 | -34,58418 | -,21582 |
| | Varianzen sind nicht gleich | | | -2,074 | 27,036 | ,024 | ,048 | -17,40000 | 8,38904 | -34,61181 | -,18819 |

**Abbildung 13.7:** Ergebnis des t-Tests bei unabhängigen Stichproben mit *ergebnis* als Testvariable und mit *geschlecht* als Gruppierungsvariable (*T-UNABHÄNGIG1.SAV*) (Output-Auszug)

Sie hätten den **Welch-Test** im Übrigen auch direkt aufrufen können (wenn Sie sich ziemlich sicher sind, dass Varianzheterogenität vorliegt), und zwar über ANALYSIEREN|MITTELWERTE VERGLEICHEN UND PROPORTIONEN|EINFAKTORIELLE VARIANZANALYSE mit den Einträgen von *ergebnis* bei ABHÄNGIGE VARIABLEN und *geschlecht* bei FAKTOR sowie anschließend bei OPTIONEN durch Setzen eines Häkchens bei WELCH (siehe Abbildung 13.8 zum Ergebnis des Welch-Tests). Dies geht mit folgender SPSS-Syntax einher:

```
ONEWAY ergebnis BY geschlecht
    /STATISTICS WELCH
    /MISSING ANALYSIS.
```

**ANOVA**

ergebnis

| | Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|
| Zwischen den Gruppen | 2270,700 | 1 | 2270,700 | 4,302 | ,047 |
| Innerhalb der Gruppen | 14778,975 | 28 | 527,821 | | |
| Gesamt | 17049,675 | 29 | | | |

**Robuste Testverfahren zur Prüfung auf Gleichheit der Mittelwerte**

ergebnis

| | Statistik[a] | df1 | df2 | Sig. |
|---|---|---|---|---|
| Welch | 4,302 | 1 | 27,036 | ,048 |

a. Asymptotisch F-verteilt

**Abbildung 13.8:** Welch-Test bei unabhängigen Stichproben mit *ergebnis* als Testvariable und mit *geschlecht* als Gruppierungsvariable (*T-UNABHÄNGIG1.SAV*)

## *Mann/Whitney-U-Test bei unabhängigen Stichproben*

Wenn die Voraussetzungen des vorstehend behandelten t-Tests nicht erfüllt sind, kommt als Alternative der **Mann/Whitney-U-Test** infrage. Er testet für unabhängige Stichproben bezüglich der jeweiligen sogenannten zentralen Tendenz (= Median). Dieser nichtparametrische Test – die Daten müssen also nicht normalverteilt sein – setzt lediglich ordinales Skalenniveau der Daten voraus. Ein Beispiel sind Schulnoten-Unterschiede zwischen Personen mit und ohne Migrationshintergrund.

Die **Voraussetzungen** des Mann/Whitney-U-Tests sind:

- ✔ mindestens Ordinalskalenniveau der für den Test relevanten Variablen
- ✔ eine zusätzliche Variable zur Bildung von zwei Gruppen

Der Test basiert auf der Rangordnung der Daten. Konkret werden die Daten durch Ränge ersetzt:

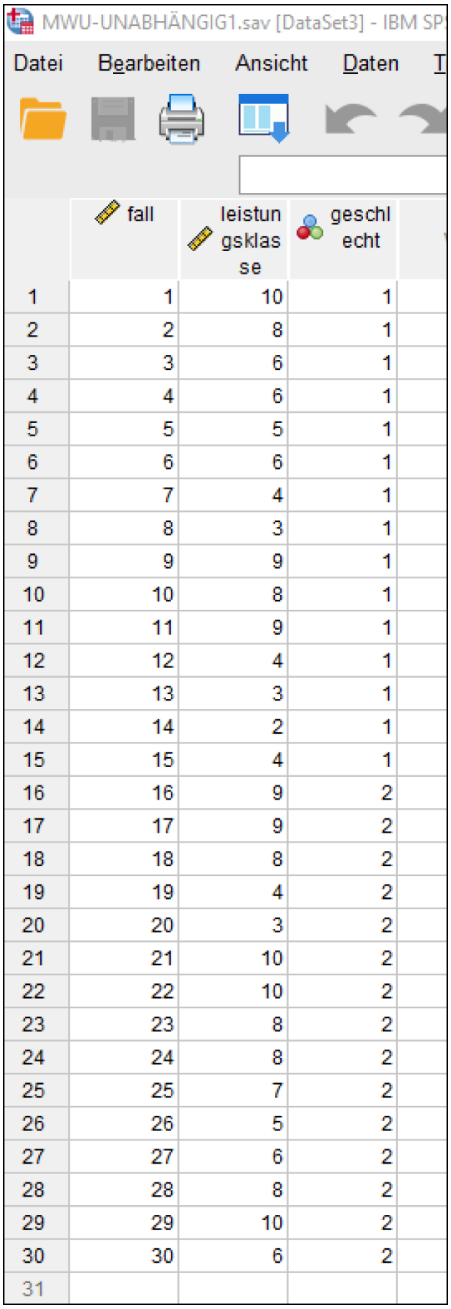
$$U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1.$$

Dabei verweist der Index 1 auf die größere Rangsumme R und der Index 2 auf die kleinere Rangsumme R. Mit $n_1$ ist demnach die Fallzahl der Gruppe mit der größeren Rangsumme und mit $n_2$ die Fallzahl der Gruppe mit der kleineren Rangsumme gemeint. $R_1$ (die Rangsumme von Gruppe 1) ist in der vorstehenden Formel also die größere Rangsumme.

Wenn $n_1 + n_2 \geq 30$ gilt, erhalten Sie als Teststatistik (mit: μ als Erwartungswert beziehungsweise arithmetisches Mittel, σ als Standardabweichung, $n_1$: Fallzahl der Gruppe mit der größeren Rangsumme, $n_2$: Fallzahl der Gruppe mit der kleineren Rangsumme):

$$u = \frac{U - \mu_U}{\sigma_U} = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}}.$$

Nachstehend wird ein Beispiel zum Mann/Whitney-U-Test auf Basis der Datendatei *MWU-UNABHÄNGIG1.SAV* erläutert (siehe Abbildung 13.9). In dieser Datei sind für verschiedene Personen in der Variablen *leistungsklasse* verschiedene Leistungsklassen, zum Beispiel wieder in Bezug auf die Weite beim Ballwurf, angegeben, die grundsätzlich von 1 (niedrigste Leistungsklasse) bis 10 (höchste Leistungsklasse) »gehen«. Es können zwei Gruppen gebildet werden: Frauen (*geschlecht* = 1) und Männer (*geschlecht* = 2).




| | fall | leistungsklasse | geschlecht |
|---|---|---|---|
| 1 | 1 | 10 | 1 |
| 2 | 2 | 8 | 1 |
| 3 | 3 | 6 | 1 |
| 4 | 4 | 6 | 1 |
| 5 | 5 | 5 | 1 |
| 6 | 6 | 6 | 1 |
| 7 | 7 | 4 | 1 |
| 8 | 8 | 3 | 1 |
| 9 | 9 | 9 | 1 |
| 10 | 10 | 8 | 1 |
| 11 | 11 | 9 | 1 |
| 12 | 12 | 4 | 1 |
| 13 | 13 | 3 | 1 |
| 14 | 14 | 2 | 1 |
| 15 | 15 | 4 | 1 |
| 16 | 16 | 9 | 2 |
| 17 | 17 | 9 | 2 |
| 18 | 18 | 8 | 2 |
| 19 | 19 | 4 | 2 |
| 20 | 20 | 3 | 2 |
| 21 | 21 | 10 | 2 |
| 22 | 22 | 10 | 2 |
| 23 | 23 | 8 | 2 |
| 24 | 24 | 8 | 2 |
| 25 | 25 | 7 | 2 |
| 26 | 26 | 5 | 2 |
| 27 | 27 | 6 | 2 |
| 28 | 28 | 8 | 2 |
| 29 | 29 | 10 | 2 |
| 30 | 30 | 6 | 2 |
| 31 | | | |

**Abbildung 13.9:** Datendatei *MWU-UNABHÄNGIG1.SAV*

Über das SPSS-Menü können Sie diesen Test wie folgt durchführen: ANALYSIEREN|NICHT PARAMETRISCHE TESTS|KLASSISCHE DIALOGFELDER|2 UNABHÄNGIGE STICHPROBEN. Sie müssen im sich dann öffnenden Dialogfenster bei MANN-WHITNEY-U-TEST ein Häkchen setzen. (Alternativ wären Sie im Übrigen zum Mann/Whitney-U-Test für zwei Stichproben auch über die Menüabfolge ANALYSIEREN|NICHT PARAMETRISCHE TESTS|KLASSISCHE DIALOGFELDER|K UNABHÄNGIGE STICHPROBEN gekommen. Auch hier führen also wieder mehrere Wege zum Ziel …)

Zusätzlich müssen Sie jetzt die Testvariable (hier: *leistungsklasse*) und zum anderen die Gruppierungsvariable (hier: *geschlecht*) auswählen, für *geschlecht* die beiden Gruppen 1 (»1«)

und 2 (»2«). Sie müssen anschließend in den relevanten Dialogfeldern auf AUSFÜHREN und OK klicken, und schon rattert der SPSS-Motor los (siehe hierzu Abbildung 13.10).

Alternativ können Sie folgende SPSS-Syntax zum »Laufen« bringen:

```
NPAR TESTS
    /M-W = leistungsklasse BY geschlecht (1 2)
    /MISSING ANALYSIS.
```

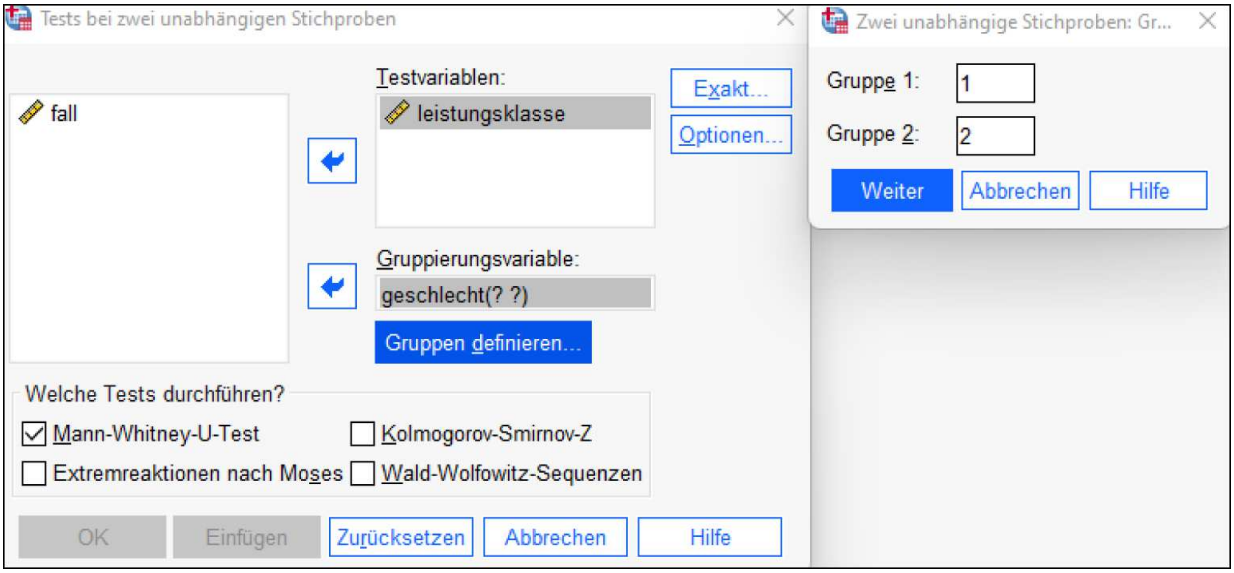



**Abbildung 13.10:** Eingabe der Testvariablen (*leistungsklasse*) und der Gruppierungsvariablen (*geschlecht*) beim Mann/Whitney-U-Test (*MWU-UNABHÄNGIG1.SAV*)

Der SPSS-Output sieht dann wie in Abbildung 13.11 aus. SPSS liefert uns hier als Ergebnis, dass die Nullhypothese einer gleichen mittleren Leistungsstärke zwischen Männern und Frauen nicht abgelehnt wird; zumindest gilt dies bis zu einer Grenze der Irrtumswahrscheinlichkeit bei 5 Prozent (da die – für den vorgenommenen einseitigen Test – berechnete Irrtumswahrscheinlichkeit hier bei 8,9 Prozent liegt, ab der erst die Nullhypothese abgelehnt wird).

## *Zwei Gruppen: Abhängigkeit*

Sind die beiden miteinander verglichenen Gruppen abhängig voneinander (zum Beispiel für zwei Zeitpunkte jeweils dieselben Personen oder zu einem bestimmten Zeitpunkt miteinander verheiratete Personen), können Sie im Rahmen eines Mittelwertvergleichs einen abhängigen Zweistichproben-Test durchführen. Alternativ existieren hier als Testvarianten ein t-Test, der Wilcoxon-Test und der Vorzeichentest (siehe gegebenenfalls nochmals Tabelle 13.1). Im Folgenden werden diese drei Tests diskutiert.

**Mann-Whitney-Test**

**Ränge**

| | geschlecht | N | Mittlerer Rang | Rangsumme |
|---|---|---|---|---|
| leistungsklasse | 1 | 15 | 12,77 | 191,50 |
| | 2 | 15 | 18,23 | 273,50 |
| | Gesamt | 30 | | |

**Teststatistiken[a]**

| | leistungsklasse |
|---|---|
| Mann-Whitney-U-Test | 71,500 |
| Wilcoxon-W | 191,500 |
| Z | -1,718 |
| Asymp. Sig. (2-seitig) | ,086 |
| Exakte Sig. [2*(1-seitige Sig.)] | ,089[b] |

a. Gruppenvariable: geschlecht

b. Nicht für Bindungen korrigiert.

Abbildung 13.11: Ergebnis des Mann/Whitney-U-Tests mit der Testvariablen *leistungsstärke* und der Gruppierungsvariablen *geschlecht* (*MWU-UNABHÄNGIG1.SAV*)

# *t-Test bei abhängigen Stichproben*

Dieser **t-Test** führt einen Mittelwertvergleich für zwei abhängige Stichproben durch. Diese Abhängigkeiten können sich zum Beispiel – wie bereits erwähnt – über Messwiederholungen (das heißt für identische Untersuchungseinheiten) oder etwa über Paarbeziehungen ergeben. Die **Voraussetzungen** für diesen Test sind:

- ✔ metrische Skalierung der für den Test relevanten Variablen
- ✔ zwei verbundene Stichproben oder Gruppen, innerhalb derer allerdings die einzelnen Messwerte unabhängig voneinander sind

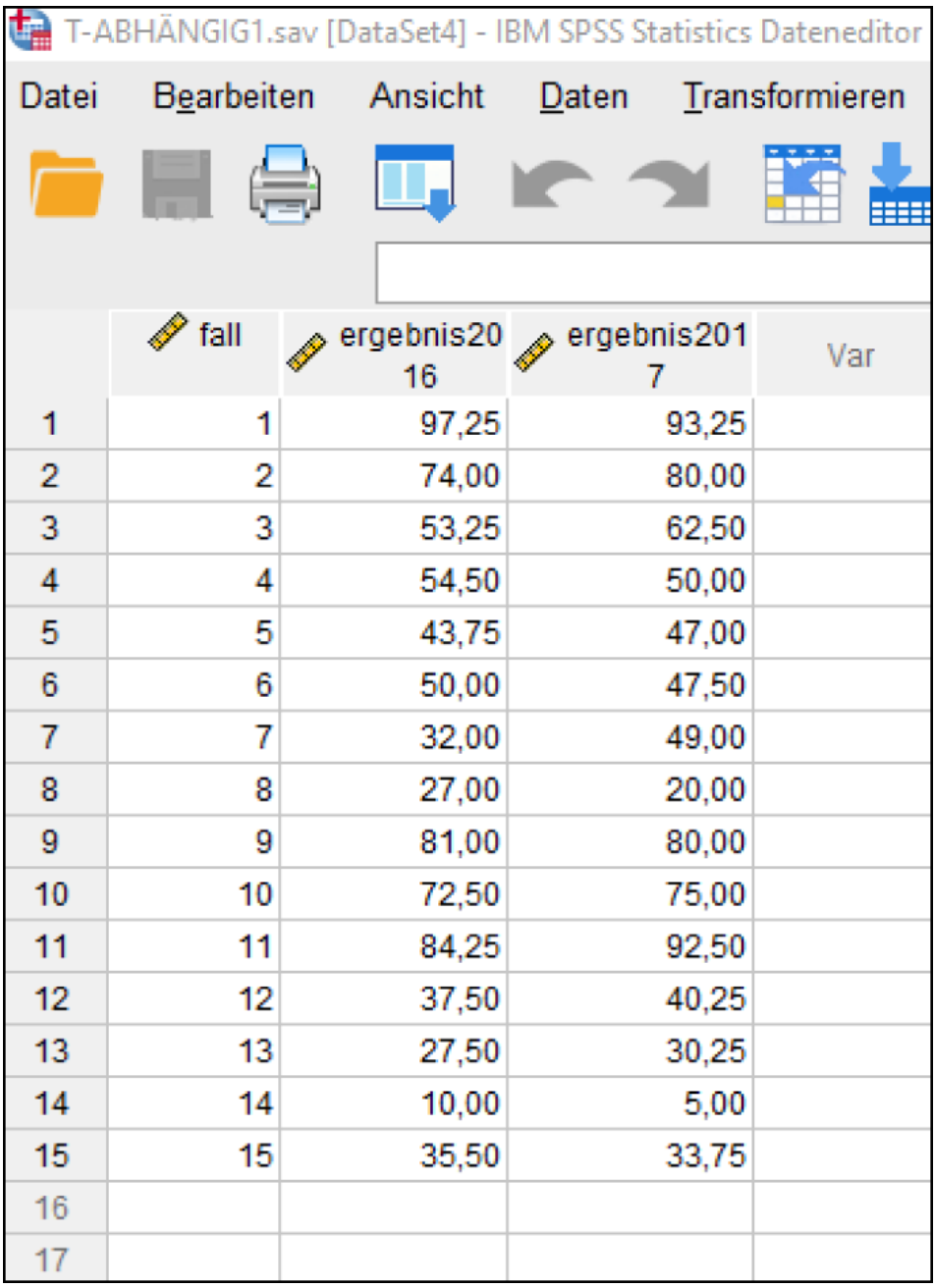
- ✔ Normalverteilung der Unterschiede zwischen den verbundenen Stichproben in der Grundgesamtheit (aber: Annahme einer näherungsweise Normalverteiltheit bei mindestens 30 Fällen möglich)

Als Teststatistik erhalten Sie (mit $\overline{X_d}$ als Unterschied zwischen den beiden arithmetischen Mittelwerten, S als Abkürzung für die Standardabweichung und n als Stichprobengröße):

$$t = \frac{\overline{X}_d}{\hat{S}_{\overline{X_d}}} = \frac{\overline{X_d}}{\sqrt{\frac{\sum_{i=1}^{n}(d_i - X_d)^2}{(n-1)\cdot n}}}.$$

Wenn $|t| > |t_{krit}|$ ist, wird die Nullhypothese verworfen. Dann liegen signifikante Unterschiede in den Durchschnittswerten zwischen den beiden miteinander verglichenen Gruppen vor.

Das folgende Beispiel basiert auf der Datendatei *T-ABHÄNGIG1.SAV*. In dieser Datendatei sind die Ballwurfweiten für zwei identische Personengruppen in zwei (früheren) Jahren (2016 und 2017) dargestellt (in den Variablen *ergebnis2016* und *ergebnis2017*). Die erste Person zum Beispiel hat 2016 97,25 Meter und 2017 93,25 Meter weit geworfen (also ganz schön weit), die zweite Person hat 2016 74,00 Meter und 2017 80,00 Meter weit geworfen und so weiter (siehe Abbildung 13.12).




| | fall | ergebnis2016 | ergebnis2017 | Var |
|---|---|---|---|---|
| 1 | 1 | 97,25 | 93,25 | |
| 2 | 2 | 74,00 | 80,00 | |
| 3 | 3 | 53,25 | 62,50 | |
| 4 | 4 | 54,50 | 50,00 | |
| 5 | 5 | 43,75 | 47,00 | |
| 6 | 6 | 50,00 | 47,50 | |
| 7 | 7 | 32,00 | 49,00 | |
| 8 | 8 | 27,00 | 20,00 | |
| 9 | 9 | 81,00 | 80,00 | |
| 10 | 10 | 72,50 | 75,00 | |
| 11 | 11 | 84,25 | 92,50 | |
| 12 | 12 | 37,50 | 40,25 | |
| 13 | 13 | 27,50 | 30,25 | |
| 14 | 14 | 10,00 | 5,00 | |
| 15 | 15 | 35,50 | 33,75 | |
| 16 | | | | |
| 17 | | | | |

**Abbildung 13.12:** Datendatei *T-ABHÄNGIG1.SAV*

Es interessiert nun, ob sich die mittlere Wurfweite 2016 von jener des Jahrs 2017 signifikant unterscheidet. Hierzu steuern Sie im SPSS-Menü diesen Test über ANALYSIEREN|MITTELWERTE UND PROPORTIONEN VERGLEICHEN|T-TEST BEI STICHPROBEN MIT PAARIGEN WERTEN an und geben in dem sich öffnenden Dialogfeld bei VARIABLE1 *ergebnis2016* und bei VARIABLE2 *ergebnis2017* ein (siehe Abbildung 13.13). Anschließend starten Sie die Ausführung dieses t-Tests durch Klicken auf die OK-Taste.

Als SPSS-Syntax müssen Sie an dieser Stelle im Grunde genommen nur schreiben:

```
T-TEST PAIRS=ergebnis2016 WITH ergebnis2017 (PAIRED)
    /CRITERIA=CI(.9500)
```

```
/MISSING=ANALYSIS.
```

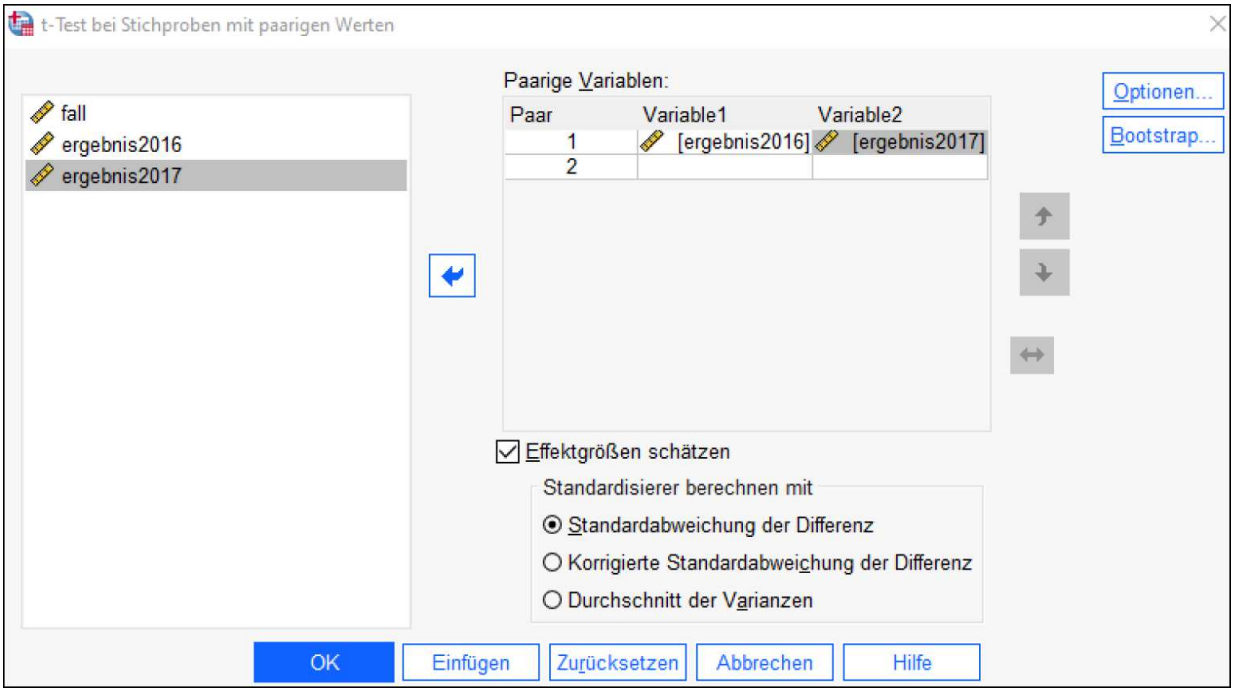



**Abbildung 13.13:** Aufruf des t-Tests bei abhängigen Stichproben mit *ergebnis2016* und *ergebnis2017* als »paarigen Variablen« (*T-ABHÄNGIG1.SAV*)

Bei entsprechendem Aufruf des hier relevanten t-Tests gewinnt man die in Abbildung 13.14 wiedergegebenen Befunde. Das eigentliche Testergebnis steht dabei in der betreffenden Abbildung ganz unten. Aus der letzten Spalte der entsprechenden Pivottabelle ist ersichtlich, dass die mittleren Unterschiede in der Wurfweite bei den einzelnen Personen im Vergleich der Jahre 2016 und 2017 nicht signifikant sind, da der in dieser Spalte angegebene Wahrscheinlichkeitswert von 31,7 Prozent relativ groß ist (und deutlich über der vielfach verwendeten, noch akzeptierten Obergrenze von 5 Prozent liegt). Die Nullhypothese gleicher mittlerer Wurfweiten wird also nicht verworfen.

Dass ein solches Testergebnis letztlich herauskommt, hätte man sich schon anhand der beiden oberen Pivottabellen in Abbildung 13.14 denken können: Zum einen ist der Mittelwertunterschied (52,00 versus 53,73 Meter) nicht allzu groß, und zum anderen beträgt die Korrelation (*Bravais/Pearson-Korrelationskoeffizient*) von *ergebnis2016* und *ergebnis2017* +0,970 und ist damit recht hoch. Dieser hohe Korrelationswert verweist auf eine große statistische »Gemeinsamkeit« beider Variablen.

**Statistik bei gepaarten Stichproben**

| | | Mittelwert | N | Std.-Abweichung | Standardfehler des Mittelwertes |
|---|---|---|---|---|---|
| Paaren 1 | ergebnis2016 | 52,0000 | 15 | 25,04924 | 6,46769 |
| | ergebnis2017 | 53,7333 | 15 | 26,37563 | 6,81016 |

**Korrelationen bei gepaarten Stichproben**

| | | N | Korrelation | Signifikanz Einseitiges p | Signifikanz Zweiseitiges p |
|---|---|---|---|---|---|
| Paaren 1 | ergebnis2016 & ergebnis2017 | 15 | ,970 | <,001 | <,001 |

**Test bei gepaarten Stichproben**

| | | Gepaarte Differenzen | | | | | | | Signifikanz | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Mittelwert | Std.-Abweichung | Standardfehler des Mittelwertes | 95% Konfidenzintervall der Differenz Unterer Wert | Oberer Wert | T | df | Einseitiges p | Zweiseitiges p |
| Paaren 1 | ergebnis2016 - ergebnis2017 | -1,73333 | 6,47003 | 1,67055 | -5,31632 | 1,84965 | -1,038 | 14 | ,159 | ,317 |

**Abbildung 13.14:** Ergebnis des t-Tests bei abhängigen Stichproben mit *ergebnis2016* und *ergebnis2017* als »paarigen Variablen« (*T-ABHÄNGIG1.SAV*)

## *Wilcoxon-Test*

Beim **Wilcoxon-Test** werden für abhängige Stichproben Unterschiede in den zentralen Tendenzen (das heißt bei den Medianen) ermittelt. Der Wilcoxon-Test wird dann angewandt, wenn die Voraussetzungen für den t-Test bei abhängigen Stichproben nicht erfüllt sind. Er ist ein nichtparametrischer Test.

Die **Voraussetzungen** des Wilcoxon-Tests sind:

- ✔ mindestens Ordinalskalenniveau der für den Test relevanten Variablen
- ✔ zwei verbundene Stichproben oder Gruppen, allerdings: voneinander unabhängige Messwertpaare innerhalb der Gruppen

Beim Wilcoxon-Test werden die Differenzen der konkreten Messwerte durch die Differenzen der Ränge ersetzt (mit n als von 0 verschiedenen Paardifferenzen; $T_+$ steht für die positiven und $T_-$ entsprechend für die negativen rangbezogenen Paardifferenzen), und es besteht folgender Zusammenhang:

$$T_+ + T_- = \frac{n \cdot (n+1)}{2}.$$

Als Teststatistik wird dann der kleinere Wert aus dem Vergleich von $T_+$ beziehungsweise von $T_-$ verwendet:

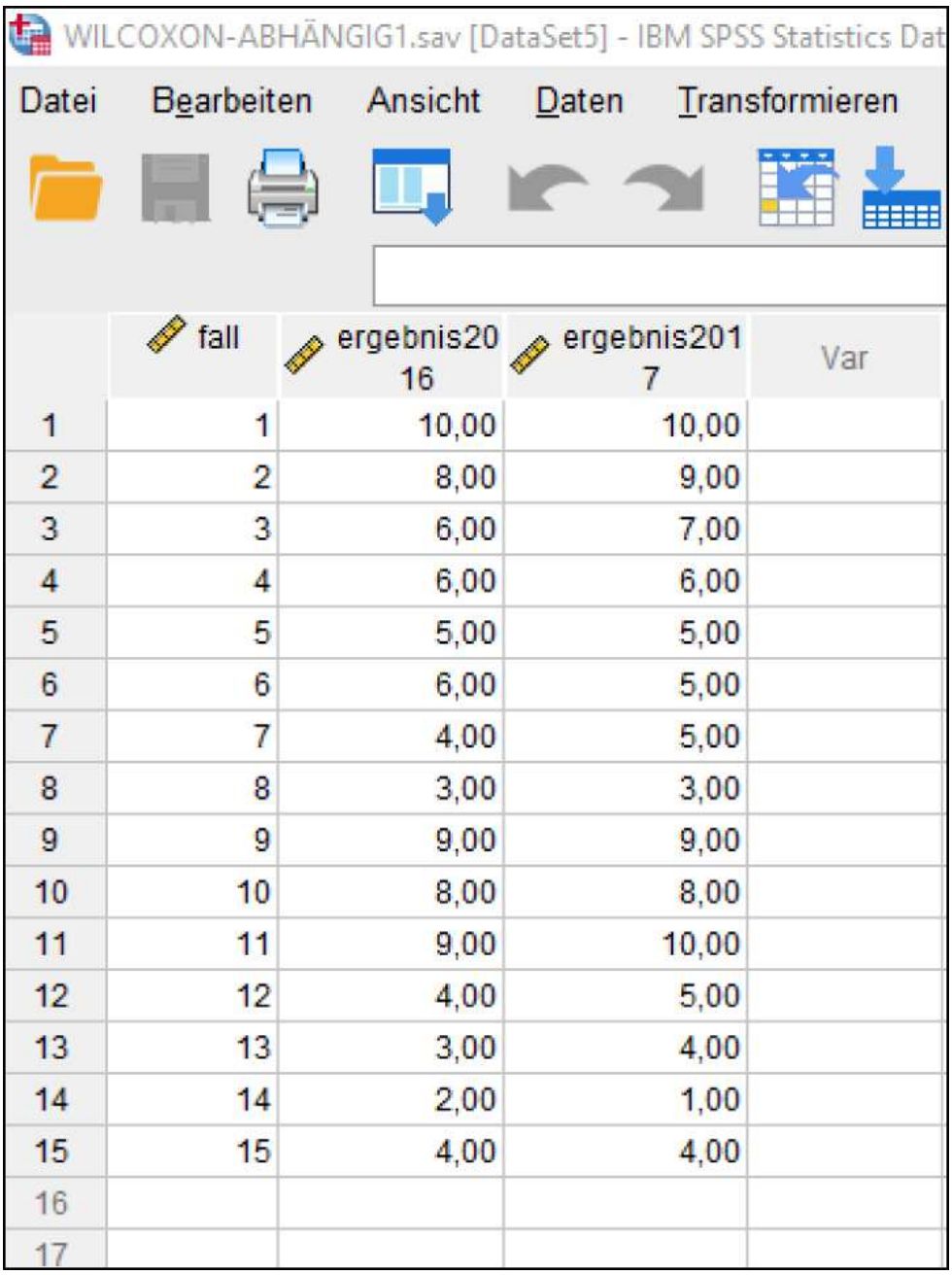
$$W = \min(T_+, T_-).$$

Wenn es keine Unterschiede gibt, ist der Erwartungswert für W: $\mu_W = \frac{n \cdot (n+1)}{4}$. Für $n \geq 30$ gilt (mit: $\mu$ als Erwartungswert und $\sigma$ als Standardabweichung):

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{W - \frac{n \cdot (n+1)}{4}}{\sqrt{\frac{n \cdot (n+1) \cdot (2 \cdot n+1)}{24}}}.$$

Als Nächstes können Sie z mit $z_{krit}$ vergleichen; Sie führen dann einen näherungsweisen Vergleich mit tabellierten Werten der Standardnormalverteilung durch. Wenn Sie die z-Transformation nicht durchführen dürfen (also für n kleiner 30), müssen Sie die kritischen W-Werte in einer Tabelle nachschlagen.

Auch beim Wilcoxon-Test möchte ich Ihnen die entsprechende Konzeption spielerisch mit einem Beispiel näherbringen, hier: auf Basis der Datendatei *WILCOXON-ABHÄNGIG1.SAV*. In dieser Datei sind in den Variablen *ergebnis2016* und *ergebnis2017* Leistungsklassen angegeben, die prinzipiell von 1 bis 10 »gehen« (siehe Abbildung 13.15).




| | fall | ergebnis2016 | ergebnis2017 | Var |
|---|---|---|---|---|
| 1 | 1 | 10,00 | 10,00 | |
| 2 | 2 | 8,00 | 9,00 | |
| 3 | 3 | 6,00 | 7,00 | |
| 4 | 4 | 6,00 | 6,00 | |
| 5 | 5 | 5,00 | 5,00 | |
| 6 | 6 | 6,00 | 5,00 | |
| 7 | 7 | 4,00 | 5,00 | |
| 8 | 8 | 3,00 | 3,00 | |
| 9 | 9 | 9,00 | 9,00 | |
| 10 | 10 | 8,00 | 8,00 | |
| 11 | 11 | 9,00 | 10,00 | |
| 12 | 12 | 4,00 | 5,00 | |
| 13 | 13 | 3,00 | 4,00 | |
| 14 | 14 | 2,00 | 1,00 | |
| 15 | 15 | 4,00 | 4,00 | |
| 16 | | | | |
| 17 | | | | |

**Abbildung 13.15:** Datendatei *WILCOXON-ABHÄNGIG1.SAV*

Im SPSS-Menü kommen Sie wie folgt zum Wilcoxon-Test: ANALYSIEREN|NICHT PARAMETRISCHE TESTS|KLASSISCHE DIALOGFELDER|ZWEI VERBUNDENE STICHPROBEN. Danach wählen Sie den Wilcoxon-Test aus und geben im nächsten Schritt *ergebnis2016* und *ergebnis2017* als Testvariablen ein (bei TESTPAARE.). Anschließend klicken Sie auf OK, und los geht die SPSS-Ausführung (siehe Abbildung 13.16)!

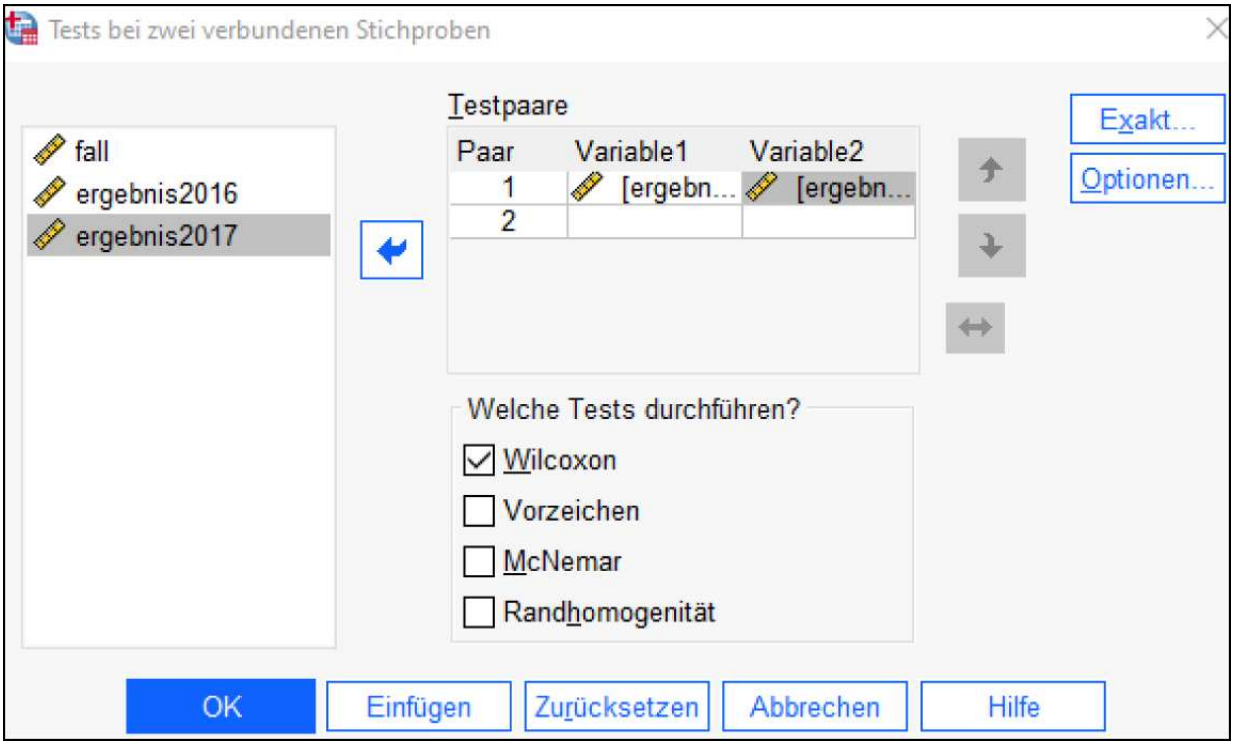



**Abbildung 13.16:** Aufruf des Wilcoxon-Tests bei abhängigen Stichproben mit *ergebnis2016* und *ergebnis2017* als Testvariablen (*WILCOXON-ABHÄNGIG1.SAV*)

Die zugehörige SPSS-Syntax lautet:

```
NPAR TESTS
    /WILCOXON = ergebnis2016 WITH ergebnis2017 (PAIRED)
    /MISSING ANALYSIS.
```

Wie aus Abbildung 13.17 hervorgeht, lehnt SPSS im vorliegenden Beispiel die Nullhypothese gleicher Mittelwerte nicht ab (weil »Asymp. Sig. (2-seitig)« = 0,157 und damit > 0,05!), sodass hier davon ausgegangen werden kann, dass sich die Wurfleistungen der einzelnen Personen (im Mittelwert) in den Jahren 2016 und 2017 nicht signifikant voneinander unterscheiden.

**Wilcoxon-Test**

**Ränge**

| | | N | Mittlerer Rang | Rangsumme |
|---|---|---|---|---|
| ergebnis2017 - ergebnis2016 | Negative Ränge | 2[a] | 4,50 | 9,00 |
| | Positive Ränge | 6[b] | 4,50 | 27,00 |
| | Bindungen | 7[c] | | |
| | Gesamt | 15 | | |

a. ergebnis2017 < ergebnis2016

b. ergebnis2017 > ergebnis2016

c. ergebnis2017 = ergebnis2016

**Teststatistiken[a]**

| | ergebnis2017 - ergebnis2016 |
|---|---|
| Z | -1,414[b] |
| Asymp. Sig. (2-seitig) | ,157 |

a. Wilcoxon-Test

b. Basiert auf negativen Rängen.

Abbildung 13.17: Ergebnis des Wilcoxon-Tests bei abhängigen Stichproben mit *ergebnis2016* und *ergebnis2017* als Testvariablen (*WILCOXON-ABHÄNGIG1.SAV*)

## *Vorzeichentest*

Eine Alternative zum Wilcoxon-Test ist der **Vorzeichentest**. Bei Letzterem werden die Differenzen in »positiv«, »negativ« und »verbunden« unterschieden. Wie gleich gezeigt wird, wird bei der Teststatistik des Vorzeichentests letztlich nur auf die positiven Paardifferenzen Bezug genommen; insoweit ist der Vorzeichentest einfacher angelegt als der eben vorgestellte Wilcoxon-Test. Auch der Vorzeichentest kann angewandt werden, wenn die Voraussetzungen des t-Tests bei abhängigen Stichproben nicht erfüllt sind.

Die Differenzen werden – wie gerade erwähnt – nach »positiv«, »negativ« und »verbunden« eingeteilt. Falls beide Variablen ähnlich verteilt sind, wird auf keine signifikanten Unterschiede zwischen positiven und negativen Differenzen gefolgert.

Die **Voraussetzungen** sind (analog zum Wilcoxon-Test):

- ✔ mindestens Ordinalskalenniveau der für den Test relevanten Variablen
- ✔ zwei verbundene Stichproben oder Gruppen

Die Anzahl der Differenzen wird beim Vorzeichentest als Teststatistik verwendet (wobei beim Vorliegen von Ties – also bei mehreren gleichen Rängen innerhalb einer Gruppe – eine Anpassung

der Teststatistik erfolgt); konkret wird auf Signifikanz mittels eines sogenannten Binomialtests geschlossen, und zwar bei n größer gleich 30 (mit einer näherungsweise standardnormalverteilten z-Größe):

$$z = \frac{|2 \cdot T_+ - n| - 1}{\sqrt[2]{n}}.$$

In der vorstehenden Formel symbolisiert $T_+$ die Anzahl der positiven Paardifferenzen und n die gesamte Fallzahl. Bei n < 30 erfolgt in SPSS ein exakter Signifikanztest.

Im SPSS-Menü kommen Sie ähnlich dem Wilcoxon-Test zum Vorzeichentest: ANALYSIEREN|NICHTPARAMETRISCHE TESTS|KLASSISCHE DIALOGFELDER|ZWEI VERBUNDENE STICHPROBEN. Auf der sich öffnenden Dialogkarte wählen Sie anstelle des Wilcoxon-Tests wie in Abbildung 13.16 hier jetzt den Vorzeichentest aus und bei VARIABLEN wieder die Variablen *ergebnis2016* und *ergebnis2017* aus der Datendatei *WILCOXON-ABHÄNGIG1.SAV* als Testvariablen.

Die SPSS-Syntax lautet:

```
NPAR TESTS
    /SIGN = ergebnis2016 WITH ergebnis2017 (PAIRED)
    /MISSING ANALYSIS.
```

Im vorliegenden Beispiel kommt das gleiche Ergebnis wie beim Wilcoxon-Test heraus: Die Nullhypothese, dass es keine mittleren Unterschiede zwischen 2016 und 2017 gibt, wird nicht verworfen, weil »Exakte Sig. (2-seitig)« = 0,289 und damit > 0,05 ist (siehe Abbildung 13.18).

## *Mehr als zwei Gruppen: Unabhängigkeit*

Bis hierhin haben Sie Mittelwerttestverfahren für maximal zwei Gruppen kennengelernt. In diesem Abschnitt wird die Betrachtung auf Testverfahren für mindestens drei Gruppen erweitert. Dargestellt werden im Folgenden zum einen die einfache Varianzanalyse und zum anderen der Testansatz nach Kruskal/Wallis.

## *Einfache Varianzanalyse ohne Messwiederholung*

Im Folgenden geht es zunächst um die **einfache Varianzanalyse ohne Messwiederholung**, das heißt: bei unabhängigen Gruppen. Sie heißt auch *einfaktorielle ANOVA* (mit: ANOVA = Analysis of Variance). Hierbei werden die arithmetischen Mittelwerte mehrerer (> 2) unabhängiger Gruppen miteinander verglichen, wobei durch eine bestimmte kategoriale Variable die jeweilige Gruppe definiert wird. Diese (Gruppierungs-)Variable heißt auch Faktor. Die einfaktorielle ANOVA ist eine Verallgemeinerung des t-Tests, wie Sie sicherlich gleich leicht erkennen.

## Vorzeichentest

**Häufigkeiten**

| | | N |
|---|---|---|
| ergebnis2017 - ergebnis2016 | Negative Differenzen[a] | 2 |
| | Positive Differenzen[b] | 6 |
| | Bindungen[c] | 7 |
| | Gesamt | 15 |

a. ergebnis2017 < ergebnis2016

b. ergebnis2017 > ergebnis2016

c. ergebnis2017 = ergebnis2016

**Teststatistiken[a]**

| | ergebnis2017 - ergebnis2016 |
|---|---|
| Exakte Sig. (2-seitig) | ,289[b] |

a. Vorzeichentest

b. Verwendetete Binomialverteilung.

**Abbildung 13.18:** Ergebnis des Vorzeichentests bei abhängigen Stichproben mit *ergebnis2016* und *ergebnis2017* als Testvariablen (*WILCOXON-ABHÄNGIG1.SAV*)

Inhaltlich geht es bei der einfaktoriellen ANOVA um die Varianzzerlegung der abhängigen Variablen in die Varianz innerhalb der Gruppen und in die Varianz zwischen den Gruppen.

Die **Voraussetzungen** der einfaktoriellen ANOVA sind:

- ✔ metrische Skalierung der für den Test relevanten Variablen
- ✔ ein kategorialer Faktor
- ✔ Unabhängigkeit der Gruppen voneinander

- ✔ grundsätzlich: Normalverteilung der abhängigen Variablen in den Gruppen (aber bei $n \geq 30$ Fälle über den Zentralen Grenzwertsatz (asymptotische) Normalverteilung grundsätzlich unterstellbar)
- ✔ Varianzhomogenität der Gruppen in der Grundgesamtheit (*Levene-Test*!)

Die sogenannte F-Testgröße ist:

$$F_{Modell} = \frac{MS_{zwischen}}{MS_{innerhalb}}$$

$$\text{mit: } MS_{zwischen} = \frac{\sum_{g=1}^{G} K_g \cdot \left(\overline{X_g} - \overline{X}\right)^2}{G-1}$$

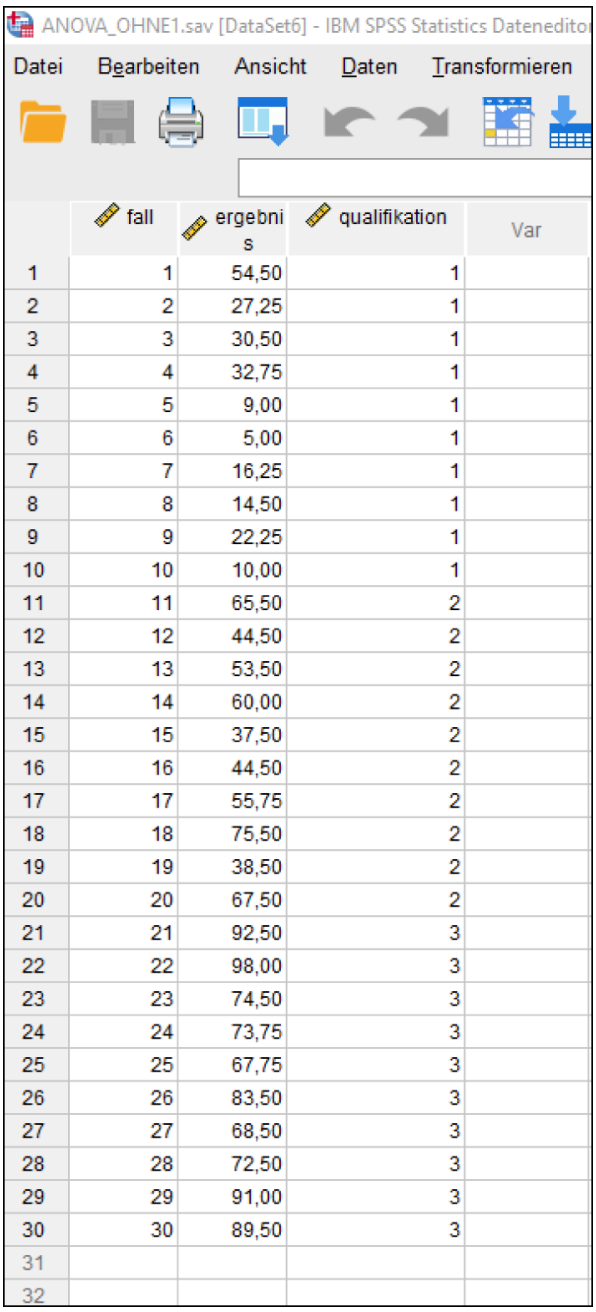
$$\text{und } MS_{innerhalb} = \frac{\sum_{g=1}^{G}\sum_{k=1}^{K}\left(X_{gk} - \overline{X_g}\right)^2}{K_{total} - G}.$$

Vorstehend kennzeichnet G die Anzahl der Gruppen, $K_g$ die Anzahl der jeweiligen Gruppenmitglieder, $K_{total}$ die Anzahl aller statistischen Einheiten (hier: Personen), $\overline{X}$ den arithmetischen Mittelwert über alle Personen und alle Gruppen hinweg, $\overline{X_g}$ den arithmetischen Mittelwert in einer Gruppe g und $X_{gk}$ den X-Wert des k-ten Gruppenmitglieds in der g-ten Gruppe.

Ferner ist $MS_{zwischen}$ das Maß für die durch die Gruppenaufteilung erklärte Varianz (zwischen den Gruppen) und $MS_{innerhalb}$ das Maß für die nicht durch die Gruppenaufteilung erklärte Varianz (also: für die Varianz innerhalb der Gruppen).

Je höher der F-Wert ist, desto mehr Varianz wird durch die Gruppenzugehörigkeit erklärt. Daraus folgt, dass die Nullhypothese dann abzulehnen ist (beziehungsweise der Test signifikant ist), wenn der $F_{Modell}$-Wert größer als $F_{krit}$ ist.

Das nachstehende Beispiel nimmt Bezug auf die Datendatei *ANOVA_OHNE1.SAV*. In ihr sind für drei Qualifikationsgruppen 1–3 (1: niedrig qualifiziert, 2: mittel qualifiziert, 3: hoch qualifiziert), die in der Variablen *qualifikation* abgebildet sind, die in einer Klausur erzielten Punktezahlen enthalten – Letztere in der Variablen *ergebnis* (siehe Abbildung 13.19).




| | fall | ergebnis | qualifikation | Var |
|---|---|---|---|---|
| 1 | 1 | 54,50 | 1 | |
| 2 | 2 | 27,25 | 1 | |
| 3 | 3 | 30,50 | 1 | |
| 4 | 4 | 32,75 | 1 | |
| 5 | 5 | 9,00 | 1 | |
| 6 | 6 | 5,00 | 1 | |
| 7 | 7 | 16,25 | 1 | |
| 8 | 8 | 14,50 | 1 | |
| 9 | 9 | 22,25 | 1 | |
| 10 | 10 | 10,00 | 1 | |
| 11 | 11 | 65,50 | 2 | |
| 12 | 12 | 44,50 | 2 | |
| 13 | 13 | 53,50 | 2 | |
| 14 | 14 | 60,00 | 2 | |
| 15 | 15 | 37,50 | 2 | |
| 16 | 16 | 44,50 | 2 | |
| 17 | 17 | 55,75 | 2 | |
| 18 | 18 | 75,50 | 2 | |
| 19 | 19 | 38,50 | 2 | |
| 20 | 20 | 67,50 | 2 | |
| 21 | 21 | 92,50 | 3 | |
| 22 | 22 | 98,00 | 3 | |
| 23 | 23 | 74,50 | 3 | |
| 24 | 24 | 73,75 | 3 | |
| 25 | 25 | 67,75 | 3 | |
| 26 | 26 | 83,50 | 3 | |
| 27 | 27 | 68,50 | 3 | |
| 28 | 28 | 72,50 | 3 | |
| 29 | 29 | 91,00 | 3 | |
| 30 | 30 | 89,50 | 3 | |
| 31 | | | | |
| 32 | | | | |

**Abbildung 13.19:** Datendatei *ANOVA_OHNE1.SAV*

Im SPSS-Menü gelangen Sie zur einfaktoriellen ANOVA über ANALYSIEREN|ALLGEMEINES LINEARES MODELL|UNIVARIAT. In der sich öffnenden Dialogbox geben Sie bitte im Rahmen des vorliegenden Beispiels *ergebnis* bei ABHÄNGIGE VARIABLE und *qualifikation* bei FESTE FAKTOREN ein (siehe Abbildung 13.20).

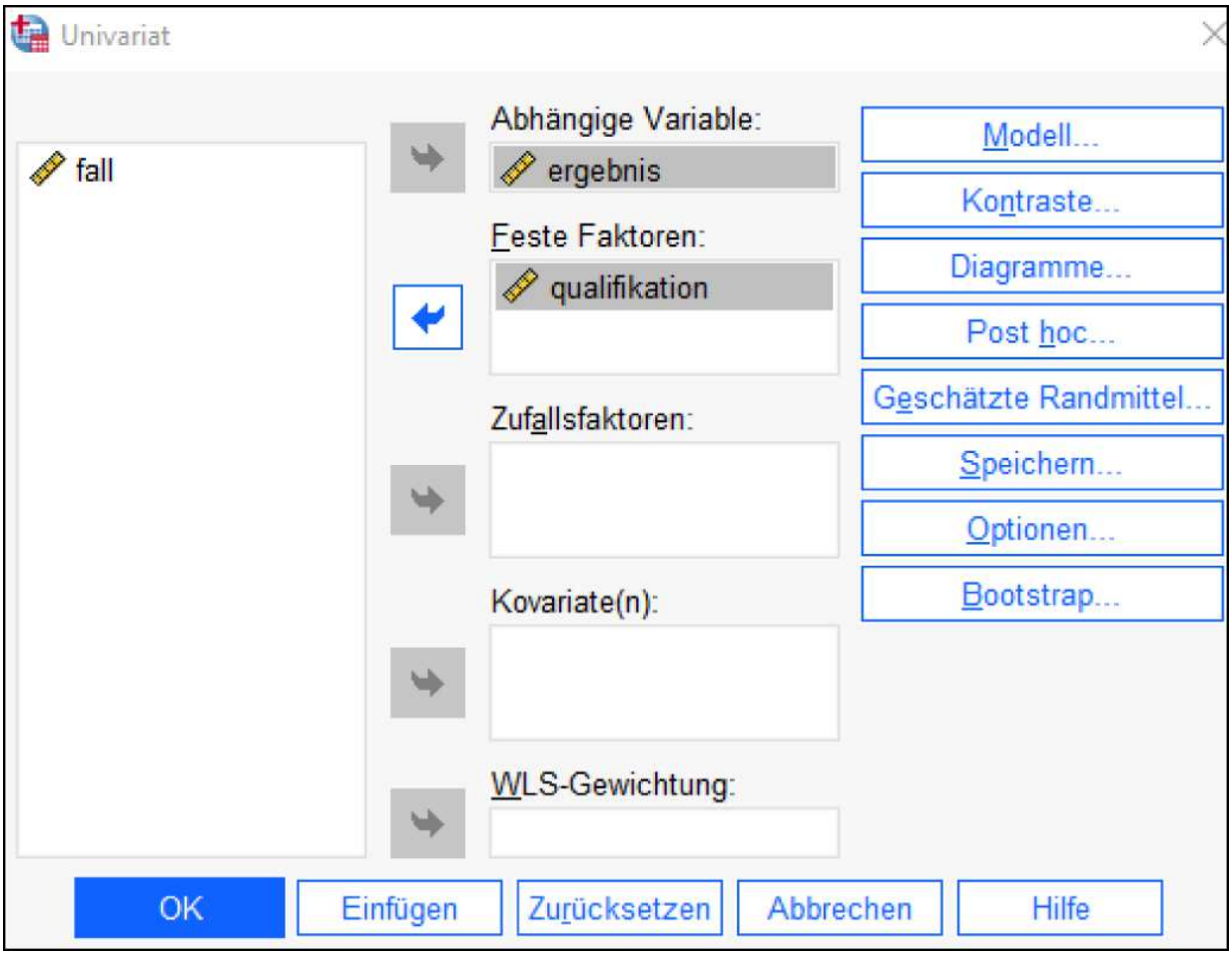



**Abbildung 13.20:** Aufruf der einfaktoriellen ANOVA für *ergebnis* mit *qualifikation* als Faktor (*ANOVA_OHNE1.SAV*)

Wenn Sie nicht nur wissen möchten, ob es Mittelwertunterschiede zwischen den Gruppen gibt, sondern welcher Art sie sind, sollten Sie in der Dialogbox UNIVARIAT unter POST HOC die BONFERRONI-Korrektur anklicken (siehe hierzu Abbildung 13.21).

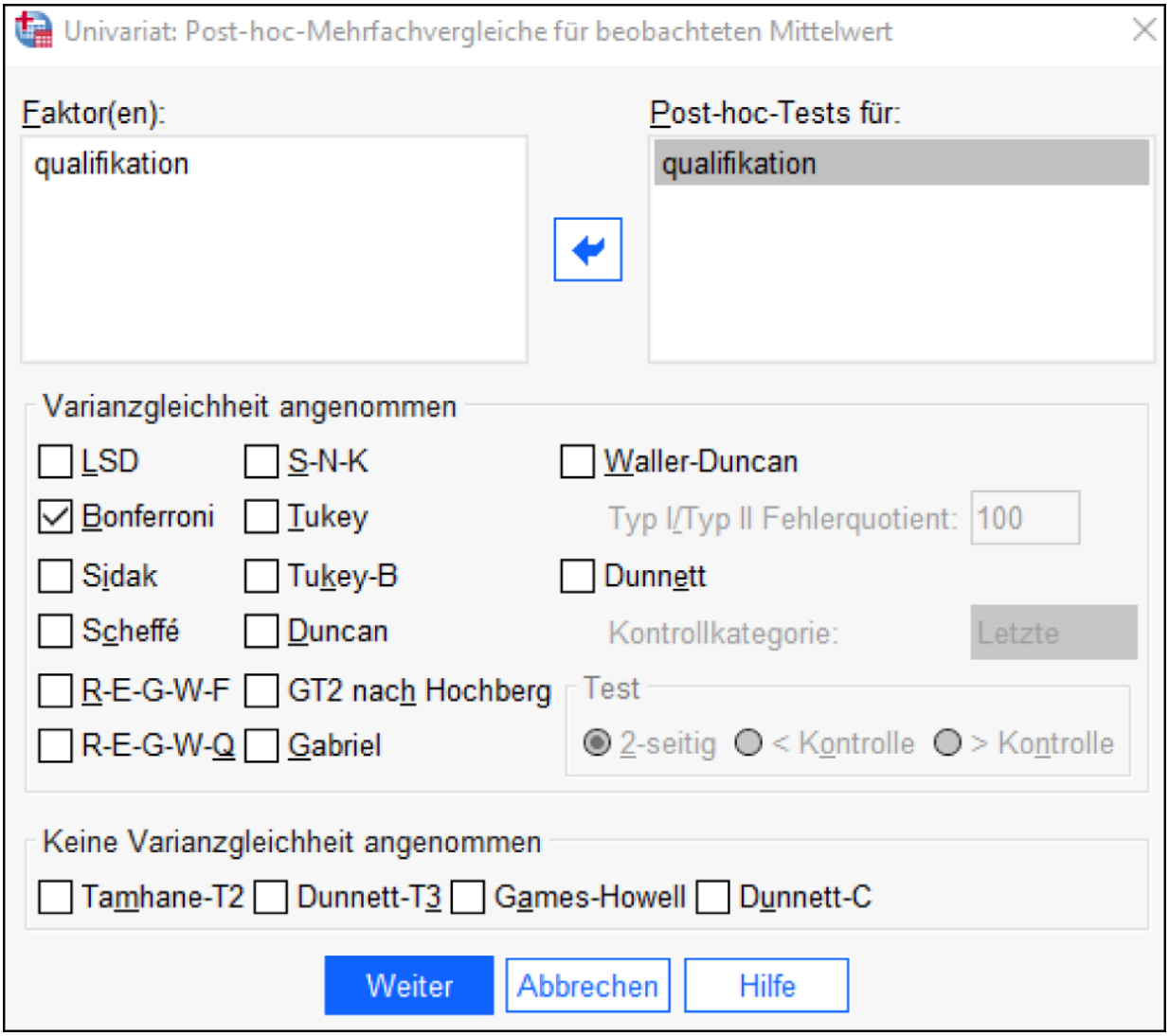



**Abbildung 13.21:** Zusätzlicher Aufruf der BONFERRONI-Korrektur bei der einfaktoriellen ANOVA für *ergebnis* und mit *qualifikation* als Faktor (*ANOVA_OHNE1.SAV*)

Die SPSS-Syntax lautet hier:

```
UNIANOVA ergebnis BY qualifikation
    /METHOD = SSTYPE (3)
    /INTERCEPT = INCLUDE
    /POSTHOC = qualifikation (BONFERRONI)
    /CRITERIA = ALPHA(0.05).
    /DESIGN = qualifikation.
```

Mit den vorstehenden Angaben wird der in Abbildung 13.22 dargestellte Output in SPSS erzeugt. Aus der mittleren Pivottabelle kann abgelesen werden, dass sich die (arithmetischen) Mittelwerte zwischen den drei Qualifikationsgruppen signifikant voneinander unterscheiden. Ergänzend gibt die untere Pivottabelle – dank der zusätzlich ausgewählten Bonferroni-Korrektur – an, *wie* die drei Gruppen voneinander differieren: Diese paarweisen Gruppenvergleiche sind allesamt signifikant (da in der »Signifikanz«-Spalte jeweils »< 0,001« steht). Dabei kommen höhere Leistungswerte in der obersten Qualifikationsgruppe gegenüber den beiden anderen Qualifikationsgruppen sowie höhere Leistungswerte in der mittleren Qualifikationsgruppe gegenüber der unteren zum

Ausdruck. Zwischen der obersten und der untersten Qualifikationsgruppe etwa wird im Beispiel ein mittlerer Unterschied von 58,95 Punkten festgehalten.

**Zwischensubjektfaktoren**

| | | N |
|---|---|---|
| qualifikation | 1 | 10 |
| | 2 | 10 |
| | 3 | 10 |

**Tests der Zwischensubjekteffekte**

Abhängige Variable: ergebnis

| Quelle | Typ III Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|
| Korrigiertes Modell | 17420,579[a] | 2 | 8710,290 | 51,469 | <,001 |
| Konstanter Term | 82818,802 | 1 | 82818,802 | 489,376 | <,001 |
| qualifikation | 17420,579 | 2 | 8710,290 | 51,469 | <,001 |
| Fehler | 4569,306 | 27 | 169,234 | | |
| Gesamt | 104808,688 | 30 | | | |
| Korrigierte Gesamtvariation | 21989,885 | 29 | | | |

a. R-Quadrat = ,792 (korrigiertes R-Quadrat = ,777)

**Post-Hoc-Tests**

**qualifikation**

**Mehrere Vergleiche**

Abhängige Variable: ergebnis

Bonferroni

| (I) qualifikation | (J) qualifikation | Mittelwertdifferenz (I-J) | Std.-Fehler | Sig. | 95% Konfidenzintervall Untergrenze | 95% Konfidenzintervall Obergrenze |
|---|---|---|---|---|---|---|
| 1 | 2 | -32,0750* | 5,81779 | <,001 | -46,9247 | -17,2253 |
| | 3 | -58,9500* | 5,81779 | <,001 | -73,7997 | -44,1003 |
| 2 | 1 | 32,0750* | 5,81779 | <,001 | 17,2253 | 46,9247 |
| | 3 | -26,8750* | 5,81779 | <,001 | -41,7247 | -12,0253 |
| 3 | 1 | 58,9500* | 5,81779 | <,001 | 44,1003 | 73,7997 |
| | 2 | 26,8750* | 5,81779 | <,001 | 12,0253 | 41,7247 |

Grundlage: beobachtete Mittelwerte.
Der Fehlerterm ist Mittel der Quadrate(Fehler) = 169,234.

*. Die Mittelwertdifferenz ist in Stufe 0,05 signifikant.

**Abbildung 13.22:** Ergebnis der einfaktoriellen ANOVA (inklusive Bonferroni-Korrektur) für *ergebnis* und mit *qualifikation* als Faktor (*ANOVA_OHNE1.SAV*)

# *Kruskal/Wallis-Test*

Eine Alternative zur einfaktoriellen ANOVA ist der **Kruskal/Wallis-Test**. Er testet auf Unterschiede in den zentralen Tendenzen (= bezüglich der Mediane) für mehrere (> 2)

unabhängige Gruppen. Seine Anwendung empfiehlt sich, wenn die Voraussetzungen für die einfache Varianzanalyse nicht erfüllt sind.

Er ist ein nichtparametrischer, das heißt »voraussetzungsfreier« Test. Die Daten müssen also zum Beispiel nicht normalverteilt sein. Sie müssen überdies lediglich ordinalskaliert sein. Dieser Test kann auch für kleine Stichproben recht gut durchgeführt werden.

Die **Voraussetzungen** des Kruskal/Wallis-Tests sind:

- ✔ mindestens Ordinalskalenniveau der für den Test relevanten Variablen
- ✔ eine weitere Variable zur Gruppenbildung

Der Kruskal/Wallis-Test basiert auf der Rangbildung. Die Messwerte werden also durch Ränge ersetzt.

