UNIVERSITÄT OSNABRÜCK

Einführung in die statistische Daten-analyse mit SPSS* für Geographinnen und Geographen

Programm-version 20

Stand: 13.10.2014

*SPSS ist ein eingetragenes Warenzeichen der International Business Machines Corporation (IBM)

von
Carsten Felgentreff
Frank Westholt

GEOS
Institut für Geographie
Universität Osnabrück

Osnabrück, 2014

Institut für Geographie
Seminarstraße 19 a/b
D-49074 Osnabrück

ABBILDUNGSVERZEICHNIS

VORWORT

Die Umstellung auf Bachelor-Studiengänge hat es mit sich gebracht, dass die Veranstaltungen zur Einführung in die Statistik vielerorts einsemestrig abgehalten werden müssen. Für die Studierenden des 2-Bach-Bachelors in Osnabrück legt die Modulbeschreibung für die Lehreinheit Geographie recht umfassende Qualifikationsziele fest:

„Im methodischen Basismodul Fachmethodik I sollen die Studierenden kritische Vertrautheit mit ausgewählten Methoden der deskriptiven und schließenden Statistik erlangen:
- Einblick in Rolle und Stellung statistischer Verfahren in der Geographie
- Kenntnis der Möglichkeiten und Grenzen sowie Stärken und Schwächen der verschiedenen Verfahren
- Fähigkeit, die erlernten Kenntnisse mit Hilfe von Programmsystemen umzusetzen und anzuwenden
- Befähigung zur Beurteilung von Ergebnissen quantitativer Forschung sowie zur Methodenauswahl bei eigenen Untersuchungen

Methodenkompetenzen: Informationsgewinnung und –verarbeitung speziell quantitativer Daten, IT-Kompetenz, kritisches Methodenbewusstsein
Sozialkompetenzen: Kommunikationskompetenz
Selbstkompetenzen: Leistungsbereitschaft, Zuverlässigkeit, Genauigkeit“
(Modulbeschreibungen für die Lehreinheit Geographie vom 5. Juni 2014, S. 9-10).

Am Ende dieses Semesters sollen die Absolventen also versiert sein in möglichst vielen statistischen Verfahren der quantifizierenden empirischen Forschung, Daten interpretieren können, Datenanalysen verstehen, kritische Vertrautheit erlangt haben mit den Fallstricken der Teststatistik – und das ungeachtet überaus heterogener Lerngruppen, in denen einzelne seit vielen Semestern im Zweitfach Mathematik studieren, andere hingegen felsenfest davon überzeugt sind, noch niemals in ihrem Leben einem Summenzeichen begegnet zu sein. Es gibt Berührungsängste aller Art; die Motivation, sich auf mathematische Beweisführungen einzulassen, tendiert gegen Null und nicht nur Lehramtskandidaten fragen sich und ihre Umgebung (halblaut), ob sie den Anpassungstest auf Normalverteilung wirklich noch einmal in ihrem Leben brauchen werden.

Es liegt nahe, dass zur Zielerreichung mehrgleisig gefahren werden muss. In der Vorlesung steht die Vermittlung wesentlicher Grundbegriffe und -prinzipien im Vordergrund. Das geht nicht ohne einige Formeln, gerechnet wird mit dem Taschenrechner – wie in der Abschlussklausur. Uni- und bivariate Verteilungen werden charakterisiert, Zusammenhangsmaße vorgestellt, Grundlagen der Schätzstatistik vermittelt und die Grundprinzipien statistischer Test an verschiedenen Beispielen veranschaulicht, häufig mit Übungsaufgaben verbunden. Die Teilnehmerinnen und Teilnehmer erhalten nicht nur das der Veranstaltung zugrundeliegende Vorlesungsskript ausgehändigt, sondern auch wöchentlich eine Vielzahl von Powerpoint-Folien, die den Stoff der Vorlesung visualisieren.