Die Teststatistik ist (mit: $R_i$ als Rangsumme der i-ten Gruppe, K als Anzahl der Gruppen, n als Gesamtfallzahl, $n_i$ als Fallzahl der i-ten Gruppe):

$$H = \frac{12}{n \cdot (n+1)} \cdot \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3 \cdot (n+1)$$

beziehungsweise bei vorhandenen Ties ($t_j$), also bei gleichen Rängen innerhalb einer Gruppe:

$$H_{korr} = \frac{H}{1 - \frac{\sum_{j=1}^{m} (t_j^3 - t_j)}{n^3 - n}}.$$

Es ergibt sich – je nach Stichprobengröße – eine asymptotische Chi-Quadrat-Verteilung, und zwar wenn (im Sinne von Faustregeln)

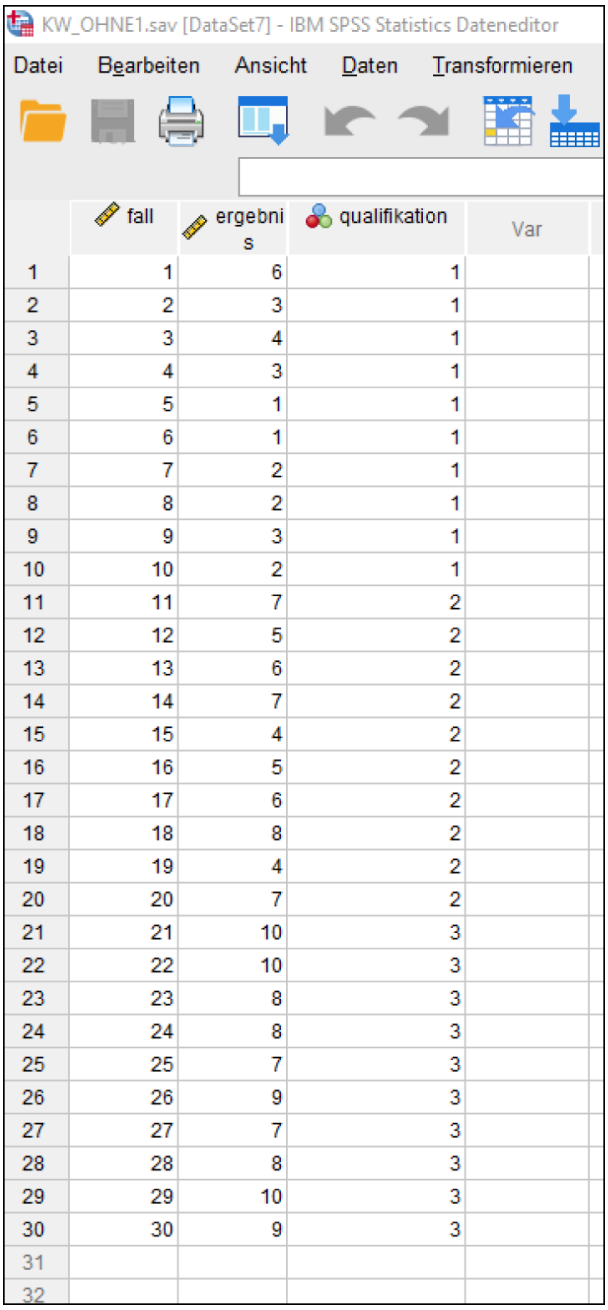
- ✔ mehr als fünf Gruppen vorliegen,
- ✔ fünf Gruppen und in jeder Gruppe mindestens vier Fälle vorliegen,
- ✔ vier Gruppen und in jeder Gruppe mindestens fünf Fälle vorliegen oder
- ✔ drei Gruppen und in jeder Gruppe mindestens neun Fälle vorliegen.

Bei diesen Konstellationen kann der Test auf Basis des Vergleichs mit einer kritischen Chi-Quadrat-Größe erfolgen. In unserem gleich diskutierten Beispielfall ist der letztgenannte Fall gegeben: drei Gruppen mit mindestens neun Fällen in jeder Gruppe (konkret: jeweils zehn Fälle).

Auch beim Kruskal/Wallis-Test empfehlen sich Post-hoc-Tests, um zu bestimmen, welche Gruppen sich signifikant voneinander unterscheiden. Eine Möglichkeit hierzu bietet SPSS über ANALYSIEREN|NICHT PARAMETRISCHE TESTS|UNABHÄNGIGE STICHPROBEN|DUNN-BONFERRONI-TESTS.

Der Kruskal/Wallis-Test wird im Folgenden auf den Datensatz *KW-OHNE1.SAV* angewandt (siehe Abbildung 13.23). In gewisser Analogie zum bei der einfaktoriellen ANOVA verwendeten Datensatz *ANOVA_OHNE1.SAV* enthält auch *KW-OHNE1.SAV* in der Variablen *qualifikation* drei

Qualifikationsgruppen, in der Variablen *ergebnis* jetzt aber nicht mehr (metrisch skalierte) Punktezahlen, sondern verschiedene (ordinalskalierte) Leistungsklassen.



KW_OHNE1.sav [DataSet7] - IBM SPSS Statistics Dateneditor

Datei Bearbeiten Ansicht Daten Transformieren

| | fall | ergebnis | qualifikation | Var |
|---|---|---|---|---|
| 1 | 1 | 6 | 1 | |
| 2 | 2 | 3 | 1 | |
| 3 | 3 | 4 | 1 | |
| 4 | 4 | 3 | 1 | |
| 5 | 5 | 1 | 1 | |
| 6 | 6 | 1 | 1 | |
| 7 | 7 | 2 | 1 | |
| 8 | 8 | 2 | 1 | |
| 9 | 9 | 3 | 1 | |
| 10 | 10 | 2 | 1 | |
| 11 | 11 | 7 | 2 | |
| 12 | 12 | 5 | 2 | |
| 13 | 13 | 6 | 2 | |
| 14 | 14 | 7 | 2 | |
| 15 | 15 | 4 | 2 | |
| 16 | 16 | 5 | 2 | |
| 17 | 17 | 6 | 2 | |
| 18 | 18 | 8 | 2 | |
| 19 | 19 | 4 | 2 | |
| 20 | 20 | 7 | 2 | |
| 21 | 21 | 10 | 3 | |
| 22 | 22 | 10 | 3 | |
| 23 | 23 | 8 | 3 | |
| 24 | 24 | 8 | 3 | |
| 25 | 25 | 7 | 3 | |
| 26 | 26 | 9 | 3 | |
| 27 | 27 | 7 | 3 | |
| 28 | 28 | 8 | 3 | |
| 29 | 29 | 10 | 3 | |
| 30 | 30 | 9 | 3 | |
| 31 | | | | |
| 32 | | | | |

**Abbildung 13.23:** Datendatei *KW-OHNE1.SAV*

So, und jetzt rufen Sie bitte in SPSS den Kruskal/Willis-Test via ANALYSIEREN|NICHT PARAMETRISCHE TESTS|KLASSISCHE DIALOGFELDER|K UNABHÄNGIGE STICHPROBEN oder alternativ, wie nachfolgend, über ANALYSIEREN|NICHT PARAMETRISCHE TESTS|UNABHÄNGIGE STICHPROBEN auf und geben bei EINSTELLUNGEN an, dass Sie den Kruskal/Wallis-Test durchführen möchten (siehe Abbildung 13.24).

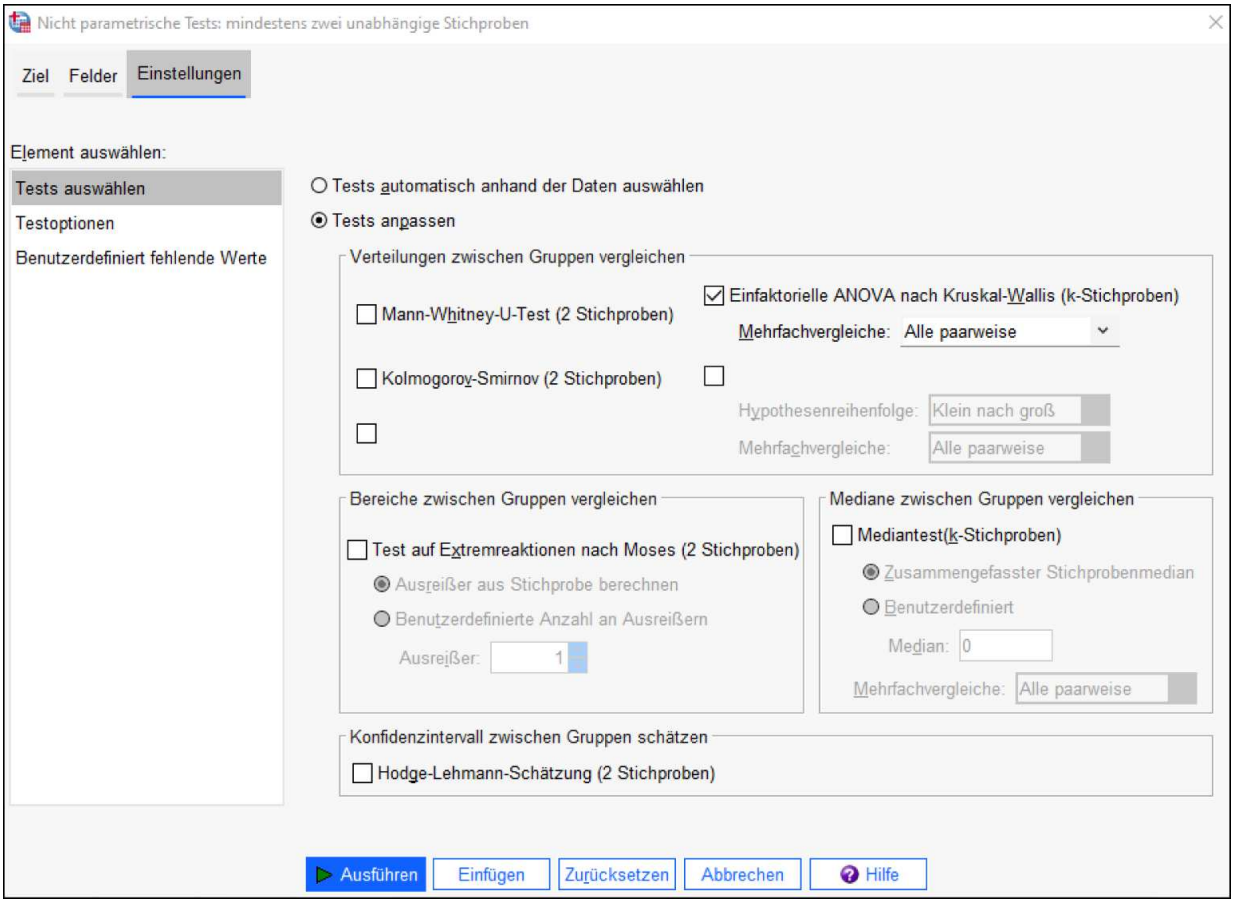



**Abbildung 13.24:** Aufruf des Kruskal/Wallis-Tests bei unabhängigen Stichproben (hier in Bezug auf *KW-OHNE1.SAV*)

Unter FELDER müssen Sie nun bei TESTVARIABLE nur noch *ergebnis* und bei GRUPPEN *qualifikation* eingeben und anschließend auf AUSFÜHREN klicken: Los geht's mit dem Kruskal/Wallis-Test (siehe Abbildung 13.25)!

Mit der vorstehenden Menüabfolge korrespondiert folgende SPSS-Syntax:

```
NPTESTS
    /INDEPENDENT TEST (ergebnis) GROUP (qualifikation) KRUSKAL_WALLIS
(COMPARE=PAIRWISE)
    /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
    /CRITERIA ALPHA=0.05 CILEVEL=95.
```

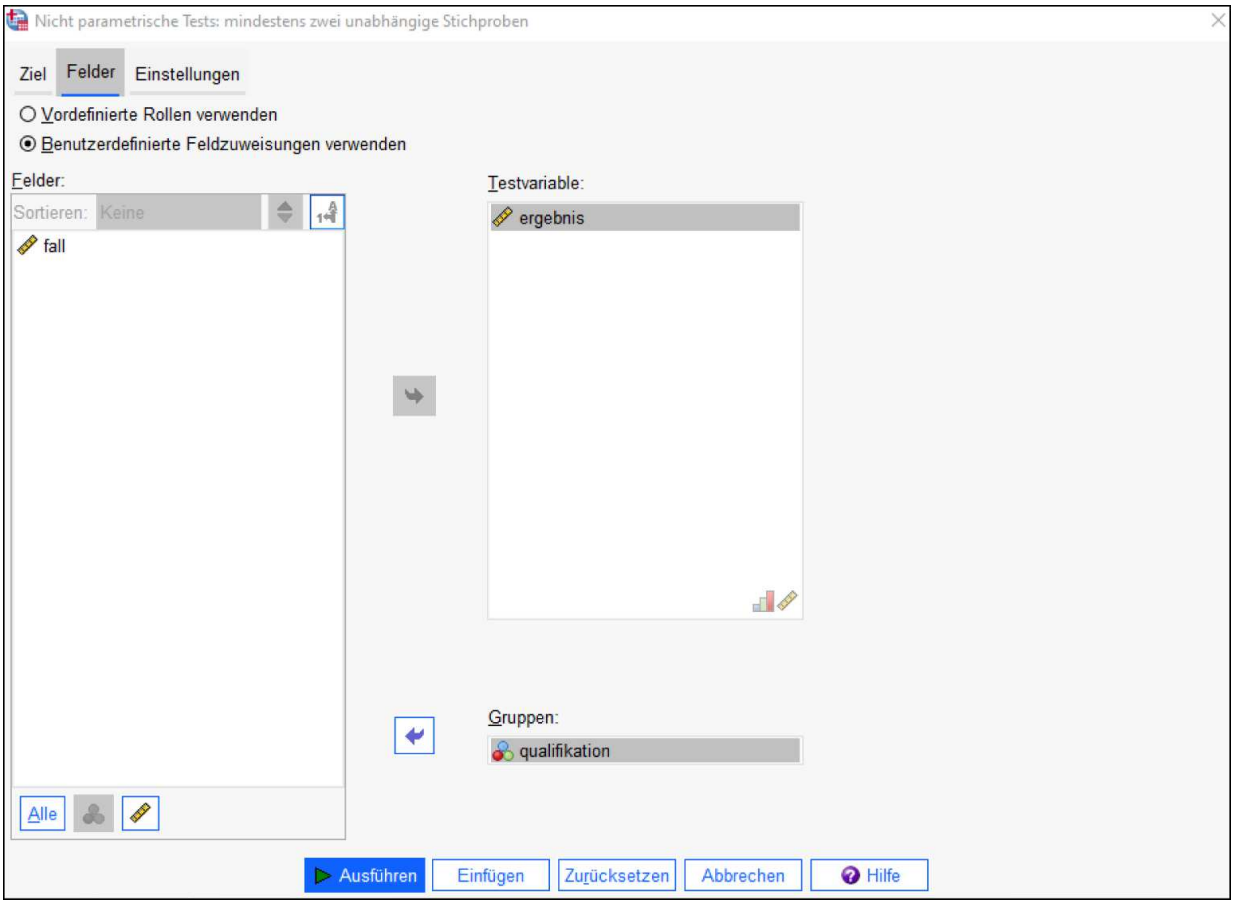



**Abbildung 13.25:** Aufruf des Kruskal/Wallis-Tests bei unabhängigen Stichproben mit *ergebnis* als Testvariable und *qualifikation* als Gruppierungsvariable (*KW-OHNE1.SAV* )

Im vorliegenden Beispiel wird die Nullhypothese, dass es keine Mittelwertunterschiede zwischen den Gruppen gibt beziehungsweise dass diese vernachlässigbar sind, abgelehnt. Das geht aus Abbildung 13.26 hervor.

| Zusammenfassung des Kruskal-Wallis-Tests bei unabhängigen Stichproben | |
|---|---|
| Gesamtzahl | 30 |
| Teststatistik | 22,953[a] |
| Freiheitsgrad | 2 |
| Asymptotische Sig. (zweiseitiger Test) | <,001 |

a. Die Teststatistik wird für Bindungen angepasst.

**Abbildung 13.26:** Ergebnis des Kruskal/Wallis-Tests bei unabhängigen Stichproben mit *ergebnis* als Testvariable und *qualifikation* als Gruppierungsvariable (*KW-OHNE1.SAV*) (Output-Auszug)

## Mehr als zwei Gruppen: Abhängigkeit

Es gibt auch für Mittelwertvergleiche von mehr als zwei Gruppen, die verbunden (also abhängig voneinander) sind, Testverfahren. Von diesen werden nachfolgend die einfache Varianzanalyse mit Messwiederholung und der Friedman-Test vorgestellt.

## Einfache Varianzanalyse mit Messwiederholung

Bei der **einfachen Varianzanalyse mit Messwiederholung**, das heißt für verbundene Gruppen, wird geprüft, ob sich die (arithmetischen) Mittelwerte mehrerer abhängiger Gruppen voneinander unterscheiden.

Die üblichen **Test-Voraussetzungen** sind unter anderem:

- ✔ metrische Skalierung der für den Test relevanten Variablen
- ✔ Normalverteilung der abhängigen Variablen innerhalb der Messzeitpunkte (wobei möglichst mindestens 25 Fälle pro Messzeitpunkt vorliegen sollen)

Es findet eine Varianzzerlegung statt, wobei die Teststatistik ziemlich kompliziert aussieht:

$$F = \frac{K \cdot \sum_{k=1}^{K} (\overline{X_g} - \overline{X})^2 / (G-1)}{\sum_{k=1}^{K} \sum_{g=1}^{G} (X_{kg} - \overline{X_g} - \overline{X_k} + \overline{X}^2 / [(K-1) \cdot (G-1)]}.$$

In dieser Formel steht $\overline{X_g}$ für den arithmetischen Mittelwert im Messzeitpunkt g, $\overline{X_k}$ für den arithmetischen Mittelwert einer statistischen Einheit (zum Beispiel einer Person) über alle Messzeitpunkte hinweg, $\overline{X}$ für den arithmetischen Gesamtmittelwert über alle Messzeitpunkte hinweg, $X_{kg}$ für den X-Wert einer statistischen Einheit (zum Beispiel einer Person) im Messzeitpunkt g, G also hier für die Anzahl der Messzeitpunkte und K für die gesamte Fallzahl.

So kompliziert die vorstehende Formel ist, desto einfacher ist die Entscheidungsregel des betreffenden Tests: Wenn $F > F_{krit}$ ist, ist der Test signifikant. Da die obige Formel aber, wie geschrieben, ziemlich kompliziert ist, habe ich die nachstehenden Betrachtungen vereinfacht, und zwar derart, dass für die einzelnen Gruppenvergleiche jeweils separate (t-)Tests durchgeführt werden. Ich bitte um Nachsicht, aber: warum die Welt komplizierter machen, als sie ohnehin schon ist?

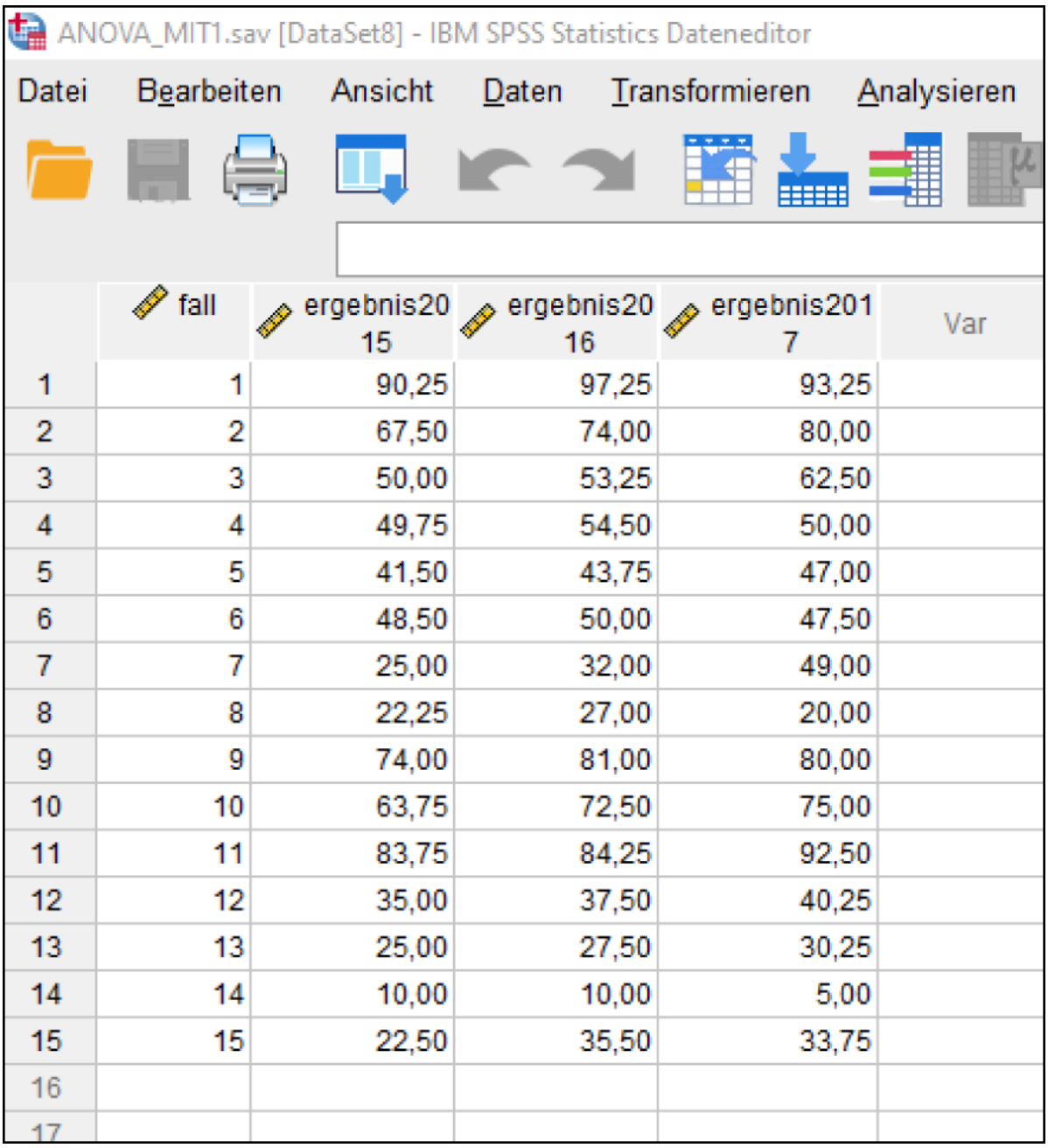
Das folgende Beispiel bezieht sich auf *ANOVA_MIT1.SAV* mit den Variablen *ergebnis2015, ergebnis2016* und *ergebnis2017*. In diesen Ergebnisvariablen sind für jeweils dieselben Personen die Ballwurfergebnisse in Metern abgespeichert (siehe Abbildung 13.27).

Im SPSS-Menü rufen Sie die einzelnen t-Tests bitte folgendermaßen auf: ANALYSIEREN|MITTELWERTE UND PROPORTIONEN VERGLEICHEN|T-TEST BEI STICHPROBEN MIT PAARIGEN WERTEN und geben im sich öffnenden Dialogfeld die drei Gruppenvergleiche ein (siehe Abbildung 13.28). Nochmals: Sie führen hier einfach drei t-Tests bei verbundenen Gruppen durch.

Die zugehörige SPSS-Syntax lautet:

```
T-TEST PAIRS=ergebnis2015 ergebnis2015 ergebnis2016 WITH ergebnis2016
ergebnis2017 ergebnis2017 (PAIRED)
    /CRITERIA=CI(.9500)
    /MISSING=ANALYSIS.
```

Im unteren Teil von Abbildung 13.29 stehen die eigentlichen Testergebnisse im vorliegenden Beispielfall. Während hierbei die Ergebnisunterschiede zwischen 2015 und 2016 sowie zwischen 2015 und 2017 als signifikant zu bezeichnen sind, gilt dies offenkundig für den Ergebnisvergleich zwischen 2016 und 2017 nicht. Die individuellen Wurfunterschiede sind demnach zwischen 2015 und 2016 sowie zwischen 2015 und 2017 signifikant unterschiedlich gewesen, zwischen 2016 und 2017 hingegen nicht. Anhand der in Abbildung 13.29 auch angegebenen durchschnittlichen Wurfweiten in der obersten Pivottabelle unter »Mittelwert« kann gesagt werden, dass die Werfer 2016 und 2017 signifikant weiter als 2015 geworfen haben, dass sich aber zwischen 2016 und 2017 keine signifikante Leistungssteigerung ergeben hat (bei einem leicht höheren Durchschnitt für 2017 gegenüber 2016).




| | fall | ergebnis2015 | ergebnis2016 | ergebnis2017 | Var |
|---|---|---|---|---|---|
| 1 | 1 | 90,25 | 97,25 | 93,25 | |
| 2 | 2 | 67,50 | 74,00 | 80,00 | |
| 3 | 3 | 50,00 | 53,25 | 62,50 | |
| 4 | 4 | 49,75 | 54,50 | 50,00 | |
| 5 | 5 | 41,50 | 43,75 | 47,00 | |
| 6 | 6 | 48,50 | 50,00 | 47,50 | |
| 7 | 7 | 25,00 | 32,00 | 49,00 | |
| 8 | 8 | 22,25 | 27,00 | 20,00 | |
| 9 | 9 | 74,00 | 81,00 | 80,00 | |
| 10 | 10 | 63,75 | 72,50 | 75,00 | |
| 11 | 11 | 83,75 | 84,25 | 92,50 | |
| 12 | 12 | 35,00 | 37,50 | 40,25 | |
| 13 | 13 | 25,00 | 27,50 | 30,25 | |
| 14 | 14 | 10,00 | 10,00 | 5,00 | |
| 15 | 15 | 22,50 | 35,50 | 33,75 | |
| 16 | | | | | |
| 17 | | | | | |

**Abbildung 13.27**: Datendatei *ANOVA_MIT1.SAV*

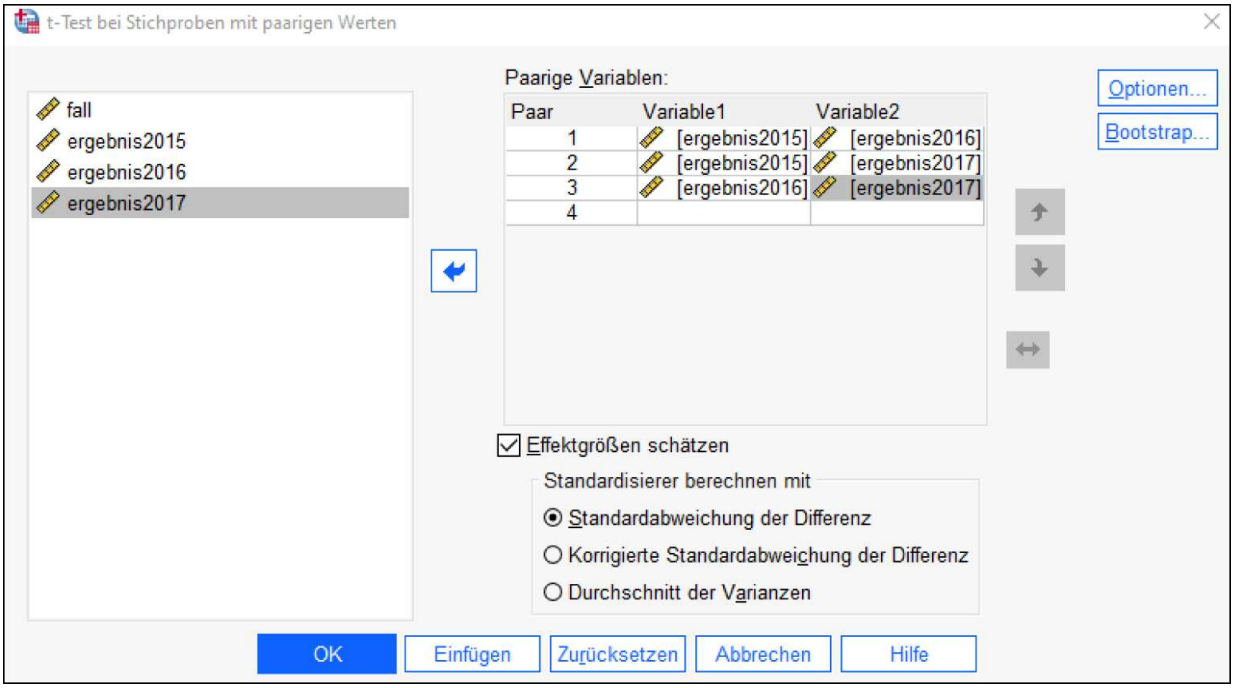



**Abbildung 13.28:** Aufruf von mehreren t-Tests bei Messwiederholung und drei Gruppenvergleichen (*ANOVA_MIT1.SAV*)

**Statistik bei gepaarten Stichproben**

| | | Mittelwert | N | Std.-Abweichung | Standardfehler des Mittelwertes |
|---|---|---|---|---|---|
| Paaren 1 | ergebnis2015 | 47,2500 | 15 | 24,48451 | 6,32187 |
| | ergebnis2016 | 52,0000 | 15 | 25,04924 | 6,46769 |
| Paaren 2 | ergebnis2015 | 47,2500 | 15 | 24,48451 | 6,32187 |
| | ergebnis2017 | 53,7333 | 15 | 26,37563 | 6,81016 |
| Paaren 3 | ergebnis2016 | 52,0000 | 15 | 25,04924 | 6,46769 |
| | ergebnis2017 | 53,7333 | 15 | 26,37563 | 6,81016 |

**Korrelationen bei gepaarten Stichproben**

| | | N | Korrelation | Signifikanz Einseitiges p | Signifikanz Zweiseitiges p |
|---|---|---|---|---|---|
| Paaren 1 | ergebnis2015 & ergebnis2016 | 15 | ,990 | <,001 | <,001 |
| Paaren 2 | ergebnis2015 & ergebnis2017 | 15 | ,961 | <,001 | <,001 |
| Paaren 3 | ergebnis2016 & ergebnis2017 | 15 | ,970 | <,001 | <,001 |

**Test bei gepaarten Stichproben**

| | | Gepaarte Differenzen | | | | | T | df | Signifikanz | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Mittelwert | Std.-Abweichung | Standardfehler des Mittelwertes | 95% Konfidenzintervall der Differenz Unterer Wert | Oberer Wert | | | Einseitiges p | Zweiseitiges p |
| Paaren 1 | ergebnis2015 - ergebnis2016 | -4,75000 | 3,50637 | ,90534 | -6,69176 | -2,80824 | -5,247 | 14 | <,001 | <,001 |
| Paaren 2 | ergebnis2015 - ergebnis2017 | -6,48333 | 7,31526 | 1,88879 | -10,53439 | -2,43227 | -3,433 | 14 | ,002 | ,004 |
| Paaren 3 | ergebnis2016 - ergebnis2017 | -1,73333 | 6,47003 | 1,67055 | -5,31632 | 1,84965 | -1,038 | 14 | ,159 | ,317 |

**Abbildung 13.29:** Ergebnis von drei t-Tests bei Messwiederholung und drei Gruppenvergleichen (*ANOVA_MIT1.SAV*) (Output-Auszug)

# *Friedman-Test*

Beim **Friedman-Test** geht es um die Unterschiede in den zentralen Tendenzen (= in den Medianwerten) bei abhängigen Stichproben. Dieser nichtparametrische Test kann angewandt werden, wenn die Voraussetzungen der einfaktoriellen Varianzanalyse mit Messwiederholung nicht erfüllt sind.

Die **Test-Voraussetzungen** sind beim Friedman-Test:

- ✔ mindestens Ordinalskalenniveau der für den Test relevanten Variablen
- ✔ verbundene Stichproben, aber Unabhängigkeit der Messwerte innerhalb der Gruppen

Die Messwerte werden durch Ränge beziehungsweise Rangsummen ersetzt.

Die Teststatistik nach Friedman lautet (mit: K als Stichprobengröße, G als Anzahl der Messzeitpunkte und $R_g$ als Rangsumme im Messzeitpunkt g):

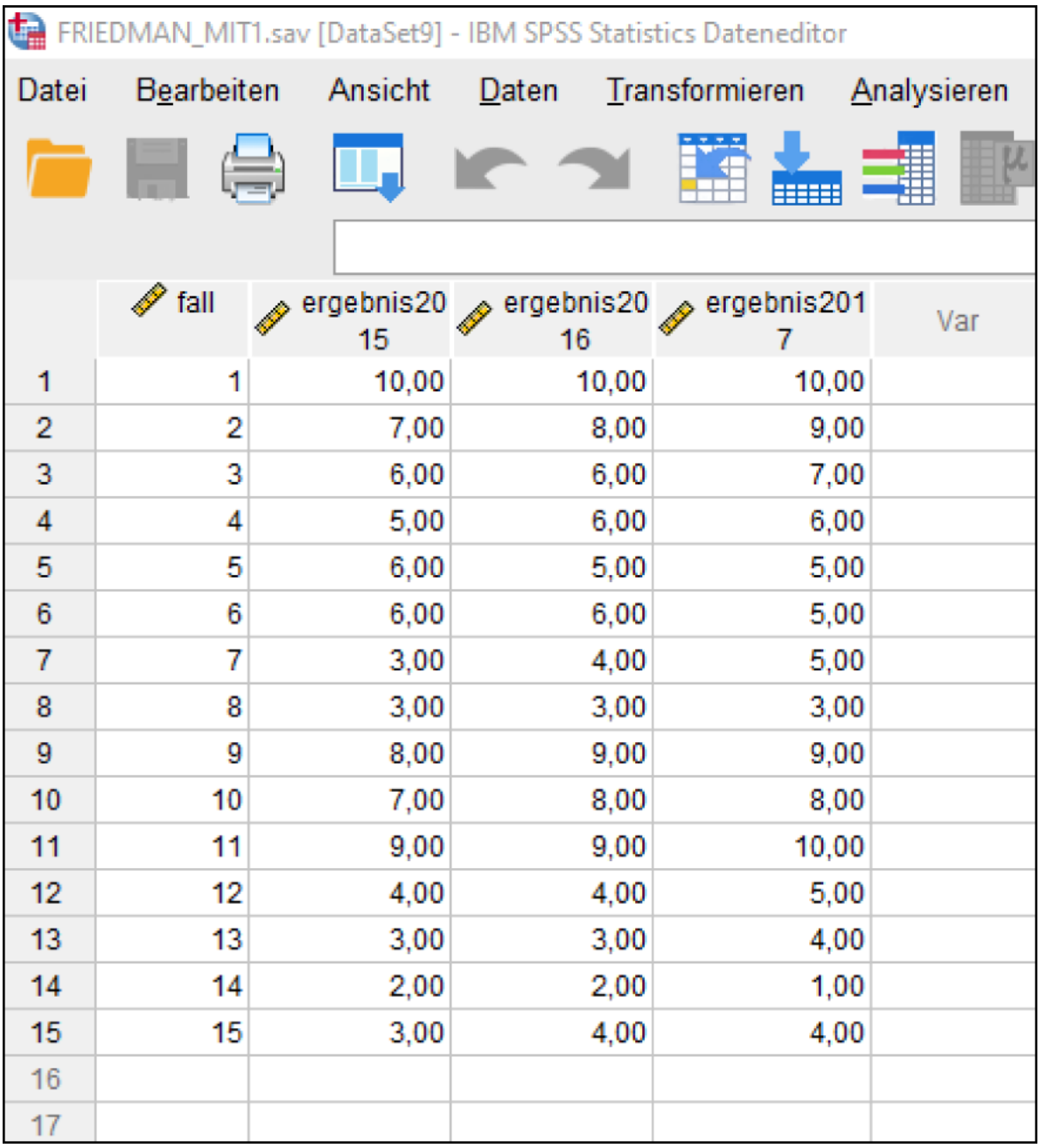
$$\chi^2_{\text{Friedman}} = \frac{12}{K \cdot G \cdot (G+1)} \cdot \sum_{g=1}^{G} R_g^2 - 3 \cdot K \cdot (G+1)$$

beziehungsweise beim Vorliegen von Ties, das heißt von mehreren gleichen Rängen zu einem Messzeitpunkt (mit: J = Anzahl der Ties und $t_j$ als Anzahl der Werte auf Rang j):

$$\chi^2_{\text{Friedman(korr)}} = \frac{\chi^2_{\text{Friedman}}}{1 - \frac{1}{K \cdot G \cdot (G^2 - 1)} \cdot \sum_{j=1}^{J} (t_j^3 - t_j)}.$$

Anschließend findet ein Vergleich mit $\chi^2_{\text{krit}}$ statt.

Der Friedman-Test wird nachfolgend auf Basis der Datendatei *FRIEDMAN-MIT1.SAV* illustriert. Analog zum Beispieldatensatz aus dem vorhergehenden Abschnitt geht es auch hier für drei (frühere) Jahre (2015–2017) für dieselben Personen um deren Ballwurfergebnisse, dieses Mal aber nicht metrisch skaliert (in Metern), sondern ordinalskaliert (in Leistungsklassen), wie Abbildung 13.30 zeigt.




| | fall | ergebnis2015 | ergebnis2016 | ergebnis2017 | Var |
|---|---|---|---|---|---|
| 1 | 1 | 10,00 | 10,00 | 10,00 | |
| 2 | 2 | 7,00 | 8,00 | 9,00 | |
| 3 | 3 | 6,00 | 6,00 | 7,00 | |
| 4 | 4 | 5,00 | 6,00 | 6,00 | |
| 5 | 5 | 6,00 | 5,00 | 5,00 | |
| 6 | 6 | 6,00 | 6,00 | 5,00 | |
| 7 | 7 | 3,00 | 4,00 | 5,00 | |
| 8 | 8 | 3,00 | 3,00 | 3,00 | |
| 9 | 9 | 8,00 | 9,00 | 9,00 | |
| 10 | 10 | 7,00 | 8,00 | 8,00 | |
| 11 | 11 | 9,00 | 9,00 | 10,00 | |
| 12 | 12 | 4,00 | 4,00 | 5,00 | |
| 13 | 13 | 3,00 | 3,00 | 4,00 | |
| 14 | 14 | 2,00 | 2,00 | 1,00 | |
| 15 | 15 | 3,00 | 4,00 | 4,00 | |
| 16 | | | | | |
| 17 | | | | | |

**Abbildung 13.30**: Datendatei *FRIEDMAN-MIT1.SAV*

Der Friedman-Test findet sich im SPSS-Menü unter ANALYSIEREN|NICHT PARAMETRISCHE TESTS|KLASSISCHE DIALOGFELDER|K VERBUNDENE STICHPROBEN. Auf dem sich dann öffnenden Dialogblatt müssen Sie den Friedman-Test einfach nur ebenso wie Ihre gewünschten Testvariablen auswählen (siehe Abbildung 13.31).

Die SPSS-Syntax lautet hier:

```
NPAR TESTS
    /FRIEDMAN=ergebnis2015 ergebnis2016 ergebnis2017
    /MISSING LISTWISE.
```

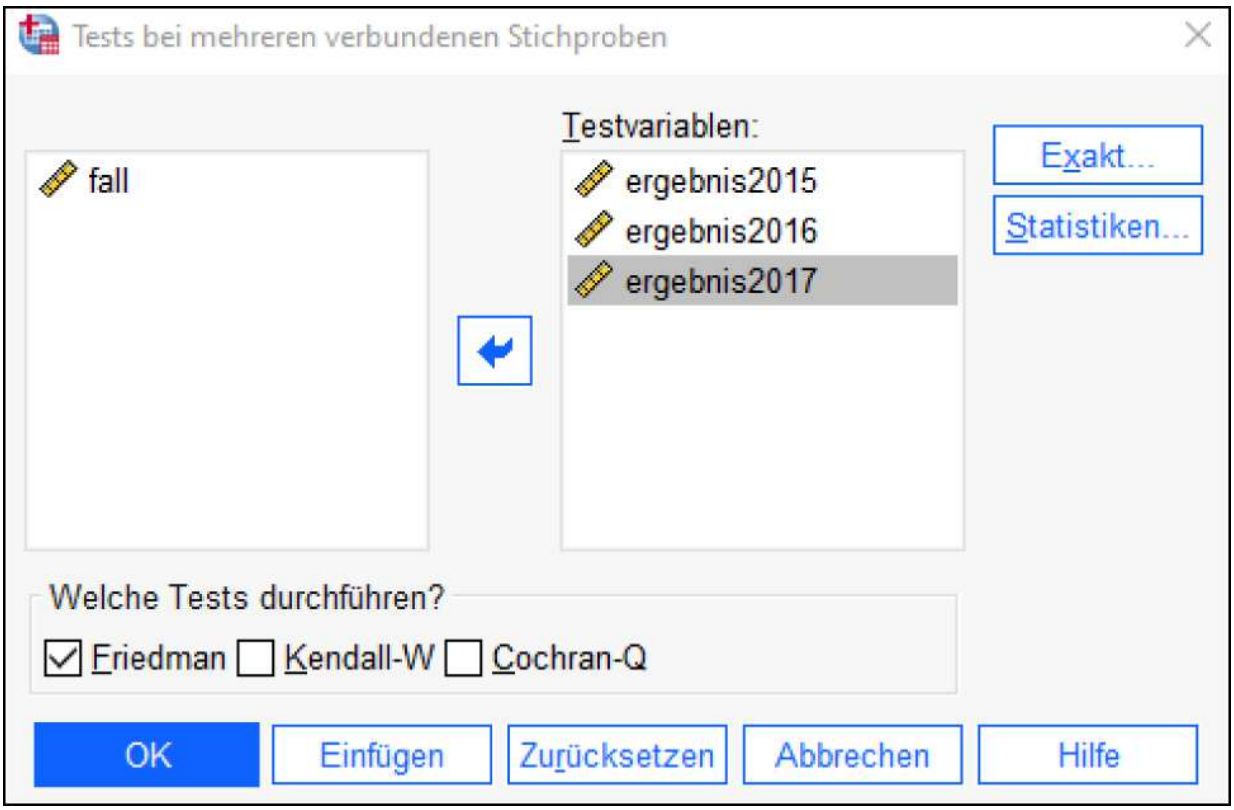



**Abbildung 13.31:** Aufruf des Friedman-Tests mit den Testvariablen *ergebnis2015*, *ergebnis2016* und *ergebnis2017* (*FRIEDMAN-MIT1.SAV*)

Im vorliegenden Beispielfall ergibt sich in der Rubrik der (asymptotischen) Signifikanz ein Wert von 3,9 Prozent (siehe Abbildung 13.32). Legt man hier den üblichen 5-Prozent-Signifikanzmaßstab an, kann gesagt werden, dass die (mittleren) rangbezogenen Ergebnisse der drei Jahre sich signifikant voneinander unterscheiden. (Es kann an dieser Stelle aber nur diese grobe Aussage getroffen werden, weil keine paarweisen Vergleiche im zuletzt behandelten Beispiel vorgenommen wurden.)

## Friedman-Test

**Ränge**

| | Mittlerer Rang |
|---|---|
| ergebnis2015 | 1,60 |
| ergebnis2016 | 2,03 |
| ergebnis2017 | 2,37 |

**Teststatistiken[a]**

| | |
|---|---|
| N | 15 |
| Chi-Quadrat | 6,488 |
| df | 2 |
| Asymp. Sig. | ,039 |

a. Friedman-Test

**Abbildung 13.32:** Ergebnis des Friedman-Tests mit den Testvariablen *ergebnis2015*, *ergebnis2016* und *ergebnis2017* (*FRIEDMAN-MIT1.SAV* )

# Kapitel 14

# Tests II: Glockenkurve – oder was?

## IN DIESEM KAPITEL

Chi-Quadrat-Verteilungstest

Kolmogorov/Smirnov-Verteilungstest

Shapiro/Wilk-Normalverteilungstest

Kapitel 14 setzt sich mit verschiedenen Verteilungstests auseinander. Es werden im Einzelnen der Chi-Quadrat-Verteilungstest, der Kolmogorov/Smirnov-Verteilungstest und der Shapiro/Wilk-Normalverteilungstest und ihre Anwendungsmöglichkeiten in SPSS vorgestellt.

## *Vorbemerkung*

Den nachfolgend vorgestellten **Verteilungstests** liegen unterschiedliche **Verteilungsannahmen** zugrunde:

- ✔ Der Chi-Quadrat- und der Kolmogorov/Smirnov-Verteilungstest können prinzipiell auf beliebige (Häufigkeits-)Verteilungen bezogen werden.
- ✔ Wie der Name schon andeutet, bezieht sich der Shapiro/Wilk-Normalverteilungstest hingegen speziell auf die Normalverteilung, die – wegen ihrer konkreten Erscheinungsform – umgangssprachlich auch als »Glockenkurve« bezeichnet wird.

Bei allen diskutierten Tests lautet die **Nullhypothese** in SPSS jeweils grundsätzlich: »Es liegt eine bestimmte Verteilung vor.« Wird die Nullhypothese beim jeweiligen Test verworfen, können Sie davon ausgehen (mit einer bestimmten Irrtumswahrscheinlichkeit), dass diese bestimmte Verteilungsform nicht vorliegt. Wünschen Sie sich die betreffende Verteilung, so müssen Sie bei den Verteilungstests an der Nichtablehnung der Nullhypothese interessiert sein (und das ist ein Unterschied zu den anderen in diesem Buch vorgestellten Tests)!

# *Chi-Quadrat-Verteilungstest*

Hier wird die empirische Häufigkeitsverteilung einer kategorialen Variablen mit einer theoretischen Verteilung verglichen, und zwar im SPSS-Menü über ANALYSIEREN|NICHT PARAMETRISCHE TESTS|KLASSISCHE DIALOGFELDER|CHI-QUADRAT.

Sie können alternativ aber auch als SPSS-Syntax allgemein Folgendes eingeben:

```
NPAR TESTS
    /CHISQUARE = variable
    /EXPECTED = wert1 wert2 wert3 …
    /STATISTICS DESCRIPTIVES
    /MISSING ANALYSIS.
```

Als theoretische Verteilung wird zum Beispiel die Normalverteilung unterstellt. Dann müssen unter EXPECTED in der Syntax die gemäß Normalverteilung erwarteten Häufigkeiten eingetragen werden.

## *Voraussetzungen*

Die Voraussetzungen dieses oft auch als *Chi-Quadrat-Anpassungstest* bezeichneten Tests sind unter anderem, dass eine kategoriale Variable vorliegt und dass in dem Fall, in dem die Fallzahlen pro Kategorie kleiner als 5 Fälle sind, die betreffenden Kategorien zusammengefasst werden. Nur unter diesen Voraussetzungen ist der Test erlaubt.

Die zugehörige **Teststatistik** lautet:

$$\chi^2 = \sum_{j=1}^{k} \frac{\left(h_{bj} - h_{ej}\right)^2}{h_{ej}}.$$

In der vorstehenden Formel bezeichnet k die Anzahl der Kategorien, und $h_{bj}$ steht für die beobachteten und $h_{ej}$ für die – gemäß der in der Nullhypothese angenommenen Verteilungsform – erwarteten absoluten Häufigkeiten.

Diese Teststatistik wird mit einem tabellierten, dem kritischen Chi-Quadrat-Wert (bei einer vorgegebenen Irrtumswahrscheinlichkeit) verglichen. Ist die Teststatistik größer als der kritische Wert, wird gefolgert, dass die in der Nullhypothese angenommene Verteilungsform (mit einer bestimmten Irrtumswahrscheinlichkeit) nicht vorliegt. Ein relativ hoher Wert der Teststatistik deutet demnach darauf hin, dass die in der Nullhypothese unterstellte Verteilung nicht angemessen ist.

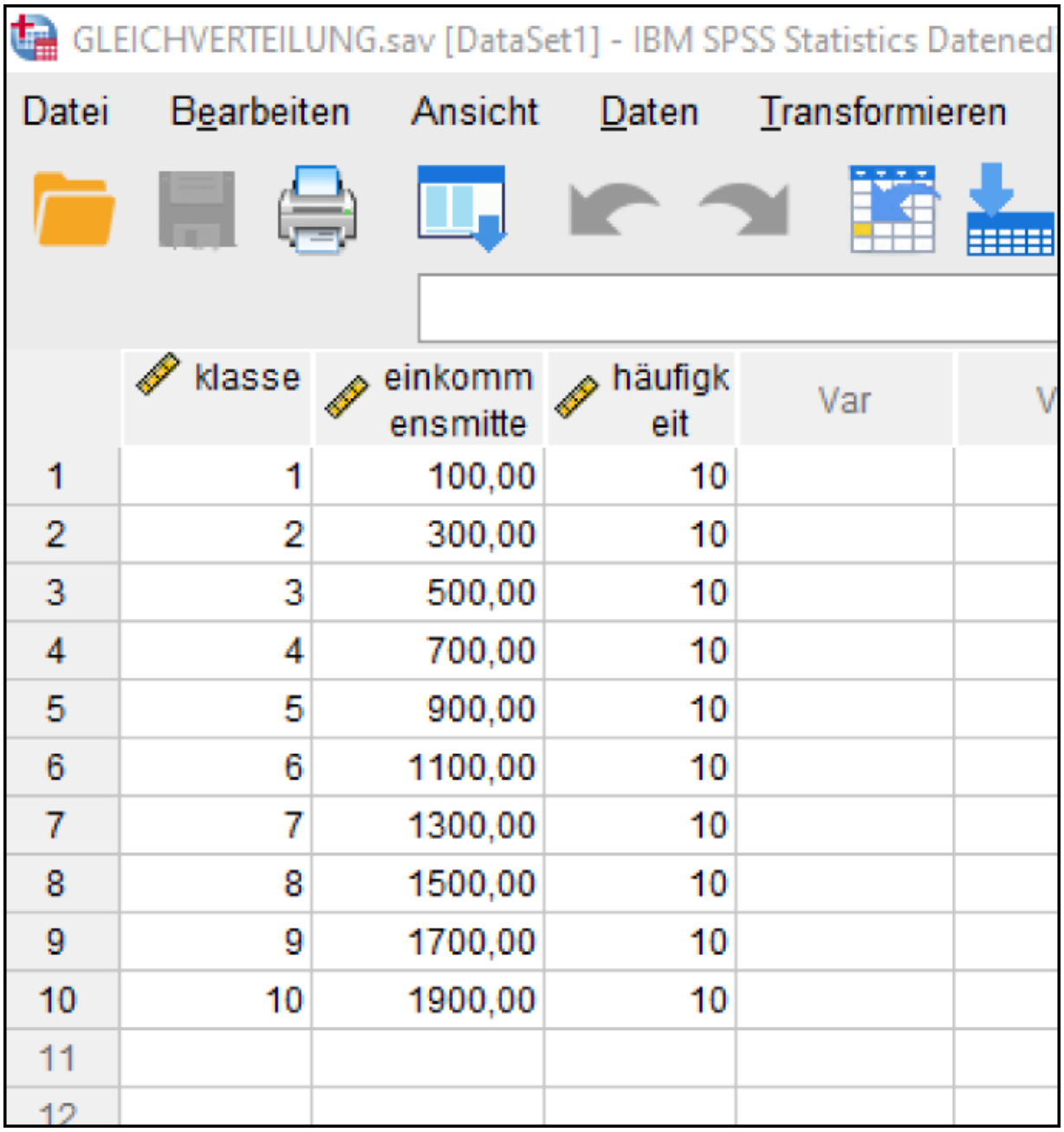
Lautet die Nullhypothese also zum Beispiel: »Die Variable ist normalverteilt.« und wird diese im vorstehenden Sinne verworfen, dann bedeutet das, dass aufgrund des Testergebnisses von keiner Normalverteilung ausgegangen wird. Ist die Normalverteilung sozusagen »erwünscht«, dann sollte das Testergebnis also möglichst nicht signifikant sein!

Die vorstehend genannte »Erwünschtheit« einer bestimmten Verteilungsform – zum Beispiel der Normalverteilung – kann sich etwa aus einer vorgeschalteten grafischen oder einer tabellarischen Betrachtung der betreffenden Variablen über (ANALYSIEREN|DESKRIPTIVE STATISTIKEN|HÄUFIGKEITEN) ergeben haben.

## *Beispiel 1*

Wie eingangs zu diesem Abschnitt erwähnt, ist der Chi-Quadrat-Verteilungstest für kategoriale beziehungsweise für klassifizierte Daten konzipiert. In einem ersten Beispiel soll von einer Gleichverteilung ausgegangen werden. Hierzu öffnen Sie bitte die Datendatei *GLEICHVERTEILUNG.SAV*. In ihr sind zehn Einkommensklassen mit

den jeweiligen Klassenmitten und den zugehörigen absoluten Häufigkeiten dargestellt (siehe Abbildung 14.1). Jeder Klassenmitte ist dabei als absolute Häufigkeit der Wert 10, das heißt jeweils zehn Fälle, zugeordnet. Damit liegt ganz offenkundig eine perfekte Gleichverteilung über alle zehn Klassen vor. Wenn das keine Gleichverteilung ist, was soll dann, bitte schön, eine Gleichverteilung sein?




| | klasse | einkommensmitte | häufigkeit | Var | V |
|---|---|---|---|---|---|
| 1 | 1 | 100,00 | 10 | | |
| 2 | 2 | 300,00 | 10 | | |
| 3 | 3 | 500,00 | 10 | | |
| 4 | 4 | 700,00 | 10 | | |
| 5 | 5 | 900,00 | 10 | | |
| 6 | 6 | 1100,00 | 10 | | |
| 7 | 7 | 1300,00 | 10 | | |
| 8 | 8 | 1500,00 | 10 | | |
| 9 | 9 | 1700,00 | 10 | | |
| 10 | 10 | 1900,00 | 10 | | |
| 11 | | | | | |
| 12 | | | | | |

**Abbildung 14.1:** Datendatei *GLEICHVERTEILUNG.SAV*

Damit Sie aus der in Abbildung 14.1 angegebenen Häufigkeitstabelle einen in SPSS gut zu verarbeitenden Individualdatensatz machen können, müssen Sie die Werte der Variablen *einkommensmitte* mit den Werten der Häufigkeitsvariablen *häufigkeit* gewichten, menügesteuert über DATEN|FÄLLE GEWICHTEN oder in Syntaxform:

```
WEIGHT BY häufigkeit.
```

Anschließend stoßen Sie den Chi-Quadrat-Anpassungstest über ANALYSIEREN|NICHT PARAMETRISCHE TESTS|KLASSISCHE DIALOGFELDER|CHI-QUADRAT an. Es öffnet sich ein Dialogfeld, in das Sie die Variable, deren Verteilung überprüft werden soll (*einkommensmitte*), eingeben und außerdem die Voreinstellung ERWARTETE WERTE|ALLE KATEGORIEN GLEICH stehen lassen, da Sie ja auf Gleichverteilung testen möchten (siehe hierzu Abbildung 14.2).

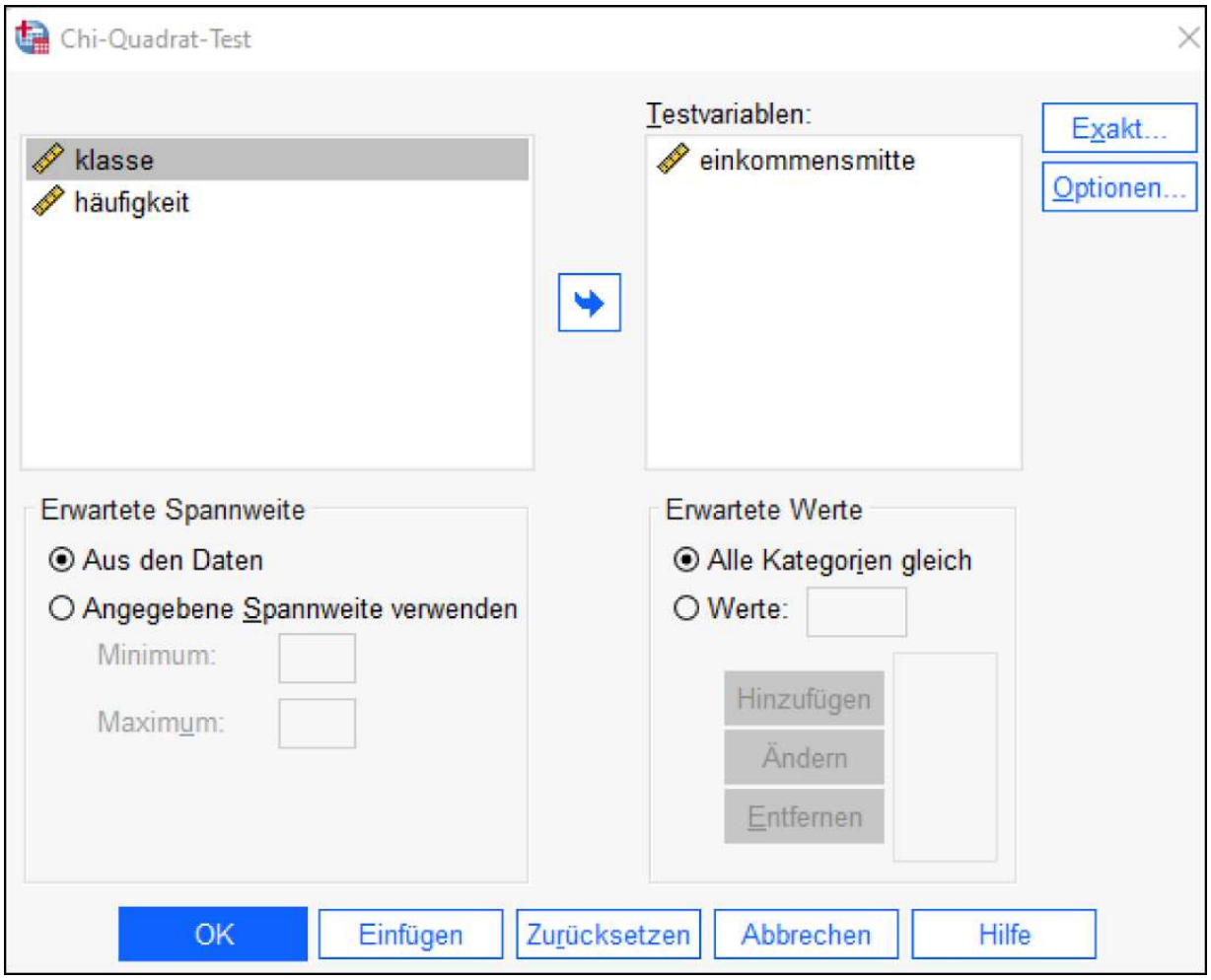



**Abbildung 14.2:** Aufruf des Chi-Quadrat-Verteilungstests auf Gleichverteilung für die Variable *einkommensmitte* (*GLEICHVERTEILUNG.SAV*)

Die entsprechende SPSS-Syntax lautet in diesem Beispiel:

```
NPAR TESTS
    /CHISQUARE=einkommensmitte
    /EXPECTED=EQUAL
    /MISSING ANALYSIS.
```

Lassen Sie diese Syntax »laufen«, erhalten Sie den in Abbildung 14.3 wiedergegebenen SPSS-Output. Zunächst weist SPSS tabellarisch die Unterschiede zwischen den beobachteten und den gemäß Gleichverteilungsannahme erwarteten Werten der Variablen *einkommensmitte* aus. Erwartungsgemäß unterscheiden sich die betreffenden Werte nicht voneinander, sodass der $\chi^2$-Testwert gleich 0 ist. Folgerichtig gibt SPSS in der Spalte für die (asymptotische)

Signifikanz den Wert 1,000 an. Die Nullhypothese der Gleichverteilung wird für *einkommensmitte* nicht verworfen. Anders formuliert: Der Chi-Quadrat-Verteilungstest funktioniert in diesem Beispielfall mit einer im Datensatz konstruierten Gleichverteilung der Daten für *einkommensmitte*.

## *Beispiel 2*

Lassen Sie uns nun noch testen, ob SPSS eine Ungleichverteilung der Werte einer ausgewählten Variablen via Chi-Quadrat-Verteilungstest erkennt. Zu diesem Zweck betrachten Sie bitte die in Abbildung 14.4 dargestellte Datendatei *UNGLEICHVERTEILUNG.SAV* und achten auf die Verteilung der Variablen *einkommensmitte*.

Dass die Häufigkeitsverteilung von *einkommensmitte* keineswegs einer Gleichverteilung folgt, wird schon beim Blick auf Abbildung 14.4 deutlich. Um die betreffende beobachtete Häufigkeitsverteilung noch besser erkennen zu können, finden Sie in Abbildung 14.5 das Histogramm zu den elf Einkommensklassen von *UNGLEICHVERTEILUNG.SAV*. Nachdem (wieder) die in *UNGLEICHVERTEILUNG.SAV* enthaltene Häufigkeitstabelle über die Gewichtung mit der Variablen *häufigkeiten* in das erforderliche SPSS-Format übertragen worden ist, ist das betreffende Histogramm via GRAFIKEN|HISTOGRAMM erstellt worden, wobei in der HISTOGRAMM-Dialogbox noch zusätzlich bei NORMALVERTEILUNGSKURVE ANZEIGEN ein Häkchen gesetzt worden ist. Es ergibt sich das Bild einer symmetrischen Verteilung, die der eingezeichneten Normalverteilungskurve (»Glockenkurve«) zumindest nicht unähnlich ist.

**Chi-Quadrat-Test**

**Häufigkeiten**

**einkommensmitte**

| Beobachtetes N | Erwartete Anzahl | Residuum |
|---|---|---|

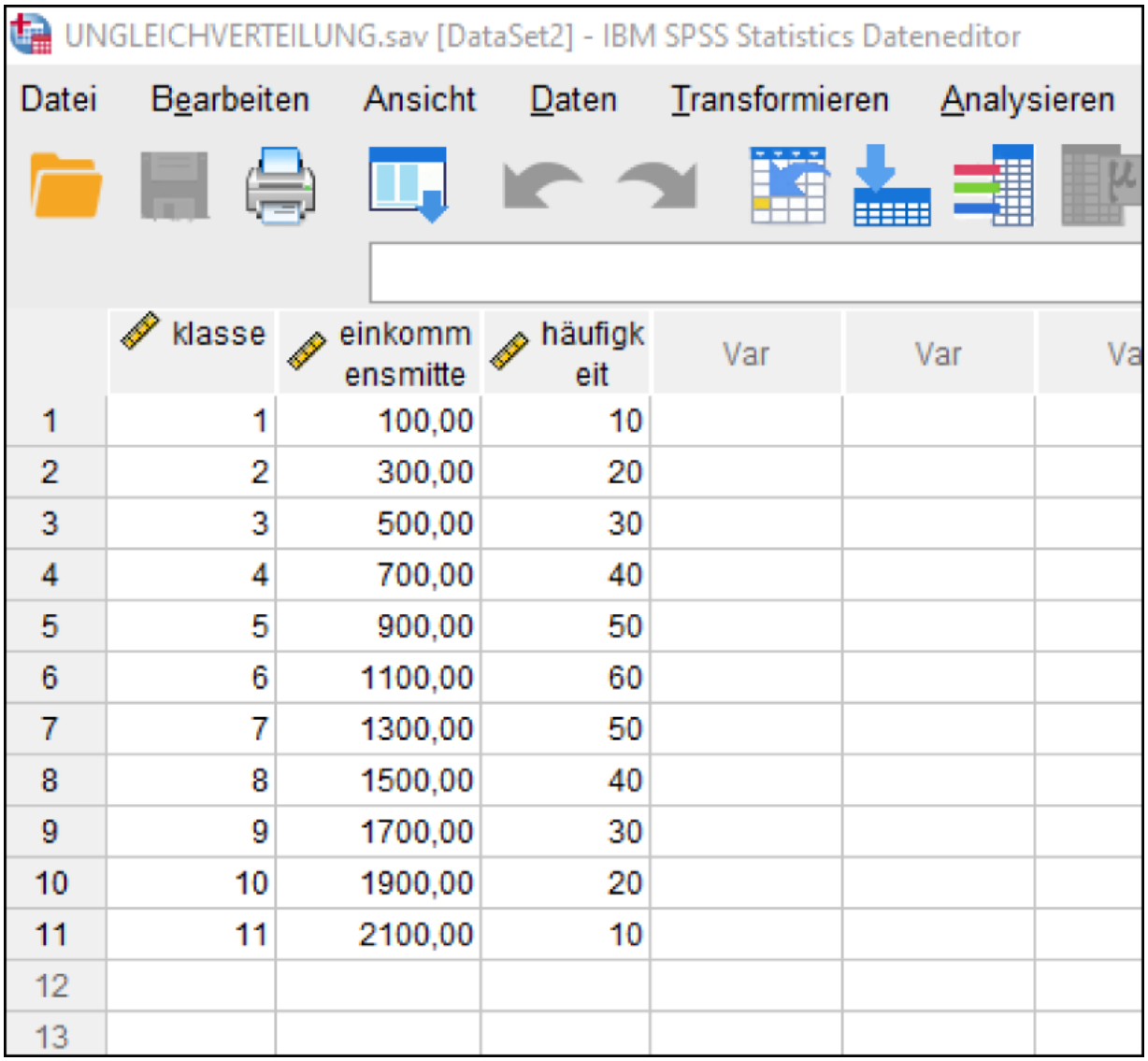
| | N | Anzahl | Residuum |
|---|---|---|---|
| 100,00 | 10 | 10,0 | ,0 |
| 300,00 | 10 | 10,0 | ,0 |
| 500,00 | 10 | 10,0 | ,0 |
| 700,00 | 10 | 10,0 | ,0 |
| 900,00 | 10 | 10,0 | ,0 |
| 1100,00 | 10 | 10,0 | ,0 |
| 1300,00 | 10 | 10,0 | ,0 |
| 1500,00 | 10 | 10,0 | ,0 |
| 1700,00 | 10 | 10,0 | ,0 |
| 1900,00 | 10 | 10,0 | ,0 |
| Gesamt | 100 | | |

**Teststatistiken**

| | einkommensmitte |
|---|---|
| Chi-Quadrat | ,000[a] |
| df | 9 |
| Asymp. Sig. | 1,000 |

a. Bei 0 Zellen (0,0%) werden weniger als 5 Häufigkeiten erwartet. Die kleinste erwartete Zellenhäufigkeit ist 10,0.

**Abbildung 14.3:** Ergebnis des Chi-Quadrat-Verteilungstests auf Gleichverteilung für die Variable *einkommensmitte* (*GLEICHVERTEILUNG.SAV* )




| | klasse | einkommensmitte | häufigkeit | Var | Var | Va |
|---|---|---|---|---|---|---|
| 1 | 1 | 100,00 | 10 | | | |
| 2 | 2 | 300,00 | 20 | | | |
| 3 | 3 | 500,00 | 30 | | | |
| 4 | 4 | 700,00 | 40 | | | |
| 5 | 5 | 900,00 | 50 | | | |
| 6 | 6 | 1100,00 | 60 | | | |
| 7 | 7 | 1300,00 | 50 | | | |
| 8 | 8 | 1500,00 | 40 | | | |
| 9 | 9 | 1700,00 | 30 | | | |
| 10 | 10 | 1900,00 | 20 | | | |
| 11 | 11 | 2100,00 | 10 | | | |
| 12 | | | | | | |
| 13 | | | | | | |

**Abbildung 14.4:** Datendatei *UNGLEICHVERTEILUNG.SAV*

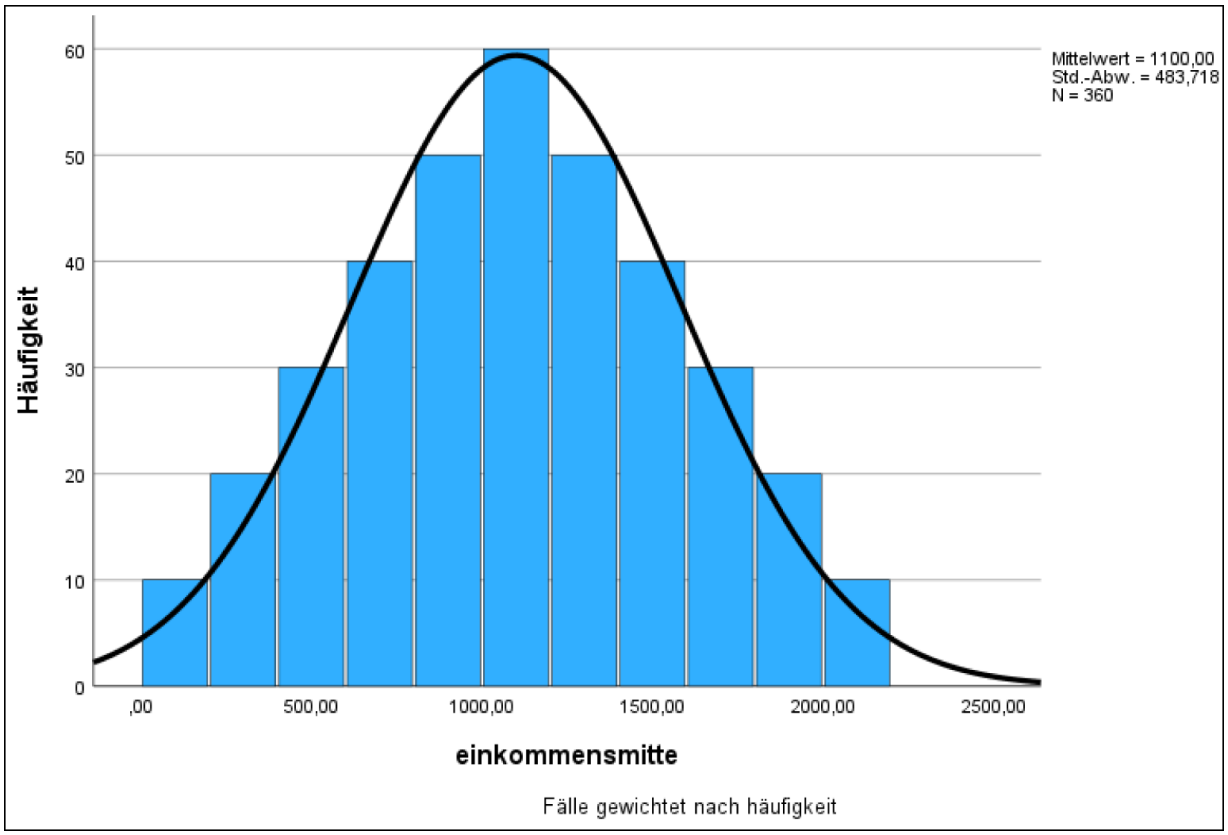



**Abbildung 14.5:** Histogramm bezüglich der klassifizierten Variablen *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*) mit eingezeichneter Normalverteilungskurve

## *Test auf Gleichverteilung in Beispiel 2*

Lassen Sie uns trotzdem den Chi-Quadrat-Verteilungstest für *einkommensmitte* so durchführen, wie bereits beim ersten Beispiel im vorhergehenden Abschnitt angegeben (mit ALLE KATEGORIEN GLEICH), jetzt aber nicht auf der Datenbasis von *GLEICHVERTEILUNG.SAV*, sondern von *UNGLEICHVERTEILUNG.SAV*. Als Ergebnis erhalten Sie, dass die Gleichverteilungs-Nullhypothese abgelehnt wird (siehe Abbildung 14.6, in der bei »Asymp(totische) Sig(nifikanz)« »< 0,001« steht). Alles andere wäre aber auch ein »Skandal« gewesen!

## *Test auf Normalverteilung in Beispiel 2*

Abbildung 14.5 hat ja Anlass zur Hypothese gegeben, dass die Variable *einkommensmitte* in *UNGLEICHVERTEILUNG.SAV* einer Normalverteilung folgt. Einen entsprechenden Test auf Normalverteilung können Sie auch mithilfe des Chi-Quadrat-

Verteilungstests durchführen, und zwar genauso wie einen Test auf eine beliebig andere Verteilungsform als einer Gleich- oder Normalverteilung. Allerdings ist das entsprechende Testen beim Chi-Quadrat-Verteilungstest in SPSS doch einigermaßen kompliziert. Sie müssen nämlich die jeweils erwarteten Häufigkeiten in der Chi-Quadrat-Dialogbox manuell eingeben, das heißt: auf der Grundlage der jeweiligen Wahrscheinlichkeits- beziehungsweise hier: Dichtefunktion selbst berechnen. In Abbildung 14.7 habe ich für die Variable *einkommensmitte* in *UNGLEICHVERTEILUNG.SAV*, das heißt: für die elf Einkommensklassen dort, die gemäß Normalverteilungs-Dichtefunktion berechneten theoretischen absoluten Häufigkeiten bei ERWARTETETE WERTE selbst eingetragen. Hierbei wurden für die beiden Normalverteilungsparameter, den Erwartungswert und die Standardabweichung, die bereits in Abbildung 14.5 angegebenen Werte von 1.100 beziehungsweise von 483,718 verwendet. Welche elf Werte genau eingegeben worden sind, können Sie auszugsweise anhand von Abbildung 14.7, aber auch anhand der etwas weiter unten stehenden Syntax erkennen, also: 8,30; 12,75; 25,01 und so weiter.

**Chi-Quadrat-Test**

**Häufigkeiten**

**einkommensmitte**

| | Beobachtetes N | Erwartete Anzahl | Residuum |
|---|---|---|---|
| 100,00 | 10 | 32,7 | -22,7 |
| 300,00 | 20 | 32,7 | -12,7 |
| 500,00 | 30 | 32,7 | -2,7 |
| 700,00 | 40 | 32,7 | 7,3 |
| 900,00 | 50 | 32,7 | 17,3 |
| 1100,00 | 60 | 32,7 | 27,3 |
| 1300,00 | 50 | 32,7 | 17,3 |
| 1500,00 | 40 | 32,7 | 7,3 |

| | | | |
|---|---|---|---|
| 1700,00 | 30 | 32,7 | -2,7 |
| 1900,00 | 20 | 32,7 | -12,7 |
| 2100,00 | 10 | 32,7 | -22,7 |
| Gesamt | 360 | | |

**Teststatistiken**

| | einkommensmitte |
|---|---|
| Chi-Quadrat | 86,111[a] |
| df | 10 |
| Asymp. Sig. | <,001 |

a. Bei 0 Zellen (0,0%) werden weniger als 5 Häufigkeiten erwartet. Die kleinste erwartete Zellenhäufigkeit ist 32,7.

**Abbildung 14.6:** Ergebnis des Chi-Quadrat-Verteilungstests auf Gleichverteilung für die Variable *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV* )

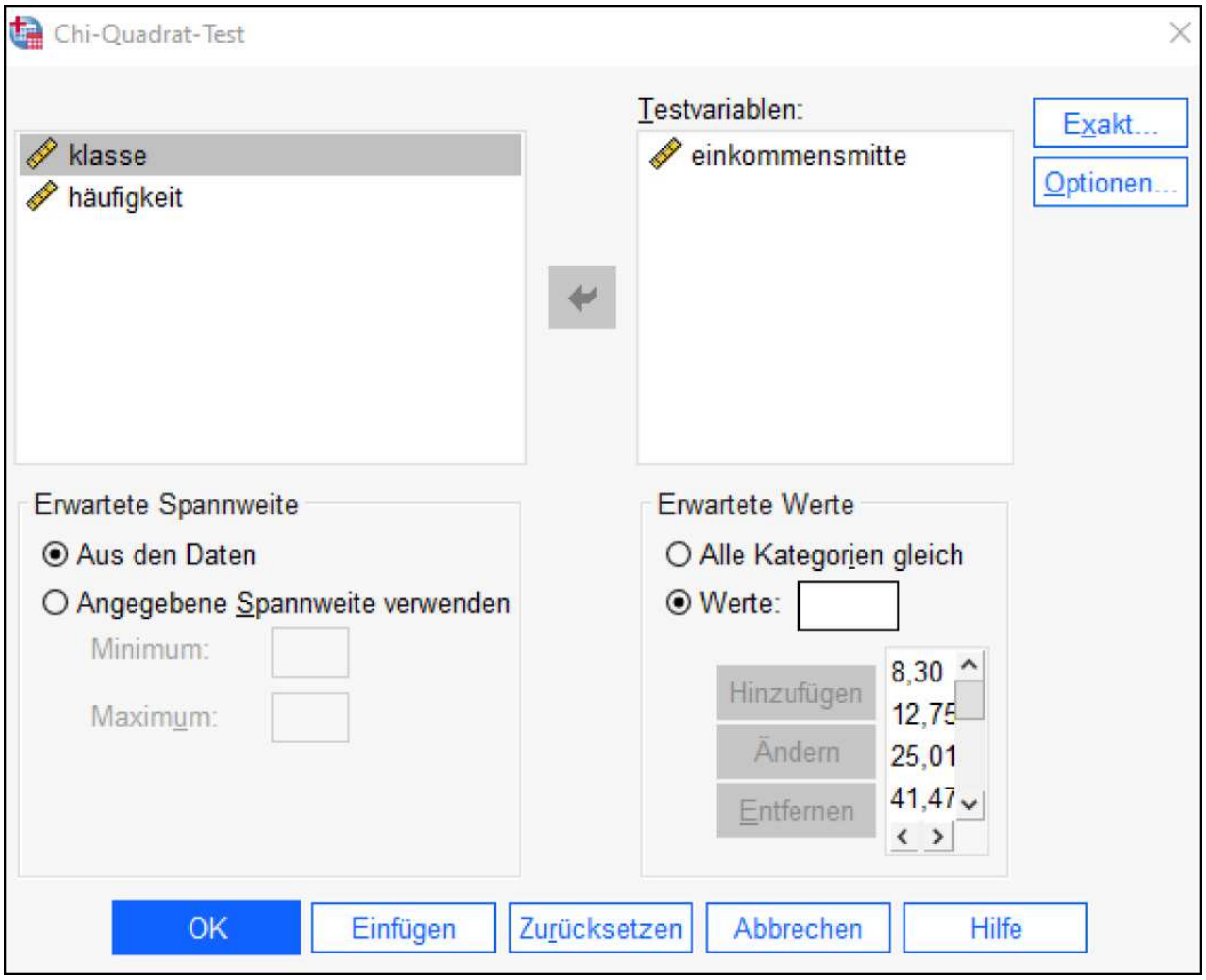



**Abbildung 14.7:** Aufruf des Chi-Quadrat-Verteilungstests auf Normalverteilung für die Variable *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*)

Allerdings zeigt sich an dieser Stelle eine gewisse Schwäche im Rahmen des Chi-Quadrat-Verteilungstests: Die selbst einzutragenden Häufigkeiten sind mehr oder weniger ungenau und können daher das Testergebnis verzerren!

An Syntax müssen Sie in diesem Fall Folgendes eingeben:

```
NPAR TESTS
    /CHISQUARE=einkommensmitte
    /EXPECTED=8.30 12.75 25.01 41.47 58.09 68.76 58.09 41.47 25.01 12.75 8.30
    /MISSING ANALYSIS.
```

Die Resultate, die die Ausführung dieser Syntax mit sich bringt, finden sich in Abbildung 14.8. Wie ersichtlich, wird die Nullhypothese der

Normalverteilung für *einkommensmitte* in *UNGLEICHVERTEILUNG.SAV* bei gängigen Irrtumswahrscheinlichkeiten (bis zu 5 Prozent) nicht abgelehnt; erst ab einer Irrtumswahrscheinlichkeit von 15,5 Prozent würde diese Nullhypothese abgelehnt werden. Sie können also im zweiten Beispiel hier auf Basis des Chi-Quadrat-Verteilungstests von einer (zumindest näherungsweisen) Normalverteilung der Variablen einkommensmitte ausgehen (natürlich nur bei einer bestimmten Irrtumswahrscheinlichkeit).

Ein (großes) Problem des **Chi-Quadrat-Verteilungstests** besteht darin, dass bei einem vergleichsweise großen Stichprobenumfang die Nullhypothese bezüglich einer bestimmten Verteilung der Zufallsvariablen im Regelfall abgelehnt wird. Der Grund hierfür ist, dass in einem solchen Fall der $\chi^2$-Testwert sehr hoch wird. Diese Grundschwäche einer $\chi^2$-verteilten Größe wurde bereits in Kapitel 8 bei der Behandlung des Chi-Quadrat-Kontingenzkoeffizienten dargelegt.

**Chi-Quadrat-Test**

**Häufigkeiten**

**einkommensmitte**

| | Beobachtetes N | Erwartete Anzahl | Residuum |
|---|---|---|---|
| 100,00 | 10 | 8,3 | 1,7 |
| 300,00 | 20 | 12,8 | 7,3 |
| 500,00 | 30 | 25,0 | 5,0 |
| 700,00 | 40 | 41,5 | -1,5 |
| 900,00 | 50 | 58,1 | -8,1 |
| 1100,00 | 60 | 68,8 | -8,8 |
| 1300,00 | 50 | 58,1 | -8,1 |

| | | | |
|---|---|---|---|
| 1500,00 | 40 | 41,5 | -1,5 |
| 1700,00 | 30 | 25,0 | 5,0 |
| 1900,00 | 20 | 12,8 | 7,3 |
| 2100,00 | 10 | 8,3 | 1,7 |
| Gesamt | 360 | | |

**Teststatistiken**

| | einkommensmitte |
|---|---|
| Chi-Quadrat | 14,406[a] |
| df | 10 |
| Asymp. Sig. | ,155 |

a. Bei 0 Zellen (0,0%) werden weniger als 5 Häufigkeiten erwartet. Die kleinste erwartete Zellenhäufigkeit ist 8,3.

**Abbildung 14.8:** Ergebnis des Chi-Quadrat-Verteilungstests auf Normalverteilung für die Variable *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*)

Verdoppeln Sie im oben diskutierten zweiten Beispiel die Besetzungshäufigkeiten sowohl bei der empirischen als auch bei der theoretischen Verteilung (bezüglich der empirischen Verteilung ist dies in *UNGLEICHVERTEILUNG2.SAV* abgespeichert, und die erwarteten Häufigkeitswerte verdoppeln sich zu 16,60; 25,50; 50,02 und so weiter), erhöht sich die Teststatistik $\chi^2$ von 14,406 auf 29,547. (Dass hier keine exakte Verdopplung der Teststatistik $\chi^2$ stattfindet, ist auf Rundungsungenauigkeiten bei den erwarteten Häufigkeitswerten zurückzuführen.) Wie aus Abbildung 14.9 ersichtlich, wird jetzt die Nullhypothese der Normalverteilung verworfen, ohne dass sich an der zugrunde liegenden Verteilungsform strukturell irgendetwas verändert hat.

**Chi-Quadrat-Test**

## Häufigkeiten

**einkommensmitte**

| | Beobachtetes N | Erwartete Anzahl | Residuum |
|---|---|---|---|
| 100,00 | 20 | 16,6 | 3,4 |
| 300,00 | 40 | 25,0 | 15,0 |
| 500,00 | 60 | 50,1 | 9,9 |
| 700,00 | 80 | 83,0 | -3,0 |
| 900,00 | 100 | 116,3 | -16,3 |
| 1100,00 | 120 | 137,6 | -17,6 |
| 1300,00 | 100 | 116,3 | -16,3 |
| 1500,00 | 80 | 83,0 | -3,0 |
| 1700,00 | 60 | 50,1 | 9,9 |
| 1900,00 | 40 | 25,5 | 14,5 |
| 2100,00 | 20 | 16,6 | 3,4 |
| Gesamt | 720 | | |

**Teststatistiken**

| | einkommensmitte |
|---|---|
| Chi-Quadrat | 29,547[a] |
| df | 10 |
| Asymp. Sig. | ,001 |

a. Bei 0 Zellen (0,0%) werden weniger als 5 Häufigkeiten erwartet. Die kleinste erwartete Zellenhäufigkeit ist 16,6.

**Abbildung 14.9:** Ergebnis des Chi-Quadrat-Verteilungstests auf Normalverteilung für die Variable *einkommensmitte* bei Fallzahlenverdoppelung (*UNGLEICHVERTEILUNG2.SAV*)

# *Kolmogorov/Smirnov-Verteilungstest*

Eine Alternative zum vorstehend thematisierten Chi-Quadrat-Verteilungstest ist der Kolmogorov/Smirnov-Verteilungstest. Bei diesem Test muss – im Unterschied zum Chi-Quadrat-Verteilungstest – die jeweilige Zellenbesetzung der erwarteten absoluten Häufigkeiten nicht unbedingt mindestens fünf Fälle betragen.

Bezüglich des Erwartungswerts μ und der Standardabweichung σ werden beim Kolmogorov/Smirnov-Test für die Teststatistik die gleichen Werte angenommen wie bei den erhobenen Daten.

Sie erhalten den Kolmogorov/Smirnov-Verteilungstest über das SPSS-Menü mit ANALYSIEREN|NICHT PARAMETRISCHE TESTS|KLASSISCHE DIALOGFELDER|K-S BEI EINER STICHPROBE. Wollen Sie hierbei auf Normalverteilung testen, klicken Sie bitte naheliegenderweise NORMAL an.

Beim Kolmogorov/Smirnov-Verteilungstest wird die **Teststatistik** aus der empirischen und der theoretischen Verteilung gebildet. (F symbolisiert in den beiden folgenden Formeln die jeweilige Verteilungsfunktion.) Die maximale diesbezügliche absolute Abweichung ist beim Kolmogorov/Smirnov-Test die Teststatistik $D_{prüf}$:

$$D_{prüf} = \text{Max} \left| F_{theor}(X) - F_{emp}(X) \right|.$$

Beim Kolmogorov/Smirnov-Test liegen für die empirische Verteilung in der Regel klassifizierte Daten vor, sodass sich in diesem Fall die empirische Verteilungsfunktion $F_{emp}(X)$ als Treppenfunktion darstellt. Da diese mit einer stetigen theoretischen Verteilungsfunktion $F_{theor}(X)$ verglichen wird, werden die genannten Abweichungen an den Sprungstellen der empirischen Verteilungsfunktion berechnet. Dies geschieht zum einen an der Unter- und zum anderen an der Obergrenze einer Klasse, sodass Sie für jede Klasse zwei Abweichungen erhalten:

$$d_i^{(o)} = \left| F_{theor}(X_i) - F_{emp}(X_i) \right|$$

und

$$d_i^{(u)} = \left| F_{theor}(X_i) - F_{emp}(X_{i-1}) \right|.$$

Die kritischen Werte für den Kolmogorov/Smirnov-Test hängen allein vom Stichprobenumfang und dem geforderten Signifikanzniveau ab. Sie finden sich in Tabellenform zahlreich im Internet. Für einen Stichprobenumfang von mindestens 30 Fällen gilt für den kritischen Wert beim Kolmogorov/Smirnov-Test folgende Formel, in der α für die Irrtumswahrscheinlichkeit und n für die Fallzahl in Ihrer Datendatei stehen:

$$D_{krit} = \frac{\sqrt[2]{\ln\left(\frac{2}{\alpha}\right)}}{\sqrt[2]{2 \cdot n}}.$$

## *Beispiel 1*

Auch beim Kolmogorov/Smirnov-Verteilungstest soll – wie oben beim Chi-Quadrat-Verteilungstest – zunächst die Gleichverteilungs-Nullhypothese für die Variable *einkommensmitte* aus dem (in Abbildung 14.1 dargelegten) Datensatz *GLEICHVERTEILUNG.SAV* getestet werden. Hierzu rufen Sie bitte den Kolmogorov/Smirnov-Test – wie bereits erwähnt – über ANALYSIEREN|NICHT PARAMETRISCHE TESTS|KLASSISCHE DIALOGFELDER|K-S BEI EINER STICHPROBE auf (nachdem Sie natürlich Ihre Daten wieder mit *häufigkeit* gewichtet haben). Es öffnet sich ein Dialogfenster, in dem Sie bei TESTVERTEILUNG von den vier Möglichkeiten ein Häkchen bei GLEICHVERTEILUNG setzen (siehe Abbildung 14.10).

Der vorstehenden Menüabfolge liegt folgende Syntax zugrunde:

```
NPAR TESTS
    /K-S(UNIFORM)=einkommensmitte
    /MISSING ANALYSIS
    /KS_SIM CIN(99) SAMPLES(10000).
```

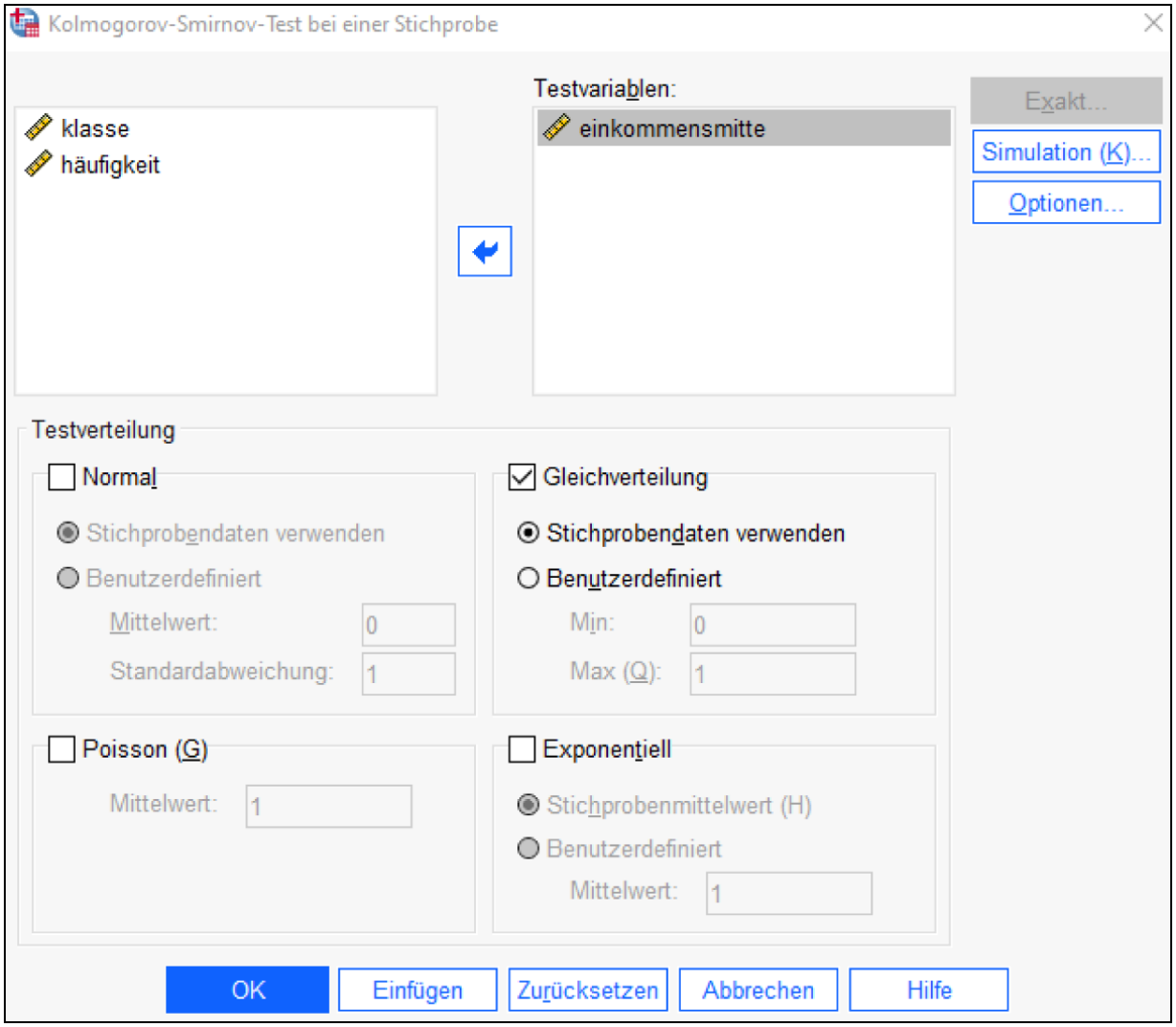



**Abbildung 14.10:** Aufruf des Kolmogorov/Smirnov-Verteilungstests auf Gleichverteilung für die Variable *einkommensmitte* (*GLEICHVERTEILUNG.SAV*)

Im vorliegenden Beispiel erbringt der Kolmogorov/Smirnov-Verteilungstest das in Abbildung 14.11 wiedergegebene Ergebnis, allerdings nicht so eindrucksvoll wie oben der Chi-Quadrat-Verteilungstest. Der Kolmogorov/Smirnov-Verteilungstest verweist aber doch ziemlich klar in Bezug auf die Variable *einkommensmitte* aus dem Datensatz *GLEICHVERTEILUNG.SAV* darauf, dass die Nullhypothese der Gleichverteilung nicht verworfen wird. Der Grund dafür ist, dass die Überschreitungswahrscheinlichkeit immerhin 25,2 Prozent beträgt – und damit deutlich oberhalb der üblichen Signifikanz-Obergrenze von 5 Prozent liegt.

**Kolmogorov-Smirnov-Test bei einer Stichprobe**

| | | | einkommensmitte |
|---|---|---|---|
| N | | | 100 |
| Parameter der Gleichverteilung[a,b] | Minimum | | 100,00 |
| | Maximum | | 1900,00 |
| Extremste Differenzen | Absolut | | ,100 |
| | Positiv | | ,100 |
| | Negativ | | -,100 |
| Teststatistik | | | ,100 |
| Monte-Carlo-Signifikanz (2-seitig)[c] | Sig. | | ,252 |
| | 99% Konfidenzintervall | Untergrenze | ,241 |
| | | Obergrenze | ,263 |

a. Die zu testende Verteilung ist eine Gleichverteilung.

b. Aus den Daten berechnet.

c. Lilliefors-Methode auf der Basis von 10000 Monte-Carlo-Stichproben mit Startwert 2000000.

**Abbildung 14.11:** Ergebnis des Kolmogorov/Smirnov-Verteilungstests auf Gleichverteilung für *einkommensmitte* (*GLEICHVERTEILUNG.SAV*)

## *Beispiel 2*

Auch beim Kolmogorov/Smirnov-Verteilungstest wird nachfolgend für die Variable einkommensmitte aus dem Datensatz *UNGLEICHVERTEILUNG.SAV* sowohl auf Gleich- als auch auf Normalverteilung getestet. Zunächst verwenden Sie in diesem Zusammenhang wieder *häufigkeit* als Gewichtungsvariable. Anschließend rufen Sie die beiden genannten Testvarianten wieder über ANALYSIEREN|NICHT PARAMETRISCHE TESTS|KLASSISCHE DIALOGFELDER|K-S BEI EINER STICHPROBE auf und setzen dieses Mal bei TESTVERTEILUNG von den vier Möglichkeiten sowohl bei GLEICHVERTEILUNG als auch bei NORMAL ein Häkchen (siehe Abbildung 14.12).

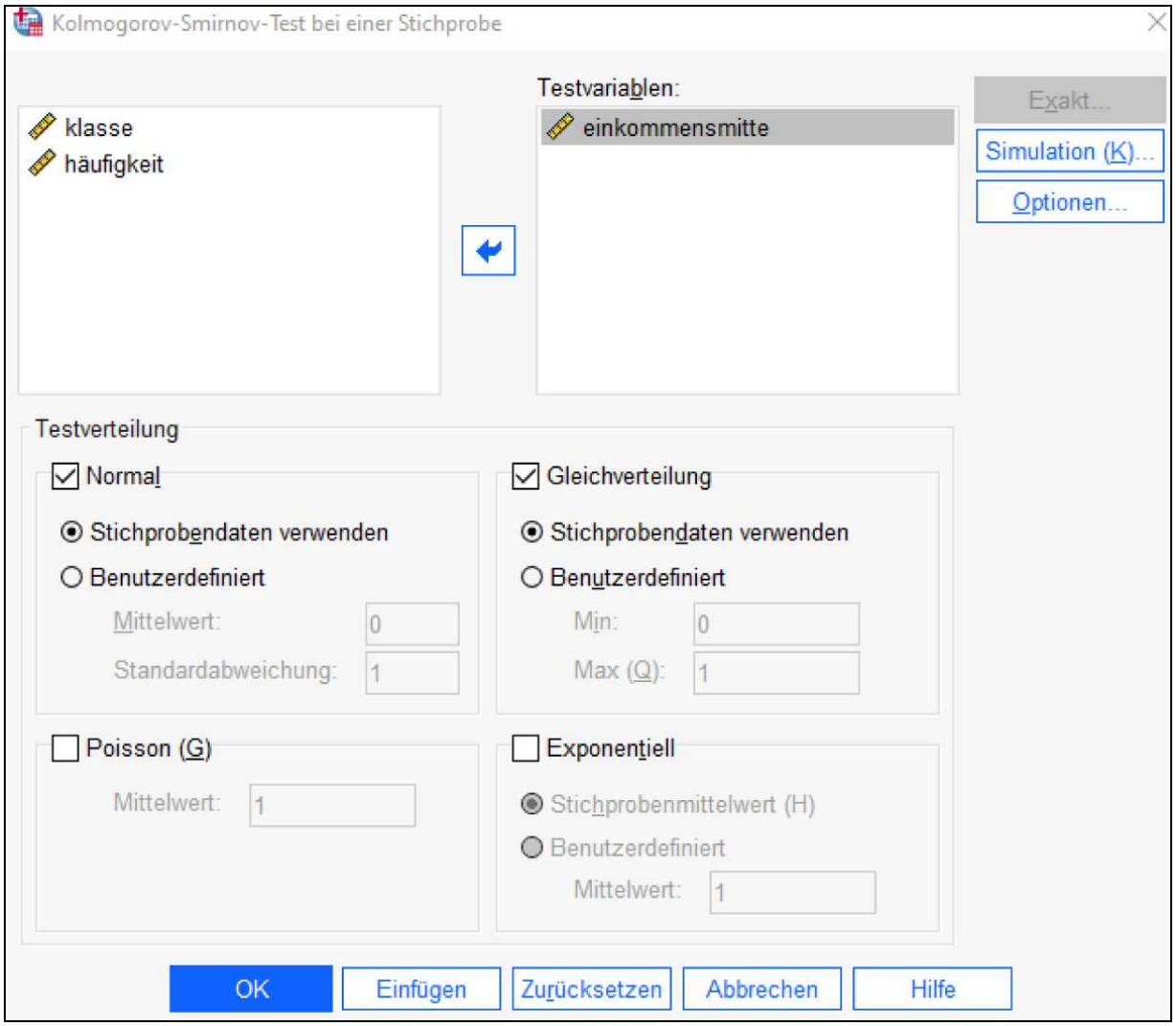



**Abbildung 14.12:** Aufruf des Kolmogorov/Smirnov-Verteilungstests auf Gleich- und Normalverteilung für *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*)

Das Vorstehende geht mit dieser Syntax einher:

```
NPAR TESTS
    /K-S(NORMAL)=einkommensmitte
    /K-S(UNIFORM)=einkommensmitte
    /MISSING ANALYSIS
    /KS_SIM CIN(99) SAMPLES(10000).
```

Als Ergebnisse (siehe Abbildung 14.13) erhalten Sie, dass für *einkommensmitte* aus dem Datensatz *UNGLEICHVERTEILUNG.SAV* sowohl die Gleichverteilungs- als auch die Normalverteilungsannahme verworfen werden. Dies kommt jeweils dadurch zum Ausdruck, dass in der »Signifikanz (2-seitig)«-Zeile jeweils »< 0,001« steht.

**Kolmogorov-Smirnov-Test bei einer Stichprobe**

| | | | einkommensmitte |
|---|---|---|---|
| N | | | 360 |
| Parameter der Normalverteilung[a,b] | Mittelwert | | 1100,0000 |
| | Std.-Abweichung | | 483,71819 |
| Extremste Differenzen | Absolut | | ,083 |
| | Positiv | | ,083 |
| | Negativ | | -,083 |
| Teststatistik | | | ,083 |
| Asymp. Sig. (2-seitig)[c] | | | <,001 |
| Monte-Carlo-Signifikanz (2-seitig)[d] | Sig. | | <,001 |
| | 99% Konfidenzintervall | Untergrenze | ,000 |
| | | Obergrenze | ,000 |

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

c. Signifikanzkorrektur nach Lilliefors.

d. Lilliefors-Methode auf der Basis von 10000 Monte-Carlo-Stichproben mit Startwert 334431365.

**Kolmogorov-Smirnov-Anpassungstest 2**

| | | | einkommensmitte |
|---|---|---|---|
| N | | | 360 |
| Parameter der Gleichverteilung[a,b] | Minimum | | 100,00 |
| | Maximum | | 2100,00 |
| Extremste Differenzen | Absolut | | ,133 |
| | Positiv | | ,133 |
| | Negativ | | -,133 |
| Teststatistik | | | ,133 |
| Monte-Carlo-Signifikanz (2-seitig)[c] | Sig. | | <,001 |
| | 99% Konfidenzintervall | Untergrenze | ,000 |
| | | Obergrenze | ,000 |

a. Die zu testende Verteilung ist eine Gleichverteilung.

b. Aus den Daten berechnet.

c. Lilliefors-Methode auf der Basis von 10000 Monte-Carlo-Stichproben mit Startwert 334431365.

**Abbildung 14.13:** Ergebnis des Kolmogorov/Smirnov-Verteilungstests auf Gleich- und Normalverteilung für *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*)

Bezüglich der Normalverteilungsannahme kommt demnach in unserem Beispiel der Kolmogorov/Smirnov-Verteilungstest zu einem anderen Ergebnis als oben der Chi-Quadrat-Anpassungstest. Grundsätzlich verweist dies auf gewisse Unschärfen beziehungsweise Unvollkommenheiten beim statistischen Testen.

# *Shapiro/Wilk-Normalverteilungstest*

Um die Normalverteilungsannahme zu überprüfen, können Sie über ANALYSIEREN|DESKRIPTIVE STATISTIKEN|EXPLORATIVE DATENANALYSE in der Subdialogbox DIAGRAMME bei NORMALVERTEILUNGSDIAGRAMM MIT TESTS ein Häkchen setzen (siehe Abbildung 14.14). Auch auf diese Weise rufen Sie den Kolmogorov/Smirnov-Verteilungstest auf. Zusätzlich wird hierdurch auch noch der in der Wissenschaft durchaus verbreitete Shapiro/Wilk-Normalverteilungstest in die Wege geleitet. Die vorstehende Menüabfolge bezieht sich wieder auf die Variable *einkommensmitte* aus der Datendatei *UNGLEICHVERTEILUNG.SAV* (wobei Sie wieder alle Fälle mittels *häufigkeit* zuvor gewichten müssen).

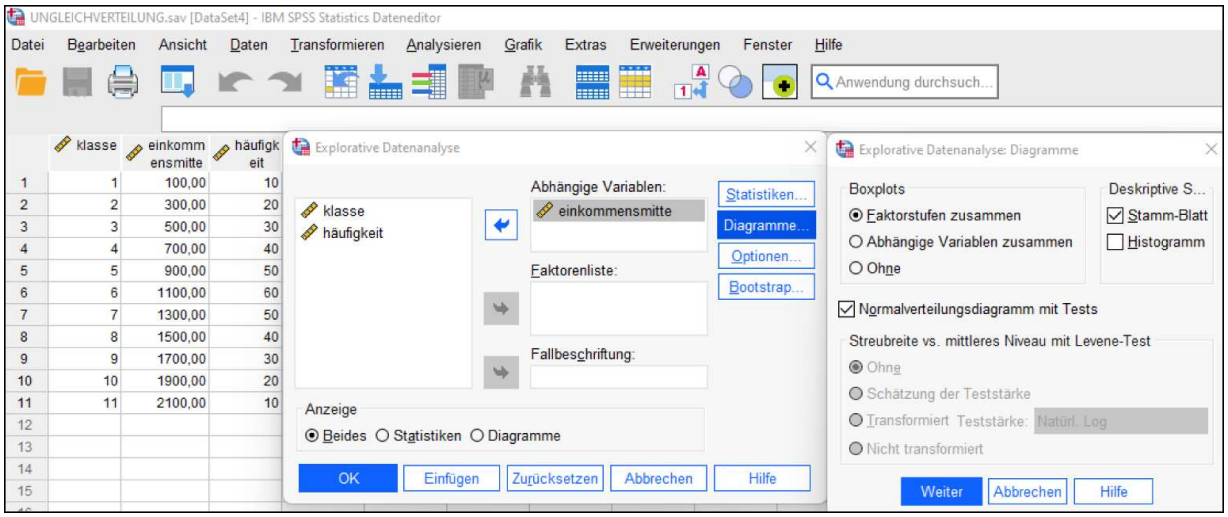



**Abbildung 14.14:** Aufruf des Kolmogorov/Smirnov- und des Shapiro/Wilk-Verteilungstests auf Normalverteilung für *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*)

Die zugehörige (umfassende) Syntax ist:

```
EXAMINE VARIABLES=einkommensmitte
    /PLOT BOXPLOT STEMLEAF NPPLOT
    /COMPARE GROUPS
    /STATISTICS DESCRIPTIVES
    /CINTERVAL 95
    /MISSING LISTWISE
    /NOTOTAL.
```

Wie der nachstehende Ergebnisauszug in Abbildung 14.15 zeigt, wird für *einkommensmitte* aus *UNGLEICHVERTEILUNG.SAV* sowohl beim Kolmogorov/Smirnov- als auch beim Shapiro/Wilk-Verteilungstest die Normalverteilungsannahme verworfen. Unter »Signifikanz« wird ja sowohl beim Kolmogorov/Smirnov- als auch beim Shapiro/Wilk-Test ausgewiesen, ob die Nullhypothese abzulehnen ist oder nicht. Wenn die Signifikanz in Form der ausgewiesenen Überschreitungswahrscheinlichkeit sehr klein ist, werden Sie die Nullhypothese (hier: die These der Normalverteiltheit) verwerfen. Dies ist hier bei beiden Tests der Fall.

Der Shapiro/Wilk-Test bezieht sich im Unterschied zu den anderen vorstehend behandelten Verteilungstests ausschließlich auf die Normalverteilung. Inhaltlich ist der Shapiro/Wilk-Schätzwert SW definiert als Quotient aus der gemäß der Referenz(normal)verteilung erwarteten Varianz $\sigma^2$ und dem Schätzwert für die tatsächliche Varianz in der Stichprobe $S^2$ (n = Stichprobengröße):

**Tests auf Normalverteilung**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistik | df | Signifikanz | Statistik | df | Signifikanz |
| einkommensmitte | ,083 | 360 | <,001 | ,975 | 360 | <,001 |

a. Signifikanzkorrektur nach Lilliefors

**Abbildung 14.15:** Ergebnisauszug bezüglich des Kolmogorov/Smirnov- und des Shapiro/Wilk-Verteilungstests auf Normalverteilung für *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*) (Output-Auszug)

$$SW = \frac{\sigma^2}{(n-1)\cdot S^2}.$$

Wegen der Quotientenbildung und der Quadrate im Zähler und Nenner wird der kritische Wert für den betreffenden Signifikanztest einer Tabelle zur F-Verteilung entnommen.

# Kapitel 15

# Tests III: Was hängt mit wem zusammen?

## IN DIESEM KAPITEL

Chi-Quadrat-Kontingenztest

Test bezüglich der Rangkorrelation nach Spearman

Test bezüglich der Korrelation nach Bravais/Pearson

In Kapitel 15 geht es um die Aussagekraft statistischer Zusammenhänge zwischen zwei Variablen. Für die bereits aus Kapitel 8 bekannten Kennziffern des Chi-Quadrat-Kontingenzkoeffizienten, des Spearman'schen Rangkorrelationskoeffizienten und des Bravais/Pearson-Korrelationskoeffizienten werden Tests präsentiert, mit deren Hilfe eine Aussage getroffen werden kann, ob der entsprechend festgestellte Koeffizientenwert statistisch signifikant, also statistisch aussagekräftig ist oder nicht.

## *Vorbemerkung*

Im Folgenden werden für drei **Assoziations-** beziehungsweise **Korrelationsmaße** Tests auf Signifikanz der Koeffizientenwerte durchgeführt, und zwar für:

- ✔ den Chi-Quadrat-Kontingenzkoeffizienten,
- ✔ den Spearman'schen Rangkorrelationskoeffizienten und
- ✔ den Bravais/Pearson-Korrelationskoeffizienten.

Dies sind für verschiedene Skalenniveaus die an dieser Stelle jeweils verbreitetsten Kennziffern. Den Chi-Quadrat-Kontingenzkoeffizienten können Sie bereits für nominalskalierte Variablen berechnen. Der Spearman'sche Rangkorrelationskoeffizient erfordert mindestens Ordinalskalenniveau bei den einbezogenen Variablen und der Bravais/Pearson-Korrelationskoeffizient sogar metrisches Skalenniveau bei beiden Variablen (wobei der Bravais/Pearson-Korrelationskoeffizient in der Praxis allerdings mitunter auch für die Korrelation zwischen einer metrischen Variablen und einem 0/1-Dummy berechnet wird).

Die **Nullhypothese** lautet für die einzelnen Maße in SPSS jeweils: »Der Koeffizientenwert beläuft sich auf null.« Dies bedeutet, dass die Nullhypothese keine Assoziation beziehungsweise Korrelation zwischen den betrachteten Variablen behauptet. Kann die Nullhypothese entsprechend beim jeweiligen Test verworfen werden, können Sie von einer vorhandenen Assoziation beziehungsweise Korrelation zwischen den betreffenden Variablen ausgehen und folglich sagen: »Die gemessene Assoziation beziehungsweise Korrelation zwischen den betrachteten Variablen ist signifikant (von null verschieden).«

# *Test für den Chi-Quadrat-Kontingenzkoeffizienten*

Beim **Chi-Quadrat-Kontingenzkoeffizienten** befasst man sich vom Grundsatz her mit der Frage, ob zwischen insgesamt zwei nominalskalierten und/oder kategorialen Variablen ein statistischer Zusammenhang besteht. Entsprechend fragt man sich bei dem hiermit gekoppelten Test, ob dieser Zusammenhang signifikant ist oder nicht.

Das folgende Beispiel beruht auf der Datendatei *GESCHL_FARBE.SAV* und den darin enthaltenen Variablen *geschlecht* und *augenfarbe*. Wenn Sie für diesen Zusammenhang ein Testergebnis erhalten wollen, müssen Sie im SPSS-Menü ANALYSIEREN|DESKRIPTIVE STATISTIKEN|KREUZTABELLEN anklicken (siehe Abbildung 15.1) beziehungsweise als SPSS-Syntax eingeben:

```
CROSSTABS
    /TABLES=geschlecht BY augenfarbe
    /FORMAT=AVALUE TABLES
    /STATISTICS=CHISQ
    /CELLS=COUNT EXPECTED
    /COUNT ROUND CELL.
```

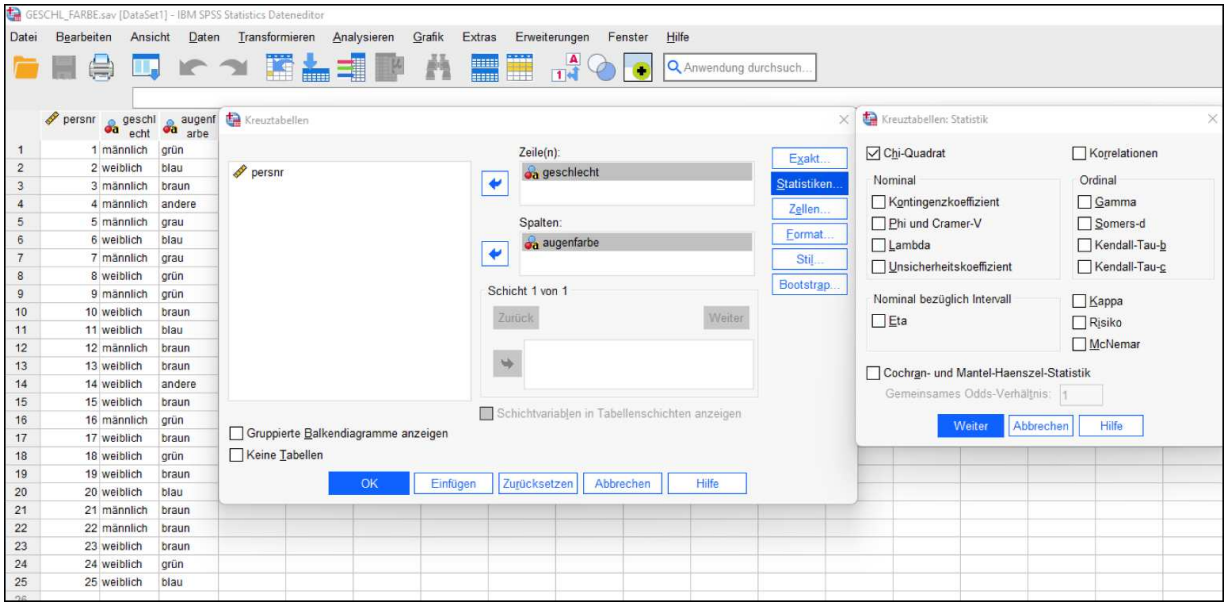



**Abbildung 15.1:** Aufruf des Chi-Quadrat-Tests (*GESCHL_FARBE.SAV*)

Mit der vorstehenden Syntax lassen Sie sich neben den beobachteten auch noch die erwarteten (absoluten) Häufigkeiten ausgeben. Dies geschieht deshalb, damit Sie neben der *Kreuz-* auch noch die *Indifferenztabelle* bei statistischer Unabhängigkeit erhalten.

Die **Voraussetzungen** für den Chi-Quadrat-Kontingenzkoeffizententest lauten:

- ✔ nominalskalierte beziehungsweise kategoriale Variablen
- ✔ Stichprobengröße: möglichst mehr als 50 Fälle (bei weniger als 20 Fällen: exakter Test nach Fisher, zwischen 20 und 50 Fällen: Korrektur nach Yates; Yates-Korrektur: -0,5 von jedem Summanden – also von jedem Ausdruck bei der Summenbildung – des Chi-Quadrat-Tests)
- ✔ erwartete Zellenhäufigkeiten: mehr als 5 Fälle (ansonsten: Test nach Fisher)

Der hier diskutierte Chi-Quadrat-Kontingenzkoeffizententest ist bitte nicht mit dem in Kapitel 14 behandelten Chi-Quadrat-Verteilungstest zu verwechseln. Bei dem letztgenannten Test geht es um eine Beurteilung der Signifikanz einer bestimmten (Häufigkeits-)Verteilungsform, während der Chi-Quadrat-Kontingenzkoeffizententest sich auf die Signifikanzbeurteilung hinsichtlich der Assoziationsmessziffer Chi-Quadrat-Kontingenzkoeffizient bezieht.

Die **Chi-Quadrat-Teststatistik** sieht so aus:

$$\chi^2 = \sum_{i=1}^{N} \frac{(h_{i,b} - h_{i,e})^2}{h_{i,e}}.$$

In dieser Formel gibt $h_{i,b}$ die beobachteten und die $h_{i,e}$ die (bei statistischer Unabhängigkeit) erwarteten absoluten Häufigkeiten an.

Im vorliegenden Beispiel liefert SPSS den in Abbildung 15.2 dargelegten Output. Dabei beinhaltet die obere Pivottabelle sowohl die *Kreuz-* als auch die *Indifferenztabelle* für die Beziehung zwischen *geschlecht* und *augenfarbe*.

Der geschätzte Testwert für den Chi-Quadrat-Kontingenzkoeffizienten lautet 6,667, und er ist – verglichen mit für die Chi-Quadrat-Verteilung tabellierten Werten – nicht-signifikant von null verschieden. Dies geht aus der Spalte »Asymptotische Signifikanz (zweiseitig)« hervor. Dort steht in Abbildung 15.2 ein Wert von 0,155; das heißt: Erst ab einer Irrtumswahrscheinlichkeit von 15,5 Prozent würde die Nullhypothese der statistischen Unabhängigkeit zwischen den Variablen *geschlecht* und *augenfarbe* abgelehnt.

Die Chi-Quadrat-Teststatistik in Höhe von 6,667 hat SPSS dabei aus der Gegenüberstellung der Werte in der Kreuz- und in der Indifferenztabelle berechnet:

$$\chi^2 = \frac{(1-0{,}8)^2}{0{,}8} + \frac{(0-2)^2}{2} + \frac{(4-4)^2}{4} + \frac{(2-0{,}8)^2}{0{,}8} + \frac{(3-2{,}4)^2}{2{,}4} + \frac{(1-1{,}2)^2}{1{,}2} + \frac{(5-3)^2}{3} + \frac{(6-6)^2}{6} + \frac{(0-1{,}2)^2}{1{,}2} + \frac{(3-3{,}6)^2}{3{,}6} = 6{,}667.$$

**geschlecht * augenfarbe Kreuztabelle**

| | | | augenfarbe | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | andere | blau | braun | grau | grün | Gesamt |
| geschlecht | männlich | Anzahl | 1 | 0 | 4 | 2 | 3 | 10 |
| | | Erwartete Anzahl | ,8 | 2,0 | 4,0 | ,8 | 2,4 | 10,0 |
| | weiblich | Anzahl | 1 | 5 | 6 | 0 | 3 | 15 |
| | | Erwartete Anzahl | 1,2 | 3,0 | 6,0 | 1,2 | 3,6 | 15,0 |
| Gesamt | | Anzahl | 2 | 5 | 10 | 2 | 6 | 25 |
| | | Erwartete Anzahl | 2,0 | 5,0 | 10,0 | 2,0 | 6,0 | 25,0 |

**Chi-Quadrat-Tests**

| | Wert | df | Asymptotische Signifikanz (zweiseitig) |
|---|---|---|---|
| Pearson-Chi-Quadrat | 6,667[a] | 4 | ,155 |
| Likelihood-Quotient | 9,100 | 4 | ,059 |
| Anzahl der gültigen Fälle | 25 | | |

a. 9 Zellen (90,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist ,80.

**Abbildung 15.2:** Chi-Quadrat-Kontingenzkoeffiziententest für den Zusammenhang zwischen *geschlecht* und *augenfarbe* (*GESCHL_FARBE.SAV*)

# *Test für den Rangkorrelationskoeffizienten nach Spearman*

Beim **Rangkorrelationskoeffizienten nach Spearman** geht es um einen Zusammenhang zwischen zwei Variablen, die mindestens Ordinalskalenniveau aufweisen.

Für diesen Zusammenhang wird untersucht, ob eine gleichsinnige (= positive) oder eine gegenläufige (= negative) Korrelation vorliegt.

Der Rangkorrelationskoeffizient nach Spearman ist grundsätzlich wie folgt definiert (der – der Abwechslung wegen – leicht unterschiedliche Ausdruck zur aus Kapitel 8 bekannten Formel ergibt sich lediglich durch Umformung im Nenner):

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n}.$$

In dieser Formel gibt $d_i$ die Rangdifferenzen zwischen den beiden betrachteten Variablen an, und n steht für die Fallzahl. Der Spearman'sche Rangkorrelationskoeffizient ist dabei zwischen −1 und +1 definiert. Bei $r_S = -1$ liegt eine vollständig-negative Rangkorrelation zwischen den beiden Variablen vor, bei $r_S = +1$ hingegen eine vollständig-positive Rangkorrelation. Bei $r_S = 0$ besteht (bezogen auf die jeweiligen Ränge beziehungsweise Platzierungen) kein Zusammenhang.

Im Folgenden beziehe ich mich auf das aus Kapitel 8 bekannte »Speerwurf-Beispiel«. In diesem Beispiel sind für fünf Speerwerferinnen in den (früheren) Jahren 2016 und 2017 jeweils die Rangfolgen bei den betreffenden Wurfweiten angegeben, und zwar in den Variablen *weite2016* und *weite2017*. Die entsprechenden Daten finden Sie in der Datendatei *SPEERWERFEN.SAV*. Für dieses Beispiel ist in Kapitel 8 ein Spearman'scher Rangkorrelationskoeffizient in Höhe von +0,600 berechnet worden. Nachstehend wird nun geprüft, ob diese ausgewiesene positive (Rang-)Korrelation signifikant ist, das heißt, ob die Nullhypothese der statistischen Unabhängigkeit (= keine Korrelation) zwischen *weite2016* und *weite2017* verworfen werden kann.

Hierzu rufen Sie bitte über das SPSS-Menü den Rangkorrelationskoeffizienten nach Spearman wie folgt auf: ANALYSIEREN|KORRELATION|BIVARIAT, wobei Sie noch zusätzlich (und bitte nur) bei SPEARMAN ein Häkchen setzen (siehe Abbildung 15.3).

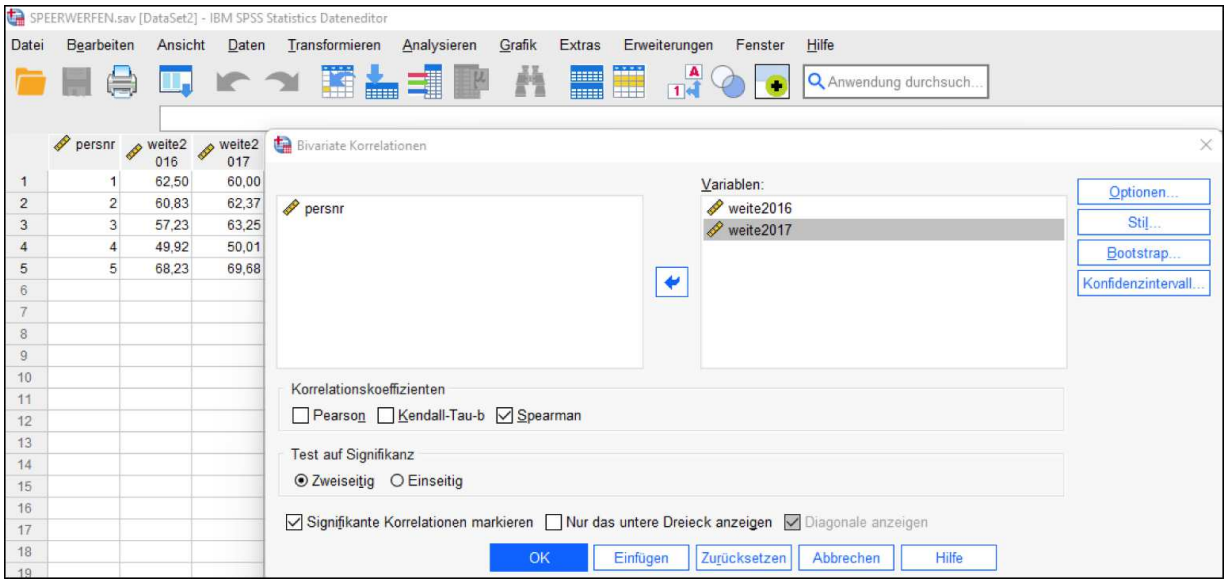



**Abbildung 15.3:** Aufruf des Rangkorrelationskoeffizienten nach Spearman (*SPEERWERFEN.SAV*)

Die zugehörige Syntax ist:

```
NONPAR CORR
    /VARIABLES=weite2016 weite2017
    /PRINT=SPEARMAN TWOTAIL NOSIG
    /MISSING=PAIRWISE.
```

Für den Signifikanztest bezüglich des Spearman'schen Rangkorrelationskoeffizienten ($r_S$) berechnen Sie bitte folgende Testgröße (mit n als Symbol für die Fallzahl):

$$t = \frac{r_s}{\sqrt[2]{\frac{\left(1-r_s^2\right)}{n-2}}}.$$

Diesen Testwert vergleichen Sie in einem nächsten Schritt mit tabellierten Werten der t-Verteilung bei bestimmten vorgegebenen Irrtumswahrscheinlichkeiten. Ist der Testwert – als Betrag – größer als der tabellierte t-Wert, kann die Nullhypothese der statistischen Unabhängigkeit verworfen werden.

Im vorliegenden Beispiel berechnen Sie als Testwert:

$$t = \frac{0{,}600}{\sqrt[2]{\frac{\left(1-0{,}600^2\right)}{5-2}}} = +1{,}299.$$

Dies ist erfahrungsgemäß ein recht niedriger Testwert und lässt befürchten, dass SPSS die festgestellte positive Korrelation zwischen *weite2016* und *weite2017* als nicht-signifikant ausweist. Genau dies passiert auch (siehe Abbildung 15.4). Erst ab einer Irrtumswahrscheinlichkeit von 28,5 Prozent würde bei diesem zweiseitigen Test die Nullhypothese der statistischen Unabhängigkeit verworfen.

**Korrelationen**

| | | | weite2016 | weite2017 |
|---|---|---|---|---|
| Spearman-Rho | weite2016 | Korrelationskoeffizient | 1,000 | ,600 |
| | | Sig. (2-seitig) | . | ,285 |
| | | N | 5 | 5 |
| | weite2017 | Korrelationskoeffizient | ,600 | 1,000 |
| | | Sig. (2-seitig) | ,285 | . |
| | | N | 5 | 5 |

**Abbildung 15.4:** Test für den Zusammenhang zwischen *weite2016* und *weite2017* beim Spearman'schen Rangkorrelationskoeffizienten (*SPEERWERFEN.SAV*)

# *Test für den Korrelationskoeffizienten nach Bravais/Pearson*

Im Rahmen des **Korrelationskoeffizienten nach Bravais/Pearson** wird der lineare Zusammenhang zweier metrisch skalierter Variablen erfasst (beziehungsweise in der statistischen Praxis auch mitunter: zwischen einer metrisch skalierten Variablen und einem 0/1-Dummy). Wie bei der

Rangkorrelation nach Spearman wird untersucht, ob es sich um eine gleichsinnige (= positive) oder um eine gegenläufige (= negative) Korrelation handelt (beim Bravais/Pearson-Korrelationskoeffizienten aber ausschließlich auf lineare Zusammenhänge bezogen).

Wird ein linearer Zusammenhang durch einen entsprechenden Test verworfen, heißt das nicht, dass zwischen den beiden betrachteten Variablen notwendigerweise kein Zusammenhang besteht, da ja eventuell ein nichtlinearer (zum Beispiel ein u-förmiger) Zusammenhang zwischen den beiden Variablen vorliegen kann.

Der Korrelationskoeffizient nach Bravais/Pearson ($r_{BP}$) ist für die beiden Variablen X und Y folgendermaßen definiert (mit n wieder als (Stichproben-)Fallzahl und den X- beziehungsweise Y-Größen mit einem Strich »auf dem Kopf« als den arithmetischen Mittelwerten der beiden Variablen; es handelt sich hier im Übrigen um die ausformulierte Fassung der aus Kapitel 8 bekannten Definition für den Korrelationskoeffizienten nach Bravais/Pearson):

$$r_{BP} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\cdot\left(Y_i - \bar{Y}\right)}{\sqrt[2]{\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right]\cdot\left[\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2\right]}}.$$

Ähnlich wie beim Rangkorrelationskoeffizienten nach Spearman gilt beim Korrelationskoeffizienten nach Bravais/Pearson (jetzt aber auf lineare Zusammenhänge bezogen), dass bei $r_{BP} = -1$ eine vollständig-negative lineare Korrelation zwischen den beiden Variablen vorliegt, bei $r_{BP} = +1$ hingegen eine vollständig-positive lineare Korrelation und dass bei $r_{BP} = 0$ kein linearer Zusammenhang zwischen den beiden Variablen besteht.

Erneut beziehe ich mich nachfolgend auf ein aus Kapitel 8 bekanntes »Speerwurf-Beispiel«, in dem für fünf Speerwerferinnen für die Jahre 2016 und 2017 jeweils die Wurfweiten festgehalten sind (in den

Variablen *weite2016* und *weite2017*). Der zugehörige Datensatz lautet *SPEERWERFEN2.SAV*. In Kapitel 8 ist für dieses Beispiel ein Bravais/Pearson-Korrelationskoeffizient in Höhe von +0,942 berechnet worden. Im Folgenden prüfen Sie bitte, ob diese ausgewiesene positive Korrelation signifikant ist, das heißt, ob die Nullhypothese der statistischen Unabhängigkeit (= keine (lineare) Korrelation) zwischen *weite2016* und *weite2017* verworfen werden kann.

Auch hier gelangen Sie über ANALYSIEREN|KORRELATIONEN|BIVARIAT im SPSS-Menü zur entsprechenden Korrelation, wobei Sie jetzt aber bei PEARSON (und nur dort) ein Häkchen setzen (siehe Abbildung 15.5).

Sie können den Korrelationskoeffizienten nach Bravais/Pearson auch über ein kleines SPSS-Syntaxprogramm aufrufen:

```
CORRELATIONS
    /VARIABLES=weite2016 weite2017
    /PRINT=TWOTAIL NOSIG
    /MISSING=PAIRWISE.
```

Für den Signifikanztest bezüglich des Bravais/Pearson-Korrelationskoeffizienten müssen Sie als Testgröße berechnen (mit n als Symbol für die Fallzahl):

$$t = \frac{r_{BP} \cdot \sqrt[2]{(n-2)}}{\sqrt[2]{1-r_{BP}^2}}.$$

Diesen Testwert vergleichen Sie in einem nächsten Schritt mit tabellierten Werten der t-Verteilung bei bestimmten vorgegebenen Irrtumswahrscheinlichkeiten. Ist der Testwert – als Betrag – größer als der tabellierte t-Wert, wird die Nullhypothese der (linearen) statistischen Unabhängigkeit zwischen *weite2016* und *weite2017* verworfen.

Im vorliegenden Beispiel erhalten Sie als Testwert folgendes Resultat:

$$t = \frac{0{,}942 \cdot \sqrt[2]{(5-2)}}{\sqrt[2]{1-0{,}942^2}} = +4{,}862.$$

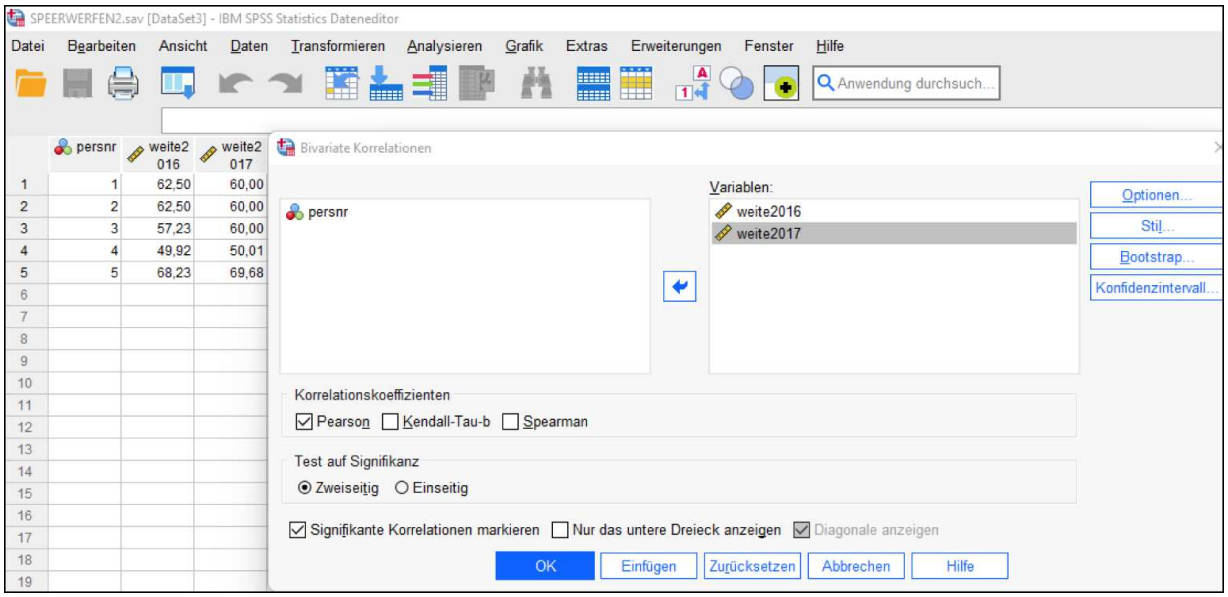



**Abbildung 15.5:** Aufruf des Korrelationskoeffizienten nach Bravais/Pearson (*SPEERWERFEN2.SAV*)

Dieser Testwert ist relativ hoch, was die Hoffnung nährt, dass die Nullhypothese der statistischen Unabhängigkeit zwischen *weite2016* und *weite2017* verworfen werden kann. Und genau dies passiert (bei einem zweiseitigen Test) in SPSS: Wie Abbildung 15.6 zeigt, ist die betreffende (positive) Korrelation bei einer Irrtumswahrscheinlichkeit von 1,7 Prozent signifikant. Man kann sagen, dass die betreffende Korrelation zwar nicht hochsignifikant, aber doch immerhin »ziemlich« signifikant ist.