Parallel findet wöchentlich eine Übung im Computerpool statt. Hier besteht die Möglichkeit, in kleineren Lerngruppen den Stoff der Vorlesung zu rekapitulieren. Zugleich wird hier mit dem Statistikprogramm SPSS gearbeitet, mit dem viele Absolventen auch in der späteren Berufspraxis konfrontiert werden. Die diesbezüglichen Lern- und Lehrprozesse zu unterstützen ist Anliegen des vorliegenden Skripts. Dabei wird an verschiedenen Stellen weiter ausgeholt als in der zur Verfügung stehenden Unterrichts- und Kontaktzeit möglich, anderes mag – für sich betrachtet – wie ein Kochrezept erscheinen, das im wesentlichen aus der Nennung der notwendigen Klicks in der Menüoberfläche der Programmumgebung besteht. Zusammengenommen mit den Ausführungen in der Vorlesung, dem separaten Vorlesungsskript sowie den mündlichen Erläuterungen und Praxisbeispielen in der Übung soll es nicht nur das Schritt-für-Schritt-Nachvollziehen der Analyse ermöglichen, sondern zugleich die umfassende Einsicht in die behandelten Denkweisen und Methoden. Auch Teilnehmer mit lückenhaften mathematischen Vorkenntnissen sollen begreifen können, was sich hinter Computeralgorithmen verbirgt, bevor sie sich auf ein Computerprogramm verlassen. Was genau eine Standardabweichung (etwa in der Einheit ‚Quadratsekunden'!) ist, ist zwar immer noch schwer vorstellbar, wenn man sie mit dem Taschenrechner berechnet hat, aber dem Verständnis, was wie in diese Kennzahl einfließt, wofür sie steht und was sie ausdrückt, ist die Beherrschung des Rechenweges allemal zuträglich.

Für ein solches mehrgleisiges Vorgehen bei der Vermittlung des Lernstoffs sprechen nicht zuletzt auch die Erkenntnisse der experimentellen Lernforschung, der zufolge der nur gehörte Stoff hochgradig vergänglich ist. Die Lernkurve steigt, wenn der Stoff auch visuell vermittelt wird. Wird darüber zusätzlich diskutiert, steigt die Erfolgsrate weiter an. Am wirkungsvollsten ist aber die praktische Anwendung des Erlernten (Wellenreuther, 2010). Dieser will das vorliegende Skript Vorschub leisten.

Das Skript wäre nicht entstanden, wenn Frank Westholt, der mehrere Jahre als Lehrbeauftragter und Tutor zu besagter Veranstaltung fungierte, nicht beharrlich Übungsaufgaben aufbereitet, systematisch gesammelt und durch eigene Beispiele ergänzt hätte. In der Endphase kamen zudem die bemerkenswerten Fähigkeiten von Karin Schumacher und Dorit Heckeroth zum Tragen, deren Unterstützung die Fertigstellung wesentlich erleichtert hat. Den Genannten wie auch zahlreichen nichtgenannten Studierenden, die durch ihre Fragen und Kritik in besagten Veranstaltungen meine eigenen Einsicht in die Materie befördert haben, gebührt mein aufrichtiger Dank.

Carsten Felgentreff

Modulbeschreibungen für die Lehreinheit Geographie vom 5. Juni 2014. Universität Osnabrück: Fachbereich Kultur- und Geowissenschaften. Osnabrück. http://www-old.uni-osnabrueck.de/ordnungen/Modulbeschreibungen_Geographie_2014-06.pdf (09.10.2014)

Wellenreuther, M. 2010: Lehren und Lernen – aber wie? Empirisch-experimentelle Forschungen zum Lehren und Lernen im Unterricht. 5. Aufl. Schneider-Verl. Hohengehren: Baltmansweiler.