**Korrelationen**

| | | weite2016 | weite2017 |
|---|---|---|---|
| weite2016 | Pearson-Korrelation | 1 | ,942* |
| | Sig. (2-seitig) | | ,017 |
| | N | 5 | 5 |
| weite2017 | Pearson-Korrelation | ,942* | 1 |
| | Sig. (2-seitig) | ,017 | |
| | N | 5 | 5 |

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

**Abbildung 15.6:** t-Test für den Zusammenhang zwischen *weite2016* und *weite2017* beim Bravais/Pearson-Korrelationskoeffizienten (*SPEERWERFEN2.SAV*)

# Teil V

# SPSS-Syntax für Fortgeschrittene

# IN DIESEM TEIL …

- ✔ Koppelung mit R und Python
- ✔ Schleifenkonstruktionen
- ✔ Makros
- ✔ Beispielhafte Syntaxprogramme

# Kapitel 16

# Koppelung mit R und Python

## IN DIESEM KAPITEL

R, Python und SPSS

R-Anwendungsbeispiel: Robuste Regression

In Kapitel 16 geht es um die Beziehungen zwischen R und Python einerseits und SPSS andererseits. Durch diese Koppelungen können in SPSS weitere statistische Analysen durchgeführt werden. Beispielhaft wird in diesem Zusammenhang auf die robuste Regression Bezug genommen.

## *Zu den R- und den Python-Essentials*

**R** ist eine statistikorientierte Programmiersprache, die auf S basiert. Die sogenannten **R-Essentials** stellen für SPSS 28 eine Integrationslösung dar in Form einer Schnittstelle zwischen R und SPSS, die auch für SPSS 29 Gültigkeit hat. Die Berücksichtigung von R innerhalb von SPSS erscheint sinnvoll, weil R einige Erweiterungen gegenüber SPSS bietet, zum Beispiel:

- ✔ die robuste Regression
- ✔ die polychonische Korrelation
- ✔ die Rasch-Itemanalyse

Etwas weiter unten in diesem Kapitel wird in Bezug auf derartige Erweiterungen die robuste Regression thematisiert.

**Python** ist eine Programmiersprache, deren Elemente mit SPSS mitgeliefert werden. Einige statistische Erweiterungen in SPSS erfordern die Bezugnahme auf Python.

Wenn Sie nur an grundlegenden statistischen Berechnungen interessiert sind, sind Sie mit dem SPSS-Standard-Lieferumfang gut bedient. Für weitergehende statistische Arbeiten kann die Bezugnahme auf R und/oder Python hingegen nutzbringend sein. Insofern sind die Ausführungen dieses Kapitels eher an statistisch Fortgeschrittene adressiert.

Die für SPSS relevanten **R-Essentials** sind prinzipiell leicht über https://github.com/IBMPredictiveAnalytics/R_Essentials_Statistics/releases erhältlich. Dort finden Sie leider für SPSS 29 (noch) kein R-Plug-in (zumindest galt dies bis zur Drucklegung dieser Neuauflage), aber immerhin ein solches für SPSS 28, das Sie bitte auswählen (siehe Abbildung 16.1).

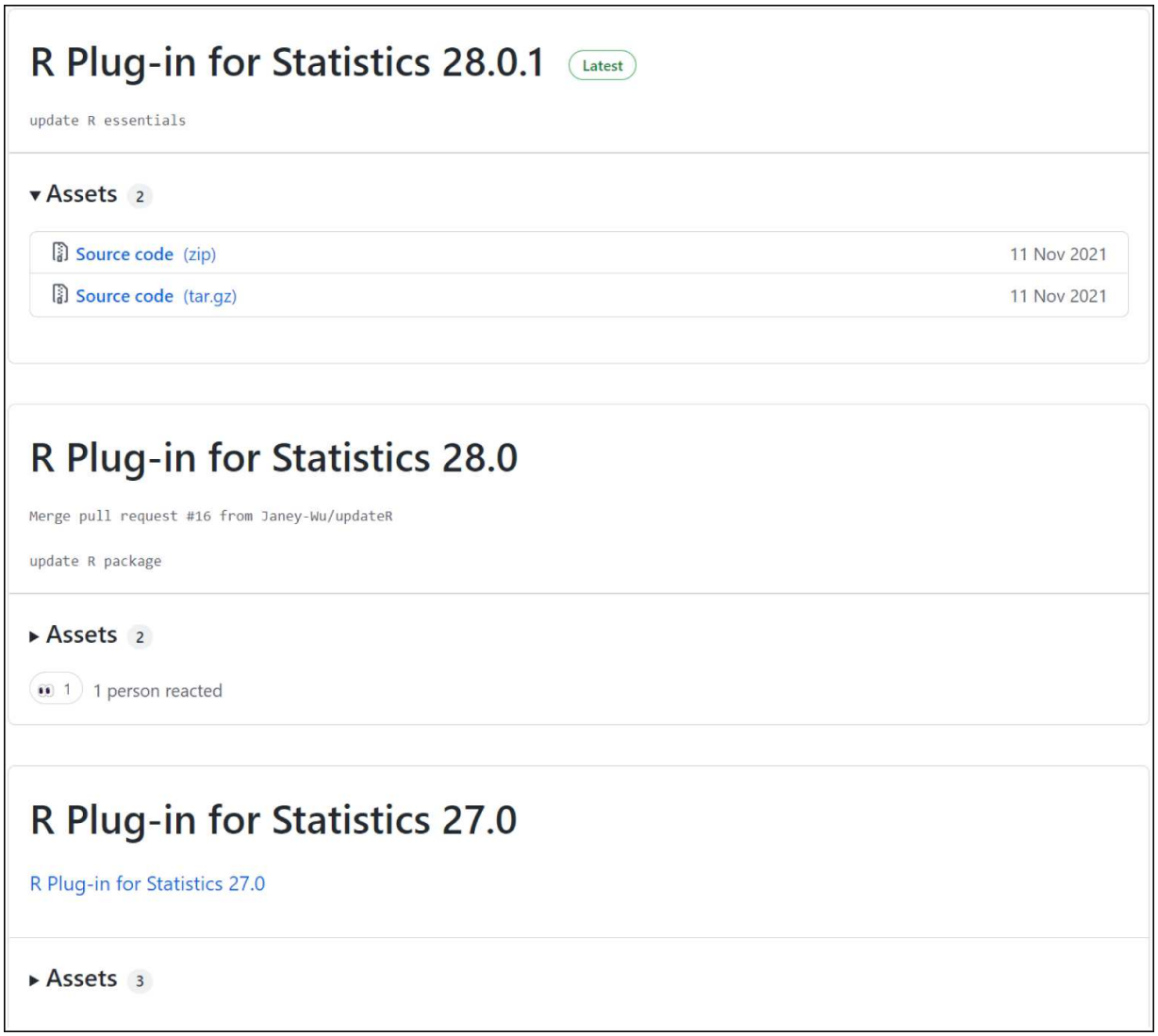



**Abbildung 16.1:** Download-Möglichkeiten für die R-Essentials aus dem Internet (https://github.com/IBMPredictiveAnalytics/R_Essentials_Statistics/releases)

Wenn Sie nun auf »Source code (zip)« klicken, laden Sie eine auf R bezogene ZIP-Datei herunter (siehe Abbildung 16.2).

**Abbildung 16.2:** ZIP-Datei für R-Essentials (SPSS 28, auch für SPSS 29 passend)

Zuvor müssen Sie auf Ihrem Rechner natürlich noch R (zum Beispiel – wie bei mir – Version 4.2.1) installieren. Mit SPSS 29 ist R zum Beispiel in der genannten Version 4.2.1 gut kompatibel. Der Desktop-Button für R 4.2.1 ist in Abbildung 16.3 dargestellt.

**Abbildung 16.3:** Desktop-Button für R 4.2.1

Klicken Sie auf diesen Button, öffnet sich R beziehungsweise das komfortablere *RStudio* für die betreffende Version (siehe Abbildung 16.4). RStudio ist eine grafische Benutzeroberfläche für R-Anwendungen, auf Englisch als RGui abgekürzt (mit Gui = **G**raphical **u**ser interface). Ihre Arbeit mit R wird zwar durch RStudio erleichtert, allerdings ist RStudio etwa im Vergleich zum SPSS-Menü meiner Ansicht nach benutzerunfreundlicher. Sie können RStudio kostenfrei im Internet von der Webseite

http://www.rstudio.com/products/rstudio/download/

herunterladen.

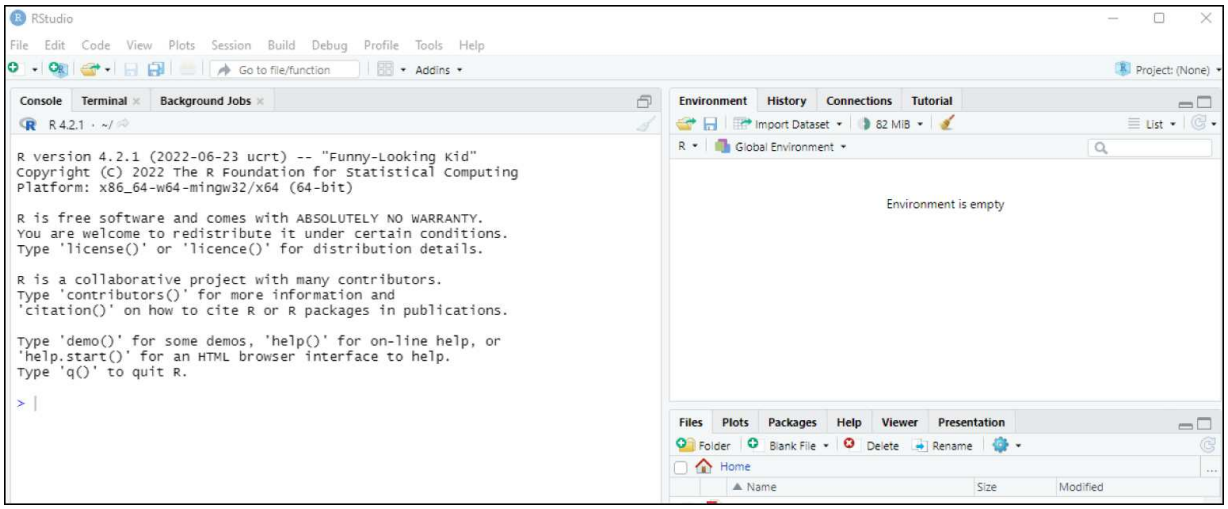



**Abbildung 16.4:** RStudio-Maske (hier für R 4.2.1)

Wenn Sie nun – wie eben angedeutet – die R-Essentials für SPSS 28 (und damit auch für SPSS 29) aus dem Internet heruntergeladen haben, können Sie diese Essentials installieren. Im Rahmen des betreffenden Installationsprozesses werden die R-Essentials sowohl mit Ihrer ausgewählten R-Version als auch mit SPSS verbunden, und R steht Ihnen innerhalb des Benutzerrahmens von SPSS zur Verfügung. Das war ja auch das Ziel!

Im Übrigen können Sie grundsätzlich auch umgekehrt in R auf SPSS-Arbeitsdateien zugreifen. Hierbei beachten Sie bitte, dass in R die Groß- und Kleinschreibung essenziell ist. Während in SPSS große Datensätze zudem nur teilweise in den Arbeitsspeicher (RAM) geschrieben werden, gilt dies in R für die betreffenden Datensätze komplett. Des Weiteren müssen Sie in diesem Zusammenhang unter anderem beachten, dass systemfehlende Werte in SPSS generell als »SYSMIS« deklariert werden, während hierbei in R zwischen numerischen Variablen (mit dem Wert NaN = **N**ot **a** **N**umber) und alphanumerischen Variablen (NA = **N**ot **A**vailable) zu unterscheiden ist.

Da manche R-Essentials die **Python**-Skriptsprache benötigen (zum Beispiel das Erweiterungskommando für die heterogene Korrelation), ist auch die Installation von **Python-Essentials** in SPSS zu empfehlen.

Python-Essentials können Sie bei der SPSS-Installation gleich mitinstallieren; sie sind im Lieferumfang von SPSS standardmäßig enthalten.

# *Statistische Einbindung von R in SPSS*

Unter ERWEITERUNGEN|ERWEITERUNGSHUB können Sie sich auf der Registerkarte INSTALLIERT einen Überblick über in SPSS vorhandene R-Anwendungen (und auch über Python-basierte Anwendungen) verschaffen (siehe Abbildung 16.5).

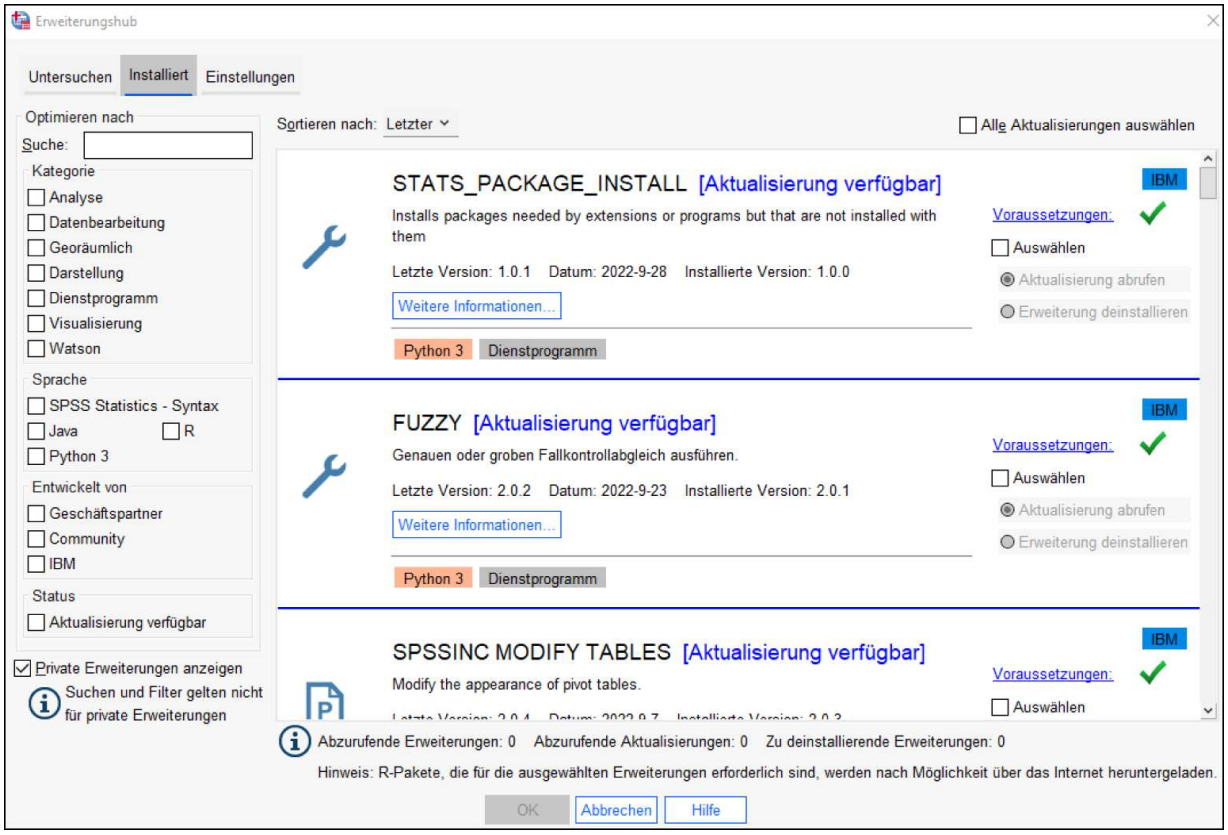



**Abbildung 16.5:** In SPSS installierte (R- beziehungsweise) Python-Anwendungen

Suchen Sie konzentriert nach bereits installierten R-Anwendungen, führen Sie dies im ERWEITERUNGSHUB durch, indem Sie unter INSTALLIERT das Häkchen bei R setzen. Wie Sie anhand von

Abbildung 16.6 erkennen, war bei mir gar keine R-Anwendung installiert (und auch die in Abbildung 16.5 angegebenen Anwendungen sind ja allesamt Python-basierte Anwendungen).

Wie können Sie grundsätzlich nach R-Anwendungen suchen, die Sie dann in SPSS einbetten? Dies tun Sie bitte weiterhin im ERWEITERUNGSHUB, jetzt aber nicht unter INSTALLIERT, sondern unter UNTERSUCHEN. Ich habe in diesem Reiter nach der R-Anwendung RREGRIDGE gesucht, und siehe da: Diese Anwendung ist gefunden worden (siehe Abbildung 16.7).

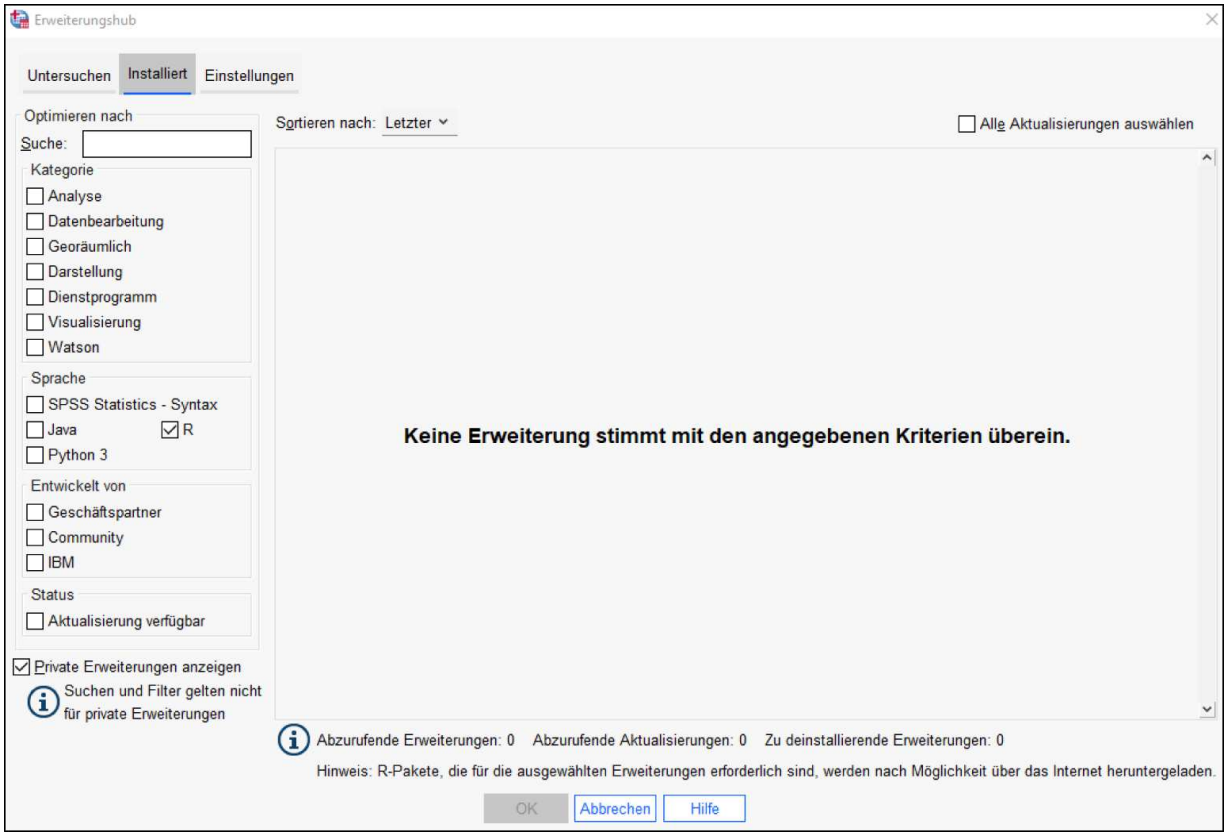



**Abbildung 16.6:** Suche nach in SPSS installierten R-Anwendungen

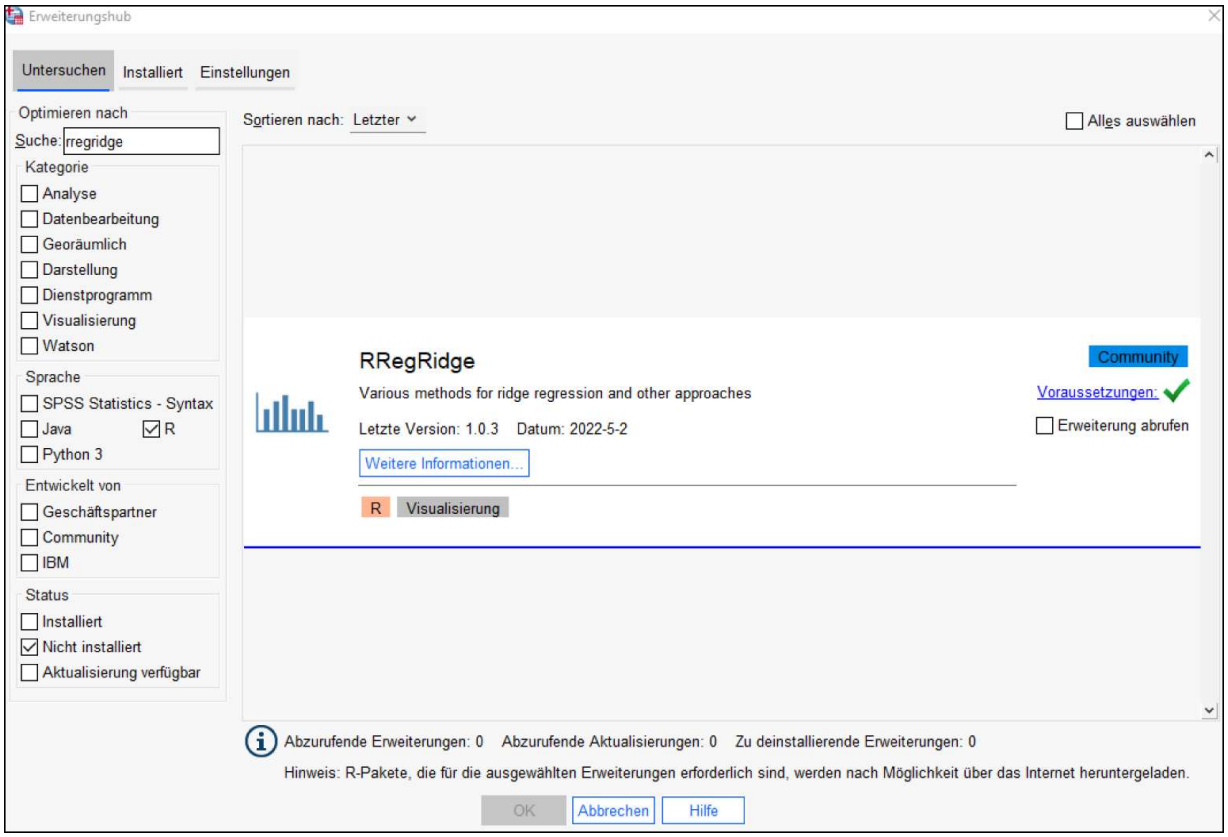



**Abbildung 16.7:** Im SPSS-Erweiterungshub (unter UNTERSUCHEN) gefundene R-Anwendung RREGRIDGE (Beispiel)

Für RREGRIDGE können Sie durch einen Klick auf VORAUSSETZUNGEN prüfen, ob Ihr System für diese Anwendung ausreichend konfiguriert ist. Im vorliegenden Fall ist eine SPSS-Version größer gleich 21 erforderlich, was hier ja gilt (SPSS 29), wie Sie Abbildung 16.8 entnehmen können.

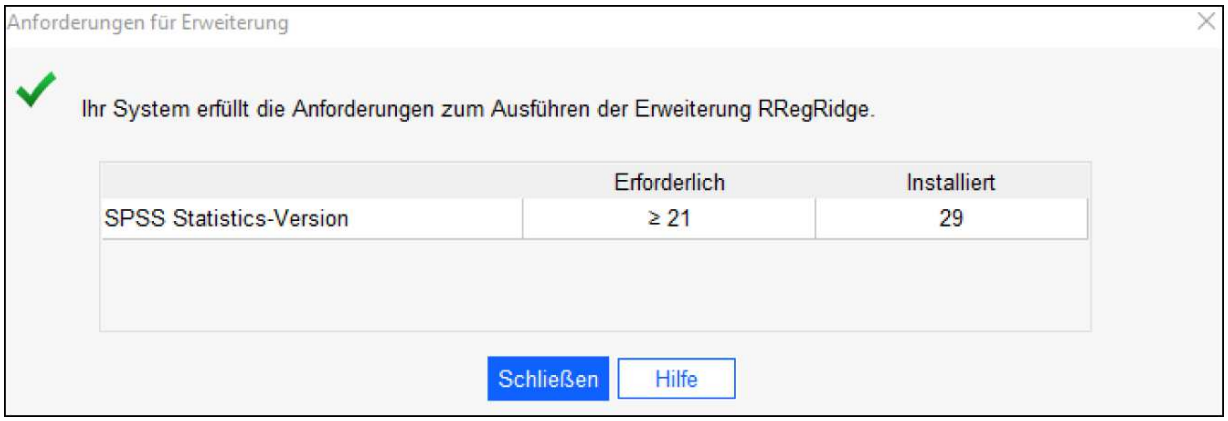



**Abbildung 16.8:** Voraussetzungen für die Funktion RREGRIDGE (Beispiel)

In Abbildung 16.9 ist nun bezüglich der R-Anwendung RREGRIDGE noch bei ERWEITERUNG ABRUFEN ein Häkchen gesetzt worden, damit diese R-Anwendung innerhalb von SPSS genutzt werden kann.

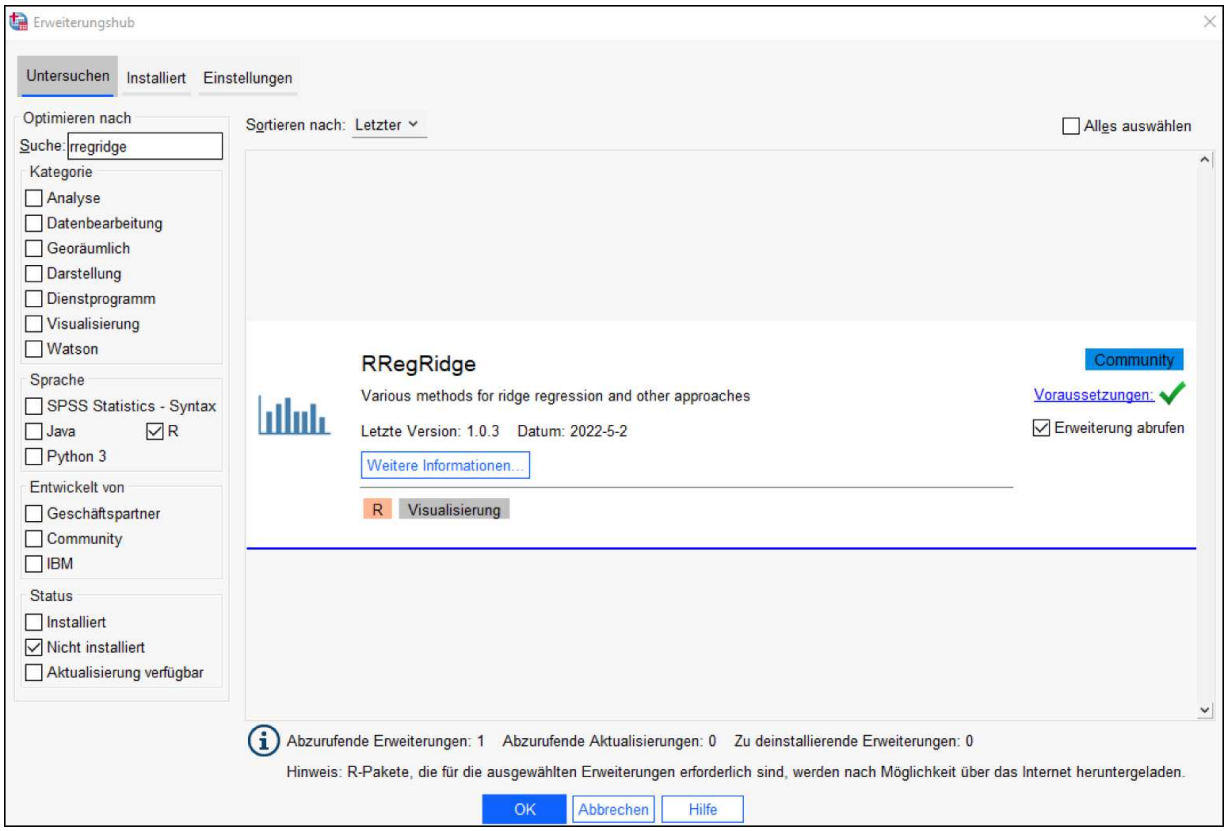



**Abbildung 16.9:** Abruf der R-Erweiterung RREGRIDGE (Beispiel)

Wie aus Abbildung 16.10 zu erkennen ist, ist diese R-Erweiterung erfolgreich in SPSS installiert worden. Diese spezielle Anwendung steht jetzt im SPSS-Menü unter ANALYSIEREN|REGRESSION zu Ihrer Verfügung (siehe Abbildung 16.11), und Sie könnten jetzt mit dieser Anwendung die sogenannte Ridge-Regression durchführen. Diese ist ein alternativer Schätzansatz gegenüber der in Kapitel 9 skizzierten einfachen multiplen Regressionsanalyse auf Basis der Kleinst-Quadrate-Methode.

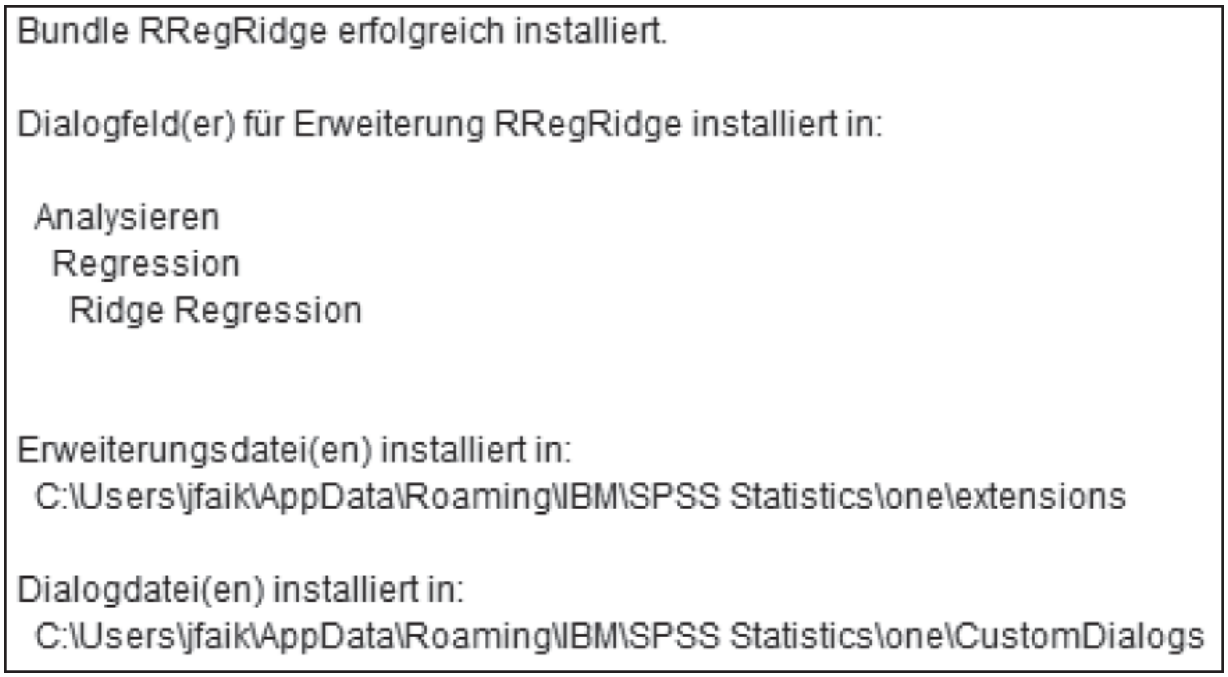



**Abbildung 16.10:** Erfolgreiche Installation der R-Erweiterung RREGRIDGE (Beispiel)

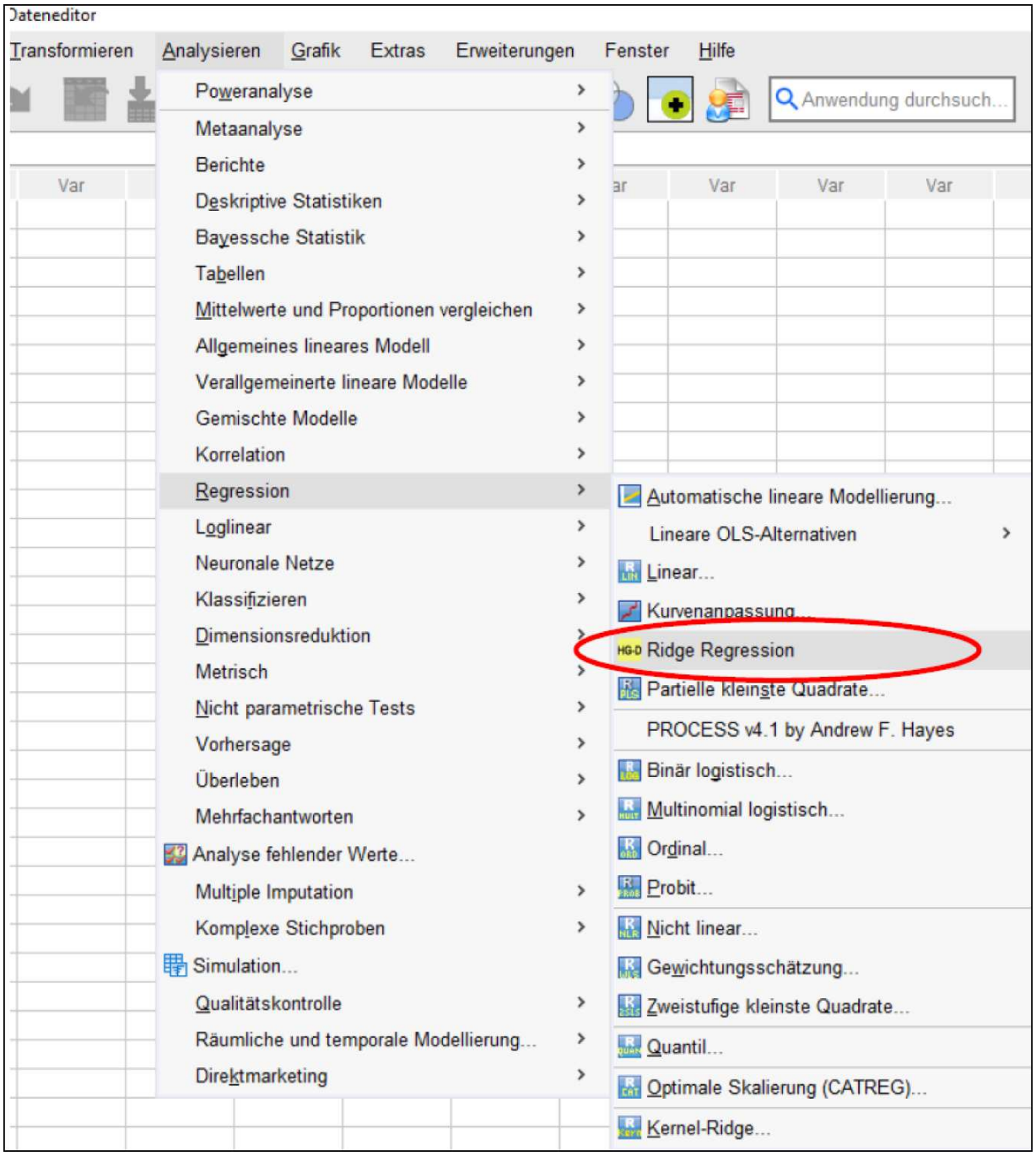



**Abbildung 16.11:** Aufruf von RREGRIDGE (Beispiel) in SPSS

Alternativ beziehungsweise ergänzend zum vorstehend dargelegten Integrationsansatz von R-Anwendungen in das SPSS-Menü können Sie R-Syntax auch in SPSS-Syntax einbauen. Das heißt: Mit R können Sie in SPSS Outputtabellen erstellen, indem Sie in ein SPSS-Syntaxfenster einen R-Syntaxblock einbinden:

```
BEGIN PROGRAM R
…
END PROGRAM.
```

Für die Ausführung von R-Anwendungen sind in der Regel sogenannte *statistische Pakete* erforderlich. Sind diese noch nicht in SPSS über R-Essentials enthalten, müssen Sie sich diese Pakete aus dem Internet besorgen. Dies geht allerdings sehr einfach über ein kleines SPSS-Syntaxprogramm mit R-Bezug (hier bezogen auf das R-Paket »MASS«; als kleiner Service meinerseits finden Sie das nachfolgende Syntaxfile unter *R.SPS*):

```
BEGIN PROGRAM R.
install.packages(´MASS´)
END PROGRAM.
```

Bitte achten Sie darauf (worauf bereits zu Beginn dieses Kapitels hingewiesen worden ist), dass Sie die R-Befehle kleinschreiben. Im Unterschied zu SPSS ist R bekanntlich case-sensitiv, unterscheidet also zwischen Groß- und Kleinschreibung.

Auf die vorstehend geschilderte Weise wird dieses R-Paket in SPSS integriert (siehe Abbildung 16.12).

```
BEGIN PROGRAM R.
install.packages('MASS')
END PROGRAM.
Installiere Paket nach 'C:/Users/jfaik/AppData/Roaming/IBM/SPSS Statistics/one/R42/renv/library/R-4.2/x86_64-w64-mingw32'
(da 'lib' nicht spezifiziert)
versuche URL 'https://cloud.r-project.org/bin/windows/contrib/4.2/MASS_7.3-58.1.zip'
Content type 'application/zip' length 1173901 bytes (1.1 MB)
==================================================
downloaded 1.1 MB
Paket 'MASS' erfolgreich ausgepackt und MD5 Summen abgeglichen
Die heruntergeladenen Binärpakete sind in
                C:\Users\jfaik\AppData\Local\Temp\spss35584\RtmpWyKuQi\downloaded_packages
```

**Abbildung 16.12:** Installation eines R-Pakets (hier von MASS) innerhalb von SPSS

# *R-Anwendungsbeispiel: Robuste Regression*

Im Folgenden wird mit der **robusten Regression** eine über die SPSS-Schnittstelle aufrufbare R-Anwendung thematisiert. Dies geschieht auf Basis der Datendatei *LINEAREREGRESSION_GERADE7.SAV* und in Gegenüberstellung mit der bereits aus Kapitel 9 bekannten *linearen Regression*.

Wie in dem eben genannten Kapitel 9 bereits ausgeführt, weist die lineare Regression bei problematischen Einzelwerten (das heißt bei Ausreißerwerten, bei Messfehlern und dergleichen) mehr oder weniger große Schwächen auf. Weniger anfällig für derartige problematische Einzelwerte ist demgegenüber die robuste Regression, bei der ein besonderer Schätzansatz (sogenannte *M-Schätzer*) zur Anwendung kommt, der zumindest den Einfluss von Fällen mit großen Residuen begrenzt. Dies geschieht durch die Vergabe von Gewichten in dem betreffenden (iterativen) Schätzansatz, und zwar dergestalt, dass Fälle mit großen Residuen ein kleineres Gewicht bei der Schätzung erhalten als Fälle mit relativ kleinen Residuen. Insoweit handelt es sich bei der robusten Regression um eine Spielart der in Kapitel 9 ebenfalls diskutierten WLS-Regression (WLS = Weighted Least Squares).

Für die erwähnte Datendatei *LINEAREREGRESSION_GERADE7.SAV* wird eine robuste Regressionsbeziehung zwischen *ausgaben* (abhängige Variable) und *einkommen* (unabhängige Variable) geschätzt. Dies wird über die Einbindung von R-Syntax in SPSS-Syntax bewerkstelligt. Hierbei müssen Sie berücksichtigen, dass die Schätzung einer robusten Regression in R auf das (oben bereits erwähnte) statistische Paket MASS Bezug nimmt. Die entsprechende Syntax sieht – bei geöffneter SPSS-Datendatei *LINEAREREGRESSION_GERADE7.SAV* – wie folgt aus (siehe hierzu auch das Syntaxfile *R-SYNTAX.SPS*):

```
BEGIN PROGRAM R.
library (´MASS´)
meinedaten <- spssdata.GetDataFromSPSS( )
fm.rlm <- rlm (ausgaben ~ einkommen, data = meinedaten)
summary (fm.rlm)
END PROGRAM.
```

Lassen Sie diese SPSS/R-Syntax »laufen«, erhalten Sie das in Abbildung 16.13 wiedergegebene Schätzergebnis. Die solcherart berechneten Werte

lauten (jeweils auf drei Nachkommastellen gerundet) für das absolute Glied 152,917 und für die Steigung 0,067.

```
BEGIN PROGRAM R.
library ("MASS")
meinedaten <- spssdata.GetDataFromSPSS( )
fm.rlm <- rlm (ausgaben ~ einkommen, data = meinedaten)
summary (fm.rlm)
END PROGRAM.
Warnmeldung:
Paket 'MASS' wurde unter R Version 4.2.1 erstellt
Call: rlm(formula = ausgaben ~ einkommen, data = meinedaten)
Residuals:
    Min      1Q  Median      3Q     Max
-137.594 -29.953  -1.989  30.047  96.076
Coefficients:
            Value   Std. Error t value
(Intercept) 152.9166  22.9330    6.6680
einkommen     0.0673   0.0191    3.5175
Residual standard error: 44.56 on 18 degrees of freedom
```

**Abbildung 16.13:** Koeffizientenwerte für den Regressionszusammenhang zwischen *ausgaben* und *einkommen* auf Basis einer robusten Regression und eines SPSS/R-Syntaxprogramms (*LINEAREREGRESSION_GERADE7.SAV*)

Demgegenüber ergeben sich über die entsprechende SPSS/R-Syntax für eine lineare Regression zwischen *ausgaben* und *einkommen* – erneut auf Basis der Datendatei *LINEAREREGRESSION_GERADE7.SAV* – für das absolute Glied (auf drei Nachkommastellen gerundet) 161,684 Einheiten und für die Steigung (ebenfalls auf drei Nachkommastellen gerundet) 0,056 als Schätzergebnisse (siehe linker Teil von Abbildung 16.14). Dies sind – sicherlich für Sie nicht unerwartet – dieselben Koeffizientenwerte wie bei einem »reinen« SPSS-Vorgehen über ANALYSIEREN|REGRESSION|LINEAR im rechten Teil von Abbildung 16.14, allerdings beim SPSS/R-Vorgehen in einem

schlichteren Output-Design im Vergleich zur wunderschönen SPSS-Pivottabelle im rechten Teil von Abbildung 16.14 …

```
BEGIN PROGRAM R.
meinedaten <- spssdata.GetDataFromSPSS( )
fm.lm <- lm (ausgaben ~ einkommen, data = meinedaten)
summary (fm.lm)
END PROGRAM.
Call:
lm(formula = ausgaben ~ einkommen, data = meinedaten)
Residuals:
    Min      1Q  Median      3Q     Max
-123.714  -24.400    0.703   29.904  104.293
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 161.68421   25.30865   6.388 5.14e-06 ***
einkommen     0.05602    0.02113   2.651   0.0162 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 54.48 on 18 degrees of freedom
Multiple R-squared:  0.2808,    Adjusted R-squared:  0.2409
F-statistic: 7.029 on 1 and 18 DF,  p-value: 0.01624
```

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten: Regressionskoeffizient B | Nicht standardisierte Koeffizienten: Std.-Fehler | Standardisierte Koeffizienten: Beta | T | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Konstante) | 161,684 | 25,309 | | 6,388 | <,001 |
| | einkommen | ,056 | ,021 | ,530 | 2,651 | ,016 |

a. Abhängige Variable: ausgaben

**Abbildung 16.14:** Koeffizientenwerte für den Regressionszusammenhang zwischen *ausgaben* und *einkommen* auf Basis einer linearen Regression und eines SPSS/R-Syntaxprogramms (links) beziehungsweise mittels SPSS (rechts; jeweils *LINEAREREGRESSION_GERADE7.SAV*)

Die SPSS/R-Syntax im vorstehenden Beispiel sieht im Übrigen für die durchgeführte lineare Regression so aus (siehe hierzu auch *R-SYNTAX_1.SPS*):

```
BEGIN PROGRAM R.
meinedaten <- spssdata.GetDataFromSPSS( )
fm.lm <- lm (ausgaben ~ einkommen, data = meinedaten)
summary (fm.lm)
END PROGRAM.
```

# Kapitel 17

# Schleifen zum schnelleren Programmieren

## IN DIESEM KAPITEL

Allgemeines zu Schleifen

DO-REPEAT-Schleifen

LOOP-Konstruktionen

DO-IF-Befehl

In Kapitel 17 geht es um die Möglichkeit der Schleifenbildung. Hierdurch können Sie schneller, das heißt »effizienter« programmieren. Es gibt im Wesentlichen zwei Schleifenarten: DO-REPEAT-Schleifen und LOOP-Konstruktionen. Auf diese beiden Schleifenarten wird im Folgenden ausführlich eingegangen. Darüber hinaus können Sie mit dem DO-IF-Befehl analoge Vereinfachungen im Vergleich zu den beiden Schleifenarten in Ihren SPSS-Programmen herbeiführen.

## *Allgemeines zu Schleifen*

Haben Sie sich mit der SPSS-Syntax einigermaßen vertraut gemacht, werden Sie vielleicht festgestellt haben, dass sich in den betreffenden Programmen manche Befehle wiederholen. Trotz der Copy-and-Paste- beziehungsweise trotz der EINFÜGEN-Möglichkeit kann so etwas schon ziemlich nerven!

Abhilfe schafft hier die Automatisierung entsprechender Befehlsabfolgen in Form von **Schleifen.**

Stellen Sie sich zum Beispiel vor, Sie wollten 50 Variablen gleich null setzen. Dann müssten Sie 50-mal den COMPUTE-Befehl verwenden (das heißt: 50-mal COMPUTE *variable_i* = 0 mit i = 1, 2, …, 50). Das sind 50 Programmzeilen, die Ihr Syntaxprogramm etwas unleserlich werden lassen, zumal Sie möglicherweise noch weitere solcher »Massen«-Befehle in Ihrem Syntaxprogramm haben.

# DO-REPEAT-Schleifen

Um sich das Leben leichter zu machen, das heißt, um Zeit zu sparen, können Sie jetzt für diese 50 Variablen eine Platzhalter-Variable definieren, den COMPUTE-Befehl nur einmal angeben und ihn innerhalb einer **DO-REPEAT-Schleife** 50-mal ablaufen lassen. Konkret verwenden Sie in der nachfolgenden Syntax die Platzhaltervariable *a* für die 50 Variablen *var1* bis *var50*. Natürlich können Sie anstelle von *a* einen anderen Platzhalter verwenden und die Variablen auch anders als *var1* bis *var50* bezeichnen. Wichtig ist eigentlich nur, dass die Variablen fortlaufend gekennzeichnet (»indiziert«) werden (das heißt hier: von 1 bis 50).

## COMPUTE-Fall

Mit dem Start-Befehl DO REPEAT starten Sie die Schleife für diese 50 Variablen. Innerhalb der Schleife geben Sie an, welcher Befehl beziehungsweise welche Prozedur auf diese 50 Variablen (stellvertretend ist bekanntlich an dieser Stelle *a* angegeben) angewandt werden soll. Mit END REPEAT beenden Sie die Schleife, die in Ihrem Fall 50-mal durchlaufen wird. Alternativ hätten Sie diese 50 Zeilen in Ihr Syntaxprogramm eingeben müssen:

```
COMPUTE var1 = 0.
COMPUTE var2 = 0.
COMPUTE var3 = 0.
…
COMPUTE var48 = 0.
COMPUTE var49 = 0.
COMPUTE var50 = 0.
```

Offenkundig ist die gewählte Schleifen-Konstruktion kompakter und daher übersichtlicher:

```
DO REPEAT a = var1 to var50.
COMPUTE a=0.
END REPEAT.
EXECUTE.
```

## *RECODE-Fall (selbe Variable)*

Ein anderes Beispiel für die Verwendung einer DO-REPEAT-Schleife ist die **Umskalierung** einer ordinalen (oder auch einer kategorialen) Größe. Stellen Sie sich zum Beispiel vor, dass Sie nach der individuellen Zufriedenheit in verschiedenen Lebensbereichen fragen und dies in fünf Variablen *var1* bis *var5* abbilden: zum Beispiel in *var1* die allgemeine Lebenszufriedenheit, in *var2* die Zufriedenheit mit der Gesundheit, in *var3* die Zufriedenheit mit dem Haushaltseinkommen, in *var4* die Zufriedenheit mit der Wohnsituation und in *var5* die Zufriedenheit mit der familiären Situation. Stellen Sie sich bitte des Weiteren vor, dass die betreffenden fünf Variablen jeweils folgende drei Ausprägungen haben: 1 = (sehr) zufrieden, 2 = mittelmäßig zufrieden und 3 = (sehr) unzufrieden. International soll aber eine Skalierung »genau andersherum« üblich sein, also: 1 = (sehr) unzufrieden, 2 = mittelmäßig zufrieden und 3 = (sehr) zufrieden.

Um sozusagen international anschlussfähig zu sein, beschließen Sie, Ihre Variablenskalierungen an diese internationale Vereinbarung anzupassen. Dies bewerkstelligen Sie leicht mit dem RECODE-Befehl. Hierzu müssen Sie jeweils nur die unterste und die oberste Ausprägung verändern, wenn Sie die neue Skalierung in den bereits bestehenden Variablen *var1* bis *var5* abbilden möchten. Die mittlere Ausprägung »2« bleibt ja unverändert, muss also in diesem Fall nicht verändert werden (als Platzhalter-Variable ist nachfolgend *b* verwendet worden):

```
DO REPEAT b=var1 to var5.
RECODE b (1=3) (3=1).
END REPEAT.
EXECUTE.
```

Angenommen, es sind drei Personen nach ihrer Zufriedenheit in den fünf Bereichen befragt worden und Person 1 hat in der ursprünglichen

Skalierung immer 1 (»(sehr) zufrieden«), Person 2 immer 2 (»mittelmäßig zufrieden«) und Person 3 immer 3 (»(sehr) unzufrieden«) angegeben. Dann sieht der Ausgangsdatensatz *ANGST1.SAV* so aus wie in Abbildung 17.1.

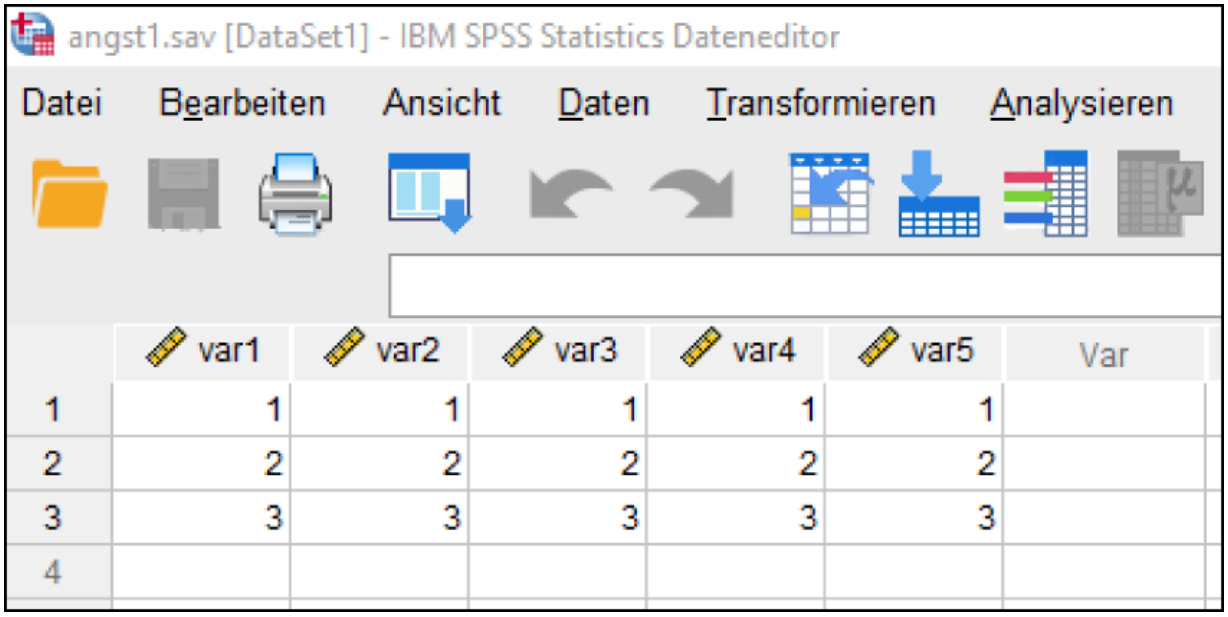



**Abbildung 17.1:** Ausgangsdatensatz *ANGST1.SAV* vor der Umskalierung von bestehenden Variablen

Nach der oben stehenden Umcodierung (1 → 3; 3 → 1) muss nach Durchlaufen der angegebenen DO-REPEAT-Schleife bei Person 1 nunmehr bei allen fünf Variablen eine 3 und bei Person 3 bei allen fünf Variablen eine 1 stehen. Bei Person 2 ändert sich nichts: Die 2 bleibt hier stehen. Führen Sie die angegebene DO-REPEAT-Schleife aus, verändert sich der Datensatz in der Tat im vorstehenden Sinne (siehe hierzu Abbildung 17.2).

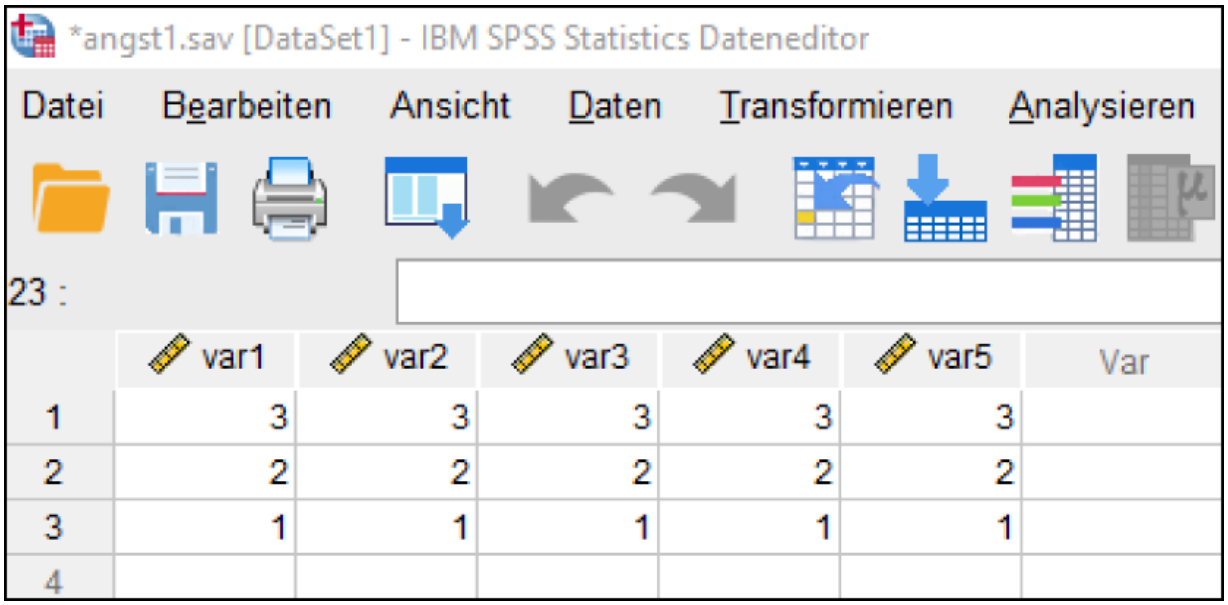



**Abbildung 17.2:** Recodierter Datensatz (Umskalierung in denselben Variablen) (*ANGST1.SAV*)

## *RECODE-Fall (neue Variablen)*

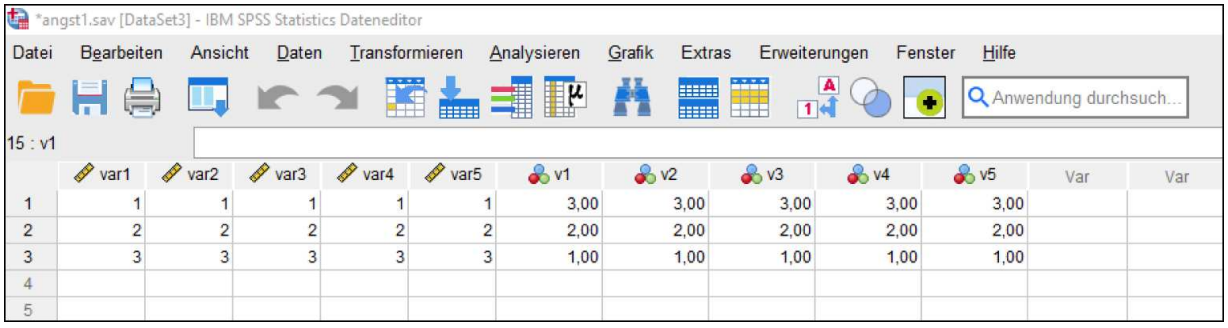
Sie hätten die eben diskutierten Umskalierungen aber auch in fünf neuen Variablen *v1* bis *v5* vornehmen können. Hierzu definieren Sie in der DO-REPEAT-Schleife zunächst die alten Variablen (nachfolgend im Variablenbereich *old* für die alten Variablen *var1* bis *var5*) und danach – nach einem Slash – die neuen Variablen (nachfolgend im Variablenbereich *new* für die neuen Variablen *v1* bis *v5*). Anschließend führen Sie mittels des RECODE-Befehls die Umcodierungen von den alten in die neuen Variablen durch, ehe Sie mit END REPEAT die DO-REPEAT-Schleife abschließen.

Wenn Sie in einer neuen Variablen einen recodierten Wert einer alten Variablen abspeichern, müssen Sie unbedingt für alle Ausprägungen der alten Variablen die neuen Ausprägungen in der neuen Variablen angeben, auch wenn sich der Ursprungswert nicht verändert. In unserem Fall verändert sich die Ausprägung 2 ja nicht. Im Unterschied zum Recodieren in ein und derselben Variablen müssen Sie aber hier beim RECODE-Befehl die Recodierung 2 → 2 angeben, damit die 2 in die neue Variable übertragen wird. Würden Sie dies nicht angeben, würde die neue Variable beim alten Wert 2 in der neuen Variablen ein Missing enthalten, die Umskalierung würde also unvollständig verlaufen.

Die Recodierungs-Schleife lautet demnach im vorliegenden Beispiel für den (Ursprungs-)Datensatz *ANGST1.SAV* mit Recodierungen in neuen Variablen (siehe *DO-REPEAT_3.SPS*):

```
DO REPEAT old=var1 to var5
/new=v1 to v5.
RECODE old (1=3) (2=2) (3=1) INTO new.
END REPEAT.
EXECUTE.
```

Sie erhalten mit dieser Schleifenkonstruktion einen neuen Datensatz, der zusätzlich zu den alten Variablen *var1* bis *var5* jetzt auch die neuen Variablen *v1* bis *v5* enthält (siehe Abbildung 17.3). (Sie finden den in Abbildung 17.3 dargestellten Datensatz auch in der Datei *ANGST2.SAV*.)




| | var1 | var2 | var3 | var4 | var5 | v1 | v2 | v3 | v4 | v5 | Var | Var |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 3,00 | 3,00 | 3,00 | 3,00 | 3,00 | | |
| 2 | 2 | 2 | 2 | 2 | 2 | 2,00 | 2,00 | 2,00 | 2,00 | 2,00 | | |
| 3 | 3 | 3 | 3 | 3 | 3 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | | |
| 4 | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | |

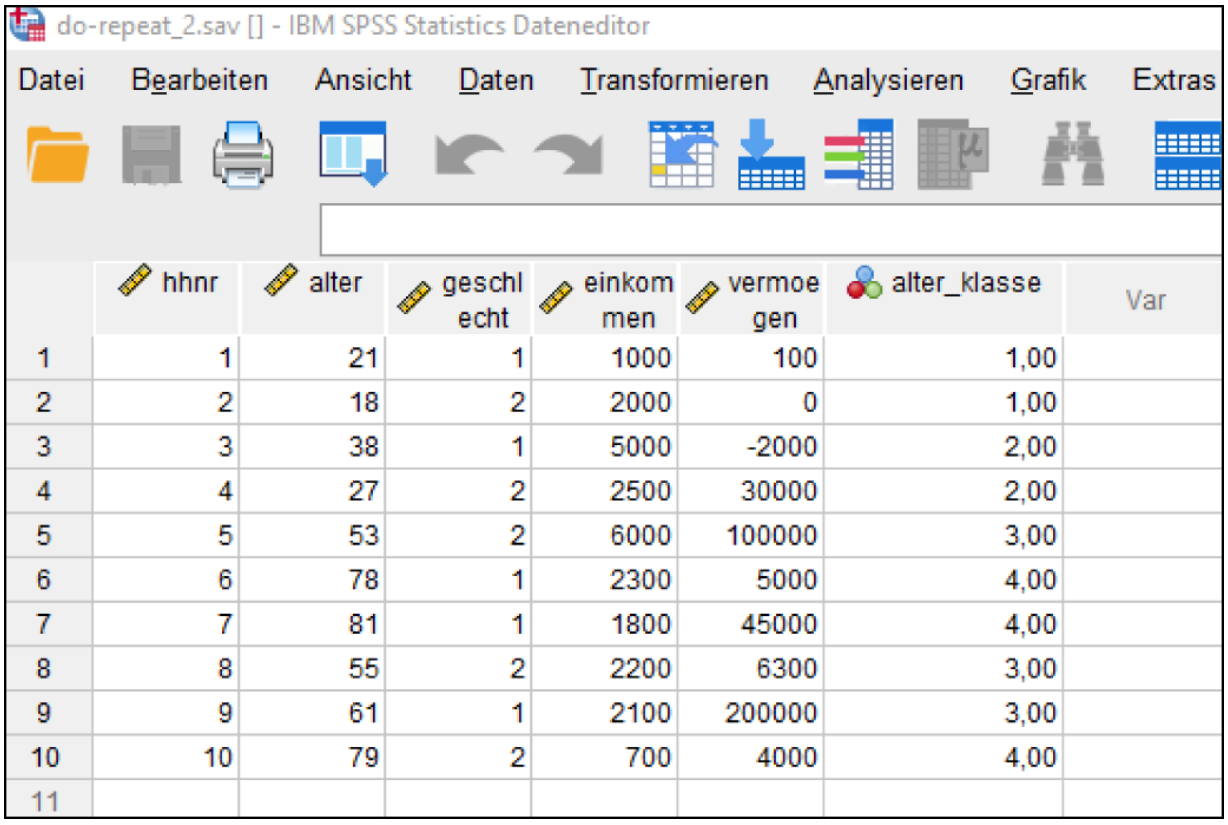
**Abbildung 17.3:** Recodierter Datensatz (in neue Variablen) (*ANGST1.SAV*)

## *Klassenbildung*

Hilfreich ist der DO-REPEAT-Befehl auch bei Klassenbildung, wenn Sie zum Beispiel eine metrische Variable (nachfolgend als Beispiel: eine Altersvariable in Lebensjahren) kategorial, das heißt in Klassen, abbilden möchten. Auch hier setzen Sie – bei Bezugnahme auf *DO-REPEAT1.SAV* – nach DO REPEAT einen Platzhalter (hier: *alt*) für die Ursprungsvariable und danach – nach einem Schrägstrich – einen Platzhalter (hier: *neu*) für die neue Variable. Nach diesen anfänglichen Setzungen bilden Sie bitte im Datensatz *DO-REPEAT1.SAV* unter Verwendung der Platzhalter *alt* und *neu* in unserem Beispiel vier Altersklassen: 1 = bis 24 Jahre, 2 = 25 bis 49 Jahre, 3 = 50 bis 74 Jahre und 4 = 75 Jahre und älter (siehe zum Syntaxprogramm übrigens *DO-REPEAT_4.SPS*).

```
GET FILE='C:\SPSS-DATEIEN\DO-REPEAT1.SAV'.

DO REPEAT alt=alter
/neu=alter_klasse.
IF(alt>=0 and alt<25) neu =1.
IF(alt>=25 and alt<50) neu =2.
IF(alt>=50 and alt<75) neu =3.
IF(alt>=75) neu =4.
END REPEAT.
EXECUTE.

SAVE OUTFILE='C:\SPSS-DATEIEN\DO-REPEAT2.SAV'.
```

Das Ergebnis dieser Klassenbildung findet sich in der Variablen *alter_klasse* im als *DO-REPEAT2.SAV* abgespeicherten Datenfile wieder und sieht wie in Abbildung 17.4 aus. Der Vorteil dieser DO-REPEAT-Schleife ergibt sich dann, wenn Sie für unterschiedliche Variablen die gleichen Klassenbildungen vornehmen müssen. Dann müssen Sie in der betreffenden Schleife einfach nur nach *alt* und nach *neu* die Variablennamen austauschen (oder die betreffende DO-REPEAT-Schleife kopieren und die Variablenzuordnungen in der kopierten Schleife entsprechend verändern).




| | hhnr | alter | geschlecht | einkommen | vermoegen | alter_klasse | Var |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 21 | 1 | 1000 | 100 | 1,00 | |
| 2 | 2 | 18 | 2 | 2000 | 0 | 1,00 | |
| 3 | 3 | 38 | 1 | 5000 | -2000 | 2,00 | |
| 4 | 4 | 27 | 2 | 2500 | 30000 | 2,00 | |
| 5 | 5 | 53 | 2 | 6000 | 100000 | 3,00 | |
| 6 | 6 | 78 | 1 | 2300 | 5000 | 4,00 | |
| 7 | 7 | 81 | 1 | 1800 | 45000 | 4,00 | |
| 8 | 8 | 55 | 2 | 2200 | 6300 | 3,00 | |
| 9 | 9 | 61 | 1 | 2100 | 200000 | 3,00 | |
| 10 | 10 | 79 | 2 | 700 | 4000 | 4,00 | |
| 11 | | | | | | | |

**Abbildung 17.4:** Klassenbildung in DO-REPEAT-Schleife (*DO-REPEAT2.SAV*)

# *LOOP-Konstruktion*

Eine Alternative zu DO-REPEAT-Schleifen ist der **LOOP-Befehl**. Auch bei ihm wird mit Indizes (Kennzeichnungen) gearbeitet.

Vor dem LOOP-Befehl steht ein **VECTOR-Befehl**. Mit diesem VECTOR-Befehl werden in einem Vektor (das heißt in einer Aneinanderreihung von Variablen) die Variablen angegeben, die dann in der LOOP-Sequenz bearbeitet werden.

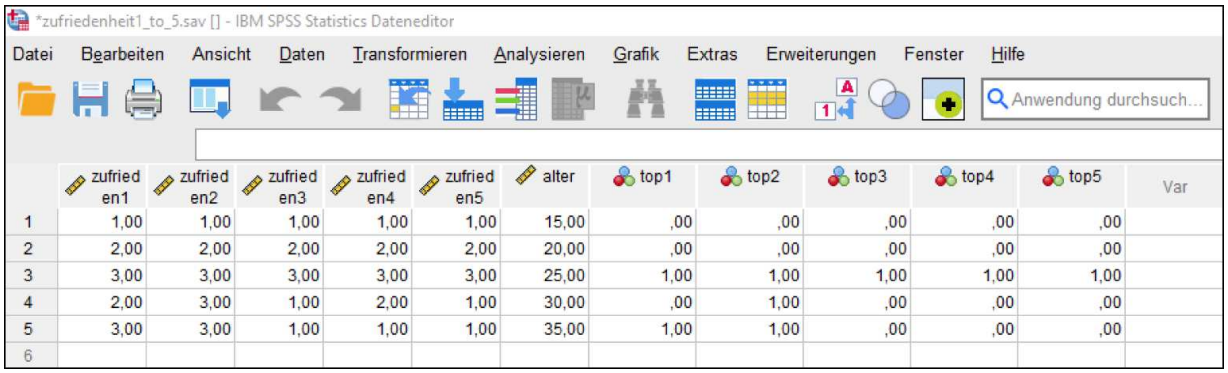
Allgemein gilt dabei, dass dem vor dem LOOP-Befehl stehenden Vektor nur in einem Datensatz direkt nebeneinanderstehende Variablen beziehungsweise gleichlautende Variablen mit Index angehören dürfen. Um Schwierigkeiten bei der Zuordnung zu vermeiden, sollten Sie auf »Nummer sicher« gehen und die erstgenannte Variante der direkt nebeneinander im Datensatz stehenden Variablen wählen.

## *Beispiel 1*

Im folgenden Beispiel der Datendatei *ZUFRIEDENHEIT1_TO_5.SAV* wird für jede der im Vektor *zufrieden* enthaltenen Variablen *zufrieden1* bis *zufrieden5* bei Erreichen des jeweiligen Höchstwerts 3 eine Variable *top1* bis *top5* mit der jeweiligen Ausprägung 1 kreiert, andernfalls mit der jeweiligen Ausprägung 0. Die LOOP-Sequenz wird dabei für jede Variable einmal, insgesamt also fünfmal, durchlaufen. Die **Zählvariable** ist in diesem Beispiel *#i*. Sie kennzeichnet die insgesamt fünf Durchgänge. Im ersten Durchgang ist *#i* gleich 1 und die IF-Abfragen in der LOOP-Sequenz lauten: »IF(*zufrieden1* = 3) *top1* = 1.« und »IF(*zufrieden1* <= 2) *top1*=0.«. Demzufolge haben Sie als IF-Abfragen im zweiten Durchgang »IF(*zufrieden2* = 3) *top2* = 1.« und »IF(*zufrieden2* <= 2) *top2* = 0.« und so weiter (siehe hierzu das Syntaxfile *LOOP1.SPS*).

```
VECTOR zufrieden=zufrieden1 TO zufrieden5 / top(5).
LOOP #i=1 TO 5.
IF (zufrieden(#i) = 3) top(#i)=1.
IF (zufrieden(#i) <=2) top(#i)=0.
END LOOP.
EXECUTE.
```

Das Ergebnis finden Sie in Abbildung 17.5. Offenkundig ist der Datensatz um die Variablen *top1* bis *top5* erweitert worden. Diese Variablen nehmen überall dort, wo eine der Variablen zufrieden1 bis *zufrieden5* eine 3 aufweist, eine 1 an und überall dort, wo eine der Variablen *zufrieden1* bis *zufrieden5* eine 1 oder eine 2 als Ausprägung hat, eine 0.

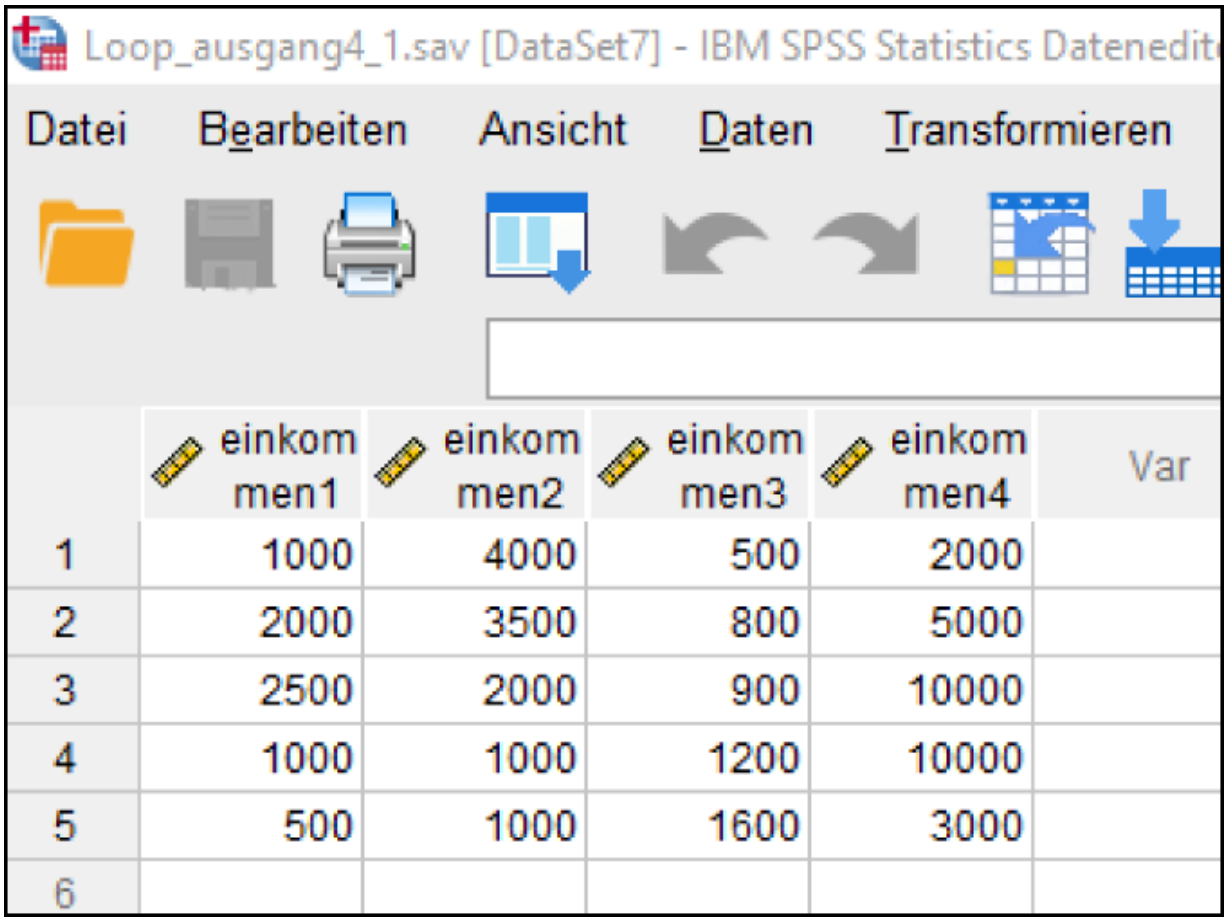



| | zufrieden1 | zufrieden2 | zufrieden3 | zufrieden4 | zufrieden5 | alter | top1 | top2 | top3 | top4 | top5 | Var |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 15,00 | ,00 | ,00 | ,00 | ,00 | ,00 | |
| 2 | 2,00 | 2,00 | 2,00 | 2,00 | 2,00 | 20,00 | ,00 | ,00 | ,00 | ,00 | ,00 | |
| 3 | 3,00 | 3,00 | 3,00 | 3,00 | 3,00 | 25,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | |
| 4 | 2,00 | 3,00 | 1,00 | 2,00 | 1,00 | 30,00 | ,00 | 1,00 | ,00 | ,00 | ,00 | |
| 5 | 3,00 | 3,00 | 1,00 | 1,00 | 1,00 | 35,00 | 1,00 | 1,00 | ,00 | ,00 | ,00 | |
| 6 | | | | | | | | | | | | |

**Abbildung 17.5:** Bildung von neuen Variablen mittels LOOP-Sequenz (*ZUFRIEDENHEIT1_TO_5.SAV*)

## *Beispiel 2*

Ein weiteres Beispiel für die LOOP-Konstruktion ist das nachfolgende. In diesem Beispiel liegen Einkommensinformationen in vier Regionen von jeweils fünf Personen vor (zum Beispiel 1.000 Euro bei *einkommen1* in Zeile 1 als Einkommen der ersten Person in der ersten Region). Die entsprechende Datenstruktur findet sich in Abbildung 17.6 für die Datendatei *LOOP_AUSGANG4_1.SAV*.

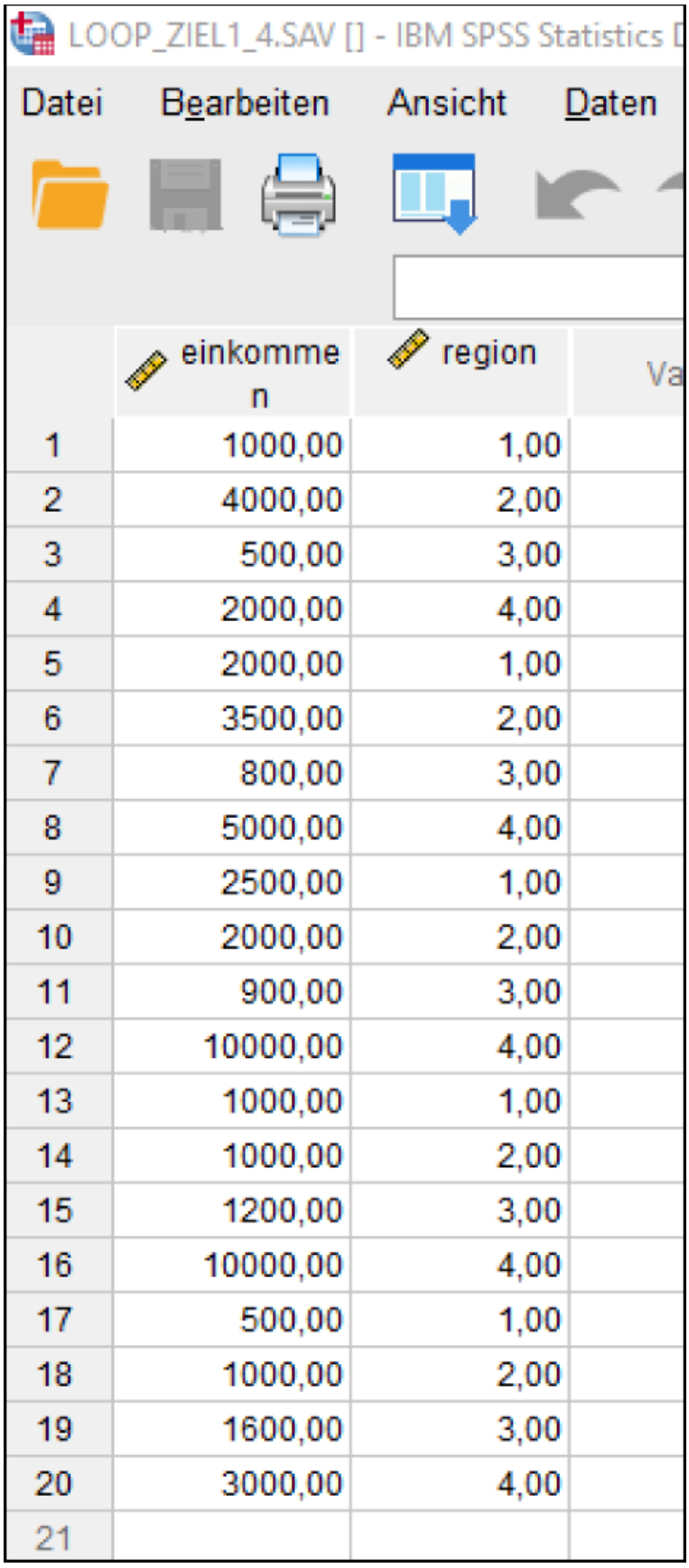



| | einkommen1 | einkommen2 | einkommen3 | einkommen4 | Var |
|---|---|---|---|---|---|
| 1 | 1000 | 4000 | 500 | 2000 | |
| 2 | 2000 | 3500 | 800 | 5000 | |
| 3 | 2500 | 2000 | 900 | 10000 | |
| 4 | 1000 | 1000 | 1200 | 10000 | |
| 5 | 500 | 1000 | 1600 | 3000 | |
| 6 | | | | | |

**Abbildung 17.6:** Datensatz mit vier Variablen vor LOOP-Prozedur (*LOOP_AUSGANG4_1.SAV*)

Dieser Datensatz soll nun so verändert werden, dass aus den vier Einkommensvariablen *einkommen1* bis *einkommen4* für die vier Regionen zwei Variablen werden, und zwar eine Variable *einkommen*, in der die insgesamt 20 Einkommenswerte (= 5 Personen · 4 Einkommenswerte pro Person) stehen, und eine Variable *region*, deren Ausprägungen 1 bis 4 Auskunft über die jeweilige Region geben. Der betreffende, veränderte Datensatz *LOOP_ZIEL1_4.SAV* ist in Abbildung 17.7 zu sehen.

Wie ist jetzt diese neue Datendatei erzeugt worden? Sie hätten bei diesen kleinen Fallzahlen (jeweils fünf Fälle pro Region; insgesamt 20 Fälle) die Ausgangsdaten für die Einkommen einfach in eine neue Variable *einkommen* kopieren und anschließend die Variable *region* manuell »bestücken« können.




| | einkommen | region | Va |
|---|---|---|---|
| 1 | 1000,00 | 1,00 | |
| 2 | 4000,00 | 2,00 | |
| 3 | 500,00 | 3,00 | |
| 4 | 2000,00 | 4,00 | |
| 5 | 2000,00 | 1,00 | |
| 6 | 3500,00 | 2,00 | |
| 7 | 800,00 | 3,00 | |
| 8 | 5000,00 | 4,00 | |
| 9 | 2500,00 | 1,00 | |
| 10 | 2000,00 | 2,00 | |
| 11 | 900,00 | 3,00 | |
| 12 | 10000,00 | 4,00 | |
| 13 | 1000,00 | 1,00 | |
| 14 | 1000,00 | 2,00 | |
| 15 | 1200,00 | 3,00 | |
| 16 | 10000,00 | 4,00 | |
| 17 | 500,00 | 1,00 | |
| 18 | 1000,00 | 2,00 | |
| 19 | 1600,00 | 3,00 | |
| 20 | 3000,00 | 4,00 | |
| 21 | | | |

**Abbildung 17.7:** Datensatz mit zwei Variablen nach LOOP-Prozedur (*LOOP_ZIEL1_4.SAV*)

Bei einer größeren Fallzahl ist das aber ein mühsames Unterfangen. Als Alternative bietet sich eine LOOP-Konstruktion an wie die nachfolgend angegebene.

Nach dem Öffnen der in Abbildung 17.6 dargestellten Datendatei wird hierbei ein Vektor *r* definiert, der die auf zwei Variablen zu verdichtenden Variablen *einkommen1* bis *einkommen4* enthält. Außerdem wird für die vier Indizes der Einkommensvariablen, die die Regionen 1 bis 4 symbolisieren, zu Beginn der LOOP-Konstruktion als Zählvariable *region* definiert, wobei *region* eben von 1 bis 4 – also von der ersten bis zur vierten Region – »läuft«. Als Befehle werden innerhalb von LOOP die Einkommen in *einkommen* – als Aneinanderreihen der vier ursprünglichen vier Einkommensvariablen – ebenso wie *region* in der Datei *LOOP_ZIEL1_4.SAV* gespeichert.

Dieses Abspeichern erfolgt hier mit XSAVE und nicht mit SAVE. Der Grund dafür ist, dass bei XSAVE der Speichervorgang erst abgeschlossen ist, wenn ein EXECUTE folgt (oder eine Prozedur wie GET FILE). Innerhalb der LOOP-Konstruktion wenden Sie deshalb XSAVE an, weil ja aus den ursprünglichen vier Einkommensvariablen die insgesamt 20 Einkommenswerte ausgelesen und schrittweise nacheinander in der genannten Zieldatei zwischengespeichert werden. Dabei werden die Einkommenswerte zeilenweise über die vier ursprünglichen Einkommensvariablen hinweg in die neue Einkommensvariable *einkommen* eingelesen.

Nachdem Sie via KEEP festgelegt haben, welche Variablen in Ihre Zieldatei kommen, beenden Sie die Schleifenkonstruktion. Mit EXECUTE lösen Sie das endgültige Speichern der Zieldatei aus, und mit GET FILE können Sie sich die Zieldatei anschauen. Fertig! (Der EXECUTE-Befehl ist im Übrigen in der nachfolgenden Syntax – die im Syntaxfile *LOOP2.SPS* enthalten ist – deshalb entbehrlich, weil durch

den GET-FILE-Befehl bekanntlich das endgültige Speichern der Zieldatei veranlasst wird.)

```
GET FILE='C:\SPSS-DATEIEN\LOOP_AUSGANG4_1.SAV'.
VECTOR r=einkommen1 TO einkommen4.
LOOP region=1 TO 4.
COMPUTE einkommen=r(region).
XSAVE OUTFILE='C:\SPSS-DATEIEN\LOOP_ZIEL1_4.SAV'
/KEEP einkommen region.
END LOOP.
EXECUTE.
GET FILE='C:\SPSS-DATEIEN\LOOP_ZIEL1_4.SAV'.
```

LOOP-Konstruktionen in der hier vorgestellten Form ermöglichen es Ihnen »nur«, Datentransformationen innerhalb einer Schleife durchzuführen. Demgegenüber sind bei LOOP-Makros auch statistische Prozeduren zulässig. Dazu gibt es im bald folgenden Abschnitt »Grundlagen der Makroprogrammierung« in Kapitel 18 mehr Informationen.

# DO-IF-Befehl

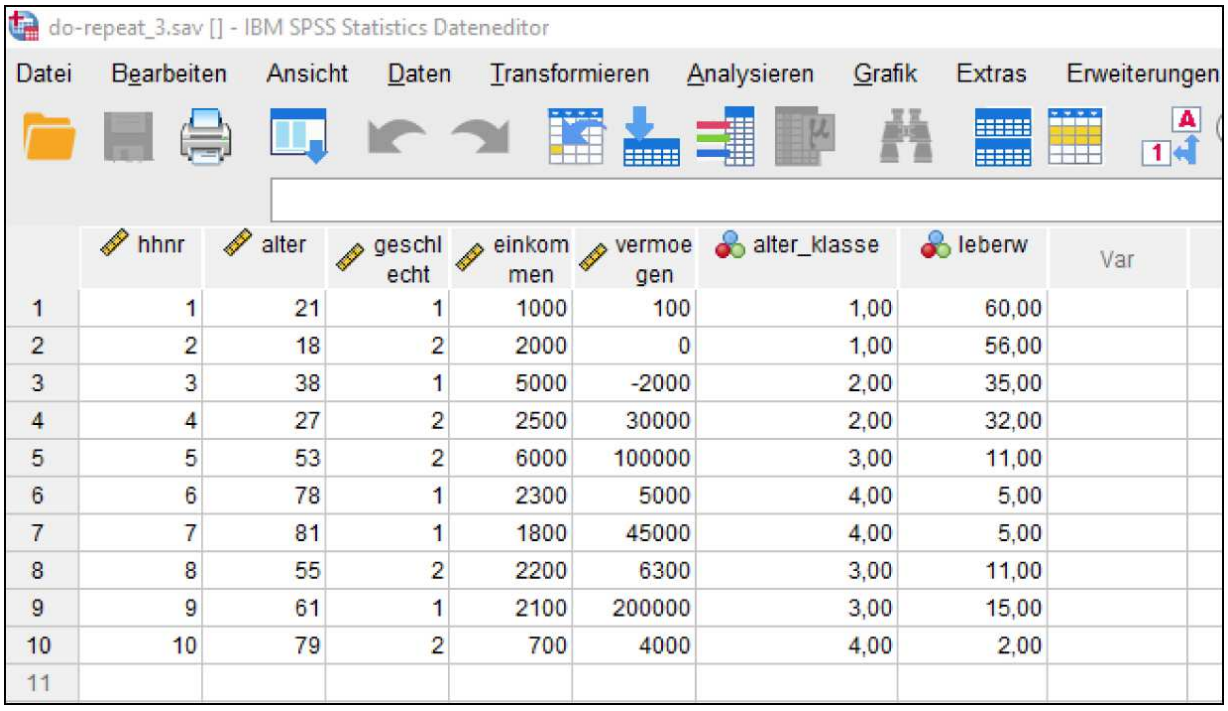
Vereinfachungen in Ihren Syntaxprogrammen erzielen Sie auch mit dem **DO-IF-Befehl**. Bei bestimmten Fragestellungen ist eine Fallunterscheidung (für disjunkte Teilmengen, also für solche Teilmengen ohne gemeinsame Schnittmenge) sinnvoll, die Sie über eine entsprechende Bedingung bewerkstelligen. Im Menü-Kontext könnten Sie über den Menüpunkt FALLS bei TRANSFORMIEREN eine solche Bedingung formulieren. Dieser Vorgang ist recht umständlich, wenn Sie mit ein und derselben Bedingung mehrere Transformationen durchführen, den Transformationsvorgang also wiederholen möchten. In diesem Fall verwenden Sie besser in der SPSS-Syntaxsprache den DO-IF-Befehl.

Hierdurch können Sie zum Beispiel IF-Abfragen vereinfachen, wie auch im folgenden Beispiel auf Basis der Datendatei *DO-REPEAT2.SAV*. Bei diesem Beispiel wird zunächst die Variable *geschlecht* nach ihren beiden Ausprägungen gegliedert, und danach werden getrennt für diese

Ausprägungen (1 = weiblich und 2 = männlich) bestimmte Abfragen durchgeführt. Für die beiden Geschlechter existieren hier jeweils vier IF-Abfragen in dem Sinne, dass für die vier Altersklassen »bis 25 Jahre« (*alter_klasse* = 1), »25 bis 49 Jahre« (*alter_klasse* = 2), »50 bis 74 Jahre« (*alter_klasse* = 3) und »75 Jahre und älter« (*alter_klasse* = 4) eine (durchschnittliche) Lebenserwartung *leberw* angegeben wird (siehe hierzu das Syntaxfile *DOREPEAT_ 5.SPS*).

```
GET FILE='C:\SPSS-DATEIEN\DO-REPEAT2.SAV'.

DO IF (geschlecht=1).
IF(alter_klasse=1) leberw=60.
IF(alter_klasse=2) leberw=35.
IF(alter_klasse=3) leberw=15.
IF(alter_klasse=4) leberw=5.
END IF.

DO IF (geschlecht=2).
IF(alter_klasse=1) leberw=56.
IF(alter_klasse=2) leberw=32.
IF(alter_klasse=3) leberw=11.
IF(alter_klasse=4) leberw=2.
END IF.

EXECUTE.

SAVE OUTFILE='C:\SPSS-DATEIEN\DO-REPEAT3.SAV'.
```

Das Ergebnis dieses DO-IF-Befehls im neuen Datensatz *DO-REPEAT3.SAV* ist in Abbildung 17.8 dargelegt.




| | hhnr | alter | geschlecht | einkommen | vermoegen | alter_klasse | leberw | Var |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 21 | 1 | 1000 | 100 | 1,00 | 60,00 | |
| 2 | 2 | 18 | 2 | 2000 | 0 | 1,00 | 56,00 | |
| 3 | 3 | 38 | 1 | 5000 | -2000 | 2,00 | 35,00 | |
| 4 | 4 | 27 | 2 | 2500 | 30000 | 2,00 | 32,00 | |
| 5 | 5 | 53 | 2 | 6000 | 100000 | 3,00 | 11,00 | |
| 6 | 6 | 78 | 1 | 2300 | 5000 | 4,00 | 5,00 | |
| 7 | 7 | 81 | 1 | 1800 | 45000 | 4,00 | 5,00 | |
| 8 | 8 | 55 | 2 | 2200 | 6300 | 3,00 | 11,00 | |
| 9 | 9 | 61 | 1 | 2100 | 200000 | 3,00 | 15,00 | |
| 10 | 10 | 79 | 2 | 700 | 4000 | 4,00 | 2,00 | |
| 11 | | | | | | | | |

**Abbildung 17.8:** Datensatz nach DO-IF-Schleife (*DO-REPEAT3.SAV*)

Wie immer bei bedingten Abfragen beachten Sie in diesem Zusammenhang bitte die Regeln beziehungsweise Reihenfolgen für logische Ausdrücke:

1. Funktionen
2. Potenzieren (**)
3. Multiplikation (*), Division (/), Minus-Vorzeichen (−…)
4. Addition (+), Subtraktion (−),
5. Vergleichsoperatoren (= beziehungsweise eq, <> beziehungsweise ne, < beziehungsweise lt, <= beziehungsweise le, > beziehungsweise gt, >= beziehungsweise ge)
6. NOT
7. AND
8. OR

Ist die Priorität gleich hoch, erfolgt die Auswertung des logischen Ausdrucks von links nach rechts. Auch hier kann durch Klammersetzung eine veränderte Reihenfolge erzielt werden, und zwar in dem Sinne, dass Klammerausdrücke zuerst bearbeitet werden.

Grundsätzlich sind Sie mit dem DO-IF-(Ober-)Befehl in der Lage, mehrere (Unter-)Befehle abzuarbeiten, die allesamt die nach DO IF festgelegte Bedingung erfüllen. Mit ELSE IF können Sie danach andere, ebenfalls mehrere Befehle abarbeiten, die einer alternativen Bedingung folgen. Mit END IF schließen Sie die DO-IF-Schleife ab.

Innerhalb einer DO-IF-Schleife können Sie keine statistischen Prozeduren aufrufen, sondern – analog zu den obigen LOOP-Konstruktionen – »lediglich« Datentransformationen durchführen.

## *Beispiel 2*

Ein weiteres Beispiel für eine (umfangreiche) DO-IF-Schleifenkonstruktion findet sich nachfolgend. In diesem Beispiel mit der Datendatei *REGION.SAV* gibt es im Ausgangsdatensatz nur eine Variable, die sechs Regionen kennzeichnet (siehe Abbildung 17.9).

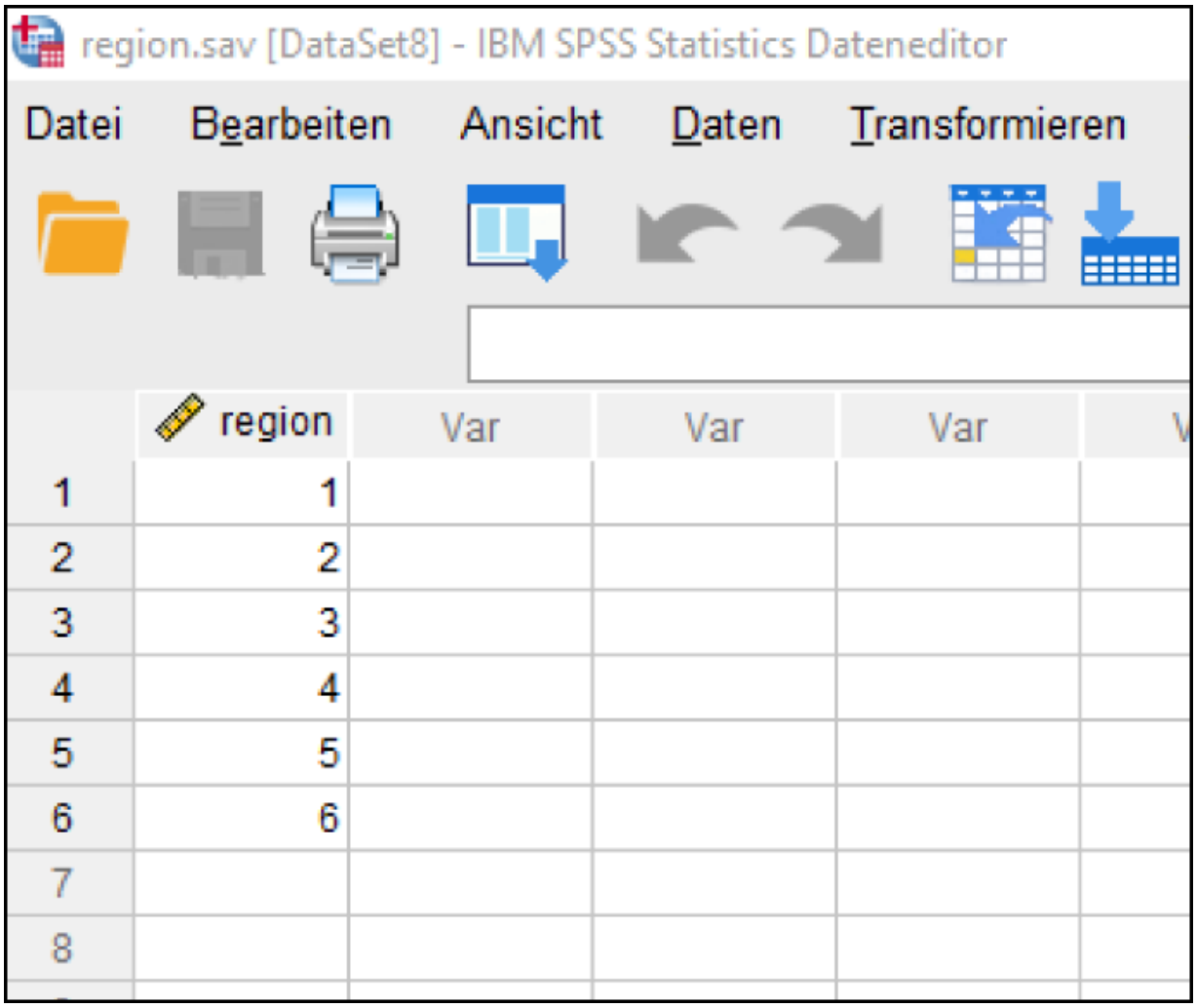



**Abbildung 17.9:** Ausgangsdatensatz *REGION.SAV* mit einer Variablen vor einer DO-IF-Schleife

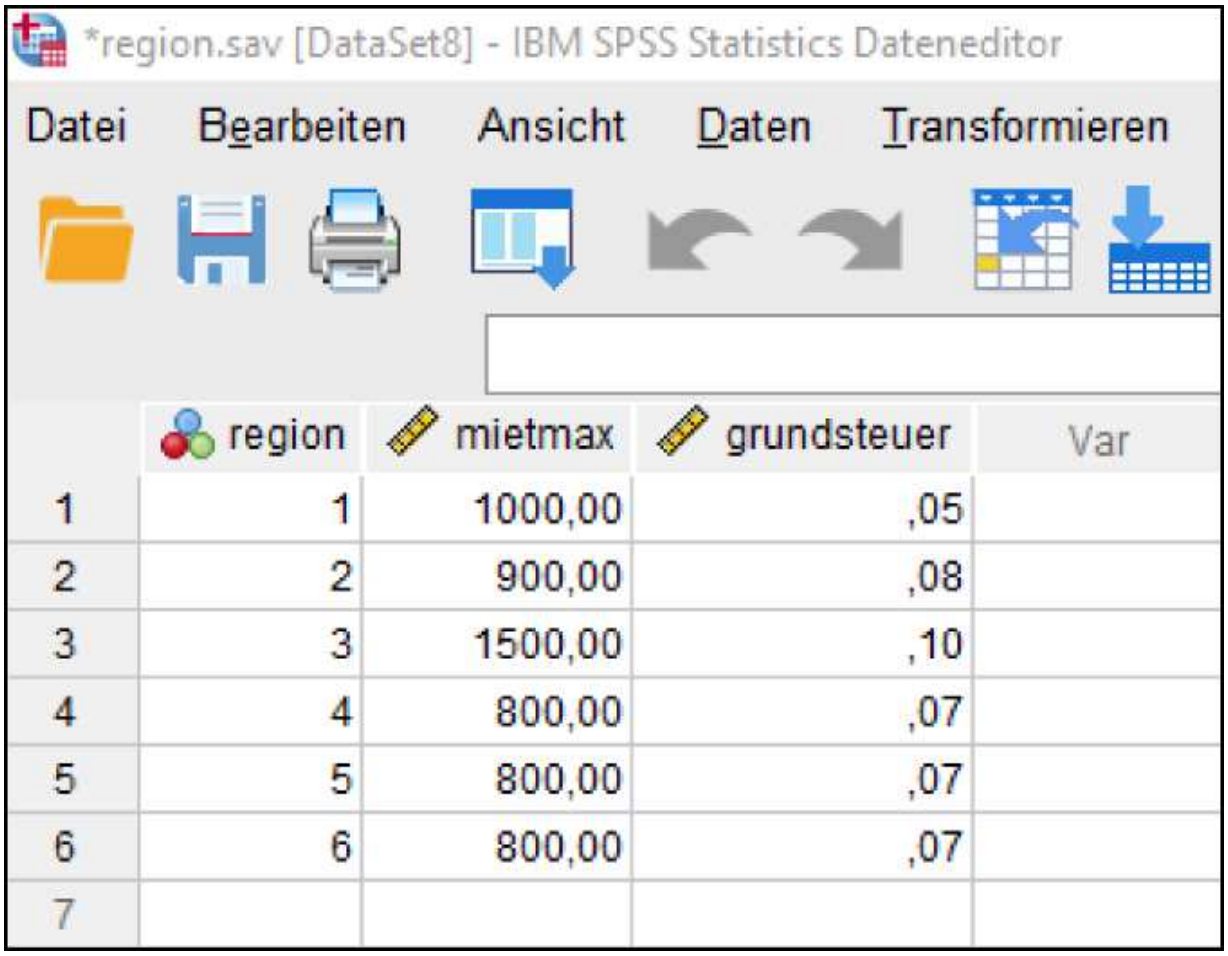
Für diese sechs Regionen sollen jetzt verschiedene Werte für zwei Variablen *mietmax* (inhaltlich zum Beispiel eine regionale monatliche Mietobergrenze für einen bestimmten Wohnungstyp) und *grundsteuer* (inhaltlich zum Beispiel der Grundsteuersatz) angegeben werden. Die entsprechenden Zuweisungen werden mit einer DO-IF-Schleife bewerkstelligt. Ausgangspunkt ist die Region 1 direkt nach dem DO-IF-Befehl. Es folgen die Regionen 2 und 3 nach ELSE-IF-Befehlen, und ganz am Ende finden Sie noch einen »profanen« ELSE-Befehl. Dieser ELSE-Befehl ist »nackt«; das heißt, er enthält keine Bedingung, muss er aber auch gar nicht, denn mit ihm werden schlicht und ergreifend alle anderen Regionen als die Regionen 1, 2 und 3 erfasst. In unserem

Beispiel beziehen sich die Setzungen nach dem »nackten« ELSE also auf die verbleibenden Regionen 4 bis 6.

Und so sieht das Ganze »in Syntax« aus (siehe auch *DO-IF_REGIONEN.SPS*):

```
DO IF (region=1).
COMPUTE mietmax=1000.
COMPUTE grundsteuer=0.05.

ELSE IF (region=2).
COMPUTE mietmax=900.
COMPUTE grundsteuer=0.08.

ELSE IF (region=3).
COMPUTE mietmax=1500.
COMPUTE grundsteuer=0.10.

ELSE.
COMPUTE mietmax=800.
COMPUTE grundsteuer=0.07.

END IF.

EXECUTE.
```

Dass Sie alles richtig gemacht haben, zeigt sich daran, dass die neue Datendatei (*REGION.SAV*) jetzt auch noch *mietmax* und *grundsteuer* als Variablen enthält (siehe Abbildung 17.10).




| | region | mietmax | grundsteuer | Var |
|---|---|---|---|---|
| 1 | 1 | 1000,00 | ,05 | |
| 2 | 2 | 900,00 | ,08 | |
| 3 | 3 | 1500,00 | ,10 | |
| 4 | 4 | 800,00 | ,07 | |
| 5 | 5 | 800,00 | ,07 | |
| 6 | 6 | 800,00 | ,07 | |
| 7 | | | | |

**Abbildung 17.10:** Datensatz *REGION.SAV* mit drei Variablen nach einer DO-IF-Schleife

## *Beispiel 3*

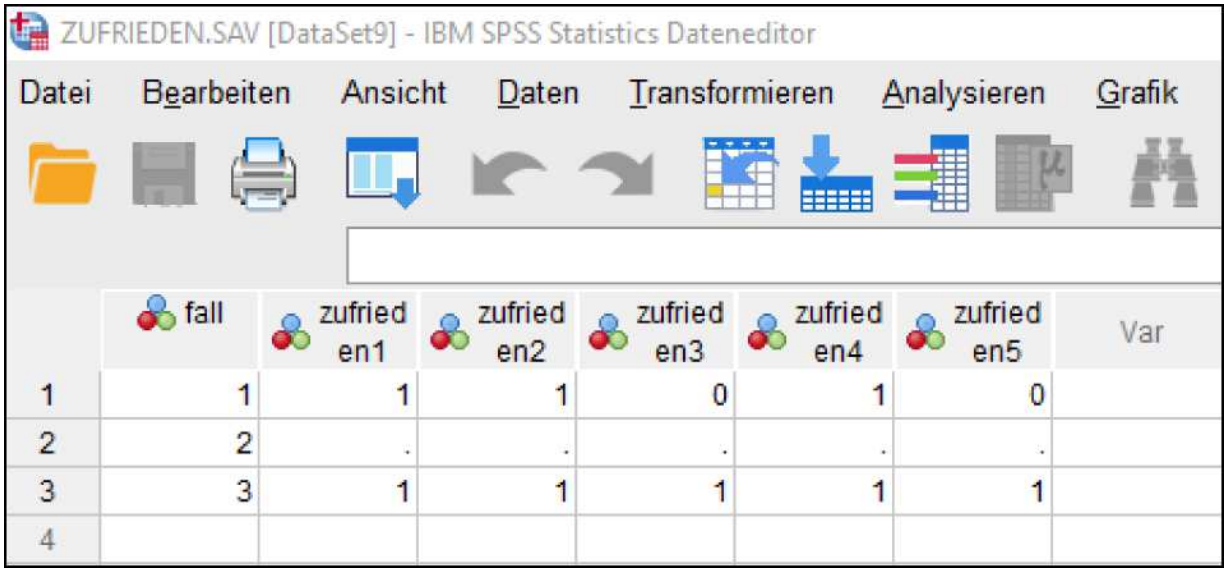
Im Menü TRANSFORMIEREN findet sich WERTE IN FÄLLEN ZÄHLEN, was dem SPSS-Befehl COUNT entspricht. Hiermit können Sie eine **Zählvariable** erstellen, und auch hierbei kann die DO-IF-Schleife hilfreich sein, wie nachfolgend gezeigt wird.

Bei einer Zählvariablen beachten Sie bitte, dass diese grundsätzlich immer einen gültigen Wert größer gleich 0 hat. Weisen alle relevanten Variablen den Wert SYSMIS auf, erhalten Sie für Ihre Zählvariable folglich nicht etwa SYSMIS, sondern 0! Das ist unschön.

Um diesen unerwünschten Fall zu umgehen, müssen Sie den NMISS-Befehl nutzen und für n Variablen *var1* bis *varn*

NMISS(*var1* TO *varn*) = 0 als Bedingung formulieren.

Lassen Sie uns den vorstehenden Sachverhalt an einem Beispiel diskutieren. Hierzu öffnen Sie bitte die Datendatei *ZUFRIEDEN.SAV*. Diese enthält für fünf Zufriedenheitsarten (zum Beispiel in Bezug auf Gesundheit, Haushaltseinkommen, Wohnsituation, familiäre Situation und Umwelt) die 0/1-Dummies *zufrieden1* bis *zufrieden5* jeweils mit 0 = unzufrieden und 1 = zufrieden. Wie Abbildung 17.11 zeigt, hat Person Nummer 2 keinerlei Angaben bei allen hier relevanten Variablen *zufrieden1* bis *zufrieden5* gemacht.




| | fall | zufrieden1 | zufrieden2 | zufrieden3 | zufrieden4 | zufrieden5 | Var |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 1 | 0 | |
| 2 | 2 | . | . | . | . | . | |
| 3 | 3 | 1 | 1 | 1 | 1 | 1 | |
| 4 | | | | | | | |

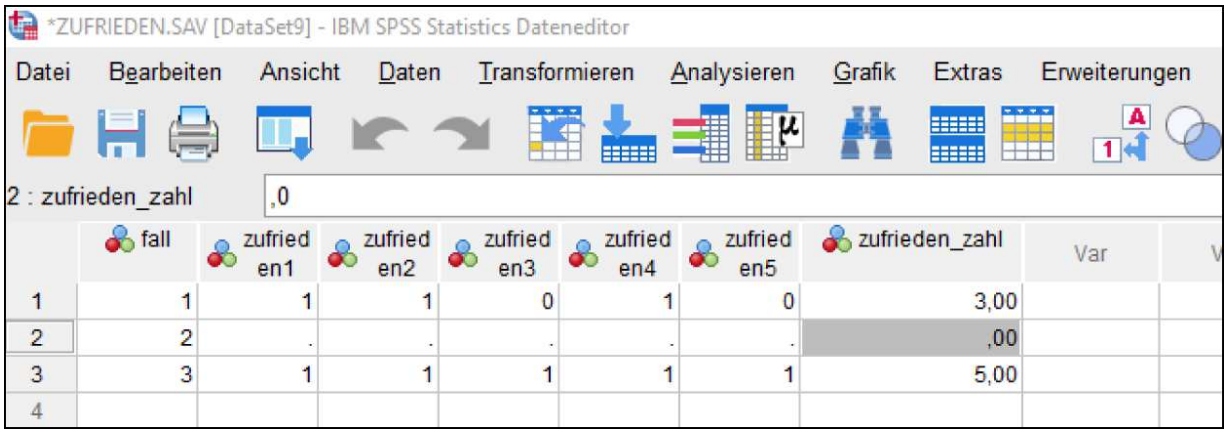
**Abbildung 17.11:** Datendatei *ZUFRIEDEN.SAV* mit fehlenden Angaben

Wählen Sie nun *zufrieden_zahl* als Zählvariable dafür, in wie vielen Kategorien eine Person zufrieden ist (das heißt, dass die betreffende Zufriedenheitsvariable den Wert 1 aufweist), würden Sie als SPSS-Befehl eingeben:

```
COUNT zufrieden_zahl = zufrieden1 zufrieden2 zufrieden3 zufrieden4
zufrieden5(1).
EXECUTE.
```

Dadurch, dass sich bei der Person 2 bei allen Zufriedenheitsvariablen keine Angaben finden lassen, ergibt sich über die vorstehende Befehlsfolge das unschöne Ergebnis von *zufrieden_ zahl* bei Person 2 »0« – und nicht »Fehlend«. Dies geht aus Abbildung 17.12 hervor.

Irrtümlicherweise könnte man also folgern, dass Person 2 in keiner Zufriedenheitskategorie zufrieden war …



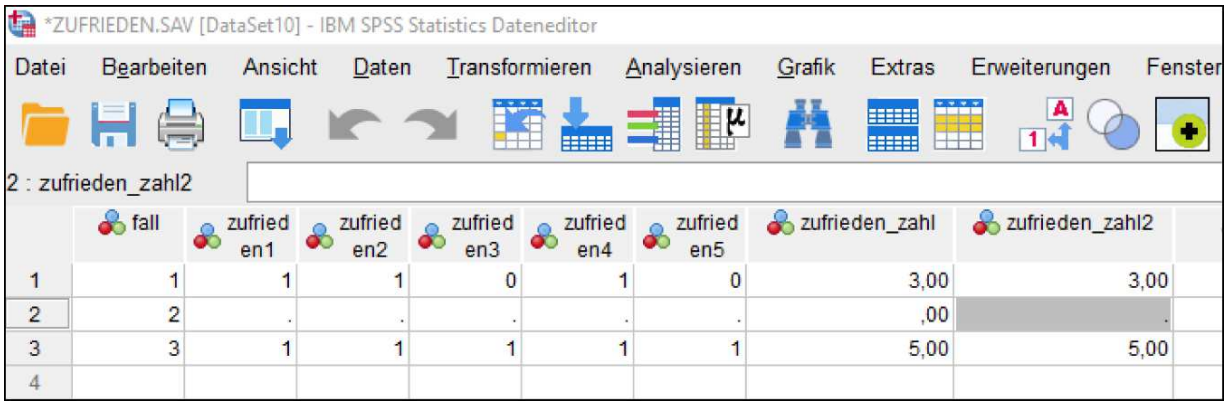

| | fall | zufrieden1 | zufrieden2 | zufrieden3 | zufrieden4 | zufrieden5 | zufrieden_zahl | Var |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 1 | 0 | 3,00 | |
| 2 | 2 | . | . | . | . | . | ,00 | |
| 3 | 3 | 1 | 1 | 1 | 1 | 1 | 5,00 | |
| 4 | | | | | | | | |

**Abbildung 17.12:** Datendatei *ZUFRIEDEN.SAV* mit fehlerhafter Zählvariablen *zufrieden_zahl*

Dieses unschöne Ergebnis können Sie vermeiden, wenn Sie den NMISS- dem COUNT-Befehl im Kontext einer DO-IF-Schleife vorschalten:

```
DO IF (NMISS(zufrieden1 TO zufrieden5) = 0).
COUNT zufrieden_zahl2 = zufrieden1 zufrieden2 zufrieden3 zufrieden4
zufrieden5(1).
END IF.
EXECUTE.
```

Führen Sie dieses kleine SPSS-Syntaxprogramm aus (für Sie im Übrigen als *ZUFRIEDEN_ZAHL.SPS* vorbereitet), erhalten Sie für die Zählvariable *zufrieden_zahl2* und für Person Nummer 2 das korrekte Ergebnis, dass ausschließlich fehlende Werte vorliegen (siehe Abbildung 17.13). So einfach kann man also mit SPSS Probleme lösen – ist das nicht toll?




| | fall | zufrieden1 | zufrieden2 | zufrieden3 | zufrieden4 | zufrieden5 | zufrieden_zahl | zufrieden_zahl2 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 1 | 0 | 3,00 | 3,00 |
| 2 | 2 | . | . | . | . | . | ,00 | . |
| 3 | 3 | 1 | 1 | 1 | 1 | 1 | 5,00 | 5,00 |
| 4 | | | | | | | | |

**Abbildung 17.13:** Datendatei *ZUFRIEDEN.SAV* mit korrekter Zählvariablen *zufrieden_zahl2*

# Kapitel 18

# Auch Makros machen das Leben leichter

## IN DIESEM KAPITEL

Grundlagen der Makroprogrammierung

Regeln der Makroprogrammierung

Einbeziehen kleiner Syntaxfiles in größere Syntaxfiles

Kapitel 18 thematisiert zunächst die Grundlagen und die Regeln der Makroprogrammierung. Dabei wird auf Makros ohne, auf solche mit Parametern sowie auch auf Makros zu Schleifen eingegangen. Schließlich wird gezeigt, wie kleine Syntaxfiles in größere eingebettet werden können.

## *Grundlagen der Makroprogrammierung: Makros ohne Parameter*

**Makros** sind kleine SPSS-Programme, die der Automatisierung sich wiederholender SPSS-Anwendungen dienen. Mit einem SPSS-Makro können Sie eine Verkürzung im Sinne einer größeren Übersichtlichkeit beziehungsweise eine Beschleunigung Ihrer SPSS-Programme erreichen.

Grundlegende Informationen zu SPSS-Makros finden sich (im dortigen Kapitel 9) bei Schendera, Christian F. G. (2005): Datenmanagement mit SPSS. Kontrollierter und beschleunigter Umgang mit Datensätzen, Texten und Werten, Berlin/Heidelberg: Springer-Verlag. Thematisch

Vergleichbares auf hohem Niveau finden Sie in Brosius, Felix (2008): SPSS-Programmierung. Effizientes Datenmanagement und Automatisierung mit SPSS-Syntax, 2. Auflage, Bonn: mitp-Verlag.

Ein Makro beginnt immer mit dem Ausdruck DEFINE, dem mit einem voranstehenden Ausrufezeichen der Makroname folgt. Ein Makroname beginnt also in SPSS mit einem Ausrufezeichen. Dies ist eigentlich auch recht praktisch, weil dadurch sofort erkennbar ist, dass es sich um ein Makro handelt. In der nächsten Befehlszeile folgt dann der eigentliche Befehl beziehungsweise die Befehlssequenz, ehe durch !ENDDEFINE (und einen Punkt am Ende) das Makro beendet wird.

Nochmals: Allgemein sagt man, dass

- ✔ ein Makro mit DEFINE beginnt,
- ✔ dem der Makroinhalt und
- ✔ dann der schließende !ENDDEFINIE-Ausdruck folgen.

Ein Makro-Beispiel ist das folgende. Es besteht sogar aus zwei Makros. Im ersten Makro !VARLIST wird eine Variablenliste (mit den Variablen *a1* bis *a100*) definiert. Für diese 100 Variablen wird im zweiten Makro jeweils der DESCRIPTIVES-Befehl angewandt. Die durch das erste Makro !VARLIST definierte Variablenliste *a1* bis *a100* wird dann im zweiten Makro, das !DESCRIP heißt, im Rahmen der DESCRIPTIVES-Prozedur aufgerufen (konkret werden dort, wie bereits angedeutet, die Variablen *a1* bis *a100* aufgerufen). Nachdem das zweite Makro formuliert worden ist (und mittels !ENDDEFINE abgeschlossen wurde), wird es mit »!DESCRIP.« aufgerufen. Konkret werden durch diesen letztgenannten Aufruf für jede der 100 Variablen deskriptive Statistiken berechnet.

```
DEFINE !VARLIST ( )
a1 TO a100
!ENDDEFINE.

DEFINE !DESCRIP ( )
DESCRIPTIVES VARIABLES=!VARLIST
!ENDDEFINE.
!DESCRIP.
```

SPSS-Makros werden in der Regel in SPSS-Syntaxprogramme eingebunden. Sie können sich ein Makro, bildlich gesprochen und vereinfacht, als die Tonaufnahme eines Lieds vorstellen, die dann »irgendwann« abgespielt wird – allerdings üblicherweise mehrfach, und zwar, um bei dem Bild der Tonaufnahme eines Lieds zu bleiben, von verschiedenen Interpreten nacheinander. In unserem vorstehenden Beispiel entspricht beim zweiten Makro die Tonaufnahme Ihrem Makro !DESCRIP, das Lied heißt »DESCRIPTIVES«, und die Interpreten sind *a1* bis *a100*.

Sie finden einige SPSS-Makros auch im Internet. Ein Beispiel hierfür ist das Makro *PROCESS*. Dieses Makro wurde von Andrew F. Hayes für die sogenannte Mediatoranalyse entwickelt. Hierunter ist eine Regressionsanalyse zu verstehen, bei der eine unabhängige Variable einen bestimmten Zusammenhang »vermittelt«. Nehmen Sie zum Beispiel an, es sei ein positiver Zusammenhang zwischen dem Alter einer Person und ihrer Ausgabenhöhe festgestellt worden. In diesem Zusammenhang könnte es nun sein, dass sozusagen nach dem Senioritätsprinzip in der betreffenden Region (oder eventuell auch in einem Unternehmen) die Einkommen mit dem Alter wachsen. Dadurch würde die Einkommensvariable zum Mittler zwischen dem Zusammenhang zwischen Alter und Ausgaben.

Wenn Sie sich das Makro PROCESS aus dem Internet heruntergeladen haben, können Sie es in SPSS 29 über ERWEITERUNGEN|EXTRAS|BENUTZERDEFINIERTES DIALOGFELD INSTALLIEREN (KOMPATIBILITÄTSMODUS) nach SPSS »bringen«, und zwar indem Sie die Datei *PROCESS.SPD* laden (siehe Abbildung 18.1).

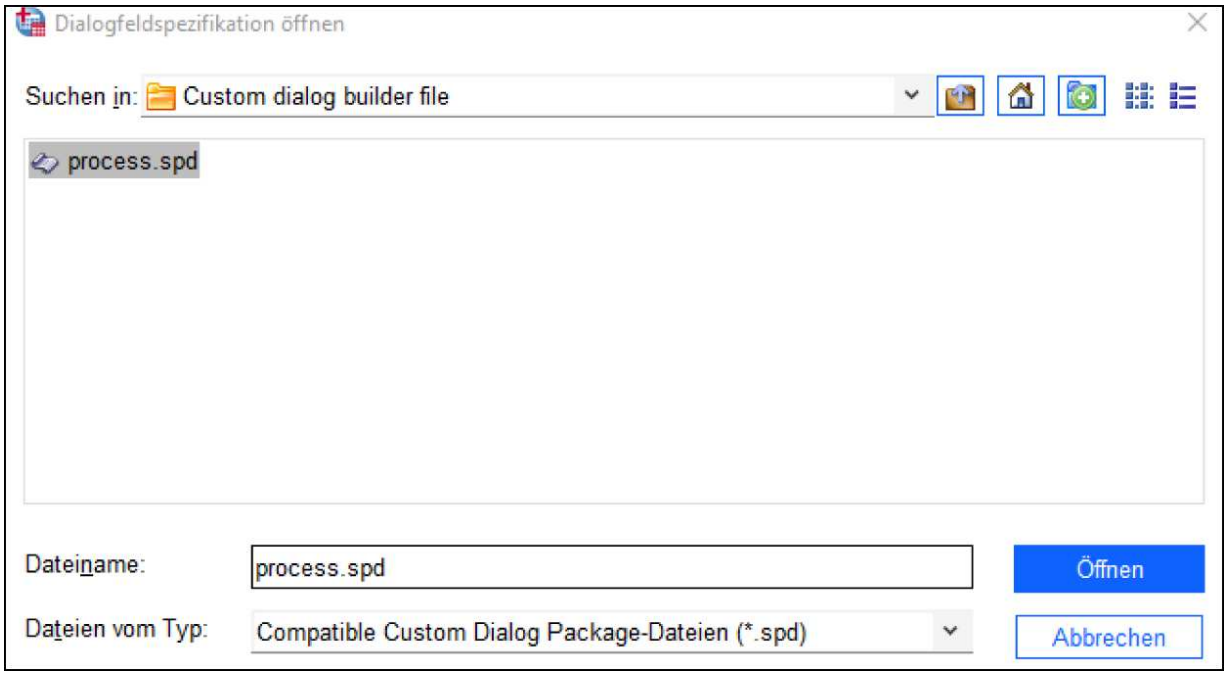



**Abbildung 18.1:** Installation des Makros PROCESS in SPSS

Anschließend fügt SPSS die Mediatoranalyse mithilfe von PROCESS in das SPSS-Menü ein, und zwar unter ANALYSIEREN|REGRESSION (siehe Abbildung 18.2). Jetzt könnten Sie mit einer Mediatoranalyse loslegen …

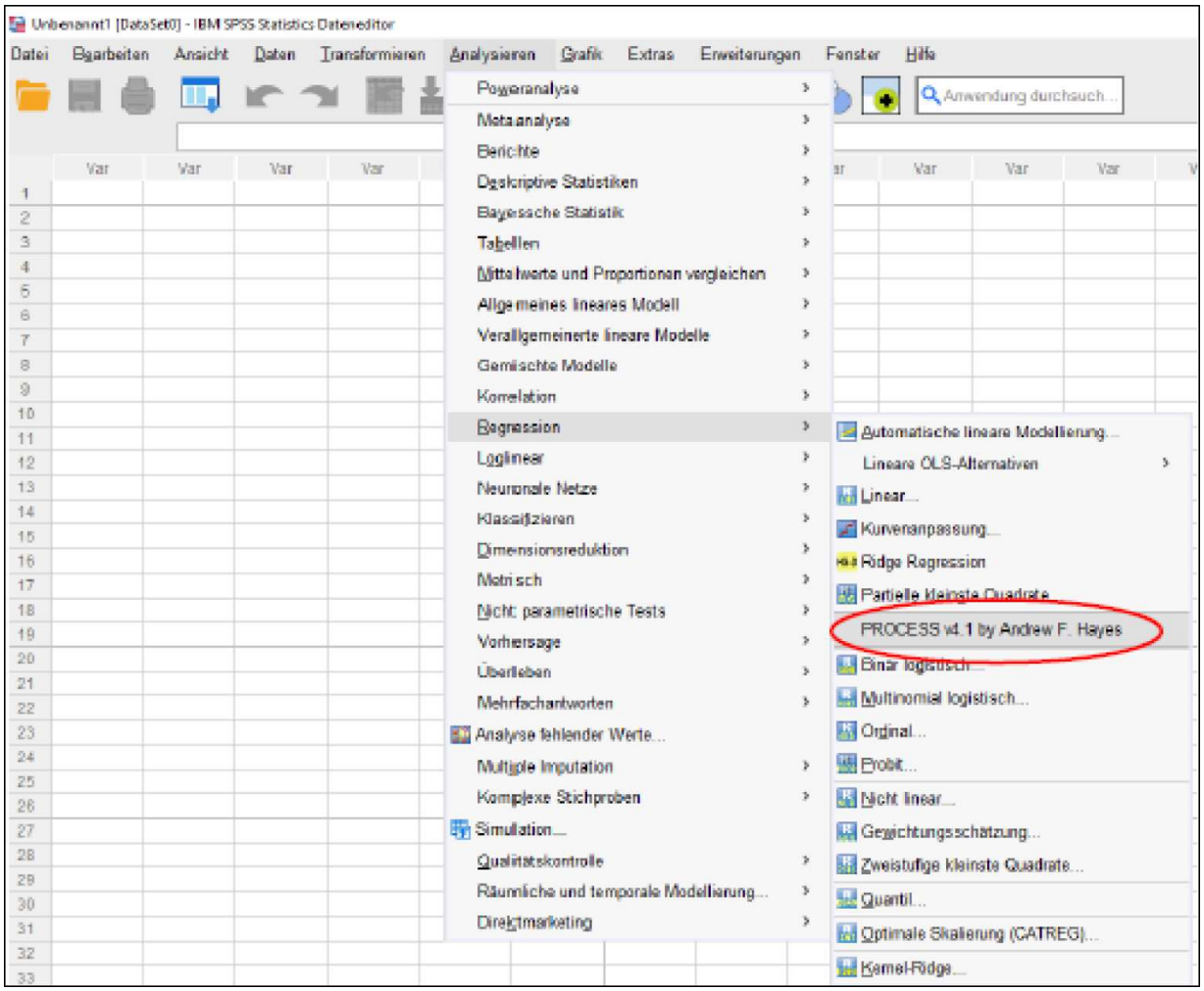



**Abbildung 18.2:** In SPSS eingebettetes Makro PROCESS

Ein von Ihnen selbst definiertes Makro steht Ihnen während Ihrer gesamten laufenden SPSS-Sitzung zur Verfügung und kann dabei jederzeit über seinen Namen aufgerufen werden. Es kann demnach während der betreffenden Sitzung allerdings auch nicht gelöscht werden. Definieren Sie ein Makro mit dem Namen eines noch aktiven Makros, wird der Makroinhalt des alten Makros mit dem Makroinhalt des neuen Makros überschrieben.

Was Sie allerdings in diesem Zusammenhang nicht dürfen, ist, innerhalb eines Makros – also im betreffenden Makroinhalt – ein weiteres Makro (mit DEFINE … !ENDDEFINIE) definieren.

# *Makros mit Parametern*

Im vorhergehend Abschnitt standen nach den Makro-Definitionen leere Klammern: »( )«. Inhaltlich heißt dies (zugegebenermaßen etwas hochgestochen formuliert), dass ein solches Makro keine Angaben im Sinne von **Parametern** enthält. Innerhalb des Klammerpaars können hingegen durchaus bestimmte Angaben (= Parameter) stehen, die dann beim Aufruf beziehungsweise bei der Verarbeitung des Makros berücksichtigt werden. Dadurch werden Makros flexibler!

Zum Beispiel können Sie eine Variablenbegrenzung bis zu einem vorgegebenen Endwert der Variablenliste in Form eines bestimmten Zeichens wie »%« angeben. Wenn Sie in diesem Sinne zum Beispiel eine Variablenliste mit fünf Variablen definieren und nach der vierten Variablen ein »%« setzen, dann werden nur die ersten vier Variablen abgearbeitet.

Hierzu nutzen Sie in der Makro-Definitionszeile für die hier so benannte Variablenliste VARLIST das in SPSS bereits existierende Makro !CHAREND ('›'%'‹'), das eben die Begrenzung Ihrer Variablenliste bei dem Zeichen »%« setzt:

```
DEFINE !DESCRIP2 (VARLIST = !CHAREND ('%'))
DESCRIPTIVES VARIABLES = !VARLIST.
!ENDDEFINE.

!DESCRIP2 VARLIST = var1 var2 var3 var4 % var5.
```

Mit dem Aufruf von !DESCRIP2 wird also die Parametersetzung wirksam, und es werden lediglich für die Variablen *var1* bis *var4* deskriptive Statistiken ausgegeben (und nicht für die nach »%« stehende Variable *var5*).

Eine verbreitete Parametersetzung funktioniert über sogenannte **Token**. Ein Token kennzeichnet die Anzahl der im Makro relevanten Parameter. Mit dem in SPSS bereits existierenden Makro !TOKENS(N) wird eine feste Anzahl n an Parametern an das von Ihnen konkret definierte Makro übergeben.

Das Beispiel von eben sieht mit !TOKENS(4) so aus, wobei Sie die (vier) ausgewählten Variablen jetzt am Ende nach VARLIST direkt benennen:

```
DEFINE !DESCRIP2 (VARLIST = !TOKENS(4))
DESCRIPTIVES VARIABLES = !VARLIST.
!ENDDEFINE.

!DESCRIP2 VARLIST = var1 var2 var3 var4.
```

Sie sollten darauf achten, dass die über !TOKENS(N) genannte Anzahl der Parameter mit der an das Makro übergebenen Parameteranzahl übereinstimmt. Bei einer Abweichung erhalten Sie nämlich eine SPSS-Fehlermeldung. Im obigen Fall ist alles in Ordnung: Die nach !TOKENS genannte Parameteranzahl (= 4) stimmt mit der in VARLIST enthaltenen Anzahl an Elementen (= 4 Variablen) überein.

Sie können die Ausführung eines Makros – die Experten sprechen von der **Makroexpansion** – im Übrigen steuern: Durch !OFFEXPAND können Sie sie ausschalten und durch !ONEXPAND (wieder) einschalten.

# Loop-Makros

In Analogie zum vorhergehenden Kapitel 17 »Schleifen zum schnelleren Programmieren« können auch mit Makros Schleifen abgearbeitet werden.