A Von der Forschungsfrage zum Datensatz

1 EINFÜHRUNG

1.1 DAS KONZEPT HINTER DIESER ANLEITUNG

„Ich glaube keiner Statistik, die ich nicht selbst gefälscht habe." Dieses Winston Churchill zugeschriebene Zitat wird oft angeführt, wenn es darum geht, unangenehme oder den eigenen Vorstellungen widersprechende statistische Auswertungen zu kritisieren. Diese Kritiker finden sich häufig in ihrer Meinung bestätigt, wenn sich ein Gutachten und das entsprechende Gegengutachten auf statistischem Wege widersprechen. Doch woher kommt diese Abneigung, wenn doch gerade die Statistik zu den zentralen Methoden gehört, mit denen Aussagen belegt werden? „Kein Monat vergeht ohne Politbarometer, Geschäftsklimaindex, Konjunkturprognosen, Konsumentenindex, etc. Viele Anleger vertrauen bei ihrer Geldanlage den Entwicklungsprognosen der Aktien im DAX und hoffen auf die Erfüllung der Prognosen der Finanzmarktökonometriker" (Cleff 2008, 1). Immer wieder wird Mangel an Statistikkenntnissen als Hauptgrund für Ablehnung oder Zurückhaltung gegenüber statistischer Verfahren vermutet. „Im Zeitalter von Standardsoftware, in dem prinzipiell ein Mausklick genügt, um eine Tabelle, eine Grafik oder sogar eine Regression zu erzeugen, wird dem Laien der Schritt zu komplizierten Anwendungen leicht gemacht. Nicht selten werden dabei Annahmen verletzt, Sachverhalte bewusst – also manipulativ – oder unbewusst verkürzt dargestellt" (Cleff 2008, 1). Krämer (2005, 10) fasst die Gründe für „falsche" Statistiken wie folgt zusammen: „Einige [Statistiken] sind bewusst manipuliert, andere nur unpassend ausgesucht. In einigen sind schon die reinen Zahlen falsch, in anderen sind die Zahlen nur irreführend dargestellt. Dann wieder werden Äpfel mit Birnen zusammengeworfen, Fragen suggestiv gestellt, Trends fahrlässig fortgeschrieben, Raten, Quoten oder Mittelwerte kunstwidrig berechnet, Wahrscheinlichkeiten vergewaltigt oder Stichproben verzerrt."

Diese Einführung hat den Anspruch, dem Leser nicht nur zu zeigen, wie man statistische Verfahren manuell auf der Tastatur ausführt, sondern soll auch erklären, wann welche Operation wirklich sinnvoll ist, um valide Ergebnisse zu erhalten. Das übergeordnete Ziel ist, nicht das simple „Knöpfe drücken" in den Mittelpunkt zu stellen, sondern dem Leser einen verständigen Umgang mit dem Statistikprogramm SPSS nahe zu bringen.

Es bleibt jedoch anzumerken, dass die folgenden Ausführungen kein vollständiges Handbuch zum SPSS-Programmpaket darstellen, sondern sich auf die Erläuterung einer Auswahl der gängigsten statistischen Operationen beschränken. Der Vorteil dieser Herangehensweise liegt in der Möglichkeit, die getätigte Auswahl in einer Breite und Anschaulichkeit zu behandeln, die auch von Anwendern als ausreichend empfunden wird, die sich nur punktuell und sporadisch mit statistischen Auswertungen auseinandersetzen. Diese Anleitung setzt jedoch voraus, dass der Anwender bereits über grundlegende Statistikkenntnisse verfügt. Aufgrund

der eingeschränkten Seitenzahl und der oftmals zum Scheitern verurteilten Bemühungen, komplizierte mathematische Prozeduren in kurzen Worten beschreiben zu wollen, wird auf mathematische Erläuterungen weitgehend verzichtet.

SPSS ist modular aufgebaut und verfügt über eine Reihe von Zusatzmodulen, die weitere Funktionen beinhalten und die Anwendung spezieller statistischer Methoden erlauben. Der Funktionsumfang des Basispaketes schließt jedoch bereits die wichtigsten Verfahren der deskriptiven und induktiven Statistik mit ein. Die vorliegende Anleitung bezieht sich deshalb auf das Basispaket von SPSS in der Programmversion 20. Genau wie in Windows kann auch in SPSS eine Aktion auf verschiedene Weise durchgeführt werden. Die folgenden Ausführungen beschränken sich auf die, aus Sicht der Autoren, zweckmäßigste Vorgehensweise.

Inhaltlich teilt sich diese Anleitung in drei Hauptkapitel:

A. Von der Forschungsfrage zum Datensatz
B. Deskriptive Datenanalyse
C. Prüfstatistik

Hauptkapitel 1 beschäftigt sich mit dem Themenkomplex „von der Datenerhebung über die Fehlerprüfung bis zur Berechnung neuer Variablen“. Im folgenden Hauptkapitel geht es dann um die gebräuchlichsten Methoden der deskriptiven Statistik inklusive tabellarischer- und grafischer Auswertungen, der Berechnung von Lage- und Streuungsparametern sowie der Darstellung und Berechnung verschiedener Korrelationsmaße. Das abschließende Kapitel C wirft einen Blick über den Tellerrand und beschäftigt sich mit Vergleichen zwischen Stichproben und der Grundgesamtheit sowie weiteren Methoden der Prüf- und Teststatistik. Die Inhalte entsprechen dabei in etwa den Inhalten einer Methoden-Veranstaltung im Fach Geographie an der Universität Osnabrück. Vor diesem Hintergrund eignet sich diese Anleitung besonders als ergänzende Lektüre zur entsprechenden Veranstaltung. Die konkreten Inhalte reichen dabei von der Erstellung eines Fragebogens bis zum fertigen Auswertungsbericht.