## Allgemeines zu !DO-Loops

Man unterscheidet zwischen:

- ✔ Index-Loops (!DO-Loops) und
- ✔ List-Processing-Loops (!DO-!IN-Loops).

Mit einem **Index-Loop** hätten Sie die im vorhergehenden Kapitel 17 stehende Anwendung COMPUTE *a* = 0 (für *a* = *var1* TO *var50*, also für

die 50 Variablen *var1* bis *var50*) so programmieren können (siehe die Syntaxdatei *SCHLEIFE.SPS*, die Sie zum Beispiel als Textbaustein für eigene Makros verwenden können):

```
DEFINE !SCHLEIFE (START=!TOKENS(1)
    /ENDE=!TOKENS(1)).
!DO !j = !START !TO !ENDE.
COMPUTE !CONCAT (var, !j) = 0.
!DOEND.
!ENDDEFINE.

!SCHLEIFE START=1 ENDE=50.
EXECUTE.
```

Beim vorstehenden Makro läuft also ein **Zähler** von 1 bis 50 (im Sinne von 50 Schleifen). Sie starten diese Schleife für ein Token, das in START abgelegt ist (= 1; siehe vorletzte Syntaxzeile), und beenden sie für ein Token, das sich in ENDE befindet (= 50; siehe ebenfalls vorletzte Syntaxzeile). Mit anderen Worten: Sie starten die Schleife beim Wert 1 und beenden sie beim Wert 50. Dies wird durch die !DO-!TO-Konstruktion bewerkstelligt.

In dieser !DO-!TO-Konstruktion definieren Sie einen Index !j (daher auch der Name »Index-Loop«), der in diesem Fall eben genau von 1 bis 50 läuft.

Inhaltlich wird der COMPUTE-Befehl 50-mal ausgeführt. Mit !CONCAT fügen Sie den Text *var* mit den Indexausprägungen 1 bis 50 zusammen, sodass 50 Variablen *var1*, *var2*, …, *var50* gebildet werden. Jede dieser 50 Variablen wird mit dem COMPUTE-Befehl hier gleich null gesetzt. Mit !DOEND wird die Schleife hier nach dem 50. Durchlauf beendet.

Sie könnten jetzt dieses Makro nutzen, um die erst einmal gleich null gesetzten 50 Variablen *var1* bis *var50* anschließend in Dummy-Variablen mit den Ausprägungen 0 und 1 zu transformieren.

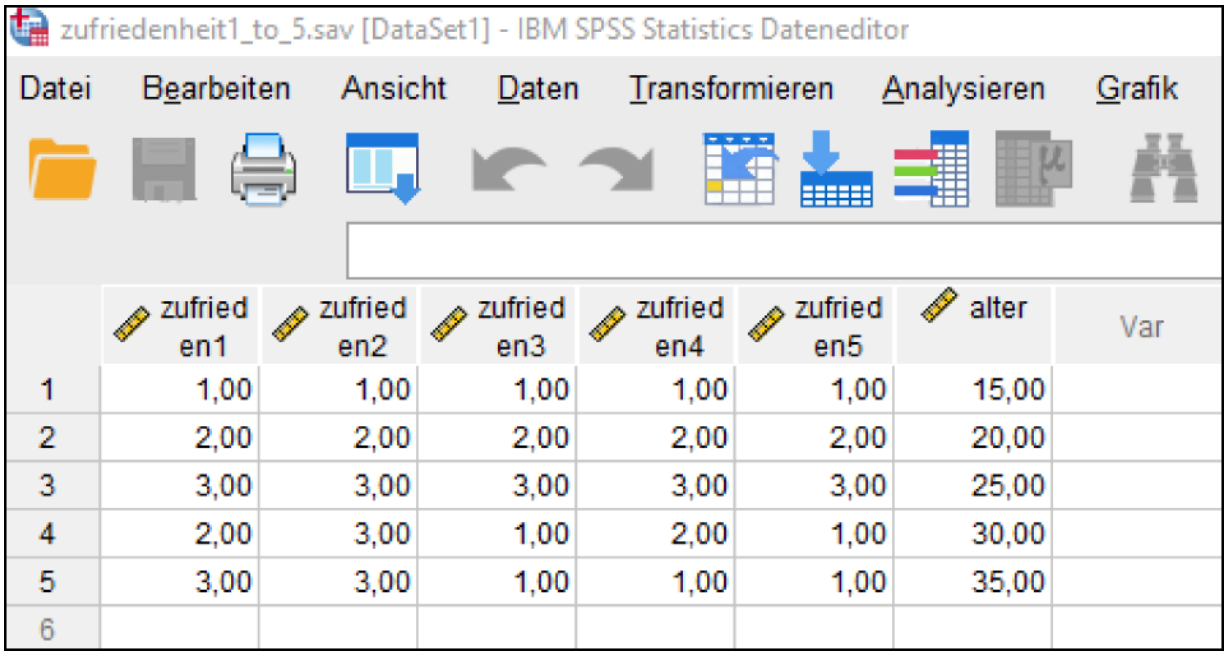
## *Ein !DO-!TO-Loop-Beispiel*

Um das Ganze übersichtlich zu halten, gehen Sie bitte von Datei *ZUFRIEDENHEIT1_TO_5.SAV* mit den dortigen fünf Variablen *zufrieden1* bis *zufrieden5* aus, die in den fünf Variablen *top1* bis *top5*

»aufgehen« sollen. Dabei sollen die Variablen *top1* bis *top5* jeweils dann den Wert 1 annehmen, wenn *zufrieden1* bis *zufrieden5* jeweils gleich 3 ist. Zunächst wird in dem folgenden Programm das Loop-Makro !SCHLEIFE definiert. Mittels des »Unter-Makros« !CONCAT und der zugehörigen COMPUTE-Anweisung wird der Wert für *top1* bis *top5* jeweils gleich 0 gesetzt. Danach wird in einem zweiten Loop-Makro !SCHLEIFE2 den Variablen *top1* bis *top5* jeweils der Wert 1 zugewiesen, wenn ihre Pendants *zufrieden1* bis *zufrieden5* den Wert 3 aufweisen (siehe das Syntaxfile *SCHLEIFE2.SPS*):

```
DEFINE !SCHLEIFE (START=!TOKENS(1)
    /ENDE=!TOKENS(1)).
!DO !j = !START !TO !ENDE.
COMPUTE !CONCAT(top, !j)=0.
!DOEND.
!ENDDEFINE.

!SCHLEIFE START=1 ENDE=5.

EXECUTE.

DEFINE !SCHLEIFE2 (START2=!TOKENS(1)
    /ENDE2=!TOKENS(1)).
!DO !k = !START2 !TO !ENDE2.
COMPUTE !CONCAT(top, !k)=(!CONCAT(zufrieden, !k)=3).
!DOEND.
!ENDDEFINE.


!SCHLEIFE2 START2=1 ENDE2=5.

EXECUTE.
```

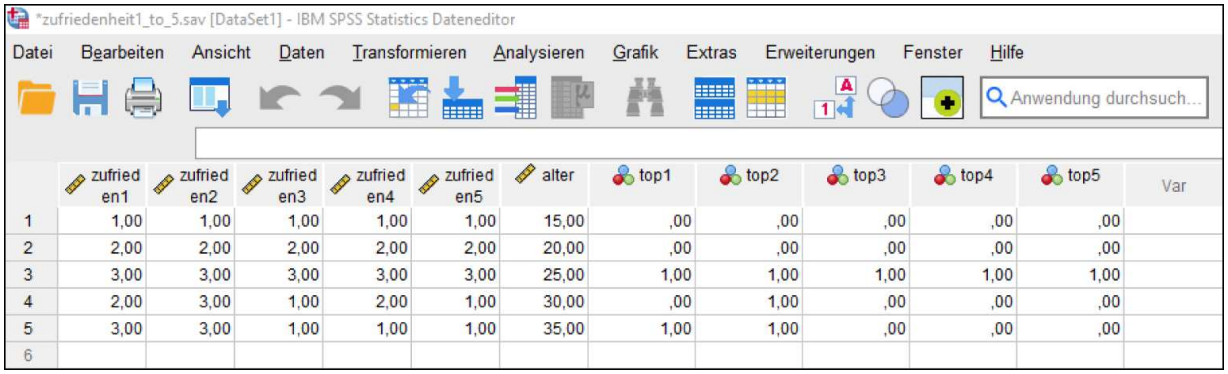
Die Ausgangsdatendatei für dieses Beispiel sieht so aus wie in Abbildung 18.3.




| | zufrieden1 | zufrieden2 | zufrieden3 | zufrieden4 | zufrieden5 | alter | Var |
|---|---|---|---|---|---|---|---|
| 1 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 15,00 | |
| 2 | 2,00 | 2,00 | 2,00 | 2,00 | 2,00 | 20,00 | |
| 3 | 3,00 | 3,00 | 3,00 | 3,00 | 3,00 | 25,00 | |
| 4 | 2,00 | 3,00 | 1,00 | 2,00 | 1,00 | 30,00 | |
| 5 | 3,00 | 3,00 | 1,00 | 1,00 | 1,00 | 35,00 | |
| 6 | | | | | | | |

**Abbildung 18.3:** Datendatei *ZUFRIEDENHEIT1_TO_5.SAV* mit fünf kategorialen Variablen

Aus dem genannten Index-Makro resultiert dann die in Abbildung 18.4 dargestellte Datendatei.




| | zufrieden1 | zufrieden2 | zufrieden3 | zufrieden4 | zufrieden5 | alter | top1 | top2 | top3 | top4 | top5 | Var |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 15,00 | ,00 | ,00 | ,00 | ,00 | ,00 | |
| 2 | 2,00 | 2,00 | 2,00 | 2,00 | 2,00 | 20,00 | ,00 | ,00 | ,00 | ,00 | ,00 | |
| 3 | 3,00 | 3,00 | 3,00 | 3,00 | 3,00 | 25,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | |
| 4 | 2,00 | 3,00 | 1,00 | 2,00 | 1,00 | 30,00 | ,00 | 1,00 | ,00 | ,00 | ,00 | |
| 5 | 3,00 | 3,00 | 1,00 | 1,00 | 1,00 | 35,00 | 1,00 | 1,00 | ,00 | ,00 | ,00 | |
| 6 | | | | | | | | | | | | |

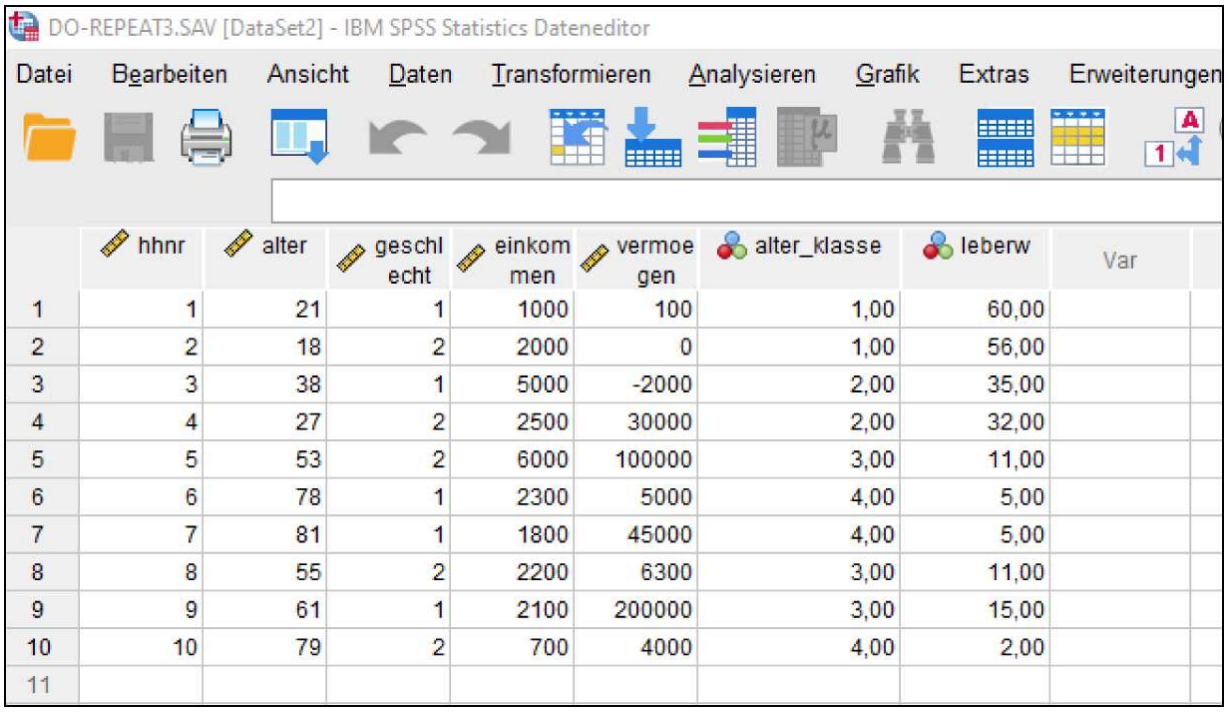
**Abbildung 18.4:** Datendatei mit fünf kategorialen Variablen und aus ihnen abgeleiteten Dummies (*ZUFRIEDENHEIT1_TO_5.SAV*)

## *!DO-!IN-Loop*

Gegenüber einem Index-Loop wird beim **List-Processing-Loop** eine Werteliste nacheinander abgearbeitet. Die Anzahl der Schleifen bestimmt sich nach der Anzahl der in einer Liste angegebenen Werte. Für ein Beispiel bezüglich eines List-Processing-Loops gehen Sie bitte von folgendem Ursprungsdatensatz aus: *DO-REPEAT3.SAV* (siehe Abbildung 18.5).

Für diesen Datensatz ist im Folgenden *!var* die Indexvariable. Das in SPSS vordefinierte Makro !CHAREND gibt hier wieder an, wann die Werteliste endet, und zwar hier bei dem Zeichen »/«. Das gleichfalls in SPSS vordefinierte Makro !CMDEND bezieht sich hier auf die Variablenliste und bedeutet, dass bis zum Ende der Variablenliste die Variablen – hier also alle Variablen – verarbeitet werden.

Konkret werden im folgenden List-Processing-Makro, über die SELECT IF-Anweisung nach der Variablen *alter_klasse* gefiltert, genauer ausgedrückt: für deren ganz unten im Makro gemäß WERTE gemeinten Ausprägungen 1, 2, 3 und 4 (also für vier Altersklassen) deskriptive Statistiken (arithmetischer Mittelwert, Standardabweichung und so weiter) ausgegeben. Diese deskriptiven Statistiken beziehen sich auf die ebenfalls ganz unten im Makro nach VARIABLEN angeführten Variablen *einkommen*, *alter*, *leberw*, *alter_klasse*, *vermoegen* und *geschlecht* (siehe das Syntaxfile *SCHLEIFE3.SPS*):




| | hhnr | alter | geschlecht | einkommen | vermoegen | alter_klasse | leberw | Var |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 21 | 1 | 1000 | 100 | 1,00 | 60,00 | |
| 2 | 2 | 18 | 2 | 2000 | 0 | 1,00 | 56,00 | |
| 3 | 3 | 38 | 1 | 5000 | -2000 | 2,00 | 35,00 | |
| 4 | 4 | 27 | 2 | 2500 | 30000 | 2,00 | 32,00 | |
| 5 | 5 | 53 | 2 | 6000 | 100000 | 3,00 | 11,00 | |
| 6 | 6 | 78 | 1 | 2300 | 5000 | 4,00 | 5,00 | |
| 7 | 7 | 81 | 1 | 1800 | 45000 | 4,00 | 5,00 | |
| 8 | 8 | 55 | 2 | 2200 | 6300 | 3,00 | 11,00 | |
| 9 | 9 | 61 | 1 | 2100 | 200000 | 3,00 | 15,00 | |
| 10 | 10 | 79 | 2 | 700 | 4000 | 4,00 | 2,00 | |
| 11 | | | | | | | | |

**Abbildung 18.5:** Datensatz *DO-REPEAT3.SAV* für Prozess-Schleifen-Makro

```
DEFINE !schleife3 (werte=!CHAREND ("/")
    /variablen=!CMDEND).
GET FILE='C:\SPSS-DATEIEN\DO-REPEAT3.SAV'.
```

```
!DO !var !IN (!WERTE).
TEMPORARY.
SELECT IF (alter_klasse=!var).
DESCRIPTIVES VARIABLES=!variablen
/STATISTICS ALL.
!DOEND.
!ENDDEFINE.

!SCHLEIFE3 WERTE= 1 2 3 4 / VARIABLEN=einkommen alter leberw alter_klasse
vermoegen geschlecht.
```

Wichtig an der vorstehenden Syntax ist, dass für jeden Schleifendurchgang das Ausgangsfile mit dem GET-FILE-Befehl den jeweiligen Berechnungen zugrunde gelegt wird. Der Grund dafür ist, dass für jede der vier Schleifen der Ausgangsdatensatz in den Arbeitsspeicher geladen wird.

Im vorstehenden Beispiel werden für vier Altersklassen getrennt und für die fünf in der Variablenliste genannten Variablen DESCRIPTIVES berechnet. Es ergibt sich der in Abbildung 18.6 angegebene Output (mit den Altersklassen 1 bis 4 von oben nach unten). In der Altersklasse 1 (»bis 24 Jahre«) beispielsweise betragen die Durchschnittswerte für das Einkommen 1.500,00 Euro, für das Alter 19,5 Jahre, die Restlebenserwartung 58,0 Jahre, die Altersklasse 1,0 (natürlich, da ja an dieser Stelle nach dieser Klasse gefiltert wurde!), das Vermögen 50,00 Euro und für das Geschlecht 1,50 (was streng genommen eigentlich nicht hätte berechnet werden dürfen, da es sich beim Geschlecht um eine nominalskalierte Variable handelt).

**Deskriptive Statistiken**

| | N | Spannweite | Minimum | Maximum | Summe | Mittelwert | | Std.-Abweichung | Varianz | Schiefe | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistik | Statistik | Statistik | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Std.-Fehler |
| einkommen | 2 | 1000 | 1000 | 2000 | 3000 | 1500,00 | 500,000 | 707,107 | 500000,000 | . | . | . | . |
| alter | 2 | 3 | 18 | 21 | 39 | 19,50 | 1,500 | 2,121 | 4,500 | . | . | . | . |
| leberw | 2 | 4,00 | 56,00 | 60,00 | 116,00 | 58,0000 | 2,00000 | 2,82843 | 8,000 | . | . | . | . |
| alter_klasse | 2 | ,00 | 1,00 | 1,00 | 2,00 | 1,0000 | ,00000 | ,00000 | ,000 | . | . | . | . |
| vermoegen | 2 | 100 | 0 | 100 | 100 | 50,00 | 50,000 | 70,711 | 5000,000 | . | . | . | . |
| geschlecht | 2 | 1 | 1 | 2 | 3 | 1,50 | ,500 | ,707 | ,500 | . | . | . | . |
| Gültige Werte (listenweise) | 2 | | | | | | | | | | | | |

**Deskriptive Statistik**

**Deskriptive Statistiken**

| | N | Spannweite | Minimum | Maximum | Summe | Mittelwert | | Std.-Abweichung | Varianz | Schiefe | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistik | Statistik | Statistik | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Std.-Fehler |
| einkommen | 2 | 2500 | 2500 | 5000 | 7500 | 3750,00 | 1250,000 | 1767,767 | 3125000,000 | . | . | . | . |
| alter | 2 | 11 | 27 | 38 | 65 | 32,50 | 5,500 | 7,778 | 60,500 | . | . | . | . |
| leberw | 2 | 3,00 | 32,00 | 35,00 | 67,00 | 33,5000 | 1,50000 | 2,12132 | 4,500 | . | . | . | . |
| alter_klasse | 2 | ,00 | 2,00 | 2,00 | 4,00 | 2,0000 | ,00000 | ,00000 | ,000 | . | . | . | . |
| vermoegen | 2 | 32000 | -2000 | 30000 | 28000 | 14000,00 | 16000,000 | 22627,417 | 512000000,00 | . | . | . | . |
| geschlecht | 2 | 1 | 1 | 2 | 3 | 1,50 | ,500 | ,707 | ,500 | . | . | . | . |
| Gültige Werte (listenweise) | 2 | | | | | | | | | | | | |

**Deskriptive Statistik**

**Deskriptive Statistiken**

| | N | Spannweite | Minimum | Maximum | Summe | Mittelwert | | Std.-Abweichung | Varianz | Schiefe | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistik | Statistik | Statistik | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Std.-Fehler |
| einkommen | 3 | 3900 | 2100 | 6000 | 10300 | 3433,33 | 1283,658 | 2223,361 | 4943333,333 | 1,728 | 1,225 | . | . |
| alter | 3 | 8 | 53 | 61 | 169 | 56,33 | 2,404 | 4,163 | 17,333 | 1,293 | 1,225 | . | . |
| leberw | 3 | 4,00 | 11,00 | 15,00 | 37,00 | 12,3333 | 1,33333 | 2,30940 | 5,333 | 1,732 | 1,225 | . | . |
| alter_klasse | 3 | ,00 | 3,00 | 3,00 | 9,00 | 3,0000 | ,00000 | ,00000 | ,000 | . | . | . | . |
| vermoegen | 3 | 193700 | 6300 | 200000 | 306300 | 102100,00 | 55926,231 | 96867,074 | 9383230000,0 | ,098 | 1,225 | . | . |
| geschlecht | 3 | 1 | 1 | 2 | 5 | 1,67 | ,333 | ,577 | ,333 | -1,732 | 1,225 | . | . |
| Gültige Werte (listenweise) | 3 | | | | | | | | | | | | |

**Deskriptive Statistik**

**Deskriptive Statistiken**

| | N | Spannweite | Minimum | Maximum | Summe | Mittelwert | | Std.-Abweichung | Varianz | Schiefe | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistik | Statistik | Statistik | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Std.-Fehler |
| einkommen | 3 | 1600 | 700 | 2300 | 4800 | 1600,00 | 472,582 | 818,535 | 670000,000 | -1,034 | 1,225 | . | . |
| alter | 3 | 3 | 78 | 81 | 238 | 79,33 | ,882 | 1,528 | 2,333 | ,935 | 1,225 | . | . |
| leberw | 3 | 3,00 | 2,00 | 5,00 | 12,00 | 4,0000 | 1,00000 | 1,73205 | 3,000 | -1,732 | 1,225 | . | . |
| alter_klasse | 3 | ,00 | 4,00 | 4,00 | 12,00 | 4,0000 | ,00000 | ,00000 | ,000 | . | . | . | . |
| vermoegen | 3 | 41000 | 4000 | 45000 | 54000 | 18000,00 | 13503,086 | 23388,031 | 547000000,00 | 1,728 | 1,225 | . | . |
| geschlecht | 3 | 1 | 1 | 2 | 4 | 1,33 | ,333 | ,577 | ,333 | 1,732 | 1,225 | . | . |
| Gültige Werte (listenweise) | 3 | | | | | | | | | | | | |

**Abbildung 18.6:** Output nach Prozess-Schleifen-Makro (*DO-REPEAT3.SAV*)

# *Einbeziehen einzelner Syntaxfiles in größere Syntaxfiles*

Mittels **INSERT-Befehl** (früher in SPSS: INCLUDE-Befehl) können Sie Programm-Module in einer separaten Datei lagern und dann später in einem »Haupt-Syntaxfile« aufrufen. Dieses Vorgehen dient in erster Linie der Übersichtlichkeit Ihres »Haupt-Syntaxfiles«. Sie müssen an

der entsprechenden Stelle des »Haupt-Syntaxfiles«, allgemein formuliert, nur Folgendes eingeben:

```
INSERT FILE = 'DATEIPFAD\DATEINAME.SPS'.
```

Gerade auch für die Einbindung vielfach genutzter Makros ist der INSERT-Befehl geeignet. Sie können INSERT-Befehle nacheinander, aber auch verschachtelt in Ihrem »Haupt-Syntaxfile« verwenden. Letzteres bedeutet, dass in Ihren »INSERT-Files« selbst INSERT-Verweise auf andere »INSERT-Files« stehen dürfen.

Des Weiteren können Sie den Symbolleisten in SPSS eigene **Schaltflächen** hinzufügen, die mit speziellen SPSS-Befehlen verknüpft sind. Eine solche Schaltfläche dient, bildlich ausgedrückt, sozusagen als »Abspieltaste« der betreffenden automatisierten Befehle.

Um Letzteres hinzubekommen, müssen Sie auf der Symbolleiste mit der rechten Maustaste auf eine beliebige Stelle klicken. Es öffnet sich ein Kontextmenü, in dem Sie bitte auf ANPASSEN klicken. Anschließend klicken Sie BEARBEITEN|NEUES SYMBOL an (siehe Abbildung 18.7).

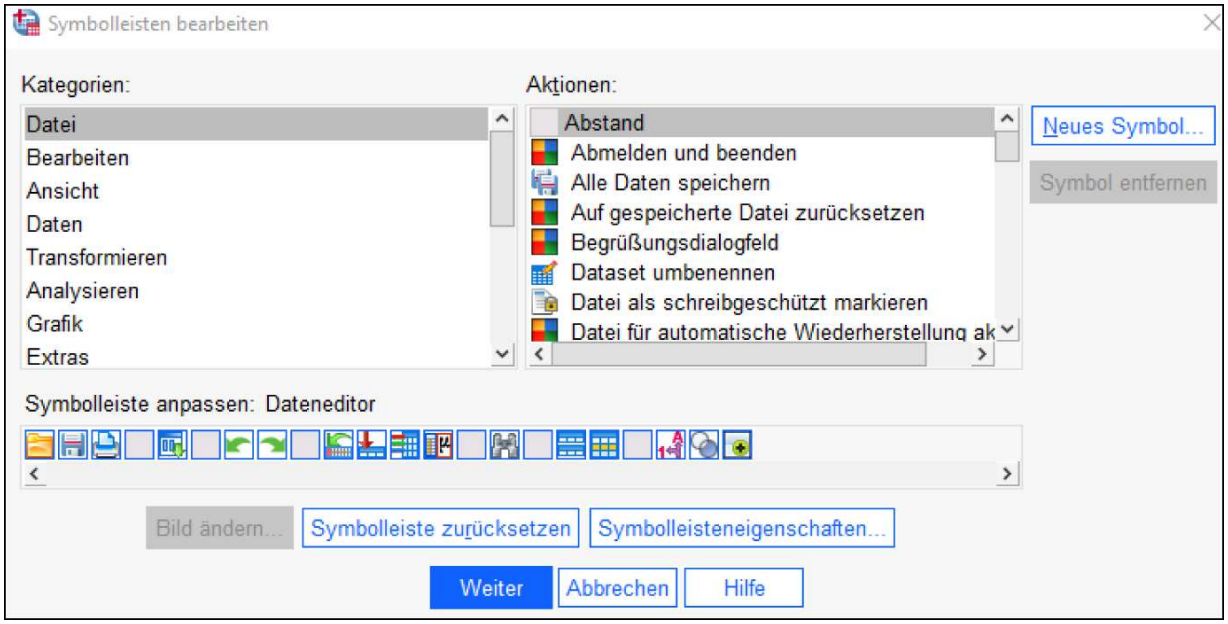



**Abbildung 18.7:** Erstellen eines Makrosymbols (1)

Dann geben Sie – als Beispiel – das aus dem vorstehenden Abschnitt bekannte Makro ein, das Sie als Syntaxdatei *SCHLEIFE3.SPS* finden. Dabei müssen Sie das betreffende Makro kurz beschreiben, und zwar im Sinne einer Namensvergabe sowie im Hinblick auf Speicherort und Namen dieses Syntaxfiles (siehe hierzu Abbildung 18.8).

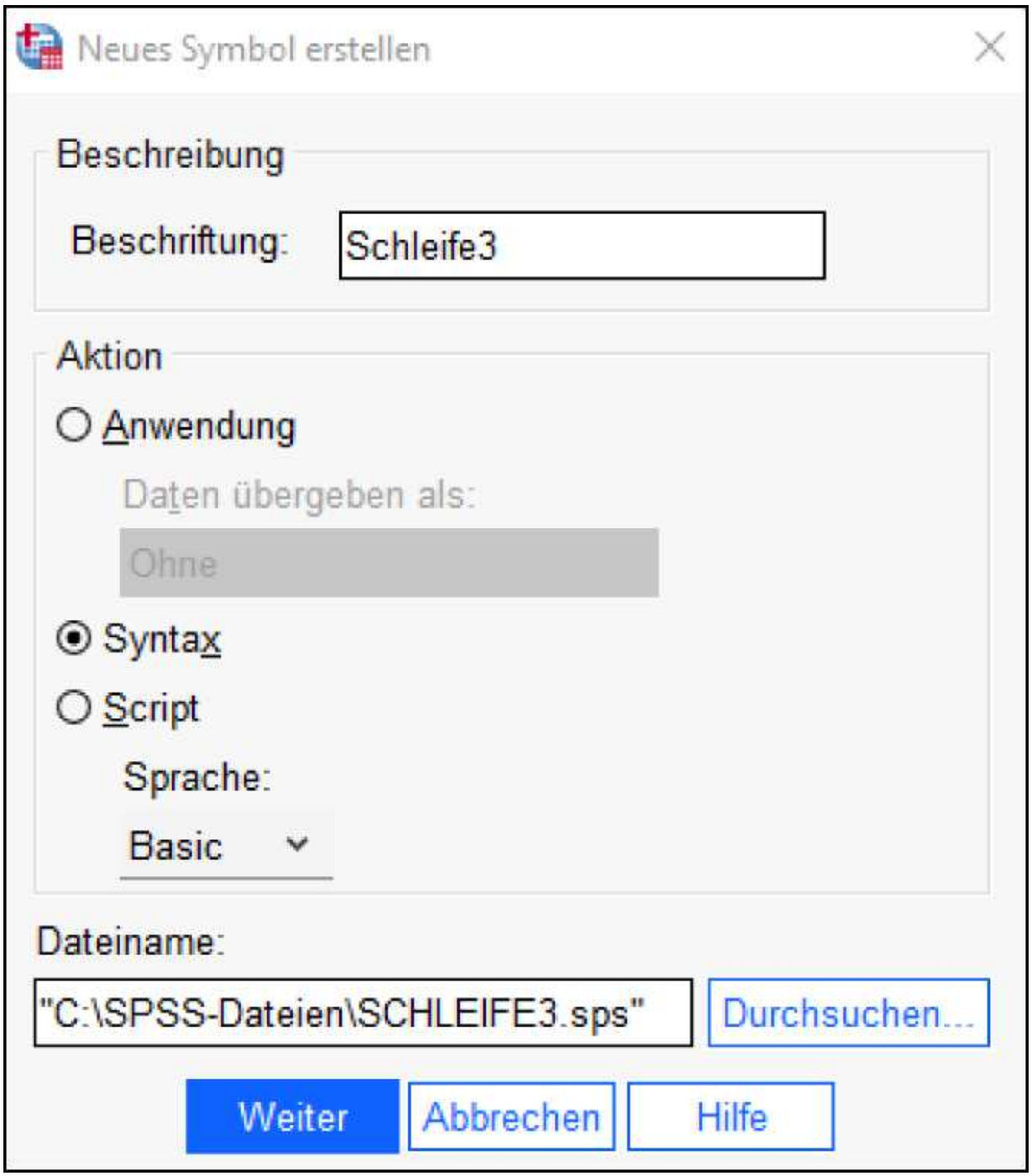



**Abbildung 18.8:** Erstellen eines Makrosymbols (2)

Nachdem Sie auf WEITER geklickt haben, müssen Sie Ihr neues Symbol »Schleife3« im Dialogfeld SYMBOLLEISTEN BEARBEITEN von AKTIONEN mit der Maus nach unten in SYMBOLLEISTEN ANPASSEN: DATENEDITOR ziehen und wieder mal auf WEITER klicken (siehe Abbildung 18.9).

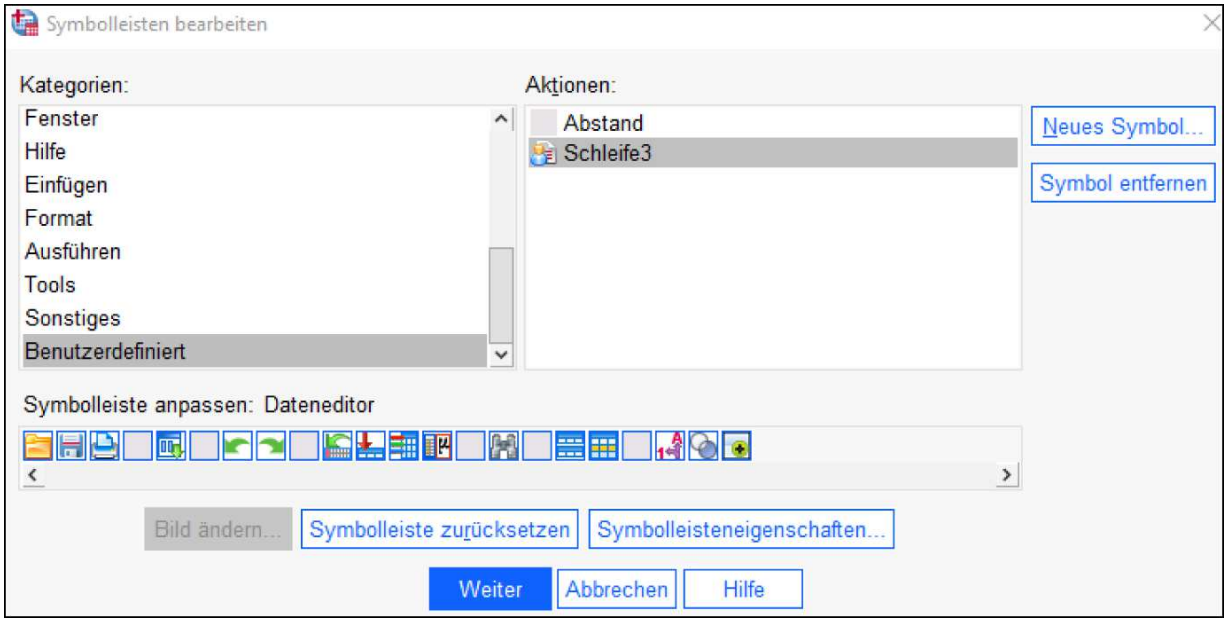



**Abbildung 18.9:** Erstellen eines Makrosymbols (3)

Wenn Sie all dies getan haben, wird Ihr neues Symbol in der Symbolleiste zu finden sein (siehe das durch die von mir eingezeichnete Ellipse hervorgehobene zweite Symbol von rechts in Abbildung 18.10). Es ist jetzt zum »Abspielen« bereit!

**Abbildung 18.10:** Erstellen eines Makrosymbols (4)

# Kapitel 19

# Beispielhafte Syntaxprogramme

## IN DIESEM KAPITEL

Berechnung von Armutsquoten

Der Gini-Koeffizient und andere Ungleichheitsmaße

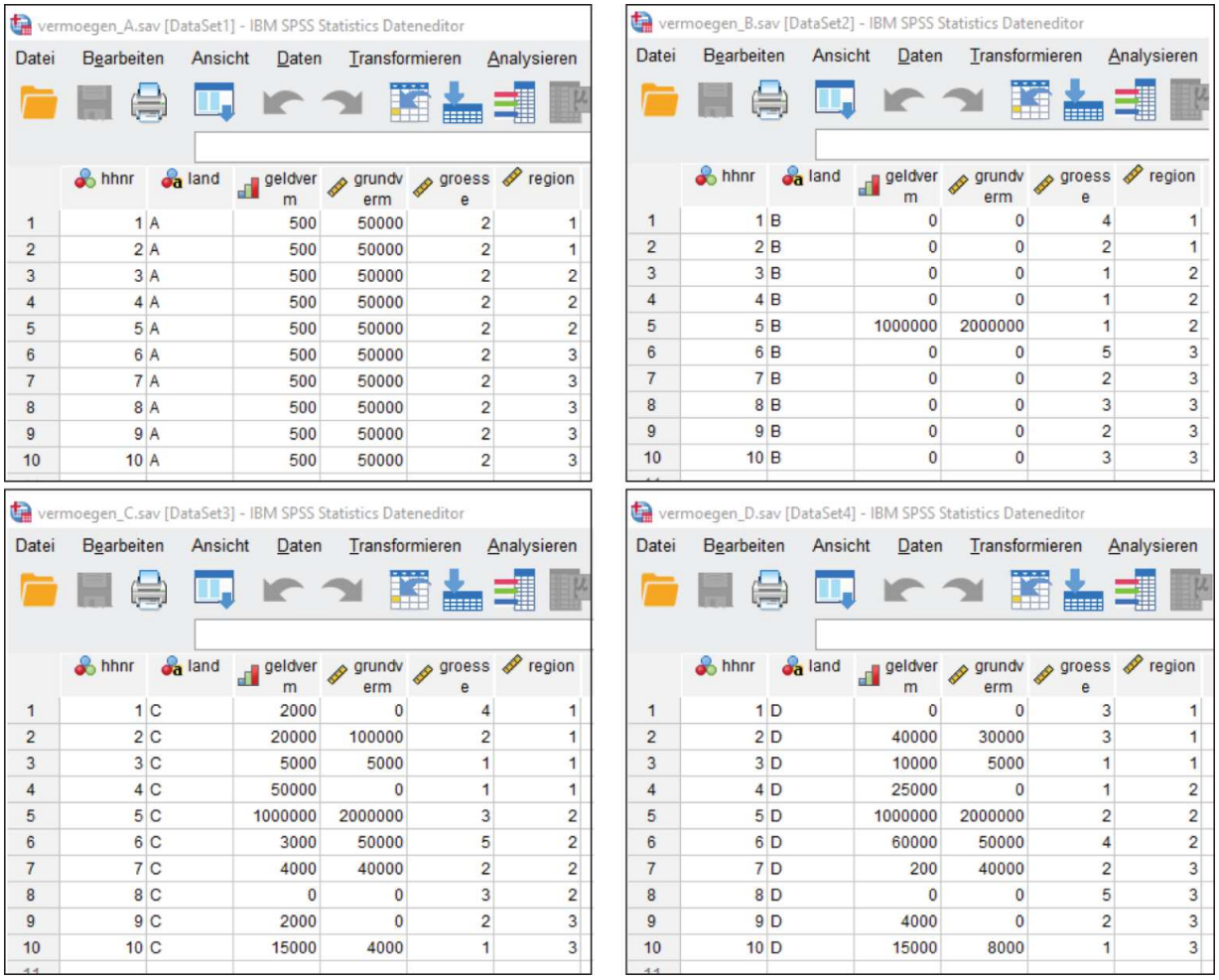
Kapitel 19 präsentiert Ihnen beispielhafte SPSS-Syntaxprogramme, und zwar aus dem Bereich der Ungleichheitsforschung. Zunächst werden sogenannte relative Armutsquoten berechnet, ehe Ungleichheitsberechnungen auf Basis des Gini-Koeffizienten und anderer Ungleichheitsmaße erfolgen. Die Berechnungen werden mittels fiktiver Datensätze durchgeführt und geben Ihnen einen Eindruck über die »Wirkmächtigkeit« von SPSS. Wie Sie unschwer erkennen werden, habe ich die Fragestellungen aus meinem Lieblingsforschungsthema gewählt, nämlich aus dem Bereich der Ungleichheitsforschung.

Dieses Kapitel stellt insofern eine Ausnahme gegenüber den anderen Kapiteln dieses Buchs dar, als Sie schon einige Grundkenntnisse aus anderen Kapiteln mitbringen sollten, um die in diesem Kapitel präsentierten SPSS-Programme verstehen zu können. Ansonsten ist es ja die Maxime des Buchs, dass man die einzelnen Buchkapitel (zumindest weitgehend) unabhängig von den anderen Buchkapiteln lesen und verstehen kann. Für das Verständnis der Zusammenhänge in diesem Kapitel wäre es schon gut, wenn Sie über gewisse, in Kapitel 5 vermittelte Kenntnisse bezüglich der SPSS-Syntax verfügen würden und sich darüber

hinaus auch schon einmal angeschaut hätten, was man unter Schleifen (Kapitel 17) und Makros (Kapitel 18) versteht.

# *Armutsquoten*

Zunächst präsentiere ich Ihnen auf der Grundlage von vier fiktiven Datendateien *VERMOEGEN_A.SAV, VERMOEGEN_B.SAV, VERMOEGEN_C.SAV* und *VERMOEGEN_D.SAV* sogenannte Armuts(risiko)quoten, und zwar bezogen auf den Wohlstandsindikator Vermögen (siehe zu den einzelnen Datendateien Abbildung 19.1). Die Datensätze sollen sich auf die vier Länder A, B, C und D beziehen.



vermoegen_A.sav [DataSet1] - IBM SPSS Statistics Dateneditor

| | hhnr | land | geldverm | grundverm | groesse | region |
|---|---|---|---|---|---|---|
| 1 | 1 | A | 500 | 50000 | 2 | 1 |
| 2 | 2 | A | 500 | 50000 | 2 | 1 |
| 3 | 3 | A | 500 | 50000 | 2 | 2 |
| 4 | 4 | A | 500 | 50000 | 2 | 2 |
| 5 | 5 | A | 500 | 50000 | 2 | 2 |
| 6 | 6 | A | 500 | 50000 | 2 | 3 |
| 7 | 7 | A | 500 | 50000 | 2 | 3 |
| 8 | 8 | A | 500 | 50000 | 2 | 3 |
| 9 | 9 | A | 500 | 50000 | 2 | 3 |
| 10 | 10 | A | 500 | 50000 | 2 | 3 |

vermoegen_B.sav [DataSet2] - IBM SPSS Statistics Dateneditor

| | hhnr | land | geldverm | grundverm | groesse | region |
|---|---|---|---|---|---|---|
| 1 | 1 | B | 0 | 0 | 4 | 1 |
| 2 | 2 | B | 0 | 0 | 2 | 1 |
| 3 | 3 | B | 0 | 0 | 1 | 2 |
| 4 | 4 | B | 0 | 0 | 1 | 2 |
| 5 | 5 | B | 1000000 | 2000000 | 1 | 2 |
| 6 | 6 | B | 0 | 0 | 5 | 3 |
| 7 | 7 | B | 0 | 0 | 2 | 3 |
| 8 | 8 | B | 0 | 0 | 3 | 3 |
| 9 | 9 | B | 0 | 0 | 2 | 3 |
| 10 | 10 | B | 0 | 0 | 3 | 3 |

vermoegen_C.sav [DataSet3] - IBM SPSS Statistics Dateneditor

| | hhnr | land | geldverm | grundverm | groesse | region |
|---|---|---|---|---|---|---|
| 1 | 1 | C | 2000 | 0 | 4 | 1 |
| 2 | 2 | C | 20000 | 100000 | 2 | 1 |
| 3 | 3 | C | 5000 | 5000 | 1 | 1 |
| 4 | 4 | C | 50000 | 0 | 1 | 1 |
| 5 | 5 | C | 1000000 | 2000000 | 3 | 2 |
| 6 | 6 | C | 3000 | 50000 | 5 | 2 |
| 7 | 7 | C | 4000 | 40000 | 2 | 2 |
| 8 | 8 | C | 0 | 0 | 3 | 2 |
| 9 | 9 | C | 2000 | 0 | 2 | 3 |
| 10 | 10 | C | 15000 | 4000 | 1 | 3 |

vermoegen_D.sav [DataSet4] - IBM SPSS Statistics Dateneditor

| | hhnr | land | geldverm | grundverm | groesse | region |
|---|---|---|---|---|---|---|
| 1 | 1 | D | 0 | 0 | 3 | 1 |
| 2 | 2 | D | 40000 | 30000 | 3 | 1 |
| 3 | 3 | D | 10000 | 5000 | 1 | 1 |
| 4 | 4 | D | 25000 | 0 | 1 | 2 |
| 5 | 5 | D | 1000000 | 2000000 | 2 | 2 |
| 6 | 6 | D | 60000 | 50000 | 4 | 2 |
| 7 | 7 | D | 200 | 40000 | 2 | 3 |
| 8 | 8 | D | 0 | 0 | 5 | 3 |
| 9 | 9 | D | 4000 | 0 | 2 | 3 |
| 10 | 10 | D | 15000 | 8000 | 1 | 3 |

**Abbildung 19.1:** Datendateien *VERMOEGEN_A.SAV*, *VERMOEGEN_B.SAV*, *VERMOEGEN_C.SAV* und *VERMOEGEN_D.SAV*

In den einzelnen Dateien stellen die einzelnen Fälle verschiedene Haushalte dar. Wie aus der zweiten Zeile des nachfolgenden Programms hervorgeht, werden die einzelnen Vermögenswerte der Haushalte mit der Variablen *groesse* gewichtet. Inhaltlich meint dies eine Gewichtung eines jeden Vermögenswerts mit der jeweiligen Haushaltsgröße (im Sinne der Personenanzahl in einem Haushalt). Mit dieser Gewichtung wird das haushaltsbezogene Vermögen jedem Haushaltsmitglied zugeordnet, sodass die Berechnungen personenbezogen stattfinden.

Konkret ist im Folgenden Vermögen als Summe aus Geldvermögen (also zum Beispiel Bargeld, Bankguthaben und so weiter) und Grundvermögen (also zum Beispiel der Wert des eigenen Hauses) zusammengesetzt. Um die unterschiedliche Größe der einzelnen Haushalte zu berücksichtigen, wird das entsprechend definierte Vermögen noch durch die jeweilige Haushaltsgröße dividiert (siehe die dritte Zeile des folgenden Syntaxprogramms). Folglich handelt es sich hier beim Vermögen um das Pro-Kopf-Haushaltsvermögen.

Mit dem AGGREGATE-Befehl wird über alle Untersuchungseinheiten hinweg (wegen der Gewichtung: über alle Personen hinweg) das mittlere (genauer ausgedrückt: das mediane) Pro-Kopf-Haushaltsvermögen berechnet und in der Variablen *verm_med* festgehalten.

Anschließend wird als Armutsgrenze (im Sinne der Abgrenzung zwischen Armen und Nichtarmen) 60 Prozent dieses medianen Pro-Kopf-Haushaltsvermögens definiert (siehe die Definition der Variablen *armutsgrenze* nachstehend). In der Variablen *arm* schließlich wird festgehalten, wer arm ist (*arm* = 1) beziehungsweise wer nicht arm ist (*arm* = 0). Über den FREQUENCIES-Befehl wird schließlich ermittelt, wie viele Arme und wie viele Nichtarme es gibt.

All die vorgenannten Befehle sind in das **Makro** !ARMUT eingebettet (und zwar stehen sie zwischen »DEFINE !ARMUT( ).« und »!ENDDEFINE.«). Dieses Makro wird dann im zweiten Teil des folgenden Programms für die vier oben genannten Datensätze jeweils zur Ausführung gebracht.

Die Syntax des nachfolgenden Programms findet sich im Übrigen in der SPSS-Syntaxdatei *ARMUT.SPS*.

```
DEFINE !ARMUT ( ).
WEIGHT BY groesse.

COMPUTE vermoegen_pk = (geldverm + grundverm)/groesse.

AGGREGATE
    /OUTFILE=* MODE=ADDVARIABLES
    /BREAK=
    /verm_med=MEDIAN(vermoegen_pk).

COMPUTE armutsgrenze=0.6 * verm_med.
COMPUTE arm=0.
IF (vermoegen_pk <= armutsgrenze) =1.
FREQUENCIES VARIABLES=arm armutsgrenze.

!ENDDEFINE.

GET FILE='C:\SPSS-DATEIEN\VERMOEGEN_A.SAV'.
!ARMUT.

GET FILE='C:\SPSS-DATEIEN\VERMOEGEN_B.SAV'.
!ARMUT.

GET FILE='C:\SPSS-DATEIEN\VERMOEGEN_C.SAV'.
!ARMUT.

GET FILE='C:\SPSS-DATEIEN\VERMOEGEN_D.SAV'.
!ARMUT.
```

Der Output für die vier fiktiven Datendateien ergibt große Unterschiede in den Armutsquoten (siehe Abbildung 19.2). Im Land A liegt keine der 20 Personen, die jeweils das gleiche Pro-Kopf-Haushaltsvermögen in Höhe von 25.250 Euro haben, unter der Armutsgrenze von 15.150 Euro; das heißt, die Armutsquote im Sinne des Anteils der (0) Armen an allen (20) Personen beträgt 0 Prozent.

Im Land B beträgt hingegen die Armutsquote 95,8 Prozent, da 23 der 24 Personen ein Pro-Kopf-Haushaltsvermögen von 0 haben, was in diesem Land der Armutsgrenze entspricht. Bezogen auf Land C lautet die Armutsquote 37,5 Prozent: In diesem Land sind 9 der insgesamt 24 Personen als arm eingestuft; das heißt, ihr Pro-Kopf-Haushaltsvermögen

beläuft sich höchstens auf den Wert der Armutsgrenze (6.360 Euro). Im Land D schließlich sind 10 der 24 Personen als arm einsortiert, was – bei einer Armutsgrenze von 12.060 Euro – einer Armutsquote von 41,7 Prozent entspricht.

C:\SPSS-DATEIEN\VERMOEGEN_A.SAV

**Statistiken**

| | | arm | armutsgrenze |
|---|---|---|---|
| N | Gültig | 20 | 20 |
| | Fehlend | 0 | 0 |

**Häufigkeitstabelle**

**arm**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | ,00 | 20 | 100,0 | 100,0 | 100,0 |

**armutsgrenze**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | 15150,00 | 20 | 100,0 | 100,0 | 100,0 |

C:\SPSS-DATEIEN\VERMOEGEN_B.SAV

**Statistiken**

| | | arm | armutsgrenze |
|---|---|---|---|
| N | Gültig | 24 | 24 |
| | Fehlend | 0 | 0 |

**Häufigkeitstabelle**

**arm**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | ,00 | 1 | 4,2 | 4,2 | 4,2 |
| | 1,00 | 23 | 95,8 | 95,8 | 100,0 |
| | Gesamt | 24 | 100,0 | 100,0 | |

**armutsgrenze**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | ,00 | 24 | 100,0 | 100,0 | 100,0 |

C:\SPSS-DATEIEN\VERMOEGEN_C.SAV

**Statistiken**

| | | arm | armutsgrenze |
|---|---|---|---|
| N | Gültig | 24 | 24 |
| | Fehlend | 0 | 0 |

**Häufigkeitstabelle**

**arm**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | ,00 | 15 | 62,5 | 62,5 | 62,5 |
| | 1,00 | 9 | 37,5 | 37,5 | 100,0 |
| | Gesamt | 24 | 100,0 | 100,0 | |

**armutsgrenze**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | 6360,00 | 24 | 100,0 | 100,0 | 100,0 |

C:\SPSS-DATEIEN\VERMOEGEN_D.SAV

**Statistiken**

| | | arm | armutsgrenze |
|---|---|---|---|
| N | Gültig | 24 | 24 |
| | Fehlend | 0 | 0 |

**Häufigkeitstabelle**

**arm**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | ,00 | 14 | 58,3 | 58,3 | 58,3 |
| | 1,00 | 10 | 41,7 | 41,7 | 100,0 |
| | Gesamt | 24 | 100,0 | 100,0 | |

**armutsgrenze**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | 12060,00 | 24 | 100,0 | 100,0 | 100,0 |

**Abbildung 19.2:** Fiktive Armutsquoten für die Länder A, B, C und D (*VERMOEGEN_A.SAV*, *VERMOEGEN_B.SAV*, *VERMOEGEN_C.SAV* und *VERMOEGEN_D.SAV*)

# *Gini-Koeffizient und andere Ungleichheitsmaße*

Ein weiteres SPSS-Syntaxbeispiel bezieht sich auf die Ungleichverteilung fiktiver Vermögensverteilungen, das heißt nicht nur auf den unteren (Armuts-)Bereich der Verteilung, wie im vorstehenden Abschnitt, sondern auf die gesamte Verteilung der Pro-Kopf-Haushaltsvermögen. Es wird wieder auf die in Abbildung 19.1 präsentierten (fiktiven) Datensätze *VERMOEGEN_A.SAV, VERMOEGEN_B.SAV, VERMOEGEN_C.SAV* und *VERMOEGEN_D.SAV* Bezug genommen.

In der Ungleichheitsforschung sind einige Kennziffern im Einsatz. Als Beispiele können genannt werden:

- ✔ der Gini-Koeffizient,
- ✔ der normierte Variationskoeffizient und
- ✔ das Theil'sche Entropiemaß.

Mit diesen Maßen wird versucht, das Ausmaß der Ungleichheit zu bestimmen. Je höher die Werte der einzelnen Maße ausfallen, desto ungleichmäßiger ist die Verteilung der zugrunde liegenden Größe X (Einkommen, Vermögen (wie hier) und dergleichen).

## *Gini-Koeffizient*

Das gebräuchlichste Ungleichheitsmaß ist wohl der **Gini-Koeffizient**, der bei vollständiger Gleichverteilung den Wert 0 aufweist. Seine Obergrenze lautet $\frac{N-1}{N}$; sie ergibt sich, wenn eine Untersuchungseinheit die gesamte Merkmalssumme (zum Beispiel die gesamte Einkommens- oder die gesamte Vermögenssumme) auf sich vereinigt und alle anderen Merkmalseinheiten nichts, aber auch gar nichts haben (N = Gesamtzahl der Untersuchungseinheiten). Mit zunehmender Fallzahl nähert sich diese Obergrenze dem Wert 1. Der

Gini-Koeffizient kann nur bei X-Ausprägungen größer gleich null berechnet werden.

Eine gebräuchliche Formel für den Gini-Koeffizienten ist:

$$GINI = \frac{2}{N^2 \cdot \bar{X}} \cdot \sum_{i=1}^{N} i \cdot X_i - \left(1 + \frac{1}{N}\right).$$

In der vorstehenden Formel repräsentiert i den Rang der X-Ausprägung und ist aufsteigend geordnet. Das heißt: Für den höchsten X-Wert ist i gleich 1, für den zweithöchsten X-Wert gleich 2 und so weiter.

## *Normierter Variationskoeffizient*

Der **normierte Variationskoeffizient** CV basiert auf dem Verhältnis von Standardabweichung S und arithmetischem Mittelwert $\bar{X}$, wobei dieses Verhältnis quadriert und anschließend durch 2 dividiert wird:

$$CV = \frac{S^2}{2 \cdot \bar{X}^2}.$$

Bei völliger Gleichverteilung ist CV gleich 0. Ansonsten gilt: Je größer CV ist, desto größer ist die durch ihn gemessene Ungleichheit. Im Gegensatz zum Gini-Koeffizienten kann CV auch für X-Werte kleiner null berechnet werden.

## *Theil'sches Entropiemaß*

Die Formel für das **Theil'sche Entropiemaß** T lautet:

$$T = \frac{1}{N \cdot \bar{X}} \cdot \left[ \sum_{i=1}^{N} X_i \cdot \ln(X_i) - \ln(\bar{X}) \cdot \sum_{i=1}^{N} X_i \right].$$

Offenkundig ist das Theil'sche Entropiemaß auf die Verarbeitung positiver X-Werte beschränkt (wegen der Logarithmierung). Bei vollständiger X-Gleichverteilung ist T gleich 0; im Falle der vollständigen Ungleichverteilung (alle bis auf einen haben nichts, und einer hat alles) beläuft sich T auf ln(N). N ist dabei die Gesamtzahl der Untersuchungseinheiten.

## *Syntaxprogramm I*

Alle vorgenannten Ungleichheitsmaße (Gini-Koeffizient, Normierter Variationskoeffizient und Theil'sches Entropiemaß) finden sich in der nachstehenden Syntax wieder.

Bei der ersten Syntax, die durch das Makro !UNGLEICH1 repräsentiert wird, werden die betreffenden Ungleichheitsmaße für die vier Länder A, B, C und D berechnet. In den zugrunde liegenden (fiktiven) Datendateien spielt die (haushaltsbezogene) Vermögensvariable die Hauptrolle, die sich jeweils aus dem Geldvermögen *geldverm* und dem Grundvermögen *grundverm* zusammensetzt. Die vier zugrunde liegenden Datendateien sind bereits in Abbildung 19.1 angegeben worden.

Über die Division durch die Haushaltsgröße *groesse* wird der Vermögenswert eines Haushalts in eine Pro-Kopf-Größe umgerechnet. Alle Werte werden mit der Haushaltsgröße gewichtet (»WEIGHT BY *groesse*.«). Auf diese Weise wird die hier betrachtete (fiktive) Vermögensverteilung auf die Personen (und nicht auf die Haushalte) bezogen.

Die oben genannten Formeln für den Gini-Koeffizienten, den normierten Variationskoeffizienten und das Theil'sche Entropiemaß sind im Makro !UNGLEICH SPSS-spezifisch umgesetzt worden. Mit dem DESCRIPTIVES-Befehl werden alle drei Ungleichheitsmaße nacheinander für die Länder A bis D berechnet.

Sie finden die Syntax des folgenden Syntaxprogramms in *UNGLEICHHEIT1.SPS*.

```
DEFINE !UNGLEICH1( ).
WEIGHT BY groesse.
COMPUTE vermoegen_pk = (geldverm + grundverm)/groesse.
*Rangbildung für Gini-Koeffizienten-Formel:

RANK VARIABLES=vermoegen_pk (a)

    /RANK
    /PRINT=NO
    /TIES=MEAN.
```

```
IF (vermoegen_pk > 0) ginisumme=rvermoeg * vermoegen_pk.
IF (vermoegen_pk > 0) theilsumme=vermoegen_pk*ln(vermoegen_pk).
IF (vermoegen_pk<=0) ginisumme=0.
IF (vermoegen_pk<=0) theilsumme=0.

AGGREGATE
    /OUTFILE=*
    /BREAK=
    /ginisumme2=SUM(ginisumme)
    /theilsumme2=SUM(theilsumme)
    /fallzahl=N(vermoegen_pk)
    /vermoegen_sum=SUM(vermoegen_pk)
    /vermoegen_mean=MEAN(vermoegen_pk)
    /vermoegen_sd=SD(vermoegen_pk).

COMPUTE gini = ((2/(fallzahl**2*vermoegen_mean))*ginisumme2) - (1+(1/
fallzahl)).
COMPUTE cv = ((vermoegen_sd)**2)/(2*vermoegen_mean**2).
COMPUTE theil = (1/(fallzahl*vermoegen_mean))*(theilsumme2-
(ln(vermoegen_mean)*vermoegen_sum)).

DESCRIPTIVES VARIABLES=gini cv theil
/STATISTICS ALL.

!ENDDEFINE.

TITLE 'PERSONELLE VERMÖGENSUNGLEICHHEIT, LAND A:'.
GET FILE='C:\SPSS-DATEIEN\VERMOEGEN_A.SAV'.
!UNGLEICH1.

TITLE 'PERSONELLE VERMÖGENSUNGLEICHHEIT, LAND B:'.
GET FILE='C:\SPSS-DATEIEN\VERMOEGEN_B.SAV'.
!UNGLEICH1.

TITLE 'PERSONELLE VERMÖGENSUNGLEICHHEIT, LAND C:'.
GET FILE='C:\SPSS-DATEIEN\VERMOEGEN_C.SAV'.
!UNGLEICH1.

TITLE 'PERSONELLE VERMÖGENSUNGLEICHHEIT, LAND D:'.
GET FILE='C:\SPSS-DATEIEN\VERMOEGEN_D.SAV'.
!UNGLEICH1.
```

Wie der nachstehende Outputauszug in Abbildung 19.3 zeigt, gibt SPSS für jedes der vier Länder die Werte für den Gini-Koeffizienten (in der Variablen *gini*), den normierten Variationskoeffizienten (in der Variablen *cv*) und das Theil'sche Entropiemaß (in der Variablen *theil*) aus.

Eine Übersicht über die in unserem Beispiel berechneten einzelnen Ungleichheitsmaße bietet Tabelle 19.1. Die niedrigste (Vermögens-)Ungleichheit wird jeweils für Land A ausgewiesen. Dort sind alle Ungleichheitskennziffern gleich 0, was jetzt aber auch nicht besonders überrascht: Betrachten Sie nämlich noch einmal den Datensatz *VERMOEGEN_A.SAV* (in Abbildung 19.1), stellen Sie fest, dass alle Untersuchungseinheiten im Land A ein gleich hohes Vermögen haben – es gibt in dem betreffenden Land also keinerlei Ungleichverteilung des Vermögens.

Die höchste Ungleichheit wird jeweils für Land B angegeben. Auch dies ist keineswegs überraschend, denn ein Blick auf den Datensatz *VERMOEGEN_B.SAV* (in Abbildung 19.1) zeigt, dass dort alle bis auf eine Untersuchungseinheit keinerlei Vermögen haben. Es herrscht im Land B also eine extreme Ungleichverteilung des Vermögens, und der Gini-Koeffizient erreicht für die insgesamt 24 Personen mit $\frac{23}{24} = 0{,}9583$ seinen größtmöglichen Wert.

PERSONELLE VERMÖGENSUNGLEICHHEIT, LAND A:

**Deskriptive Statistik**

**Deskriptive Statistiken**

| | N | Spannweite | Minimum | Maximum | Summe | Mittelwert | | Std.-Abweichung | Varianz | Schiefe | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistik | Statistik | Statistik | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Std.-Fehler |
| gini | 1 | ,00 | ,00 | ,00 | ,00 | ,0000 | . | . | . | . | . | . | . |
| cv | 1 | ,00 | ,00 | ,00 | ,00 | ,0000 | . | . | . | . | . | . | . |
| theil | 1 | ,00 | ,00 | ,00 | ,00 | ,0000 | . | . | . | . | . | . | . |
| Gültige Werte (listenweise) | 1 | | | | | | | | | | | | |

PERSONELLE VERMÖGENSUNGLEICHHEIT, LAND B:

**Deskriptive Statistik**

**Deskriptive Statistiken**

| | N | Spannweite | Minimum | Maximum | Summe | Mittelwert | | Std.-Abweichung | Varianz | Schiefe | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistik | Statistik | Statistik | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Std.-Fehler |
| gini | 1 | ,00 | ,96 | ,96 | ,96 | ,9583 | . | . | . | . | . | . | . |
| cv | 1 | ,00 | 12,00 | 12,00 | 12,00 | 12,0000 | . | . | . | . | . | . | . |
| theil | 1 | ,00 | 3,18 | 3,18 | 3,18 | 3,1781 | . | . | . | . | . | . | . |
| Gültige Werte (listenweise) | 1 | | | | | | | | | | | | |

PERSONELLE VERMÖGENSUNGLEICHHEIT, LAND C:

**Deskriptive Statistik**

**Deskriptive Statistiken**

| | N | Spannweite | Minimum | Maximum | Summe | Mittelwert | | Std.-Abweichung | Varianz | Schiefe | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistik | Statistik | Statistik | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Std.-Fehler |
| gini | 1 | ,00 | ,83 | ,83 | ,83 | ,8346 | . | . | . | . | . | . | . |
| cv | 1 | ,00 | 2,94 | 2,94 | 2,94 | 2,9412 | . | . | . | . | . | . | . |
| theil | 1 | ,00 | 1,67 | 1,67 | 1,67 | 1,6670 | . | . | . | . | . | . | . |
| Gültige Werte (listenweise) | 1 | | | | | | | | | | | | |

PERSONELLE VERMÖGENSUNGLEICHHEIT, LAND D:

**Deskriptive Statistik**

**Deskriptive Statistiken**

| | N | Spannweite | Minimum | Maximum | Summe | Mittelwert | | Std.-Abweichung | Varianz | Schiefe | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistik | Statistik | Statistik | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Std.-Fehler |
| gini | 1 | ,00 | ,87 | ,87 | ,87 | ,8686 | . | . | . | . | . | . | . |
| cv | 1 | ,00 | 4,70 | 4,70 | 4,70 | 4,7008 | . | . | . | . | . | . | . |
| theil | 1 | ,00 | 2,03 | 2,03 | 2,03 | 2,0288 | . | . | . | . | . | . | . |
| Gültige Werte (listenweise) | 1 | | | | | | | | | | | | |

**Abbildung 19.3:** (Personenbezogener) Ungleichheitsoutput (Gesamtvermögen) für die vier Länder A bis D (*VERMOEGEN_A.SAV*, *VERMOEGEN_B.SAV*, *VERMOEGEN_C.SAV*, *VERMOEGEN_D.SAV*)

Die Länder C und D liegen zwischen den beiden extremen Vermögensverteilungen in den Ländern A und B. Dabei ist die für Land D gemessene Ungleichheit des Vermögens (etwas) höher als in Land C, wie ebenfalls Tabelle 19.1 zum Ausdruck bringt.

| ***Land*** | ***Gini-Koeffizient*** | ***Normierter Variationskoeffizient*** | ***Theil'sches Entropiemaß*** |
|---|---|---|---|

| Land | Gini-Koeffizient | Normierter Variationskoeffizient | Theil'sches Entropiemaß |
|---|---|---|---|
| A | 0,0000 | 0,0000 | 0,0000 |
| B | 0,9583 | 12,0000 | 3,1781 |
| C | 0,8346 | 2,9412 | 1,6670 |
| D | 0,8686 | 4,7008 | 2,0288 |

**Tabelle 19.1:** (Personenbezogene) Ungleichheitsergebnisse (Gesamtvermögen) für die vier Länder A bis D (*VERMOEGEN_A.SAV*, *VERMOEGEN_B.SAV*, *VERMOEGEN_C.SAV*, *VERMOEGEN_D.SAV*)

# *Syntaxprogramm II*

Im nachfolgenden Syntaxprogramm wird für den (fiktiven) Datensatz des Landes D die Gesamt-, die Geld- und die Grundvermögensungleichheit berechnet. Aus Übersichtlichkeitsgründen beschränken sich die nachfolgenden Berechnungen auf den Ungleichheitsindikator des Gini-Koeffizienten. Außerdem wird im Folgenden zwischen einer gewichteten und einer ungewichteten Variante unterschieden. Bei der gewichteten Variante beziehen sich die Berechnungen auf die Personenfallzahl, bei der ungewichteten Variante auf die Haushaltsfallzahl.

Das nachstehende Syntaxprogramm findet sich in *UNGLEICHHEIT2.SPS*.

```
DEFINE !FILE( ).
GET FILE='C:\SPSS-DATEIEN\VERMOEGEN_D.SAV'.
!ENDDEFINE.


*Personenbezogene (= gewichtete) Betrachtung.

DEFINE !GINI( ).

WEIGHT BY groesse.

RANK VARIABLES = vermoegen_pk (a)
/TIES=MEAN
/RANK INTO rvermoeg.
```

```
IF (vermoegen_pk > 0) ginisumme=rvermoeg * vermoegen_pk.
IF (vermoegen_pk <= 0) ginisumme=0.

AGGREGATE
    /OUTFILE=*
    /BREAK=
    /ginisumme2=SUM(ginisumme)
    /fallzahl=N(vermoegen_pk)
    /vermoegen_mean=MEAN(vermoegen_pk).

COMPUTE gini = ((2 * ginisumme2) - ((fallzahl+1) * fallzahl *
vermoegen_mean))/(Fallzahl**2 * vermoegen_mean).

DESCRIPTIVES VARIABLES=gini
/STATISTICS ALL.

!ENDDEFINE.

TITLE 'Personenbezogene Betrachtungen'.
!FILE.
TITLE 'Gini-Koeffizient, Gesamtvermögen'.
COMPUTE vermoegen_pk=0.
COMPUTE vermoegen_pk = (geldverm + grundverm)/groesse.
!GINI.

!FILE.
TITLE 'Gini-Koeffizient, Geldvermögen'.
COMPUTE vermoegen_pk=0.
COMPUTE vermoegen_pk = geldverm/groesse.
!GINI.

!FILE.
TITLE 'Gini-Koeffizient, Grundvermögen'.
COMPUTE vermoegen_pk=0.
COMPUTE vermoegen_pk = grundverm/groesse.
!GINI.


*Haushaltsbezogene (= ungewichtete) Betrachtung.

DEFINE !GINI2( ).
WEIGHT OFF.

RANK VARIABLES = vermoegen_pk (a)
/TIES=MEAN
/RANK INTO rvermoeg.

IF (vermoegen_pk > 0) ginisumme=rvermoeg * vermoegen_pk.
```

```
IF (vermoegen_pk <= 0) ginisumme=0.

AGGREGATE
    /OUTFILE=*
    /BREAK=
    /ginisumme2=SUM(ginisumme)
    /fallzahl=N(vermoegen_pk)
    /vermoegen_mean=MEAN(vermoegen_pk).

COMPUTE gini = ((2 * ginisumme2) - ((fallzahl+1) * fallzahl *
vermoegen_mean))/(fallzahl**2 * vermoegen_mean).

DESCRIPTIVES VARIABLES =gini
/STATISTICS ALL.

!ENDDEFINE.

TITLE 'Haushaltsbezogene Betrachtungen'.
!FILE.
TITLE 'Gini-Koeffizient, Gesamtvermögen'.
COMPUTE vermoegen_pk=0.
COMPUTE vermoegen_pk = (geldverm + grundverm)/groesse.
!GINI2.

!FILE.
TITLE 'Gini-Koeffizient, Geldvermögen'.
COMPUTE vermoegen_pk=0.
COMPUTE vermoegen_pk = geldverm/groesse.
!GINI2.

!FILE.
TITLE 'Gini-Koeffizient, Grundvermögen'.
COMPUTE vermoegen_pk=0.
COMPUTE vermoegen_pk = grundverm/groesse.
!GINI2.
```

Der betreffende Output ist in Abbildung 19.4 (auszugsweise) dargestellt.

Tabelle 19.2 gibt eine Übersicht über die berechneten Ungleichheitswerte. Im Beispiel ist die Geldvermögensungleichheit jeweils geringer als die Grundvermögensungleichheit. Des Weiteren sind die haushaltsbezogenen Gini-Koeffizienten bei allen Vermögensarten kleiner als die personenbezogenen.

## *Syntaxprogramm III*

Eine weitere Abwandlung der vorstehenden Syntax findet sich nachfolgend dahin gehend, dass die (fiktive) Vermögensungleichheit im Land D für drei Regionen 1 bis 3 in diesem Land ausgewiesen wird.

Sie brauchen die nachfolgende Syntax nicht abzutippen: Sie finden sie in *UNGLEICHHEIT3.SPS*.

```
DEFINE !FILE( ).
GET FILE='C:\SPSS-DATEIEN\VERMOEGEN_D.SAV'.
!ENDDEFINE.

DEFINE !UNGLEICH3( ).
```

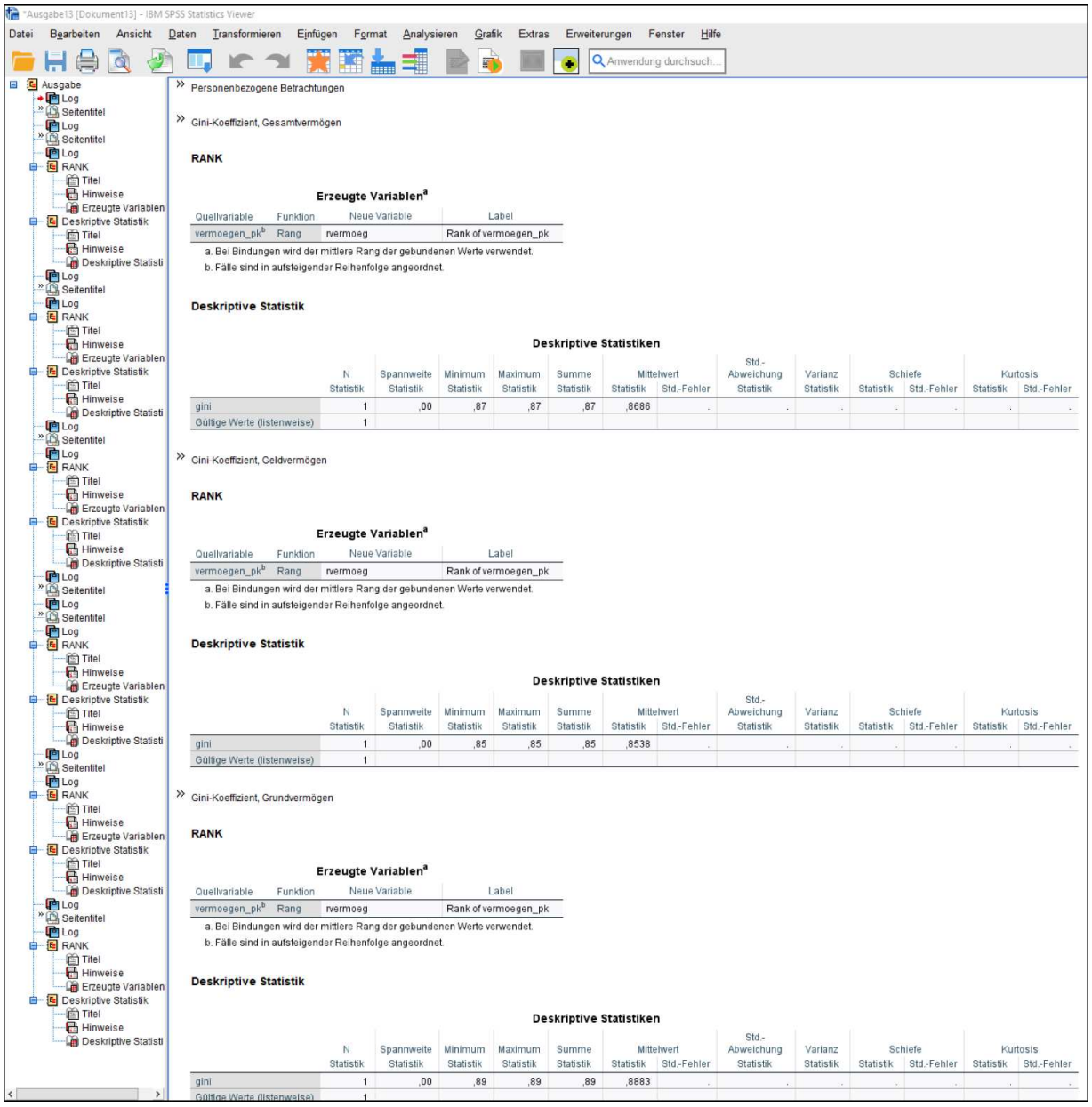



**Abbildung 19.4:** (Personen- und haushaltsbezogener) Ungleichheitsoutput (Gesamt-, Geld- und Grundvermögen) für Land D (*VERMOEGEN_D.SAV*) (Auszug)

| ***Haushaltsvermögensart (jeweils als Pro-Kopf-Vermögen)*** | ***Personenbezogener Gini-Koeffizient*** | ***Haushaltsbezogener Gini-Koeffizient*** |
|---|---|---|
| Gesamtvermögen | 0,8686 | 0,8457 |
| Geldvermögen | 0,8538 | 0,8256 |
| Grundvermögen | 0,8883 | 0,8750 |

**Tabelle 19.2:** Personen- und haushaltsbezogene Gini-Koeffizienten für die Ungleichheit des Gesamt-, des Geld- und des Grundvermögens in Land D (*VERMOEGEN_D.SAV*)

```
WEIGHT BY groesse.

COMPUTE vermoegen_pk = (geldverm + grundverm)/groesse.

*Rangbildung für Gini-Koeffizienten-Formel:
RANK VARIABLES=vermoegen_pk (a)
    /RANK
    /PRINT=NO
    /TIES=MEAN.

IF (vermoegen_pk > 0) ginisumme=rvermoeg * vermoegen_pk.
IF (vermoegen_pk <= 0) ginisumme=0.

IF (vermoegen_pk > 0) theilsumme=vermoegen_pk*ln(vermoegen_pk).
IF (vermoegen_pk<=0) theilsumme=0.

AGGREGATE
    /OUTFILE=*
    /BREAK=
    /ginisumme2=SUM(ginisumme)
    /theilsumme2=SUM(theilsumme)
    /fallzahl=N(vermoegen_pk)
    /vermoegen_sum=SUM(vermoegen_pk)
    /vermoegen_mean=MEAN(vermoegen_pk)
    /vermoegen_sd=SD(vermoegen_pk).

COMPUTE gini = ((2/(fallzahl**2*vermoegen_mean))*ginisumme2) - (1+(1/fallzahl)).
COMPUTE cv = ((vermoegen_sd)**2)/(2*vermoegen_mean**2).
COMPUTE theil = (1/(fallzahl*vermoegen_mean))*
(theilsumme2-(ln(vermoegen_mean)*vermoegen_sum)).

DESCRIPTIVES VARIABLES=gini cv theil
/STATISTICS ALL.

!ENDDEFINE.

!FILE.
SELECT IF (region=1).
TITLE 'GESAMTVERMÖGEN, REGION 1:'.
!UNGLEICH3.

!FILE.
SELECT IF (region=2).
TITLE 'GESAMTVERMÖGEN, REGION 2:'.
```

```
!UNGLEICH3.

!FILE.
SELECT IF (region=3).
TITLE 'GESAMTVERMÖGEN, REGION 3:'.
!UNGLEICH3.
```

Der betreffende Output ist aus Abbildung 19.5 ersichtlich.

GESAMTVERMÖGEN, REGION 1:

**Deskriptive Statistik**

**Deskriptive Statistiken**

| | N | Spannweite | Minimum | Maximum | Summe | Mittelwert | | Std.-Abweichung | Varianz | Schiefe | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistik | Statistik | Statistik | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Std.-Fehler |
| gini | 1 | ,00 | ,47 | ,47 | ,47 | ,4706 | . | . | . | . | . | . | . |
| cv | 1 | ,00 | ,47 | ,47 | ,47 | ,4669 | . | . | . | . | . | . | . |
| theil | 1 | ,00 | ,58 | ,58 | ,58 | ,5752 | . | . | . | . | . | . | . |
| Gültige Werte (listenweise) | 1 | | | | | | | | | | | | |

GESAMTVERMÖGEN, REGION 2:

**Deskriptive Statistik**

**Deskriptive Statistiken**

| | N | Spannweite | Minimum | Maximum | Summe | Mittelwert | | Std.-Abweichung | Varianz | Schiefe | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistik | Statistik | Statistik | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Std.-Fehler |
| gini | 1 | ,00 | ,67 | ,67 | ,67 | ,6717 | . | . | . | . | . | . | . |
| cv | 1 | ,00 | 1,29 | 1,29 | 1,29 | 1,2878 | . | . | . | . | . | . | . |
| theil | 1 | ,00 | 1,04 | 1,04 | 1,04 | 1,0358 | . | . | . | . | . | . | . |
| Gültige Werte (listenweise) | 1 | | | | | | | | | | | | |

GESAMTVERMÖGEN, REGION 3:

**Deskriptive Statistik**

**Deskriptive Statistiken**

| | N | Spannweite | Minimum | Maximum | Summe | Mittelwert | | Std.-Abweichung | Varianz | Schiefe | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistik | Statistik | Statistik | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Statistik | Statistik | Std.-Fehler | Statistik | Std.-Fehler |
| gini | 1 | ,00 | ,68 | ,68 | ,68 | ,6789 | . | . | . | . | . | . | . |
| cv | 1 | ,00 | 1,10 | 1,10 | 1,10 | 1,0991 | . | . | . | . | . | . | . |
| theil | 1 | ,00 | 1,00 | 1,00 | 1,00 | 1,0044 | . | . | . | . | . | . | . |
| Gültige Werte (listenweise) | 1 | | | | | | | | | | | | |

**Abbildung 19.5:** (Personenbezogener) Ungleichheitsoutput (Gesamtvermögen) für drei Regionen in Land D (*VERMOEGEN_D.SAV*) (Auszug)

Tabelle 19.3 bietet überdies einen schnellen Überblick über diese (fiktiven) regionalen Ungleichheitsberechnungen. Offensichtlich ist in diesem Beispiel die gemessene Ungleichheit in Region 1 am niedrigsten. Ob man die gemessene Ungleichheit in Region 2 höher oder niedriger als in Region 3 angibt, hängt vom verwendeten Ungleichheitsindikator ab: Gemäß dem Gini-Koeffizienten ist die Ungleichheit in Region 2 etwas niedriger als in Region 3, gemäß dem normierten

Variationskoeffizienten und dem Theil'schen Entropiemaß hingegen ist die gemessene Ungleichheit in Region 2 höher als in Region 3.

| *Region* | *Gini-Koeffizient* | *Normierter Variationskoeffizient* | *Theil'sches Entropiemaß* |
|---|---|---|---|
| 1 | 0,4706 | 0,4669 | 0,5752 |
| 2 | 0,6717 | 1,2878 | 1,0358 |
| 3 | 0,6789 | 1,0991 | 1,0044 |

**Tabelle 19.3:** (Personenbezogene) Ungleichheitsergebnisse (Gesamtvermögen) für die drei Regionen in Land D (*VERMOEGEN_D.SAV*)

# *Armut und Ungleichheit gleichzeitig*

Die vorstehenden Syntaxprogramme sind in der folgenden Syntax zusammengefasst. Dabei werden die vier oben verwendeten Syntaxprogramme nacheinander mittels INSERT-Befehl aufgerufen (siehe auch das Syntaxfile *ARMUT_UND_UNGLEICHHEIT.SPS*).

```
INSERT FILE = 'C:\SPSS-DATEIEN\ARMUT.SPS'.
INSERT FILE = 'C:\SPSS-DATEIEN\UNGLEICHHEIT1.SPS'.
INSERT FILE = 'C:\SPSS-DATEIEN\UNGLEICHHEIT2.SPS'.
INSERT FILE = 'C:\SPSS-DATEIEN\UNGLEICHHEIT3.SPS'.
EXECUTE.
```

Probieren Sie das vorstehende Syntaxfile doch einfach einmal aus! Und damit kommen wir zum Schluss des Buchs, und zwar zu Teil VI, das den für *für Dummies*-Bücher typischen Top-Ten-Teil enthält.

# Teil VI

# Top-10-Teil

Auf www.fuer-dummies.de finden Sie noch mehr Bücher für Dummies!

# IN DIESEM TEIL …

- ✔ Die 10 wichtigsten SPSS-Befehle
- ✔ Die 10 wichtigsten SPSS-Tipps

# Kapitel 20

# Die 10 wichtigsten SPSS-Befehle

## IN DIESEM KAPITEL

Hinweise auf die Vorteilhaftigkeit von Syntaxprogrammen

Darlegung der meines Erachtens besonders wichtigen Befehle

Syntaxprogramme ermöglichen meiner festen Überzeugung nach eine schnellere, effizientere Verarbeitung statistischer Prozeduren in SPSS. Allerdings muss man hierzu bis zu einem gewissen Grad die SPSS-Syntax beherrschen. Aus diesem Grund habe ich meine zehn wichtigsten SPSS-Befehle nachfolgend zusammengestellt, und zwar in alphabetischer Anordnung.

## *AGGREGATE*

Möchten Sie Fälle zu Gruppen zusammenfassen, bietet sich der AGGREGATE-Befehl an. Zum Beispiel könnte es Ihr Anliegen sein, die individuellen Einkommen der einzelnen Haushaltsmitglieder jeweils zum Haushaltseinkommen zusammenzufassen, in diesem Fall also von der Personen- zur Haushaltsebene zu aggregieren. Hierzu müssen Sie eine »Schlüsselvariable« im Rahmen des AGGREGATE-Befehls angeben. Dies geschieht über den BREAK-Unterbefehl. Im eben genannten Beispiel würde man vermutlich die Haushaltsnummer im Datensatz als BREAK-Variable verwenden. Geben Sie nach dem BREAK-Unterbefehl nichts ein, aggregiert SPSS alle Fälle gewissermaßen zu einem Fall. Damit würden im Beispielfall die individuellen Einkommen aller Personen im Datensatz zu einem Gesamteinkommen für den gesamten Datensatz addiert.

# *COMPUTE*

Die Bildung neuer Variablen via COMPUTE-Befehl ist in SPSS-Syntaxprogrammen das Salz in der Suppe. Beachten Sie aber bitte, dass Sie bei Ihren COMPUTE-Anweisungen einen Variablennamen für zwei oder mehr unterschiedliche Sachverhalte nicht versehentlich mehrfach vergeben. SPSS ist dann nämlich rigoros und überschreibt die vorhergehenden Variableninhalte einfach!

# *DESCRIPTIVES*

Möchten Sie sich mal schnell ein paar Statistiken wie Mittelwerte oder Streuungen ausgeben lassen, empfehle ich die DESCRIPTIVES-Anweisung. Sie müssen nur die gewünschten Variablen hinter dem DESCRIPTIVES-Befehl und die gewünschten Statistiken angeben. Am einfachsten ist es, wenn Sie nach dem STATISTICS-Unterbefehl ALL auswählen; dann »wirft« Ihnen SPSS eine ganze Latte an Statistiken für Ihre ausgewählten Variablen »aus«. Beachten Sie dabei aber unbedingt, dass manche Statistiken, wie zum Beispiel der arithmetische Mittelwert, erst ab einem bestimmten Skalenniveau (hier: ab metrischem Skalenniveau) berechnet und interpretiert werden dürfen. Möchten Sie sich einen Überblick über Ihren kompletten Datensatz und damit auch über alle Variablen verschaffen, können Sie nach dem DESCRIPTIVES-Befehl nach VARIABLES einfach ALL eingeben und damit alle Variablen des betreffenden Datensatzes auswählen. Um aber vor lauter Bäumen am Ende den Wald zu erkennen, sollten Sie in diesem Fall aus Übersichtlichkeitsgründen anstelle von ALL nur einige wenige statistische Kennziffern nach dem STATISTICS-Unterbefehl auswählen, wie zum Beispiel MIN (niedrigster Wert), MAX (höchster Wert), MEAN (arithmetisches Mittel) und STDDEV (Standardabweichung).

# *DO … IF*

Schleifen erleichtern in SPSS die Abarbeitung vergleichsweise »monotoner« Befehlsabfolgen. Die »Königs-Schleifenkonstruktion« ist

dabei die DO-IF-Schleife. Mit ihrer Hilfe können Sie unterschiedliche Bedingungen für die von Ihnen ausgewählten Variablen formulieren. Sie sollten im Regelfall alle möglichen Variablenausprägungen in die DO-IF-Schleife einbinden und dabei die »Restkategorie« mit einer gewissermaßen die Bedingungen abschließenden ELSE-Anweisung innerhalb der DO-IF-Schleife.

# *EXECUTE*

SPSS verhält sich dann etwas »bockig«, wenn nach Variablentransformationen keine statistische Prozedur folgt und Sie keinen EXECUTE-Befehl nach diesen Transformationen gesetzt haben. Dann führt SPSS diese Transformationen nämlich nicht aus, und bei den betreffenden Variablen erscheinen in Ihrem Datensatz lauter leere Felder. Setzen Sie also bitte in solchen Fällen immer den EXECUTE-Befehl. Folgt dann doch eine statistische Prozedur, ist er nicht schädlich. Es gilt also, dass Sie den EXECUTE-Befehl lieber einmal zu viel als zu wenig setzen sollten.

# *FILTER BY*

Eine Möglichkeit, Ihre Daten gefahrlos zu filtern, bietet in SPSS der FILTER-BY-Befehl. »Gefahrlos« meint hier, dass – im Unterschied zum SELECT-IF-Befehl – Ihre nicht ausgewählten Fälle nicht gelöscht, sondern lediglich markiert werden. SPSS kreiert beim FILTER-BY-Befehl eine neue 0/1-Variable *$filter*, die nach dem FILTER-BY-Befehl anzugeben ist. Die ausgewählten Fälle bekommen hierbei den Wert 1, die nicht ausgewählten Fälle den Wert 0 zugewiesen. Sie können jetzt getrennt für die beiden Gruppen der ausgewählten und der nicht ausgewählten Fälle Berechnungen durchführen und auf diese Art und Weise gruppenspezifische Unterschiede bei bestimmten Variablen entdecken. Über die FILTER-OFF-Anweisung schalten Sie den Filter wieder aus.

# *FREQUENCIES*

Mithilfe des FREQUENCIES-Befehls sind Häufigkeitsdarstellungen für ausgewählte Variablen möglich. Sinnvoll ist der Ausweis von Häufigkeitsverteilungen via FREQUENCIES-Befehl nur bei diskreten oder bei kategorialen beziehungsweise klassifizierten Variablen, und auch dann nur bei relativ wenigen Variablenausprägungen beziehungsweise bei relativ wenigen Klassen. Ansonsten wird die entsprechende Häufigkeitstabelle schnell sehr unübersichtlich! Zusätzlich zu den Häufigkeitstabellen können Sie beim FREQUENCIES-Befehl auch grafische Darstellungen anfordern: BARCHART = Balkendiagramm, PIECHART = Kreisdiagramm und HISTOGRAM = Histogramm. Mit der FREQUENCIES-Anweisung können Sie sich über den STATISTICS-Unterbefehl auch eine Reihe deskriptiver statistischer Kennziffern ausgeben lassen. Dabei sind hier mit dem genannten Unterbefehl auch statistische Kennziffern bereits ab Nominalskalenniveau (zum Beispiel MODE = Modus) beziehungsweise ab Ordinalskalenniveau (etwa MEDIAN = Median oder über den PERCENTILES-Unterbefehl Perzentile) erhältlich. Das Anzeigen einer Häufigkeitstabelle unterdrücken Sie beim FREQUENCIES-Befehl durch den Subbefehl NOTAB.

# *RECODE*

Manchmal möchte man die Inhalte diskreter oder kategorialer beziehungsweise klassifizierter Variablen verändern. Dies kann über den RECODE-Befehl bewerkstelligt werden. Sie können Recodierungen auch von einer alten in eine neue Variable vornehmen. Hierbei müssen Sie aber alle Ausprägungen der alten Variablen unbedingt berücksichtigen. Tun Sie dies nämlich nicht, würde die neue Variable beim ausgelassenen Wert der alten Variablen in der neuen Variablen ein Missing (also einen fehlenden Wert) enthalten, die Umcodierung würde also unvollständig verlaufen.

# *TEMPORARY*

Mittels des TEMPORARY-Befehls werden Variableninhalte oder die Auswahl von Fällen, wie der Befehlsname andeutet, nur kurzzeitig verändert, das heißt: nur bis zum Abschluss einer statistischen Prozedur. Die TEMPORARY-Anweisung spielt insbesondere im Zusammenhang mit dem SELECT-IF-Befehl eine große Rolle. Setzen Sie beim SELECT-IF-Befehl keine ergänzende TEMPORARY-Anweisung, werden nämlich die nicht ausgewählten Fälle dauerhaft gelöscht, was dann besonders ärgerlich ist, wenn Sie die betreffende Datendatei anschließend auch noch unter ihrem ursprünglichen Namen abspeichern und keine Sicherungskopie von der ursprünglichen Datendatei haben. Setzen Sie hingegen im Zusammenhang mit dem SELECT-IF-Befehl die TEMPORARY-Anweisung, werden nur bis zum Abschluss der entsprechenden SPSS-Prozedur die nicht ausgewählten Fälle aus den Berechnungen herausgenommen, stehen danach aber wieder für weitere Berechnungen zur Verfügung.

# *WEIGHT BY*

Möchten Sie etwa Stichprobenergebnisse möglichst repräsentativ auf die Grundgesamtheit hochrechnen, verwenden Sie am besten eine entsprechende Gewichtungsvariable, die die einzelnen Hochrechnungswerte enthält. Diese Hochrechnung Ihres Datensatzes via Gewichtungsvariable setzen Sie über den vor dieser Variablen anzugebenden WEIGHT-BY-Befehl in Gang. Über eine einfache WEIGHT-OFF-Anweisung können Sie diese Gewichtung jederzeit wieder ausschalten. Die Verwendung des WEIGHT-BY-Befehls bietet sich auch dann an, wenn Ihr Datensatz als Häufigkeitstabelle vorliegt. Dann multiplizieren (= gewichten) Sie die Variablenausprägungen mit den im Datensatz angegebenen (absoluten) Häufigkeiten. Sie müssen in diesem Fall nach WEIGHT BY lediglich die Variable angeben, die die Häufigkeitswerte beinhaltet.

# Kapitel 21

# Die 10 wichtigsten SPSS-Tipps

## IN DIESEM KAPITEL

Tipps zur Erleichterung der Arbeit mit SPSS

Tipps zum Vermeiden sinnfreier statistischer Aktionen in SPSS

Dieses letzte Kapitel soll Ihnen sozusagen noch einmal meine »letzten Weisheiten« für den Umgang mit SPSS an die Hand geben. Ich möchte dabei aber nicht besserwisserisch rüberkommen. Es handelt sich bei den folgenden Tipps um Erfahrungen, die ich im Rahmen meiner mehr oder weniger täglichen Arbeit mit SPSS gemacht habe. Die entsprechenden Tipps sollen Ihnen das Arbeiten mit SPSS erleichtern und vor allem unvorteilhafte statistische Aktionen zu vermeiden helfen. Möglicherweise haben Sie aber ganz andere Erfahrungen im Umgang mit SPSS gemacht – dann scheuen Sie sich bitte nicht, mir diese mitzuteilen! Ich bin neugierig …

## *Nutze auch mal das PDF-Benutzerhandbuch!*

Ich beginne gleich mal mit einem aus meiner Sicht sehr wichtigen Hinweis. Da bekanntlich niemand perfekt ist, kann man nicht alle Befehle und Menüpunkte in SPSS hinreichend kennen. Ab und an muss wohl jeder von uns diesbezüglich Informationen einholen. Das geht meiner Meinung nach – natürlich neben dem Blick in ein SPSS-Buch ;-) – recht gut über das umfangreiche PDF-Benutzerhandbuch von SPSS. Da finden sich alle relevanten SPSS-Befehle und noch dazu in strukturierter Form.

# *Nutze Copy-and-Paste beim Einlesen von Daten!*

Sie können zwar über den Import-Menüpunkt auch relativ einfach außerhalb von SPSS erzeugte Daten (etwa DOC(X)- oder XLS(X)-Daten) in SPSS einlesen. Bei vergleichsweise kleinen Datensätzen geht es aber wesentlich schneller, wenn Sie die Daten via Copy-and-Paste nach SPSS »transferieren«. Es gibt in diesem Zusammenhang indes eine wesentliche Ausnahme: Sind ASCII-Daten durch Semikolon voneinander getrennt, werden alle Daten pro Fall in eine einzige Variable geschrieben – in diesem (Ausnahme-) Fall also bitte auf Copy-and-Paste verzichten! Außerdem müssen Sie beim Copy-and-Paste unbedingt darauf achten, dass die Dezimalstellendarstellung Ihrer entsprechenden numerischen Variablen im Ausgangsdatenformat Ihrer Dezimalstellendarstellung in SPSS entspricht. Sind Ihre betreffenden Variablen zum Beispiel in EXCEL durch Kommata getrennt und sind bei Ihnen in SPSS Dezimalstellen durch einen Punkt getrennt, haben Sie ein Problem. Dieses ist aber leicht zu lösen, und zwar müssen Sie in diesem Fall in EXCEL einfach die Kommata durch Punkte ersetzen. Dann können Sie Ihr Copy-and-Paste betreiben.

# *Beachte grundsätzlich die gleiche Sortierung (und natürlich auch gleiche Variableninhalte) beim Daten-»Matchen«!*

Zwei Datensätze können Sie in SPSS bekanntlich entweder fall- oder variablenweise zu einem einzigen Datensatz zusammenfügen. Hierbei sollten Sie beim Zusammenfügen nach gleichen Variablen beide Datensätze gleich sortieren, also beide etwa mit Fallnummer 1 aufsteigend beginnen lassen. Ansonsten kann das Problem auftreten,

dass SPSS die Ausprägungen der neuen Variablen an die falschen Fälle anhängt. Wenn Sie zwei Datensätze fallweise zusammenfügen, müssen natürlich die gleichlautenden Variablen in beiden Datensätzen jeweils die gleichen Inhalte haben. Andernfalls haben Sie an dieser Stelle »Chaos« in Ihrem zusammengefügten neuen Datensatz.

## *Erstelle »Spiel-Datenfiles« für Syntaxprogramme!*

Gerade bei sehr großen Datensätzen und sehr umfangreichen SPSS-Programmen kann es sehr ärgerlich sein, wenn nach etlicher Zeit SPSS einen Fehler meldet beziehungsweise sogar die Programmausführung abbricht – vielleicht nur deshalb, weil Sie sich bei einer Variablenbezeichnung verschrieben oder keinen Punkt am Ende eines SPSS-Befehls gesetzt haben. Diesen Ärger können Sie vermeiden, wenn Sie für Ihr SPSS-Syntaxprogramm zum Austesten ein kleines »Spiel-Datenfile« erstellen. In dieses »Spiel-Datenfile« schreiben Sie einfach nur die für Ihre Analyse relevanten Variablen mit einigen wenigen Fällen (und auch nur mit einigen wenigen Variablenausprägungen). Wenn dann SPSS einen Fehler meldet, ist das nicht so problematisch, weil Ihr Programm wegen des kleinen »Spiel-Datenfiles« nur ein paar Sekunden »gelaufen« ist.

## *Verwende »sprechende« Variablen- und Dateibezeichnungen!*

Wenn Sie ein umfangreiches Statistik-Projekt via SPSS mit sehr vielen Variablen und sehr vielen Datensätzen verfolgen (oder sogar mehrere Statistik-Projekte gleichzeitig), können Sie relativ schnell den Überblick verlieren. Um dieser Gefahr zu begegnen, empfiehlt es sich, »sprechende« Variablen- und Dateibezeichnungen zu verwenden, die es Ihnen leicht machen, die betreffenden Variablen und Dateien den jeweiligen Inhalten Ihrer Analyse(n) zuzuordnen. Um den Überblick

über Ihre zahlreichen Aktivitäten zu behalten, ist es auch hilfreich, die Variableninhalte ausreichend zu dokumentieren. Dies können Sie gut über den VARIABLE-LABELS-Befehl in SPSS direkt tun, indem Sie die Variablen namentlich genauer kennzeichnen; zusätzlich sollten Sie über den VALUE-LABELS-Befehl auch – zumindest bei kategorialen beziehungsweise klassifizierten Variablen – die einzelnen Variablenausprägungen direkt in SPSS festhalten. Alternativ – oder noch besser: zusätzlich – sollten Sie die Inhalte Ihrer Datensätze beziehungsweise der darin enthaltenen Variablen in einem externen Dokument, also etwa in einem WORD- oder PDF-Dokument, ablegen (im Sinne eines »Datenhandbuchs«).

## *Kommentiere ausreichend viel in Syntaxprogrammen!*

Ähnlich wie die vorstehend genannten Variablen- beziehungsweise Dateibezeichnungen der Strukturierung und Nachvollziehbarkeit Ihrer Arbeit dienen, ist auf jeden Fall zu empfehlen, innerhalb Ihrer SPSS-Syntaxprogramme die einzelnen Arbeitsschritte mittels Kommentaren zu dokumentieren. Dadurch können andere Personen, mit denen Sie zusammenarbeiten, Ihre Arbeit leichter nachvollziehen. Auch Ihnen sind diese Kommentare gegebenenfalls eine wichtige Erinnerungsstütze, wenn Sie etwa längere Zeit nicht mit den betreffenden Arbeiten befasst waren.

## *Verändere Pivottabellen und Grafiken vor ihrer Präsentation!*

Die im SPSS-Viewer ausgegebenen Ergebnistabellen – die sogenannten Pivottabellen – kommen ebenso wie durch SPSS erzeugte Grafiken in der Regel zunächst einmal recht schnöde daher. Bei beruflichen Präsentationen Ihrer (SPSS-)Arbeit möchten Sie indes vielleicht durch schönere, professionellere Grafiken punkten. SPSS stellt Ihnen an dieser

Stelle eine Reihe von Überarbeitungsmöglichkeiten bezüglich Ihrer Tabellen und Grafiken zur Verfügung, die Sie nutzen können. Ich persönlich finde es aber leichter, die SPSS-Tabellen und -Grafiken mittels Copy-and-Paste nach WORD und/oder EXCEL zu »bringen« und dort zu bearbeiten, das heißt: zu formatieren. Sie müssen aber letztlich für sich selbst entscheiden, welche Bearbeitungsmöglichkeit für Sie besser geeignet ist.

## *Achte auf fehlende Werte!*

Fehlende Werte können Ihre Berechnungen in SPSS mächtig durcheinanderwirbeln; deshalb ist hier besondere Vorsicht geboten. Verändern Sie – beim RECODE-Befehl – die Kategorien einer alten Variablen zu neuen Werten einer neuen Variablen, dürfen Sie keinesfalls einen Wert der alten Variablen auslassen. Ansonsten haben Sie nämlich in der neuen Variablen einen fehlenden Wert stehen, was Sie ja vermutlich so nicht haben möchten. Ein anderes Problem mit fehlenden Werten ergibt sich beim COUNT-Befehl, wenn Sie fallweise aus mehreren 0/1-Variablen eine Auszählung vornehmen möchten, und zwar dahin gehend, wie oft die Ausprägung »1« vorkommt. Fehlt bei den betreffenden Variablen bei einem Fall ein Eintrag, so setzt SPSS im Rahmen der entsprechenden Summenbildung bei diesem Fall das Ergebnis auf »0« und nicht richtigerweise auf »Fehlend«. Dieses unerwünschte Ergebnis von »0« können Sie vermeiden, wenn Sie den NMISS- dem COUNT-Befehl vorschalten.

## *Verwende Schleifen und Makros!*

Das Verwenden von Schleifen und Makros »entschlackt« Ihre Programme dadurch, dass wiederkehrende Befehle vom Grundsatz her nur einmal – über eine Schleifenkonstruktion oder mittels eines Makros – in Ihr SPSS-Syntaxprogramm geschrieben werden müssen. Dies dient der Übersichtlichkeit Ihrer Programme. Außerdem wird die Ausführung Ihrer Programme hierdurch beschleunigt (auch wenn das vielleicht in Zeiten extrem schneller Rechner kein sehr starkes Argument mehr ist).

In Bezug auf Makros (und andere SPSS-Unterprogramme) ist im Übrigen der INSERT-Befehl hilfreich. Dank ihm wird Ihr Programm noch übersichtlicher, denn anstelle mehrzeiligen SPSS-Codes steht dann in Ihrem Programm nach dem INSERT-Befehl nur noch der Makro- beziehungsweise Unterprogrammname. Bei der Ausführung Ihrer SPSS-Syntax im Hauptprogramm werden hierbei die Befehle der genannten Makros beziehungsweise Unterprogramme aufgerufen.

# *Sei dir bewusst, dass kein »Unsinn« berechnet wird! Gebrauche dein Hirn!*

Zum Schluss möchte ich Ihnen die eigentlich wichtigste SPSS-Grundregel als Tipp mit auf Ihren weiteren (SPSS-)Weg geben: Sie sollten SPSS nicht gedankenlos gebrauchen. Sie sollten schon wissen, *was* genau sie *wie* berechnen. Hierzu benötigen Sie gewisse statistische Methodenkenntnisse. Auch sollten Sie vor der Nutzung von SPSS einen analytischen »Schlachtplan« entwickeln. Sind Sie sich diesbezüglich unsicher, sollten Sie einen Experten befragen oder in ein Statistik-Buch schauen. Vielleicht nutzen Sie ja in diesem Zusammenhang dasselbe Buch wie ich? Wie heißt dieses Buch noch gleich? Ach ja: Es ist ebenfalls im Wiley-Verlag von einem gewissen Jürgen Faik im Jahr 2015 erschienen und trägt den Titel »Statistik für Wirtschafts- und Sozialwissenschaftler«. (Sie können aber natürlich auch gerne ein anderes Statistik-Buch nutzen …)

# Abbildungsverzeichnis


Abbildung 1.1: Mitteilung nach erfolgreicher Lizenzierung (Auszug; hier bezogen auf SPSS 24)

Abbildung 1.2: Der Start-Bildschirm von SPSS 29

Abbildung 1.3: Das SPSS-Menü

Abbildung 1.4: Datei in SPSS öffnen

Abbildung 1.5: OPTIONEN bei SPSS

Abbildung 2.1: Beispiel für eine rechteckige Datentabelle (*123.SAV*)

Abbildung 2.2: Verschiedene Datentypen (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley-VCH, S. 33)

Abbildung 2.3: Zum Problem fehlender Werte bei mathematischen Variablenoperationen (*EINKOMMEN_VERMOEGEN.SAV*)

Abbildung 2.4: Manuelle Werteeingabe im Dateneditor

Abbildung 2.5: Variablenumbenennung im Variableneditor

Abbildung 2.6: Kontextmenü für FÄLLE EINFÜGEN

Abbildung 2.7: Eingefügter Fall

Abbildung 2.8: Kontextmenü für das Löschen von Fällen

Abbildung 2.9: Datendatei nach Löschen eines Falls

Abbildung 2.10: Abspeichern einer manuell erstellten Datendatei als SAV-File (*FALL.SAV*)

Abbildung 2.11: In SPSS einlesbare Datenfile-Formate

Abbildung 2.12: Semikolon-Trennung bei ASCII-Daten (DATEN_ALLE_SEM.TXT)

Abbildung 2.13: Tabulator-Trennung bei ASCII-Daten (DATEN_ALLE_TABS.TXT)

Abbildung 2.14: Öffnen einer ASCII-Datei (mit Trennzeichen) in SPSS (DATEN_ALLE_TABS.TXT)

Abbildung 2.15: Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 1

Abbildung 2.16: Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 2

Abbildung 2.17: Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 3

Abbildung 2.18: Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 3 mit Stichprobenziehung

Abbildung 2.19: Eingelesene Stichprobe (1)

Abbildung 2.20: Eingelesene Stichprobe (2)

Abbildung 2.21: Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 4

Abbildung 2.22: Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 5

Abbildung 2.23: Textassistent zum Einlesen von ASCII-Daten mit Trennzeichen – Schritt 6

Abbildung 2.24: In SPSS mittels des Textassistenten eingelesene ASCII-Daten mit Trennzeichen

Abbildung 2.25: In SPSS mittels Copy-and-Paste eingelesene ASCII-Daten (Tabulator-Trennung)

Abbildung 2.26: In SPSS mittels Copy-and-Paste eingelesene ASCII-Daten (Semikolon-Trennung)

Abbildung 2.27: ASCII-Daten mit fester Satzlänge (DATEN_ALLE_NICHTS.TXT)

Abbildung 2.28: SPSS-Warnung beim Versuch des Einlesens von ASCII-Daten mit fester Satzlänge (Unicode-Format) mittels des Textassistenten

Abbildung 2.29: Lokale Codierung von ASCII-Daten

Abbildung 2.30: Textassistent zum Einlesen von ASCII-Daten ohne Trennzeichen, mit fester Satzlänge – Schritt 1

Abbildung 2.31: Textassistent zum Einlesen von ASCII-Daten ohne Trennzeichen, mit fester Satzlänge – Schritt 2

Abbildung 2.32: Textassistent zum Einlesen von ASCII-Daten ohne Trennzeichen, mit fester Satzlänge – Schritt 4

Abbildung 2.33: Textassistent zum Einlesen von ASCII-Daten ohne Trennzeichen, mit fester Satzlänge – Schritt 5

Abbildung 2.34: In SPSS mittels Textassistenten eingelesene ASCII-Daten ohne Trennzeichen und mit fester Satzlänge

Abbildung 2.35: EXCEL-Beispieldaten (DATEN_ALLE.XLSX)

Abbildung 2.36: Bildschirm zum Einlesen von EXCEL-Daten in SPSS (DATEN_ALLE.XLSX)

Abbildung 2.37: EXCEL-Datei mit Fällen in den Spalten und Variablen in den Zeilen (DATEN_ALLE_INVERTIERT.XLSX)

Abbildung 2.38: SPSS-Anzeige für EXCEL-Datei DATEN_ALLE_INVERTIERT.XLSX mit Fällen in den Spalten und Variablen in den Zeilen

Abbildung 2.39: Zwischenschritt beim Transponieren von SPSS-Daten

Abbildung 2.40: Transponierte SPSS-Daten

Abbildung 2.41: Transponierte und um die Vorspalte bereinigte SPSS-Daten

Abbildung 3.1: SPSS-Datei mit den Fällen 1–5 (siehe *var00001*; *DATEN_1-5.SAV*)

Abbildung 3.2: SPSS-Datei mit den Fällen 6–10 (siehe *var00001*; *DATEN_6-10.SAV*)

Abbildung 3.3: Fälle hinzufügen (1)

Abbildung 3.4: Fälle hinzufügen (2)

Abbildung 3.5: Nach Fällen zusammengeführte Datei (Variante 1)

Abbildung 3.6: Nach Fällen zusammengeführte Datei (Variante 2)

Abbildung 3.7: Sortieren von Fällen in einem Datenset

Abbildung 3.8: SPSS-Datei mit den Variablen 1 und 2 (*DATEN_VAR1_2.SAV*)

Abbildung 3.9: SPSS-Datei mit den Variablen 1 und 3 (*DATEN_VAR1_3.SAV*)

Abbildung 3.10: Variablen hinzufügen (1)

Abbildung 3.11: Variablen hinzufügen (2)

Abbildung 3.12: Nach Variablen zusammengeführte Datei

Abbildung 3.13: Datei mit den Variablen 1 und 3, absteigend nach Variable 1 sortiert (*DATEN_VAR1_3A.SAV*)

Abbildung 3.14: Fehlerhafte, nach den Variablen zusammengeführte Gesamtdatei

Abbildung 3.15: Umsortieren einer Datei

Abbildung 3.16: Korrekte, nach den Variablen zusammengeführte Gesamtdatei – nach Umsortierung einer der beiden Dateien

Abbildung 3.17: Änderung der in SPSS abgebildeten Dateikennzeichnung nach Änderung eines Datenfelds

Abbildung 4.1: Beispiels-SPSS-Personendatendatei (*PERSONEN.SAV*)

Abbildung 4.2: Daten-Aggregation in SPSS starten

Abbildung 4.3: Mögliche Funktionen im Zusammenhang mit dem Aggregieren von Daten

Abbildung 4.4: Durchführen des Aggregierens in SPSS

Abbildung 4.5: Aggregierter Datensatz

Abbildung 4.6: Daten-Aggregation ohne Breakvariable

Abbildung 4.7: Ergebnis der Daten-Aggregation ohne Breakvariable

Abbildung 4.8: Ein zu formatierendes SPSS-Datenfile (Datenansicht) von *PERSONEN2.SAV*

Abbildung 4.9: Die SPSS-Variablenansicht von *PERSONEN2.SAV*

Abbildung 4.10: SPSS-Fehlermeldung bei falschem Namensanfang einer Variablen bezüglich der Variablen *1var*

Abbildung 4.11: SPSS-Fehlermeldung bei unzulässigem Sonderzeichen in einem Variablennamen bezüglich der Variablen *var-1*

Abbildung 4.12: SPSS-Fehlermeldung bei unzulässiger Verwendung eines Schlüsselworts als Variablennamen bezüglich der Variablen *all*

Abbildung 4.13: Verändern eines Variablennamens im Variableneditor

Abbildung 4.14: Verändern des Variablentyps im Variableneditor (*PERSONEN2.SAV*)

Abbildung 4.15: KOMMA-Formatierung der Variablen beitrag (*PERSONEN2.SAV*)

Abbildung 4.16: PUNKT-Formatierung der Variablen *beitrag* (*PERSONEN2.SAV*)

Abbildung 4.17: Formatierung der Variablen *beitrag* mit WISSENSCHAFTLICHE NOTATION (*PERSONEN2.SAV*)

Abbildung 4.18: DOLLAR-Formatierung der Variablen *beitrag* (*PERSONEN2.SAV*)

Abbildung 4.19: Zahlendarstellung im Dateneditor bei zu geringer Spaltenbreite (*PERSONEN4A.SAV*)

Abbildung 4.20: Skalierungsmöglichkeiten in SPSS: NOMINAL, ORDINAL, METRISCH (*PERSONEN4A.SAV*)

Abbildung 4.21: Zur ROLLE-Formatierung (*PERSONEN4A.SAV*)

Abbildung 4.22: Umcodieren in derselben Variablen am Beispiel der nominalskalierten Variablen *geschlecht* (1)

Abbildung 4.23: Umcodieren in derselben Variablen am Beispiel der nominalskalierten Variablen *geschlecht* (2)

Abbildung 4.24: Ergebnis des Umcodierens in derselben Variablen am Beispiel der nominalskalierten Variablen *geschlecht* (in *PERSONEN4A.SAV*)

Abbildung 4.25: Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (1)

Abbildung 4.26: Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (2)

Abbildung 4.27: Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (3)

Abbildung 4.28: Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (4)

Abbildung 4.29: Umcodieren in einer anderen Variablen (*PERSONEN4A.SAV*) (5)

Abbildung 4.30: Ergebnis des Umcodierens in einer anderen Variablen (*PERSONEN4A.SAV*)

Abbildung 4.31: Gleichzeitiges Umcodieren von zwei Variablen in zwei anderen Variablen (*PERSONEN4A.SAV*) (1)

Abbildung 4.32: Gleichzeitiges Umcodieren von zwei Variablen in zwei anderen Variablen (*PERSONEN4A.SAV*) (2)

Abbildung 4.33: Ergebnis des gleichzeitigen Umcodierens von zwei Variablen in zwei anderen Variablen (*PERSONEN4A.SAV*)

Abbildung 4.34: Berechnen einer neuen Variablen (*PERSONEN4A.SAV*)

Abbildung 4.35: Datendatei mit neu berechneter Variablen (*PERSONEN4A.SAV*)

Abbildung 4.36: Haushaltsbezogene Daten vor Gewichtung (in *AGGR.SAV*)

Abbildung 4.37: Dialogfeld zur Gewichtung

Abbildung 4.38: Fehlender Gewichtungswert (*AGGR_A.SAV*)

Abbildung 4.39: SPSS-Hinweis bei fehlenden Werten der Gewichtungsvariablen

Abbildung 4.40: Dialogbox für DATEN|FÄLLE AUSWÄHLEN (*AGGR.SAV*)

Abbildung 4.41: Aktivierung der Fallauswahl in der Dialogbox DATEN|FÄLLE AUSWÄHLEN (*AGGR.SAV*)

Abbildung 4.42: Angabe der Bedingung bei DATEN|FÄLLE AUSWÄHLEN (*AGGR.SAV*)

Abbildung 4.43: Datensatz mit Filtervariable *filter_$* und markierten nicht ausgewählten Fällen (*AGGR.SAV*)

Abbildung 4.44: Neuer Datensatz nur mit den ausgewählten Fällen (*AGGR_B.SAV*) (1)

Abbildung 4.45: Neuer Datensatz nur mit den ausgewählten Fällen (*AGGR_B.SAV*) (2)

Abbildung 4.46: Löschen der nicht ausgewählten Fälle (*AGGR.SAV*)

Abbildung 4.47: Beispiels-Transformation von Euro- in DM-Werte in einem Beispieldatensatz (*AGGR.SAV*)

Abbildung 4.48: Variablen-Transformation für einen gefilterten Datensatz (*AGGR.SAV*)

Abbildung 4.49: Variante »Zufallsstichprobe« bei DATEN|FÄLLE AUSWÄHLEN

Abbildung 4.50: Ergebnis einer Zufallsauswahl bei DATEN|FÄLLE AUSWÄHLEN (*AGGR.SAV*)

Abbildung 4.51: Variante »Fallbereich« bei DATEN|FÄLLE AUSWÄHLEN

Abbildung 4.52: Ergebnis einer »Fallbereichsauswahl« bei DATEN|FÄLLE AUSWÄHLEN (*AGGR.SAV*)

Abbildung 4.53: Doppelter Fall in SPSS-Datendatei *PERSONEN4.SAV*

Abbildung 4.54: Dialogfeld zur Ermittlung doppelter Fälle

Abbildung 4.55: SPSS-Ergebnis der Ermittlung doppelter Fälle (*PERSONEN4.SAV*)

Abbildung 4.56: Löschen eines doppelten Falls (*PERSONEN4.SAV*)

Abbildung 5.1: Der SPSS-Syntaxeditor

Abbildung 5.2: Beispieldatendatei *CORRELATIONS.SAV* für eine eingeschränkte Korrelationsmatrix

Abbildung 5.3: Eine eingeschränkte (syntaxgesteuerte) Korrelationsmatrix als Output (*CORRELATIONS.SAV*)

Abbildung 5.4: Eine umfangreiche (menügesteuerte) Korrelationsmatrix als Output (*CORRELATIONS.SAV*)

Abbildung 5.5: Syntax in einem Ausgabefenster (bezüglich *PERSONEN4A.SAV*)

Abbildung 5.6: Kopieren von Syntax in einem Ausgabefenster (bezüglich *PERSONEN4A.SAV*)

Abbildung 5.7: Option EINFÜGEN in einer SPSS-Dialogbox (*PERSONEN4A.SAV*)

Abbildung 5.8: Im Syntaxeditor automatisch eingefügter Befehlstext (*PERSONEN4A.SAV*)

Abbildung 5.9: Syntaxfehler im Syntaxeditor (*PERSONEN4A.SAV*)

Abbildung 5.10: OPTIONEN-Einstellungen für den Syntaxeditor

Abbildung 5.11: Syntaxfehler-Hinweis im Ausgabefenster (*PERSONEN4A.SAV*)

Abbildung 5.12: Datendateibezeichnung bei mehreren geöffneten Datendateien

Abbildung 5.13: Auswahl einer Datendatei (*CORRELATIONS.SAV*) im Syntaxeditor bei mehreren geöffneten Datendateien

Abbildung 5.14: Output eines spezifischen SPSS-Befehls (für *CORRELATIONS.SAV*)

Abbildung 5.15: Leerzeile innerhalb eines SPSS-Kommandos

Abbildung 5.16: Unvollständiger Output bei einer Leerzeile innerhalb eines SPSS-Kommandos (für *CORRELATIONS.SAV*)

Abbildung 5.17: Beendigung eines Kommentars mit einem Punkt

Abbildung 5.18: Beendigung eines Kommentars mit einer Leerzeile

Abbildung 5.19: Mehrzeiliger Kommentar

Abbildung 5.20: Leerzeile innerhalb eines mehrzeiligen Kommentars

Abbildung 5.21: Output bei einer Leerzeile innerhalb eines mehrzeiligen Kommentars (für *CORRELATIONS.SAV*)

Abbildung 5.22: Punkt als Satzzeichen innerhalb eines Kommentars

Abbildung 5.23: Problemloser Output bei einem Punkt als Satzzeichen innerhalb eines Kommentars (für *CORRELATIONS.SAV*)

Abbildung 5.24: Der SPSS-Syntaxeditor als Zugangsweg

Abbildung 5.25: Syntaxhilfe über das »Fragezeichen-Symbol« im Syntaxeditor

Abbildung 5.26: Online-Syntaxhilfe zu einem bestimmten Befehl (hier bezüglich GET FILE)

Abbildung 5.27: Öffnen eines bereits existierenden SPSS-Syntaxfiles

Abbildung 5.28: Speicherort für die SPSS-Journaldatei STATISTICS.JNL

Abbildung 5.29: Datenansicht nach Datentransformationen mit EXECUTE-Befehl unter Bezugnahme auf *CORRELATIONS.SAV* (1)

Abbildung 5.30: Datenansicht nach Datentransformationen mit EXECUTE-Befehl unter Bezugnahme auf *CORRELATIONS.SAV* (2)

Abbildung 5.31: Neue Datendatei nach Anwendung des GET-FILE-Befehls zusammen mit dem KEEP-Unterbefehl auf Basis von *PERSONEN4A.SAV*

Abbildung 5.32: Neue Datendatei nach Anwendung des GET-FILE-Befehls zusammen mit dem DROP-Unterbefehl auf Basis von *PERSONEN4A.SAV*

Abbildung 5.33: Datendatei *DATEN_ALLE_INVERTIERT.SAV* mit den Variablen in den Zeilen und den Fällen in den Spalten

Abbildung 5.34: Invertierte Datendatei *DATEN_ALLE_INVERTIERT.SAV* – FLIP-Befehl (1)

Abbildung 5.35: Invertierte Datendatei *DATEN_ALLE_INVERTIERT2.SAV* – FLIP-Befehl (2)

Abbildung 5.36: Nach Fällen gleiche, aber nach Variablen unterschiedliche zusammengeführte Datendateien in *DATEN_VAR1_2_3.SAV* – MATCH-FILES-Befehl

Abbildung 5.37: Nach Variablen gleiche, aber nach Fällen unterschiedliche zusammengeführte Datendateien in *DATEN_1-10.SAV* – ADD-FILES-Befehl

Abbildung 5.38: Verschiedene Datentransformationen mit Bezugnahme auf *PERSONEN4A.SAV* – RECODE, COMPUTE, IF

Abbildung 5.39: Zuweisung von Variablen- und Werteetiketten – VARIABLE LABELS, VALUE LABELS (auf Basis von *PERSONEN5.SAV*)

Abbildung 5.40: Häufigkeitstabelle bei Auswahl einzelner Fälle in *PERSONEN5.SAV* – SELECT IF mit TEMPORARY

Abbildung 5.41: Auswahl von Fällen auf Basis von *PERSONEN5.SAV* – SELECT IF ohne TEMPORARY

Abbildung 5.42: Filtern eines Datensatzes (*PERSONEN5.SAV*) – FILTER BY

Abbildung 5.43: Sortierung von Daten mithilfe von SORT CASES (*PERSONEN5.SAV*)

Abbildung 5.44: Häufigkeitstabellen für Männer und Frauen getrennt (auf Basis von *PERSONEN5.SAV*) – SPLIT FILE

Abbildung 5.45: Haushaltsbezogene Beitragssummen nach der Aggregation von der Personenebene (*AGGREG.SAV*)

Abbildung 5.46: Gesamtbeitragssumme nach der Aggregation von der Personenebene (*AGGREG2.SAV*)

Abbildung 5.47: Zur Anwendung des COUNT-Befehls (*EINKOMMEN_VERMOEGEN.SAV*)

Abbildung 5.48: Auflisten ausgewählter Variablen und Fälle – LIST (*PERSONEN4A.SAV*)

Abbildung 5.49: Zufallsauswahl mit dem SAMPLE-Befehl (*PERSONEN4A.SAV*)

Abbildung 5.50: Variablen- und Dateninformationen – SYSFILE INFO (*PERSONEN4A.SAV*)

Abbildung 5.51: Zu lange Titelzeile – TITLE (*PERSONEN4A.SAV*)

Abbildung 5.52: Auf zwei Zeilen verteilter Titel – TITLE (*PERSONEN4A.SAV*)

Abbildung 5.53: Statistiken im Fall mit Gewichtung – WEIGHT BY (*AGGR.SAV*)

Abbildung 5.54: Statistiken im Fall ohne Gewichtung – WEIGHT OFF (*AGGR.SAV*)

Abbildung 5.55: Ersetzen der Verzeichnisstruktur des Autors durch Ihre eigene in *BEISPIEL1.SPS*

Abbildung 5.56: Beispiel-Rohdatendatei *PERS1.SAV* (vier Variablen, Fälle 1–10)

Abbildung 5.57: Beispiel-Rohdatendatei *PERS2_FALSCH.SAV* (alle elf Variablen, Fälle 1–10, invertiert)

Abbildung 5.58: Beispiel-Rohdatendatei *PERS3.SAV* (alle Variablen, Fälle 11–20)

Abbildung 5.59: Haushaltsbezogene Datendatei *HH_ALLE2.SAV* nach Durchführen aller Aggregationen am Ende des Syntaxprogramms

Abbildung 5.60: Personenbezogene FREQUENCIES-Ergebnisse in Form von Häufigkeitstabellen im vorderen Teil von *BEISPIEL1.SPS* (für *PERS_ALLE.SAV*)

Abbildung 5.61: Personenbezogene FREQUENCIES-Ergebnisse in Form statistischer Kennziffern im vorderen Teil von *BEISPIEL1.SPS* (für *PERS_ALLE.SAV*)

Abbildung 5.62: Mit SELECT IF selektierte »einkommensschwache Haushalte« mit statistischen Kennziffern (FREQUENCIES) im hinteren Teil von *BEISPIEL1.SPS*, gewichtet (oben) und ungewichtet (unten) (für *HH_ALLE2.SAV*)

Abbildung 5.63: Mit SPLIT BY aufgeteilte Datei nach Altersgruppen, das heißt: nach dem Alter des Haushaltsvorstands mit statistischen Kennziffern auf der Haushaltsebene (FREQUENCIES) im hinteren Teil von *BEISPIEL1.SPS*, gewichtet (links) und ungewichtet (rechts) (für *HH_ALLE2.SAV*)

Abbildung 5.64: Syntax-Beispiel für AUSFÜHREN|ALLE (in Bezug auf *AGGR.SAV*)

Abbildung 5.65: Ausführlicher Output im Syntax-Beispiel bei AUSFÜHREN|ALLE (mit Gewichtung: linke Seite; ohne Gewichtung: rechte Seite) (in Bezug auf *AGGR.SAV*)

Abbildung 5.66: Syntax-Beispiel für AUSFÜHREN|AUSWAHL (in Bezug auf *AGGR.SAV*)

Abbildung 5.67: Reduzierter Output im Syntax-Beispiel bei AUSFÜHREN|AUSWAHL (in Bezug auf *AGGR.SAV*)

Abbildung 5.68: Öffnen eines Ausgabefensters nach Öffnen einer Datendatei (*123.SAV*)

Abbildung 5.69: Viewer-Einstellungen

Abbildung 5.70: Ausschalten des Syntax-Protokolls

Abbildung 5.71: SPSS-Output nach Ausschalten des Syntax-Protokolls (in Bezug auf *AGGR.SAV*)

Abbildung 5.72: Der Navigationsbereich des Viewers (in Bezug auf *AGGR.SAV*)

Abbildung 5.73: Rubrik »Titel« im Navigationsbereich des Viewers (in Bezug auf *AGGR.SAV*)

Abbildung 5.74: Rubrik »Hinweise« im Navigationsbereich des Viewers (in Bezug auf *AGGR.SAV*)

Abbildung 5.75: Eigentliche Auswertungsrubrik »Statistiken« im Navigationsbereich des Viewers (in Bezug auf *AGGR.SAV*)

Abbildung 5.76: Optionen des Ausschneidens, Verschiebens oder Kopierens im Navigationsbereich

Abbildung 5.77: Warnung im Viewer-Inhaltsbereich (in Bezug auf *PERSONEN4A.SAV*)

Abbildung 5.78: Fehlermeldung im Viewer-Inhaltsbereich (in Bezug auf *CORRELATIONS.SAV*)

Abbildung 5.79: Pivottabelle im Viewer-Inhaltsbereich (in Bezug auf *CORRELATIONS.SAV*)

Abbildung 5.80: Doppelklick zum Verändern einer Pivottabelle im Viewer-Inhaltsbereich (hier in Bezug auf *CORRELATIONS.SAV*)

Abbildung 5.81: Veränderung einer Pivottabelle im Viewer-Inhaltsbereich (hier in Bezug auf *CORRELATIONS.SAV*)

Abbildung 5.82: Erweiterte Verarbeitung-smöglichkeiten einer Pivottabelle im Viewer-Inhaltsbereich (hier in Bezug auf *CORRELATIONS.SAV*)

Abbildung 5.83: Diagramm im Viewer-Inhaltsbereich (hier in Bezug auf *CORRELATIONS.SAV*)

Abbildung 5.84: Öffnen des Diagrammeditors aus dem Viewer-Inhaltsbereich heraus (hier in Bezug auf *CORRELATIONS.SAV*)

Abbildung 5.85: Verändern eines Diagramms im Diagrammeditor

Abbildung 5.86: Abspeichern von SPSS-Output in einer EXCEL-Datei (hier in Bezug auf *CORRELATIONS.SAV*)

Abbildung 5.87: SPSS-Output in einer EXCEL-Datei (hier in Bezug auf *CORRELATIONS.SAV*)

Abbildung 5.88: Abspeichern von SPSS-Output in einer WORD-Datei (in Bezug auf *CORRELATIONS.SAV*)

Abbildung 5.89: SPSS-Output in einer WORD-Datei (hier in Bezug auf *CORRELATIONS.SAV*)

Abbildung 6.1: Fenster für HÄUFIGKEITEN (hier für *KORBTREFFER.SAV*)

Abbildung 6.2: HÄUFIGKEITEN-Output für »Basketball-Beispiel« (*KORBTREFFER.SAV*)

Abbildung 6.3: Datensatz *KORBTREFFER_F.SAV* mit zwei fehlenden Werten

Abbildung 6.4: Häufigkeitstabelle für eine Variable mit (zwei) fehlenden Werten (*KORBTREFFER_F.SAV*)

Abbildung 6.5: Häufigkeitstabelle für sehr viele (1.000) Merkmalsausprägungen einer Variablen *v1* in *DATEN1_1000.SAV* (Output-Auszug)

Abbildung 6.6: Fehleranalyse anhand einer Häufigkeitstabelle (*KORBTREFFER_F2.SAV*)

Abbildung 6.7: Suche nach einer bestimmten Variablen in einem Datensatz (*KORBTREFFER_F2.SAV*)

Abbildung 6.8: Suche nach einem bestimmten Wert einer bestimmten Variablen in einem Datensatz (*KORBTREFFER_F2.SAV*)

Abbildung 6.9: »Ausreißer-Datensatz« *AUSREISSER.SAV*

Abbildung 6.10: Häufigkeitsauszählung in SPSS mit Ausreißerwert »100« (*AUSREISSER.SAV*)

Abbildung 6.11: Starten einer Ausreißer-Analyse mithilfe von EXPLORATIVE DATENANALYSE|STATISTIK (für *AUSREISSER.SAV*)

Abbildung 6.12: Ergebnis einer tabellarischen Ausreißeranalyse mithilfe von EXPLORATIVE DATENANALYSE|STATISTIKEN (für *AUSREISSER.SAV*)

Abbildung 6.13: Ergebnis einer grafischen Ausreißer-Analyse mithilfe von EXPLORATIVE DATENANALYSE|DIAGRAMME

(*AUSREISSER.SAV*)

Abbildung 6.14: DIAGRAMME bei HÄUFIGKEITEN (in Bezug auf *KORBTREFFER.SAV*)

Abbildung 6.15: Balkendiagramm als Output (*KORBTREFFER.SAV*)

Abbildung 6.16: Kreisdiagramm als Output (*KORBTREFFER.SAV*)

Abbildung 6.17: Histogramm als Output (*KORBTREFFER.SAV*)

Abbildung 6.18: Normalverteilungskurve anzeigen lassen (hier in Bezug auf *KORBTREFFER.SAV*)

Abbildung 6.19: Histogramm und Normalverteilungskurve als Output (*KORBTREFFER.SAV*)

Abbildung 7.1: Statistische Funktion MEAN (für *KORBTREFFER4.SAV*)

Abbildung 7.2: Anwendung der MEAN-Funktion auf die Variable *korbtreffer2014_2016* (*KORBTREFFER4.SAV*)

Abbildung 7.3: Neue Variable *korbtreffer2014_2016* auf Basis der MEAN-Funktion (*KORBTREFFER4.SAV*)

Abbildung 7.4: Anwendung der SUM-Funktion auf die Variable *korbtreffer2014_2016* (*KORBTREFFER4.SAV*)

Abbildung 7.5: Neue Variable *korbtreffer2014_2016* für die SUM-Funktion (*KORBTREFFER4.SAV*)