Zu Anfang jedes Kapitels werden die Inhalte kurz in Frageform notiert, um einen schnellen Überblick (auch während der eigenen Forschungsarbeit) zu gewährleisten. Außerdem finden sich nach der Erläuterung der verschiedenen Operationen einige Übungsaufgaben, in denen man das Gelernte selbstständig umsetzen kann.

Befehlsfolgen werden hervorgehoben dargestellt und durch Pfeile miteinander verbunden, um das spätere schnelle Auffinden zu gewährleisten. Ein Beispiel wäre etwa: ***Datei → Öffnen → Daten***. Auch Schaltflächen, Variablennamen, deren Merkmalsausprägungen und andere wichtige Begriffe und Fachausdrücke werden im Folgenden kursiv und fett geschrieben.

Zum Schluss bleibt noch anzumerken, dass die in dieser Anleitung angeführten „Regeln“ und Hinweise, die insbesondere in konzeptionellen Kapiteln zur Ausgestaltung der eigenen Untersuchung angesprochen werden, mehr als Richtlinien zu verstehen sind. Die konkrete Gestaltung der eigenen Untersuchung hängt ganz wesentlich von Determinanten wie dem eigenen Erkenntnisinteresse, der Zielgruppe und dem Forschungsinstrument ab.

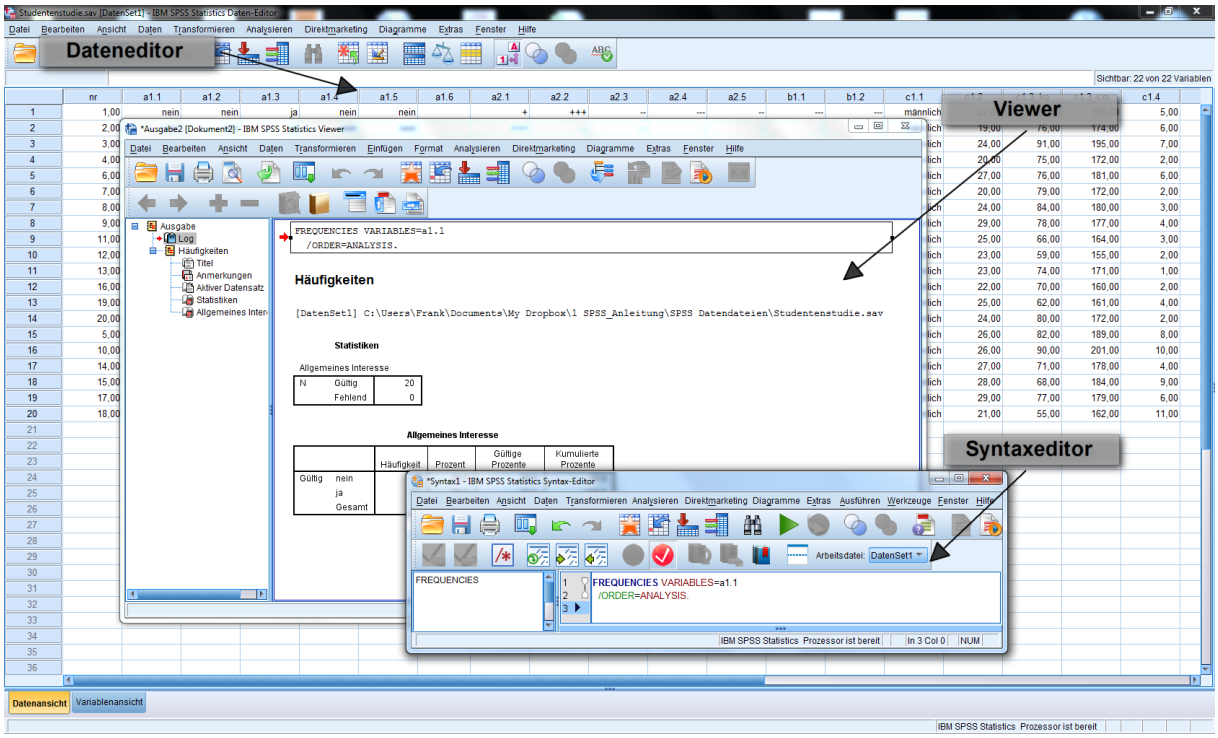
1.2 DIE ARBEITSOBERFLÄCHE VON SPSS

Mit welchen Ansichten arbeitet SPSS?