Abbildung 7.6: HÄUFIGKEITEN|STATISTIKEN für das »Basketball-Beispiel« (*KORBTREFFER.SAV*)

Abbildung 7.7: Unterdrückung der Häufigkeitstabelle bei HÄUFIGKEITEN|STATISTIKEN (*KORBTREFFER.SAV*)

Abbildung 7.8: Statistische Kennziffern (FREQUENCIES) für das »Basketball-Beispiel« (*KORBTREFFER.SAV*)

Abbildung 7.9: BOOTSTRAPPING DURCHFÜHREN bei HÄUFIGKEITEN (*KORBTREFFER.SAV*)

Abbildung 7.10: Bootstrapping-Output bei HÄUFIGKEITEN (*KORBTREFFER.SAV*)

Abbildung 7.11: Menü-Maske für DESKRIPTIVE STATISTIK (in Bezug auf *KORBTREFFER.SAV*)

Abbildung 7.12: OPTIONEN bei DESKRIPTIVE STATISTIK (hier in Bezug auf *KORBTREFFER.SAV*)

Abbildung 7.13: Output bei DESKRIPTIVE STATISTIK (*KORBTREFFER.SAV*)

Abbildung 7.14: BOOTSTRAP bei DESKRIPTIVE STATISTIKEN (in Bezug auf *KORBTREFFER.SAV*)

Abbildung 7.15: BOOTSTRAP-Output bei DESKRIPTIVE STATISTIKEN (*KORBTREFFER.SAV*)

Abbildung 7.16: Start-Dialog bei EXPLORATIVE DATENANALYSE (*KORBTREFFER.SAV*)

Abbildung 7.17: STATISTIKEN bei EXPLORATIVE DATENANALYSE (hier in Bezug auf *KORBTREFFER.SAV*)

Abbildung 7.18: Statistik-Output bei EXPLORATIVE DATENANALYSE (*KORBTREFFER.SAV*)

Abbildung 7.19: BOOTSTRAP-Output bei EXPLORATIVE DATENANALYSE (*KORBTREFFER.SAV*) (1)

Abbildung 7.20: BOOTSTRAP-Output bei EXPLORATIVE DATENANALYSE (*KORBTREFFER.SAV*) (2)

Abbildung 7.21: Um die Variable *geschlecht* erweiterte Korbtreffer-Datendatei *KORBTREFFER5.SAV*

Abbildung 7.22: MEANS-Berechnungen für *korbtreffer* in Abhängigkeit von *geschlecht* (*KORBTREFFER5.SAV*)

Abbildung 7.23: MEANS-Statistiken für *korbtreffer* in Abhängigkeit von *geschlecht* (*KORBTREFFER5.SAV*)

Abbildung 7.24: MEANS-Statistikoutput für *korbtreffer* in Abhängigkeit von *geschlecht* (*KORBTREFFER5.SAV*)

Abbildung 7.25: Die Datendatei *HARM_GEOM.SAV*

Abbildung 7.26: Geometrisches Mittel für *wachstumsrate* (MEANS) (*HARM_GEOM.SAV*)

Abbildung 7.27: Harmonisches Mittel für *km_pro_h* (MEANS) (*HARM_GEOM.SAV*)

Abbildung 7.28: BOOTSTRAP-Output bei MITTELWERTE VERGLEICHEN (*KORBTREFFER5.SAV*)

Abbildung 7.29: Alle Fälle berücksichtigen bei SUMMARIZE (hier in Bezug auf *KORBTREFFER5.SAV*)

Abbildung 7.30: SUMMARIZE-Statistikoutput für *korbtreffer* in Abhängigkeit von *geschlecht* (*KORBTREFFER5.SAV*)

Abbildung 8.1: Erweiterte »Basketball-Beispiel«-Datendatei (*KORBTREFFER2.SAV*)

Abbildung 8.2: Kreuztabellierungsanweisung für *korbtreffer* und *alter_klasse* (*KORBTREFFER3.SAV*)

Abbildung 8.3: Auswahl der beobachteten Fallzahlen in einer Kreuztabelle (*KORBTREFFER3.SAV*)

Abbildung 8.4: Kreuztabellenoutput für die beobachteten Fallzahlen (*KORBTREFFER3.SAV*)

Abbildung 8.5: Auswahl der beobachteten und der erwarteten Fallzahlen in einer Kreuztabelle (hier in Bezug auf *KORBTREFFER3.SAV*)

Abbildung 8.6: Kreuztabellenoutput für die beobachteten und die erwarteten Fallzahlen (*KORBTREFFER3.SAV*)

Abbildung 8.7: Auswahl der zeilenweise berechneten relativen Häufigkeiten in einer Kreuztabelle (hier in Bezug auf *KORBTREFFER3.SAV*)

Abbildung 8.8: Kreuztabellenoutput für die zeilenweise berechneten relativen Häufigkeiten (*KORBTREFFER3.SAV*)

Abbildung 8.9: Auswahl der spaltenweise berechneten relativen Häufigkeiten in einer Kreuztabelle (hier in Bezug auf

*KORBTREFFER3.SAV*)

Abbildung 8.10: Kreuztabellenoutput für die spaltenweise berechneten relativen Häufigkeiten (*KORBTREFFER3.SAV*)

Abbildung 8.11: Auswahl der auf die Gesamtfallzahl bezogenen relativen Häufigkeiten in einer Kreuztabelle (hier in Bezug auf *KORBTREFFER3.SAV*)

Abbildung 8.12: Kreuztabellenoutput für auf die Gesamtfallzahl bezogene relative Häufigkeiten (*KORBTREFFER3.SAV*)

Abbildung 8.13: Auswahl von absoluten und relativen Häufigkeiten in einer Kreuztabelle (hier in Bezug auf *KORBTREFFER3.SAV*)

Abbildung 8.14: Kreuztabellenoutput für absolute und relative Häufigkeiten (*KORBTREFFER3.SAV*)

Abbildung 8.15: Datendatei *GESCHL_FARBE2.SAV* mit den Variablen *geschlecht*, *augenfarbe*, *qualifikation* und *nationalität*

Abbildung 8.16: Anweisung für eine erweiterte Kreuztabelle: erste Unterteilung der Zeilenvariablen *geschlecht* (*GESCHL_FARBE2.SAV*)

Abbildung 8.17: Anweisung für eine erweiterte Kreuztabelle: zweite Unterteilung der Zeilenvariablen *geschlecht* (*GESCHL_FARBE2.SAV*)

Abbildung 8.18: Output für eine erweiterte Kreuztabelle mit zwei Unterteilungen der Zeilenvariablen *geschlecht* (*GESCHL_FARBE2.SAV*)

Abbildung 8.19: Datendatei *GESCHL_FARBE.SAV* mit den Variablen *geschlecht* und *augenfarbe*

Abbildung 8.20: Kreuztabellen-Output mit den beobachteten und den erwarteten Fallzahlen (*GESCHL_FARBE.SAV*)

Abbildung 8.21: Auswahl verschiedener Assoziationsmaße (*GESCHL_FARBE.SAV*)

Abbildung 8.22: SPSS-Output für verschiedene Assoziationsmaße (*GESCHL_FARBE.SAV*)

Abbildung 8.23: Bootstrapping für Assoziationsmaße (hier in Bezug auf *GESCHL_FARBE.SAV*)

Abbildung 8.24: Bootstrapping-Output für verschiedene Assoziationsmaße (*GESCHL_FARBE.SAV*)

Abbildung 8.25: Voraussetzung von mindestens Ordinalskalenniveau bei Korrelationsberechnungen (*GESCHL_FARBE.SAV*)

Abbildung 8.26: Datendatei *SPEERWERFEN.SAV*

Abbildung 8.27: Aufruf des Spearman'schen Rangkorrelationskoeffizienten in SPSS (*SPEERWERFEN.SAV*)

Abbildung 8.28: Wert des Rangkorrelationskoeffizienten – Beispiel 1 (*SPEERWERFEN.SAV*)

Abbildung 8.29: Datendatei *SPEERWERFEN2.SAV* (mit Mehrfachbesetzungen)

Abbildung 8.30: Wert des Rangkorrelationskoeffizienten – Beispiel 2 (*SPEERWERFEN2.SAV*)

Abbildung 8.31: Aufruf von Kendalls Tau-Koeffizienten b in SPSS (*SPEERWERFEN.SAV*)

Abbildung 8.32: Wert des Tau-b-Koeffizienten – Beispiel 1 (*SPEERWERFEN.SAV*)

Abbildung 8.33: Wert des Tau-b-Koeffizienten – Beispiel 2 (*SPEERWERFEN2.SAV*)

Abbildung 8.34: Aufruf verschiedener (weiterer) Rangkorrelationskoeffizienten in SPSS (*SPEERWERFEN2.SAV*)

Abbildung 8.35: Werte verschiedener Rangkorrelationskoeffizienten (und des Bravais/Pearson-Korrelationskoeffizienten) (*SPEERWERFEN2.SAV*)

Abbildung 8.36: Eine perfekt-positive lineare Korrelation zwischen zwei metrischen Variablen X und Y (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 153)

Abbildung 8.37: Keine lineare, aber eine nichtlineare Korrelation zwischen zwei metrischen Variablen X und Y (Quelle: Faik, Jürgen

(2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 155)

Abbildung 8.38: Aufruf des Bravais/Pearson-Korrelationskoeffizienten über das SPSS-KORRELATIONEN-Menü (*SPEERWERFEN2.SAV*)

Abbildung 8.39: Ergebnis des Bravais/Pearson-Korrelationskoeffizienten nach Aufruf über das SPSS-KORRELATIONEN-Menü (*SPEERWERFEN2.SAV*)

Abbildung 8.40: Aufruf (auch) des Bravais/Pearson-Korrelationskoeffizienten über das SPSS-KREUZTABELLEN-Menü (*SPEERWERFEN2.SAV*)

Abbildung 8.41: Ergebnis (auch) des Bravais/Pearson-Korrelationskoeffizienten nach Aufruf über das SPSS-KREUZTABELLEN-Menü (*SPEERWERFEN2.SAV*)

Abbildung 8.42: Zum Problem der verdeckten Korrelation (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 157)

Abbildung 8.43: Bootstrapping bei nominalskalierten Variablen für verschiedene Assoziationsmaße (*GESCHL_FARBE.SAV*) (1)

Abbildung 8.44: Bootstrapping bei nominalskalierten Variablen für verschiedene Assoziationsmaße (*GESCHL_FARBE.SAV*) (2)

Abbildung 8.45: Bootstrapping-Output bei nominalskalierten Variablen für verschiedene Assoziationsmaße (*GESCHL_FARBE.SAV*)

Abbildung 8.46: Bootstrapping bei mindestens ordinalskalierten Variablen für verschiedene Korrelationsmaße – KREUZTABELLEN-Menü (*SPEERWERFEN2.SAV*)

Abbildung 8.47: Bootstrapping-Output bei mindestens ordinalskalierten Variablen für verschiedene Korrelationsmaße – KREUZTABELLEN-Menü (*SPEERWERFEN2.SAV*) (Auszug)

Abbildung 8.48: Bootstrapping bei mindestens ordinalskalierten Variablen für verschiedene Korrelationsmaße – KORRELATION-Menü (*SPEERWERFEN2.SAV*)

Abbildung 8.49: Bootstrapping-Output bei mindestens ordinalskalierten Variablen für verschiedene Korrelationsmaße – KORRELATION-Menü (*SPEERWERFEN2.SAV*)

Abbildung 9.1: Lineare Regressionsfunktion und Punktewolke (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 164)

Abbildung 9.2: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE1.SAV*

Abbildung 9.3: Lineare Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE1.SAV*

Abbildung 9.4: Ergebnis der linearen Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE1.SAV* (Auszug)

Abbildung 9.5: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE2.SAV*

Abbildung 9.6: Ergebnis der linearen Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE2.SAV*

Abbildung 9.7: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE3.SAV*

Abbildung 9.8: Ergebnis der linearen Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE3.SAV*

Abbildung 9.9: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE4.SAV*

Abbildung 9.10: Ergebnis der linearen Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE4.SAV*

Abbildung 9.11: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE5.SAV*

Abbildung 9.12: Ergebnis der linearen Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE5.SAV*

Abbildung 9.13: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE6.SAV*

Abbildung 9.14: Vollständiges Ergebnis der linearen Einfachregression zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_GERADE6.SAV*

Abbildung 9.15: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_QUADRAT1.SAV*

Abbildung 9.16: Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben* sowie *einkommen* und *einkommen2* in *LINEAREREGRESSION_QUADRAT1A.SAV*

Abbildung 9.17: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* sowie *einkommen* und *einkommen2* in *LINEAREREGRESSION_QUADRAT1.SAV* (Auszug)

Abbildung 9.18: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_POTENZ1.SAV*

Abbildung 9.19: Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben_ln* und *einkommen_ln* in *LINEAREREGRESSION_POTENZ1A.SAV*

Abbildung 9.20: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben_ln* und *einkommen_ln* in *LINEAREREGRESSION_POTENZ1A.SAV*

Abbildung 9.21: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_EXPONENTIAL1.SAV*

Abbildung 9.22: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben_ln* und *einkommen* in *LINEAREREGRESSION_EXPONENTIAL1A.SAV*

Abbildung 9.23: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_LOGISTISCH1.SAV*

Abbildung 9.24: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben_logistisch* und *einkommen* in *LINEAREREGRESSION_LOGISTISCH1A.SAV*

Abbildung 9.25: Scatterplot zwischen *ausgaben* und *einkommen* in der Datendatei *LINEAREREGRESSION_DIVERSE1.SAV*

Abbildung 9.26: Anfordern verschiedener linearer Regressionsbeziehungen zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_DIVERSE1.SAV*

Abbildung 9.27: Ergebnisse verschiedener linearer Regressionsbeziehungen zwischen *ausgaben* sowie *einkommen* in *LINEAREREGRESSION_DIVERSE1.SAV* (1)

Abbildung 9.28: Ergebnisse verschiedener linearer Regressionsbeziehungen zwischen *ausgaben* sowie *einkommen* in *LINEAREREGRESSION_DIVERSE1.SAV* (2)

Abbildung 9.29: Scatterplot zwischen *ausgaben* und *alter* in der Datendatei *LINEAREREGRESSION_DUMMY.SAV*

Abbildung 9.30: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* und *alter_dummy* in *LINEAREREGRESSION_DUMMY1.SAV*

Abbildung 9.31: Ausgaben-Durchschnittswerte für zwei Altersgruppen auf Basis von *LINEAREREGRESSION_DUMMY.SAV*

Abbildung 9.32: Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben* sowie *alter_40_59* und *alter_60plus* in *LINEAREREGRESSION_DUMMY1A.SAV*

Abbildung 9.33: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* sowie *alter_40_59* und *alter_60plus* in *LINEAREREGRESSION_DUMMY1A.SAV*

Abbildung 9.34: Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben* sowie *geschl_dummy*, *alter_dummy* und *alter_geschl* in *LINEAREREGRESSION_DUMMY2.SAV*

Abbildung 9.35: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* sowie *geschl_dummy*, *alter_dummy* und *alter_geschl* in

*LINEAREREGRESSION_DUMMY2.SAV*

Abbildung 9.36: Scatterplot zwischen *ausgaben* und *jahr* in der Datendatei *LINEAREREGRESSION_ZEIT1.SAV*

Abbildung 9.37: Anfordern einer linearen Regressionsbeziehung zwischen *ausgaben* und *time* in *LINEAREREGRESSION_ZEIT1.SAV*

Abbildung 9.38: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* und *time* in *LINEAREREGRESSION_ZEIT1.SAV*

Abbildung 9.39: Anfordern verschiedener Trendfunktionen für *ausgaben* in *LINEARE_REGRESSION_ZEIT1A.SAV*

Abbildung 9.40: Ergebnis der Schätzung verschiedener Trendfunktionen für *ausgaben* in *LINEARE_REGRESSION_ZEIT1.SAV* (1)

Abbildung 9.41: Ergebnis der Schätzung verschiedener Trendfunktionen für *ausgaben* in *LINEARE_REGRESSION_ZEIT1.SAV* (2)

Abbildung 9.42: Ergebnis einer linearen Regressionsbeziehung zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_ZEIT2.SAV* (Auszug)

Abbildung 9.43: Anfordern einer linearen Regressionsbeziehung zwischen *ausg_diff* und *einkommen* in *LINEAREREGRESSION_ZEIT3.SAV*

Abbildung 9.44: Ergebnis einer linearen Regressionsbeziehung zwischen *ausg_diff* und *einkommen* in *LINEAREREGRESSION_ZEIT3.SAV* (Auszug)

Abbildung 9.45: Anfordern von Konfidenzintervallen für die geschätzten Regressionsparameter einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_GERADE6.SAV*

Abbildung 9.46: Konfidenzintervalle für die geschätzten Regressionsparameter als Ergebnis einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_GERADE6.SAV* (Ergebnisauszug)

Abbildung 9.47: BOOTSTRAP-Anfordern von Konfidenzintervallen für die geschätzten Regressionsparameter einer linearen Regression

zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_GERADE6.SAV*

Abbildung 9.48: BOOTSTRAP-Konfidenzintervalle für die geschätzten Regressionsparameter als Ergebnis einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_GERADE6.SAV* (Ergebnisauszug)

Abbildung 9.49: Anfordern der einfachen Residuen bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT.SAV*

Abbildung 9.50: Neue Variable *res_1* für die einfachen Residuen bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT.SAV*

Abbildung 9.51: Streudiagramm für die einfachen Residuen in Abhängigkeit von *einkommen* bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT.SAV*

Abbildung 9.52: Standardisierte vorhergesagte Werte der abhängigen Variablen bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT.SAV*

Abbildung 9.53: Streudiagramm für die standardisierten Residuen in Abhängigkeit von den standardisierten geschätzten Ausgabenwerten bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT.SAV*

Abbildung 9.54: Um zwei logarithmierte Variablen erweiterter Datensatz *LINEAREREGRESSION_HETEROSKEDASTIZITÄT1.SAV*

Abbildung 9.55: Streudiagramm für die standardisierten Residuen in Abhängigkeit von den standardisierten geschätzten Ausgabenwerten bei einer linearen Regression zwischen *ausgaben_ln* und *einkommen_ln* in *LINEAREREGRESSION_HETEROSKEDASTIZITÄT1.SAV*

Abbildung 9.56: Aufruf einer WLS-Schätzung für den Zusammenhang zwischen *ausgaben* und *einkommen* mit der Gewichtungsvariablen

*x_invers* auf Basis von *LINEAREREGRESSION_HETEROSKEDASTIZITÄT2.SAV*

Abbildung 9.57: Ergebnis einer WLS-Schätzung für den Zusammenhang zwischen *ausgaben* und *einkommen* mit der Gewichtungsvariablen *x_invers* auf Basis von *LINEAREREGRESSION_HETEROSKEDASTIZITÄT2.SAV*

Abbildung 9.58: Anfordern der Prüfung auf Normalverteilung in Bezug auf die Residuen bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_GERADE6.SAV*

Abbildung 9.59: Histogramm und Normalverteilung(-skurve) in Bezug auf die Residuen bei einer linearen Regression zwischen *ausgaben* und *einkommen* in *LINEAREREGRESSION_GERADE6.SAV*

Abbildung 9.60: Bivariate Korrelationen zwischen *ausgaben* einerseits sowie *einkommen*, *hhgroesse* und *vermögen* andererseits in *LINEAREREGRESSION_MULTIPEL.SAV*

Abbildung 9.61: Methoden zur Berechnung einer linearen Mehrfachregression am Beispiel von *LINEAREREGRESSION_MULTIPEL.SAV*

Abbildung 9.62: Ergebnis einer Regression zwischen *ausgaben* sowie *einkommen*, *hhgroesse* und *vermögen* in *LINEAREREGRESSION_MULTIPEL.SAV*: EINSCHLUSS beziehungsweise ENTER

Abbildung 9.63: Ergebnis einer Regression zwischen *ausgaben* sowie *einkommen*, *hhgroesse* und *vermögen* in *LINEAREREGRESSION_MULTIPEL.SAV*: SCHRITTWEISE beziehungsweise STEPWISE (Ergebnisauszug)

Abbildung 9.64: Ergebnis einer Regression zwischen *ausgaben* sowie *einkommen*, *hhgroesse* und *vermögen* in *LINEAREREGRESSION_MULTIPEL.SAV*: ENTFERNEN beziehungsweise REMOVE (Ergebnisauszug)

Abbildung 9.65: Ergebnis einer Regression zwischen *ausgaben* sowie *einkommen*, *hhgroesse* und *vermögen* in

*LINEAREREGRESSION_MULTIPEL.SAV*: RÜCKWÄRTS beziehungsweise BACKWARD (Ergebnisauszug)

Abbildung 9.66: Zur Multikollinearitätsproblematik (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 335)

Abbildung 9.67: Anfordern einer Kollinearitätsdiagnose für eine Regression zwischen *ausgaben* sowie *einkommen*, *hhgroesse* und *vermögen* in *LINEAREREGRESSION_MULTIPEL.SAV*

Abbildung 9.68: Ergebnis einer Kollinearitätsdiagnose für eine Regression zwischen *ausgaben* sowie *einkommen*, *hhgroesse* und *vermögen* in *LINEAREREGRESSION_MULTIPEL.SAV* (Ergebnisauszug)

Abbildung 9.69: Ergebnis einer Kollinearitätsdiagnose für eine Regression zwischen *ausgaben* sowie *wealth* und *hhgroesse* in *LINEAREREGRESSION_MULTIPEL1.SAV* (Ergebnisauszug)

Abbildung 9.70: Ergebnis einer Kollinearitätsdiagnose für eine Regression zwischen *ausgaben* sowie *einkommen_prokopf* und *vermögen* in *LINEAREREGRESSION_MULTIPEL2.SAV* (Ergebnisauszug)

Abbildung 10.1: Um die Variable *arm* erweiterter neuer Datensatz *REGRESSION_LOGIT1.SAV*

Abbildung 10.2: Anfordern einer Logit-Regressionsschätzung zwischen *arm* sowie *alter* und *qualifikation* in der Datendatei *REGRESSION_LOGIT1.SAV*

Abbildung 10.3: Ergebnis einer Logit-Regressionsschätzung zwischen *arm* und *alter* in der Datendatei *REGRESSION_LOGIT1.SAV* (Ergebnisauszug)

Abbildung 10.4: Um die Variable *bereich* erweiterter neuer Datensatz *REGRESSION_ORDINAL2.SAV*

Abbildung 10.5: Anfordern einer ordinalen Regressionsschätzung zwischen *bereich* und *alter* in der Datendatei

*REGRESSION_ORDINAL2.SAV*

Abbildung 10.6: Ergebnis einer ordinalen Regressionsschätzung zwischen *bereich* und *alter* in der Datendatei *REGRESSION_ORDINAL2.SAV* (Ergebnisauszug)

Abbildung 10.7: Anfordern einer multinomialen Regressionsschätzung zwischen *region* und *alter* in der Datendatei *REGRESSION_MULTINOMIAL.SAV*

Abbildung 10.8: Ergebnis einer multinomialen Regressionsschätzung zwischen *bereich* und *alter* in der Datendatei *REGRESSION_MULTINOMIAL.SAV* (Ergebnisauszug)

Abbildung 11.1: Datensatz DISKRIMINANZ1.SAV

Abbildung 11.2: Auswahl der Gruppierungsvariablen *pisa* mit Werten 1 und 2 (DISKRIMINANZ1.SAV)

Abbildung 11.3: Auswahl der unabhängigen Diskriminanzvariablen: *klassengroesse* und *schulbudget* (DISKRIMINANZ1.SAV)

Abbildung 11.4: Auswahl von Diskriminanz-Statistiken (hier in Bezug auf DISKRIMINANZ1.SAV)

Abbildung 11.5: Auswahl der (gleichen) Gruppengröße im Rahmen der Diskriminanzanalyse (in Bezug auf DISKRIMINANZ1.SAV)

Abbildung 11.6: Auswahl zu speichernder Variablen im Rahmen der Diskriminanzanalyse (hier in Bezug auf DISKRIMINANZ1.SAV)

Abbildung 11.7: Gespeicherte Variablen im Datensatz DISKRIMINANZ1.SAV

Abbildung 11.8: Grundlegende Beurteilungskennziffern für die mit DISKRIMINANZ1.SAV durchgeführte Diskriminanzanalyse

Abbildung 11.9: Kernergebnisse für die mit DISKRIMINANZ1.SAV durchgeführte Diskriminanzanalyse

Abbildung 11.10: Datendatei CLUSTER1.SAV

Abbildung 11.11: Variablenauswahl bei der zweistufigen Clusteranalyse (CLUSTER1.SAV)

Abbildung 11.12: Outputsetzungen bei der zweistufigen Clusteranalyse (CLUSTER1.SAV)

Abbildung 11.13: Um die Variable *tsc_982* ergänzte Datendatei CLUSTER1.SAV bei der zweistufigen Clusteranalyse

Abbildung 11.14: Output bei der zweistufigen Clusteranalyse (CLUSTER1.SAV)

Abbildung 11.15: Variablenauswahl mit der Vorgabe von zwei Clustern bei der K-Mittelwerte-Clusteranalyse (CLUSTER1.SAV)

Abbildung 11.16: Abzuspeichernde Variablen bei der K-Mittelwerte-Clusteranalyse (CLUSTER1.SAV)

Abbildung 11.17: Auswahl statistischer Kennziffern bei der K-Mittelwerte-Clusteranalyse (CLUSTER1.SAV)

Abbildung 11.18: Um die Variablen *qcl_1* und *qcl_2* ergänzte Datendatei CLUSTER1.SAV bei der K-Mittelwerte-Clusteranalyse

Abbildung 11.19: Output bei der K-Mittelwerte-Clusteranalyse (CLUSTER1.SAV)

Abbildung 11.20: Variablenauswahl bei der hierarchischen Clusteranalyse (CLUSTER1.SAV)

Abbildung 11.21: Auswahl eines Dendrogramms bei der hierarchischen Clusteranalyse (CLUSTER1.SAV)

Abbildung 11.22: Output der hierarchischen Clusteranalyse (CLUSTER1.SAV) (1)

Abbildung 11.23: Output der hierarchischen Clusteranalyse (CLUSTER1.SAV): Dendrogramm (2)

Abbildung 11.24: Datendatei FAKTOREN1.SAV

Abbildung 11.25: Variablenauswahl bei der Faktorenanalyse (FAKTOREN1.SAV)

Abbildung 11.26: Statistiken bei der Faktorenanalyse (FAKTOREN1.SAV)

Abbildung 11.27: Wahl der konkreten Faktorenanalyse und »Faktoren-Entscheidungsregel« (FAKTOREN1.SAV)

Abbildung 11.28: Auswahl eines Rotationsverfahrens (VARIMAX) bei der Faktorenanalyse (FAKTOREN1.SAV)

Abbildung 11.29: Faktorwerte für die einzelnen Fälle bei der Faktorenanalyse (FAKTOREN1.SAV)

Abbildung 11.30: Faktorenanalyse: Optionen (hier in Bezug auf FAKTOREN1.SAV)

Abbildung 11.31: Output der Faktorenanalyse (FAKTOREN1.SAV) (1)

Abbildung 11.32: Output der Faktorenanalyse (FAKTOREN1.SAV) (2)

Abbildung 11.33: Um zwei Variablen erweitere Datendatei FAKTOREN1.SAV im Rahmen der Faktorenanalyse

Abbildung 11.34: Variablenveränderungen in FAKTOREN1A.SAV

Abbildung 11.35: Anfordern und Erstellen einer Tabelle in SPSS (FAKTOREN1A.SAV)

Abbildung 11.36: Fallbezogene arithmetische Faktoren-Mittelwerte als Output einer Faktorenanalyse (FAKTOREN1A.SAV)

Abbildung 11.37: Anforderung eines Streudiagramms zwischen den beiden Faktoren einer Faktorenanalyse (FAKTOREN1A.SAV)

Abbildung 11.38: Unbearbeitetes Streudiagramm zwischen den beiden Faktoren (FAKTOREN1A.SAV)

Abbildung 11.39: Bearbeitetes Streudiagramm zwischen den beiden Faktoren: Linien einfügen (FAKTOREN1A.SAV)

Abbildung 11.40: Bearbeitetes Streudiagramm zwischen den beiden Faktoren: Datenbeschriftungen einfügen (FAKTOREN1A.SAV)

Abbildung 11.41: Beschriftetes und in den Nullpunkt verschobenes Streudiagramm zwischen zwei Faktoren im Rahmen der Faktorenanalyse (FAKTOREN1A.SAV)

Abbildung 12.1: Additive Zeitreihenverknüpfung: Konstante Saisonfigur (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley-VCH, S. 213)

Abbildung 12.2: Multiplikative Zeitreihenverknüpfung: Variable Saisonfigur (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley-VCH, S. 214)

Abbildung 12.3: Datendateien *ZEITREIHE_ADDITIV1.SAV* (links) und *ZEITREIHE_MULTIPLIKATIV1.SAV* (rechts)

Abbildung 12.4: Festlegung der Berechnung von gleitenden 4er-Durchschnitten für *variable_y* in *ZEITREIHE_ADDITIV1.SAV*

Abbildung 12.5: Berechnung von gleitenden 4er-Durchschnitten für *variable_y* in der neuen Variablen *variab_1* (*ZEITREIHE_ADDITIV1.SAV*)

Abbildung 12.6: Glättung der Zeitreihe von *variable_y* in *ZEITREIHE_ADDITIV1.SAV* mittels gleitender 4er-Durchschnitte

Abbildung 12.7: Definition der relevanten Datumskomponente für die Datendatei *ZEITREIHE_ADDITIV1.SAV*

Abbildung 12.8: Von SPSS in die Datendatei *ZEITREIHE_ADDITIV1.SAV* eingefügte Datumsvariablen *year_*, *quarter_* und *date_*

Abbildung 12.9: Anfordern einer additiven Zeitreihen-Komponentenzerlegung (*ZEITREIHE_ADDITIV1.SAV*)

Abbildung 12.10: Um weitere vier Variablen erweiterte Datendatei *ZEITREIHE_ADDITIV1.SAV* bei additiver Zeitreihen-Komponentenverknüpfung

Abbildung 12.11: Saisonziffern bei konstanter Saisonfigur in Bezug auf *variable_y* (*ZEITREIHE_ADDITIV1.SAV*)

Abbildung 12.12: Anfordern einer multiplikativen Zeitreihen-Komponentenzerlegung (*ZEITREIHE_MULTIPLIKATIV1.SAV*)

Abbildung 12.13: Um weitere vier Variablen erweiterte Datendatei *ZEITREIHE_MULTIPLIKATIV1.SAV* bei multiplikativer Zeitreihen-

Komponentenverknüpfung

Abbildung 12.14: Saisonziffern bei variabler Saisonfigur in Bezug auf *variable_y* in Prozent (*ZEITREIHE_MULTIPLIKATIV1.SAV*)

Abbildung 12.15: Datendatei *ÜBERLEBENSFUNKTION3.SAV*

Abbildung 12.16: Aufruf einer Sterbetafelanalyse (*ÜBERLEBENSFUNKTION3.SAV*)

Abbildung 12.17: Ergebnis einer Sterbetafelanalyse (*ÜBERLEBENSFUNKTION3.SAV*)

Abbildung 12.18: Aufruf der Kaplan/Meier-Methode (*ÜBERLEBENSFUNKTION3.SAV*) (1)

Abbildung 12.19: Aufruf der Kaplan/Meier-Methode (*ÜBERLEBENSFUNKTION3.SAV*) (2)

Abbildung 12.20: Ergebnisse einer Kaplan/Meier-Analyse (*ÜBERLEBENSFUNKTION3.SAV*) (1)

Abbildung 12.21: Ergebnisse einer Kaplan/Meier-Analyse (*ÜBERLEBENSFUNKTION3.SAV*) (2)

Abbildung 12.22: Aufruf einer Cox-Regression (*ÜBERLEBENSFUNKTION3.SAV*) (1)

Abbildung 12.23: Aufruf einer Cox-Regression (*ÜBERLEBENSFUNKTION3.SAV*) (2)

Abbildung 12.24: Ergebnisse einer Cox-Regression (*ÜBERLEBENSFUNKTION3.SAV*)

Abbildung 13.1: Beispiele für einen zweiseitigen Hypothesentest (oben) sowie für einen rechtsseitigen Hypothesentest (unten links) und einen linksseitigen Hypothesentest (unten rechts) (Quelle: Faik, Jürgen (2015): Statistik für Wirtschafts- und Sozialwissenschaftler, Weinheim: Wiley, S. 306–308)

Abbildung 13.2: Datendatei *BAUTEILE.SAV*

Abbildung 13.3: Anfordern eines t-Mittelwerttests für die Variable *größe* (*BAUTEILE.SAV*)

Abbildung 13.4: Ergebnis eines t-Mittelwerttests für *größe* (*BAUTEILE.SAV*) (Output-Auszug)

Abbildung 13.5: Datendatei *T-UNABHÄNGIG1.SAV*

Abbildung 13.6: Aufruf des t-Tests bei unabhängigen Stichproben mit *ergebnis* als Testvariable und mit *geschlecht* als Gruppierungsvariable (*T-UNABHÄNGIG1.SAV*)

Abbildung 13.7: Ergebnis des t-Tests bei unabhängigen Stichproben mit *ergebnis* als Testvariable und mit *geschlecht* als Gruppierungsvariable (*T-UNABHÄNGIG1.SAV*) (Output-Auszug)

Abbildung 13.8: Welch-Test bei unabhängigen Stichproben mit *ergebnis* als Testvariable und mit *geschlecht* als Gruppierungsvariable (*T-UNABHÄNGIG1.SAV*)

Abbildung 13.9: Datendatei *MWU-UNABHÄNGIG1.SAV*

Abbildung 13.10: Eingabe der Testvariablen (*leistungsklasse*) und der Gruppierungsvariablen (*geschlecht*) beim Mann/Whitney-U-Test (*MWU-UNABHÄNGIG1.SAV*)

Abbildung 13.11: Ergebnis des Mann/Whitney-U-Tests mit der Testvariablen *leistungsstärke* und der Gruppierungsvariablen *geschlecht* (*MWU-UNABHÄNGIG1.SAV*)

Abbildung 13.12: Datendatei *T-ABHÄNGIG1.SAV*

Abbildung 13.13: Aufruf des t-Tests bei abhängigen Stichproben mit *ergebnis2016* und *ergebnis2017* als »paarigen Variablen« (*T-ABHÄNGIG1.SAV*)

Abbildung 13.14: Ergebnis des t-Tests bei abhängigen Stichproben mit *ergebnis2016* und *ergebnis2017* als »paarigen Variablen« (*T-ABHÄNGIG1.SAV*)

Abbildung 13.15: Datendatei *WILCOXON-ABHÄNGIG1.SAV*

Abbildung 13.16: Aufruf des Wilcoxon-Tests bei abhängigen Stichproben mit *ergebnis2016* und *ergebnis2017* als Testvariablen (*WILCOXON-ABHÄNGIG1.SAV*)

Abbildung 13.17: Ergebnis des Wilcoxon-Tests bei abhängigen Stichproben mit *ergebnis2016* und *ergebnis2017* als Testvariablen (*WILCOXON-ABHÄNGIG1.SAV*)

Abbildung 13.18: Ergebnis des Vorzeichentests bei abhängigen Stichproben mit *ergebnis2016* und *ergebnis2017* als Testvariablen (*WILCOXON-ABHÄNGIG1.SAV*)

Abbildung 13.19: Datendatei *ANOVA_OHNE1.SAV*

Abbildung 13.20: Aufruf der einfaktoriellen ANOVA für *ergebnis* mit *qualifikation* als Faktor (*ANOVA_OHNE1.SAV*)

Abbildung 13.21: Zusätzlicher Aufruf der BONFERRONI-Korrektur bei der einfaktoriellen ANOVA für *ergebnis* und mit *qualifikation* als Faktor (*ANOVA_OHNE1.SAV*)

Abbildung 13.22: Ergebnis der einfaktoriellen ANOVA (inklusive Bonferroni-Korrektur) für *ergebnis* und mit *qualifikation* als Faktor (*ANOVA_OHNE1.SAV*)

Abbildung 13.23: Datendatei *KW-OHNE1.SAV*

Abbildung 13.24: Aufruf des Kruskal/Wallis-Tests bei unabhängigen Stichproben (hier in Bezug auf *KW-OHNE1.SAV*)

Abbildung 13.25: Aufruf des Kruskal/Wallis-Tests bei unabhängigen Stichproben mit *ergebnis* als Testvariable und *qualifikation* als Gruppierungsvariable (*KW-OHNE1.SAV*)

Abbildung 13.26: Ergebnis des Kruskal/Wallis-Tests bei unabhängigen Stichproben mit *ergebnis* als Testvariable und *qualifikation* als Gruppierungsvariable (*KW-OHNE1.SAV*) (Output-Auszug)

Abbildung 13.27: Datendatei *ANOVA_MIT1.SAV*

Abbildung 13.28: Aufruf von mehreren t-Tests bei Messwiederholung und drei Gruppenvergleichen (*ANOVA_MIT1.SAV*)

Abbildung 13.29: Ergebnis von drei t-Tests bei Messwiederholung und drei Gruppenvergleichen (*ANOVA_MIT1.SAV*) (Output-Auszug)

Abbildung 13.30: Datendatei *FRIEDMAN-MIT1.SAV*

Abbildung 13.31: Aufruf des Friedman-Tests mit den Testvariablen *ergebnis2015*, *ergebnis2016* und *ergebnis2017* (*FRIEDMAN-MIT1.SAV*)

Abbildung 13.32: Ergebnis des Friedman-Tests mit den Testvariablen *ergebnis2015*, *ergebnis2016* und *ergebnis2017* (*FRIEDMAN-MIT1.SAV*)

Abbildung 14.1: Datendatei *GLEICHVERTEILUNG.SAV*

Abbildung 14.2: Aufruf des Chi-Quadrat-Verteilungstests auf Gleichverteilung für die Variable *einkommensmitte* (*GLEICHVERTEILUNG.SAV*)

Abbildung 14.3: Ergebnis des Chi-Quadrat-Verteilungstests auf Gleichverteilung für die Variable *einkommensmitte* (*GLEICHVERTEILUNG.SAV*)

Abbildung 14.4: Datendatei *UNGLEICHVERTEILUNG.SAV*

Abbildung 14.5: Histogramm bezüglich der klassifizierten Variablen *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*) mit eingezeichneter Normalverteilungskurve

Abbildung 14.6: Ergebnis des Chi-Quadrat-Verteilungstests auf Gleichverteilung für die Variable *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*)

Abbildung 14.7: Aufruf des Chi-Quadrat-Verteilungstests auf Normalverteilung für die Variable *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*)

Abbildung 14.8: Ergebnis des Chi-Quadrat-Verteilungstests auf Normalverteilung für die Variable *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*)

Abbildung 14.9: Ergebnis des Chi-Quadrat-Verteilungstests auf Normalverteilung für die Variable *einkommensmitte* bei Fallzahlenverdoppelung (*UNGLEICHVERTEILUNG2.SAV*)

Abbildung 14.10: Aufruf des Kolmogorov/Smirnov-Verteilungstests auf Gleichverteilung für die Variable *einkommensmitte* (*GLEICHVERTEILUNG.SAV*)

Abbildung 14.11: Ergebnis des Kolmogorov/Smirnov-Verteilungstests auf Gleichverteilung für *einkommensmitte* (*GLEICHVERTEILUNG.SAV*)

Abbildung 14.12: Aufruf des Kolmogorov/Smirnov-Verteilungstests auf Gleich- und Normalverteilung für *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*)

Abbildung 14.13: Ergebnis des Kolmogorov/Smirnov-Verteilungstests auf Gleich- und Normalverteilung für *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*)

Abbildung 14.14: Aufruf des Kolmogorov/Smirnov- und des Shapiro/Wilk-Verteilungstests auf Normalverteilung für *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*)

Abbildung 14.15: Ergebnisauszug bezüglich des Kolmogorov/Smirnov- und des Shapiro/Wilk-Verteilungstests auf Normalverteilung für *einkommensmitte* (*UNGLEICHVERTEILUNG.SAV*) (Output-Auszug)

Abbildung 15.1: Aufruf des Chi-Quadrat-Tests (*GESCHL_FARBE.SAV*)

Abbildung 15.2: Chi-Quadrat-Kontingenzkoeffizententest für den Zusammenhang zwischen *geschlecht* und *augenfarbe* (*GESCHL_FARBE.SAV*)

Abbildung 15.3: Aufruf des Rangkorrelationskoeffizienten nach Spearman (*SPEERWERFEN.SAV*)

Abbildung 15.4: Test für den Zusammenhang zwischen *weite2016* und *weite2017* beim Spearman'schen Rangkorrelationskoeffizienten (*SPEERWERFEN.SAV*)

Abbildung 15.5: Aufruf des Korrelationskoeffizienten nach Bravais/Pearson (*SPEERWERFEN2.SAV*)

Abbildung 15.6: t-Test für den Zusammenhang zwischen *weite2016* und *weite2017* beim Bravais/Pearson-Korrelationskoeffizienten (*SPEERWERFEN2.SAV*)

Abbildung 16.1: Download-Möglichkeiten für die R-Essentials aus dem Internet

(https://github.com/IBMPredictiveAnalytics/R_Essentials_Statistics/releases)

Abbildung 16.2: ZIP-Datei für R-Essentials (SPSS 28, auch für SPSS 29 passend)

Abbildung 16.3: Desktop-Button für R 4.2.1

Abbildung 16.4: RStudio-Maske (hier für R 4.2.1)

Abbildung 16.5: In SPSS installierte (R- beziehungsweise) Python-Anwendungen

Abbildung 16.6: Suche nach in SPSS installierten R-Anwendungen

Abbildung 16.7: Im SPSS-Erweiterungshub (unter UNTERSUCHEN) gefundene R-Anwendung RREGRIDGE (Beispiel)

Abbildung 16.8: Voraussetzungen für die Funktion RREGRIDGE (Beispiel)

Abbildung 16.9: Abruf der R-Erweiterung RREGRIDGE (Beispiel)

Abbildung 16.10: Erfolgreiche Installation der R-Erweiterung RREGRIDGE (Beispiel)

Abbildung 16.11: Aufruf von RREGRIDGE (Beispiel) in SPSS

Abbildung 16.12: Installation eines R-Pakets (hier von MASS) innerhalb von SPSS

Abbildung 16.13: Koeffizientenwerte für den Regressionszusammenhang zwischen *ausgaben* und *einkommen* auf Basis einer robusten Regression und eines SPSS/R-Syntaxprogramms (*LINEAREREGRESSION_GERADE7.SAV*)

Abbildung 16.14: Koeffizientenwerte für den Regressionszusammenhang zwischen *ausgaben* und *einkommen* auf Basis einer linearen Regression und eines SPSS/R-Syntaxprogramms (links) beziehungsweise mittels SPSS (rechts; jeweils *LINEAREREGRESSION_GERADE7.SAV*)

Abbildung 17.1: Ausgangsdatensatz *ANGST1.SAV* vor der Umskalierung von bestehenden Variablen

Abbildung 17.2: Recodierter Datensatz (Umskalierung in denselben Variablen) (*ANGST1.SAV*)

Abbildung 17.3: Recodierter Datensatz (in neue Variablen) (*ANGST1.SAV*)

Abbildung 17.4: Klassenbildung in DO-REPEAT-Schleife (*DO-REPEAT2.SAV*)

Abbildung 17.5: Bildung von neuen Variablen mittels LOOP-Sequenz (*ZUFRIEDENHEIT1_TO_5.SAV*)

Abbildung 17.6: Datensatz mit vier Variablen vor LOOP-Prozedur (*LOOP_AUSGANG4_1.SAV*)

Abbildung 17.7: Datensatz mit zwei Variablen nach LOOP-Prozedur (*LOOP_ZIEL1_4.SAV*)

Abbildung 17.8: Datensatz nach DO-IF-Schleife (*DO-REPEAT3.SAV*)

Abbildung 17.9: Ausgangsdatensatz *REGION.SAV* mit einer Variablen vor einer DO-IF-Schleife

Abbildung 17.10: Datensatz *REGION.SAV* mit drei Variablen nach einer DO-IF-Schleife

Abbildung 17.11: Datendatei *ZUFRIEDEN.SAV* mit fehlenden Angaben

Abbildung 17.12: Datendatei *ZUFRIEDEN.SAV* mit fehlerhafter Zählvariablen *zufrieden_zahl*

Abbildung 17.13: Datendatei *ZUFRIEDEN.SAV* mit korrekter Zählvariablen *zufrieden_zahl2*

Abbildung 18.1: Installation des Makros PROCESS in SPSS

Abbildung 18.2: In SPSS eingebettetes Makro PROCESS

Abbildung 18.3: Datendatei *ZUFRIEDENHEIT1_TO_5.SAV* mit fünf kategorialen Variablen

Abbildung 18.4: Datendatei mit fünf kategorialen Variablen und aus ihnen abgeleiteten Dummies (*ZUFRIEDENHEIT1_TO_5.SAV*)

Abbildung 18.5: Datensatz *DO-REPEAT3.SAV* für Prozess-Schleifen-Makro

Abbildung 18.6: Output nach Prozess-Schleifen-Makro (*DO-REPEAT3.SAV*)

Abbildung 18.7: Erstellen eines Makrosymbols (1)

Abbildung 18.8: Erstellen eines Makrosymbols (2)

Abbildung 18.9: Erstellen eines Makrosymbols (3)

Abbildung 18.10: Erstellen eines Makrosymbols (4)

Abbildung 19.1: Datendateien *VERMOEGEN_A.SAV*, *VERMOEGEN_B.SAV*, *VERMOEGEN_C.SAV* und *VERMOEGEN_D.SAV*

Abbildung 19.2: Fiktive Armutsquoten für die Länder A, B, C und D (*VERMOEGEN_A.SAV*, *VERMOEGEN_B.SAV*, *VERMOEGEN_C.SAV* und *VERMOEGEN_D.SAV*)

Abbildung 19.3: (Personenbezogener) Ungleichheitsoutput (Gesamtvermögen) für die vier Länder A bis D (*VERMOEGEN_A.SAV*, *VERMOEGEN_B.SAV*, *VERMOEGEN_C.SAV*, *VERMOEGEN_D.SAV*)

Abbildung 19.4: (Personen- und haushaltsbezogener) Ungleichheitsoutput (Gesamt-, Geld- und Grundvermögen) für Land D (*VERMOEGEN_D.SAV*) (Auszug)

Abbildung 19.5: (Personenbezogener) Ungleichheitsoutput (Gesamtvermögen) für drei Regionen in Land D (*VERMOEGEN_D.SAV*) (Auszug)

# Dummies Junior – die frechen »... für Dummies« für interessierte Kids und Jugendliche

- » Projekte zum Ausprobieren, Programmieren und Experimentieren
- » Mit pädagogischem Konzept
- » Viele Abbildungen
- » Verständliche Texte
- » In Workshops erprobt

C. Ermel, N. Rosenfeld

**Spaß mit Elektronik für Dummies Junior**

1. Auflage 2020 **ISBN:** 978-3-527-71705-7

198 Seiten

**Format:** 176 mm x 240 mm

**Ladenpreis:** 15,– €*

In diesem Buch lernst du, Schaltungen für coole Gadgets aufzubauen: eine Glückwunschkarte, die leuchtet, eine blinkende Weihnachtsbaumkugel, einen klingenden Draht und anderes mehr.

M. Ponce Kärgel

**Hörspiel und Podcast selber machen für Dummies Junior**

1. Auflage 2020 **ISBN:** 978-3-527-71704-0

192 Seiten

**Format:** 176 mm x 240 mm

**Ladenpreis:** 16,– €*

Sein eigener Produzent sein, wer will das nicht? Steig ein in die Welt der Hörspiele und Podcasts. Dieses Buch zeigt dir Schritt für Schritt, wie du deine eigenen Audiobeiträge produzieren und veröffentlichen kannst.

U. Schmid, K. Weitz, M. Siebers

## Künstliche Intelligenz selber programmieren für Dummies Junior

1. Auflage 2019 **ISBN:** 978-3-527-71573-2

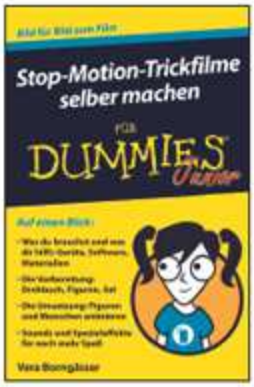
134 Seiten

**Format:** 140 mm x 216 mm

**Ladenpreis:** 12,99 €*

Mit diesem Buch verstehst du, was Künstliche Intelligenz ist. Du findest heraus, ob es Künstliche Intelligenz bereits gibt. Und das Beste: Mit Hilfe kleiner Programme erkennst du, wie durch Computerprogramme Künstliche Intelligenz entsteht

V. Borngässer




## Stop-Motion-Trickfilme selber machen für Dummies Junior

1. Auflage 2018 **ISBN:** 978-3-527-71484-1

144 Seiten

**Format:** 140 mm x 216 mm

**Ladenpreis:** 11,99 €*

Schritt für Schritt zum eigenen Stop-Motion-Video mit der richtigen Beleuchtung, krassen Geräuschen und Spezialeffekten. Hier erfährst du, wie es geht.

N. Bergner, Th. Leonhardt

## Eigene Apps programmieren für Dummies Junior

2. Auflage 2019 **ISBN:** 978-3-527-71596-1

136 Seiten

**Format:** 140 mm x 216 mm

**Ladenpreis:** 11,99 €*

Hast du Lust, deine eigene App zu programmieren, die dann auch wirklich auf einem Android-Smartphone läuft? Mit dem kostenlosen App Inventor ist das gar nicht schwer.

* Der €-Preis gilt nur für Deutschland. Preisänderungen und Irrtümer vorbehalten.

# Stichwortverzeichnis

## *A*


Ablehnungsbereich 458

ADD-FILES-Befehl 162

AGGREGATE-Befehl 170, 571

- BREAK-Unterbefehl 571
- MODE-Unterbefehl 180

Ähnlichkeitsmaß 406

Alternativhypothese 457

Analysieren 38, 157

ANOVA 333

ANOVA, einfaktorielle *siehe* Varianzanalyse, einfache 476

Ansicht 38

Anti-Image-Tabelle 420

Anwendungsdatei 35

Arbeitsdatei 179

ASCII-Daten 58

Assoziation 282

Assoziationsmaß 48, 284, 507

Asterisk 154

Ausgabeblock 197

Ausgabefenster 139, 194

Ausreißer 223, 326

Ausreißer-Analyse 223

Ausrichtung 108

Autokorrelation 320, 360, 380

# *B*

Bartlett-Test *siehe* KMO-Test 420
Bearbeiten 37
Befehl 139, 143
    wiederkehrender 136
Befehlsabfolge 136
Befehlsname 143
Befehlsprozedur 154
Befehlsschlüsselwort 143
Befehlssyntax 154
Benutzerhandbuch 39
Benutzersteuerung 135
Beobachtungspunkt 321
Beschriftung 107
Bestimmtheitsmaß 331, 375, 389
    Cox/Snells 389
    korrigiertes 332, 375
    McFaddens 389
    multiples 375
    Nagelkerkes 389
Betriebssystem 35
Bonferroni-Korrektur 479
BOOTSTRAP-Befehl 313
Bootstrapping 248, 252, 256, 261, 263, 287, 309, 364
Boxplot 225

Breakvariable 90, 171

# C

Chi-Quadrat-Teststatistik 509
Cluster 407
Clusteranalyse 406
    AIC-Kriterium 409
    BIC-Kriterium 409
    hierarchische 415
    K-Mittelwerte- 411
    zweistufige 408
Clusterkriterium 408
Codierplan 98
Codierung 51
Codierung, lokale 68
COMPUTE-Befehl 167, 201, 571
CORRELATIONS-Befehl 313
COUNT-Befehl 172, 538, 577
CROSSTABS-Befehl 269
CTABLES-Befehl 272

# D

Datei 37

Daten 38, 43
aggregieren 170
diskrete 45
eingeben 53
eingelesen 44
invertieren 159
qualitative 49
quantitative 49
rechtszensierte 447
sortieren 80
stetige 45
Datenaggregation 89
Datenansicht 37, 53, 98
Dateneditor *siehe* Datenansicht 53
Datenerhebung 44, 53
Datenfile 40
Dateninterpretation 44
Datenmanagement 157, 172
Datenskalierung 45
metrische 45, 318
nominale 45
ordinale 45
Piktogramm 108
Datenverarbeitung 44
Datumskomponente 441, 445
Definitionsbefehl *siehe* Transformationsbefehl 157
Dendrogramm 406, 416

Dependenzanalyse 398
DESCRIPTIVES-Befehl 572
Dezimalstelle 106
Diagramm 204, 227
    Balken- 227
    Kreis- 228
    Stem-Leaf- 225
Diagrammeditor 427
Dialogbox *siehe* Dialogfeld 125
Dialogfeld 36, 140, 215
Diskriminanzanalyse 397
Diskriminanzfunktion 398, 401, 404
Diskriminanzkriterium 398
d-Koeffizient 299
DO-IF-Befehl 535, 572
DO-REPEAT-Schleife 528
Durbin/Watson-Test 361
Durchschnitt *siehe* Mittelwert, arithmetischer 47
Durchschnitt, gleitender 437

# *E*

Eigenwert 420
Eingangsbildschirm 36
Einheit, statistische 43
ELSE-Befehl 572
Entropiemaß, Theils 558
ERASE-FILE-Befehl 179

Erwartungswert 320
Erweiterungen 39
EVIEWS 34
EXCEL 71
EXECUTE-Befehl 154, 572
Extension 40
Extras 38

# *F*

Faktorenanalyse 418
Fall
    doppelter 131
    einfügen 54
    löschen 55
Fallgewichtung 120
Fallstricke, SPSS-Befehle 139
Fall-Zusammenfügung 77
Fehleingabe 221
Fehlermeldung 201
Fehlerwert 320
Fenster 39
Filter 123, 166
FILTER-BY-Befehl 167, 572
FILTER-OFF-Befehl 169, 572
Filtervariable 167
FLIP-Befehl 160
Format 59

Formel, mathematische *siehe* Funktion, statistische 119
FORTRAN 98, 174
FREQUENCIES-Befehl 573
Friedman-Test 487
Funktion, statistische 233

# G

Gamma-Index 299
Gaußklammer 247
GET FILE-Befehl 157
GET-FILE-Befehl
    DROP-Unterbefehl 158
    KEEP-Unterbefehl 157
Gini-Koeffizient 557
Gleichung 337
    nichtlineare 337
Gliederungsbereich *siehe* Navigationsbereich 199
Grafik *siehe* Diagramm 206

# H

Häufigkeit 45, 214
    absolute 46, 214
    beobachtete 270
    erwartete 272
    kumulierte 214
    relative 46, 214

Häufigkeitstabelle 216, 225
eindimensionale 214
Hauptkomponentenmethode 420
Hazardrate 447, 449
Heteroskedastizität 320, 366, 380
Hilfe 39
Hinweis 195
Histogramm 229, 373
Homoskedastizität 320, 366

# *I*

IF-Befehl 163
INCLUDE-Befehl *siehe* INSERT-Befehl 549
Index-Loop 545
Indifferenztabelle 509
Information 44
Informationstechnologie 32
Informationsverarbeitung 31
Inhaltsbereich 200
INSERT-Befehl 549, 578
Installationsprozess 35
Interaktionsterm 354
Interdependenz 316
Interpolation, lineare 248
Interquartilbereich 240
Irrtumswahrscheinlichkeit 458

# J

Journaldatei 153

# K

Kaplan/Meier-Methode 449
Klassifizierungsfunktion 401
Kleinst-Quadrate-Methode 321, 374, 436
KMO-Test 420
Kollinearität *siehe* Multikollinearität 380
Kommentar 146, 154
Kommentierung *siehe* Kommentar 143
Kommunalität 423
Komponente, glatte 434
Konditionsindex 383
Konfidenzintervall 363
Kontingenz *siehe* Assoziation 282
Kontingenzkoeffizient
    C- 285
    Chi-Quadrat- 284, 508
    C-, korrigierter 285
    Phi- 285
Korrelation 288, 315
    bivariate 376
    Nonsens- 290, 316
    verdeckte 308
Korrelationskoeffizient, Bravais/Pearsons 48, 302, 472, 512

Korrelationsmaß 507
Korrelationsmatrix 137
Kovarianz 304
Kreuztabelle 268, 508–509
Kruskal/Wallis-Test 481
Kurtosis *siehe* Wölbung 242

# L

Levene-Test 466, 477
Likelihood-Funktion 389
LIMDEP 34
LIST-Befehl 172
List-Processing-Loop 547
Lizenzierung 35
LOOP-Befehl 532

# M

Makro 541, 555
Makroexpansion 545
Mann/Whitney-U-Test 467
Matchen, statistisches 161
MATCH-FILES-Befehl 161
Maximum 240, 244
Maximum-Likelihood-Methode 254, 388, 409
Median 47, 240–241, 457
Menüleiste 37
Merkmal 44

Merkmalsausprägung 44
Merkmalsträger 44
Messniveau 108
Minimum 240, 244
Mittelwert
    arithmetischer 47, 234, 240–241, 457
    geometrischer 47, 240
    harmonischer 48, 240
Modus 47, 240–241
M-Schätzer 254, 525
Multikollinearität 380

# *N*

Navigationsbereich 196
Nichtablehnungsbereich 458
NMISS-Befehl 539–540
NONPAR-CORR-Befehl 313
Nonsenskorrelation 35, 308, 316
Normalverteilung 48, 231, 241, 321, 372
Nullhypothese 457, 491, 508

# *O*

Odd ratio 389
Operation 143
Ordinary Least Squares *siehe* Kleinst-Quadrate-Methode 322
Outputfile 40, 193
Outputmodus 139

# *P*


P/P-Diagramm 373

Paar 295

- diskordantes 295
- konkordantes 295

Parameter 318, 544

Perzentil 240

Phasendurchschnittsverfahren 440

Pivottabelle 202, 577

Polynom 338

PROCESS 542

Prognose 316, 357

Prognoseziel 398

Prozedur, statistische 233

- Datenanalyse, explorative 253
- DESCRIPTIVES 250
- Fallzusammenfassung 261
- FREQUENCIES 241
- Mittelwertvergleich 257

Prüfbereich, kritischer 458

Pseudo-Bestimmtheitsmaß *siehe* Bestimmtheitsmaß 389

PSPP 35

Punktewolke 321

Python 136, 517, 520

Python-Essentials 520

# Q

Quantil 240
Quelldatei 112

# R

R 34, 136, 517
Randverteilung 271
Rangkorrelationskoeffizient 288
Rangkorrelationskoeffizient, Spearmans 48, 288, 510
RATS 34
Rechteckdatensatz 53
RECODE-Befehl 163, 573, 577
Record Linkage 162
Regel, mathematische 120

Regression 315, 317, 337, 374
bivariate 316
Cox- 452
Einfach- 317
gewichtete 369
lineare 317, 524
lineare, Voraussetzungen 318
Logit- 387
multinomiale logistische 394
multiple lineare 317, 374
multiple lineare, Voraussetzungen 374
ordinale 392
quasilineare 337
robuste 524
Regressionsberechnung
EINSCHLUSS 377, 390
ENTFERNEN 379, 390
RÜCKWÄRTS 379, 390
SCHRITTWEISE 377, 390
VORWÄRTS 380, 390
Regressionsfunktion 337, 339, 342–343
Exponential- 342
logistische 344
Potenz- 339
quadratische 337
RENAME-Befehl 179
Residuenanalyse 365

Residuum 365
R-Essentials 517–518
Rolle 109
Rotation 421
R-Paket 524
RStudio 34, 519

# S

Saisonfigur 440
    konstante 435, 440
    variable 435, 443
Saisonkomponente *siehe* Saisonfigur 434
Saisonziffer 440, 443–444, 446
SAMPLE-Befehl 173
SAS 34
SAS-Datendatei 72
Satzlänge, feste 66
SAVE-OUTFILE-Befehl 159
Scatterplot 323
Schaltfläche 550
Schiefe 48, 241–242
Schleife 527
Schlüsselvariable 81
Schwerpunkt 320
Screeplot 420
SELECT-IF-Befehl 166, 572–573
Signifikanz 213, 321, 333, 375, 461

Skalierungsniveau *siehe* Datenskalierung 46
Slash 154
SORT-CASES-Befehl 169
Spaltenbreite 108
Spannweite 48, 241, 243
Spezifikation 143
Spiel-Datenfile 142
SPLIT-FILE-Befehl 169
Sprungstelle 329
SPSS-Datenfile 65
Stamm-Blatt-Diagramm *siehe* Stem-Leaf-Diagramm 225
Standardabweichung 48, 241–242
STATA 33
STATA-Datendatei 72
Statistik 31
    als wissenschaftliche Disziplin 31
    deskriptive 213
    induktive 213
Statistik, beschreibende *siehe* Statistik, deskriptive 250
Statistik, schließende *siehe* Statistik, induktive 24
Statistikauswertung 238
Statistikprogramm 32–33
Statistikprozedur siehe Prozedur, statistische 155
Statistiksoftware 31
Statusleiste 37
Sterbetafelanalyse 447
Stichprobenziehung 61

Störgröße *siehe* Störterm 380
Störterm 375
Streudiagramm 321, 427
Summe 238
Syntax 135
Syntaxeditor 135, 151
Syntaxfile 40
Syntaxprogramm 135
SYSFILE-INFO-Befehl 174

# *T*

Tabelle 426
Tabellenkalkulationsprogramm 32
Tau-Koeffizient, Kendalls 295
TEMPORARY-Befehl 166, 573
Test
    statistischer 375
Test, statistischer 333, 457–459
    einseitiger 459
    Grundschema 459
    linksseitiger 459
    nichtparametrischer 460
    parametrischer 460
    rechtsseitiger 459
    zweiseitiger 458
    Zweistichproben- (abhängig) 469
    Zweistichproben- (unabhängig) 463

Teststatistik 458, 492, 501
Testvoraussetzung
    Chi-Quadrat-Kontingenztest 508
    Friedman-Test 487
    Kruskal/Wallis-Test 481
    Mann/Whitney-U-Test 467
    t-Test (abhängig) 470
    t-Test (unabhängig) 464
    Varianzanalyse mit Messwiederholung, einfache 485
    Varianzanalyse ohne Messwiederholung, einfache 476
    Vorzeichentest 475
    Wilcoxon-Test 473
Textassistent 58
Texteditor 136
Textfile 57
Tie 293
TITLE-Befehl 176
Token 544
Toleranzwert 382
Transformation 51, 110, 112, 337
    Variablen- 118
Transformationsbefehl 154, 157
Trend 316, 357, 434–435
Trennzeichen 57

t-Test 460
eine Variable 460
Zweistichproben- (abhängig) 470
Zweistichproben- (unabhängig) 463

# *U*

Überlebensfunktion 447, 449
Überlebensrate 447, 450
Überschreitungswahrscheinlichkeit *siehe* Signifikanz 466
Umcodieren 110
neue Variable 112
selbe Variable 110
Umskalierung 528
Unabhängigkeit, statistische 272
Unicode 66
Unterbefehl 154
Ursache-Wirkungs-Zusammenhang 308, 315
USE-ALL-Befehl 169

# *V*

V, Cramers 285
VALUE-LABELS-Befehl 165, 577

Variable 49, 143, 318
abhängige 318
Datums- 49
Dummy- 51, 318, 349, 388
latente 49
manifeste 49
numerische 49
qualitative 50
quantitative 50
String- 49
unabhängige 317
Zeit- 49
VARIABLE-LABELS-Befehl 165, 576
Variablenansicht 37, 54, 63, 96
Variablenbreite 106
Variableneditor *siehe* Variablenansicht 54
Variablenetikett 165
Variablenlabel *siehe* Variablenetikett 215
Variablenname 100
Variablentyp 103
Variablenzusammenfügung 81
Varianz 48, 241–242
Varianzanalyse, einfache
mit Messwiederholung 485
ohne Messwiederholung 476
Varianzanteil 383
Varianz-Inflations-Faktor 382

Variationskoeffizient, normierter 557
VECTOR-Befehl 532
Venn-Diagramm 381
Verteilungsannahme 491
Verteilungstest 491
    Chi-Quadrat- 498
    Kolmogorov/Smirnovs 500
    Shapiro/Wilks 505
Verzerrung 256, 287, 311
Viewer *siehe* Ausgabefenster 446
Vorzeichentest 475

# *W*

Warnung 66, 141, 195, 200
WEIGH-OFF-Befehl 178
WEIGHT-BY-Befehl 178, 574
WEIGHT-OFF-Befehl 574
Welch-Test 466
Wert 107
    fehlender 51, 107
    fehlender benutzerdefinierter 52
    fehlender systemdefinierter 52
    systemfehlender 52
Werteetikett 165
Wertelabel *siehe* Werteetikett 137
Wilcoxon-Test 473
WLS-Ansatz *siehe* Regression, gewichtete 369

Wölbung 48, 241–242

# Z

Zählvariable 172, 532, 538, 545
Zeit 316
Zeitreihe 356, 433
    saisonbereinigte 440, 443
Zeitreihenzerlegung
    additive 434, 440
    multiplikative 435, 444
Zensierung 447
Zieldatei 112
Zufallskomponente 434
Zufallsstichprobe 62, 128, 319
Zusammenhang
    bivariater 268
    multivariater 268

# WILEY END USER LICENSE AGREEMENT

Besuchen Sie www.wiley.com/go/eula, um Wiley's E-Book-EULA einzusehen.