Wie sind diese aufgebaut?

In SPSS wird in der Regel mit drei Fenstern gearbeitet: dem Dateneditor, dem Syntaxeditor und dem Datenviewer (Ausgabefenster). Alle drei werden in Abbildung 1.1. dargestellt.

Abbildung 1.1: Die verschiedenen Fenster von SPSS



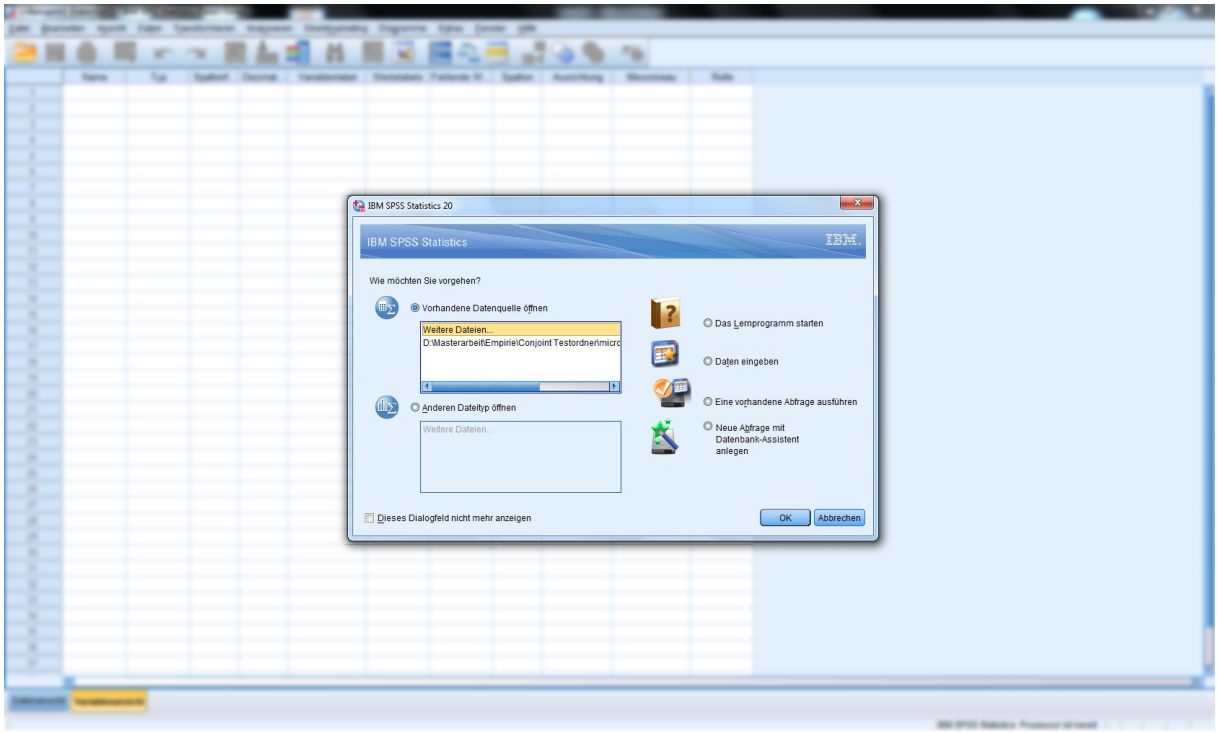
Dateneditor (erscheint immer bei Aufruf des Programms): Hier wird in der Datenansicht der Datensatz angezeigt und in der Variablenansicht die verschiedenen Variablen mit ihren jeweiligen Attributen. Beim Speichern von SPSS-Daten wird die Dateiendung .sav vorgegeben.

Viewer (erscheint automatisch nach einer Analyse): Hier werden Ergebnisse angezeigt. Dieses Fenster wird automatisch geöffnet, wenn eine Prozedur ausgeführt wird, die eine Ausgabe erzeugt. Viewer-Dateien besitzen die Endung .spv.

Syntaxeditor (Aufruf über Menü ***Datei → Neu → Syntax***): Hier werden Befehle eingegeben, die SPSS mitteilen, wie die Daten weiterverarbeitet werden sollen. Die Endung von Dateien mit SPSS-Befehlen lautet .sps. Die hierfür nötigen Kenntnisse der SPSS-Programmierung sind nicht Gegenstand der vorliegenden Einführung.

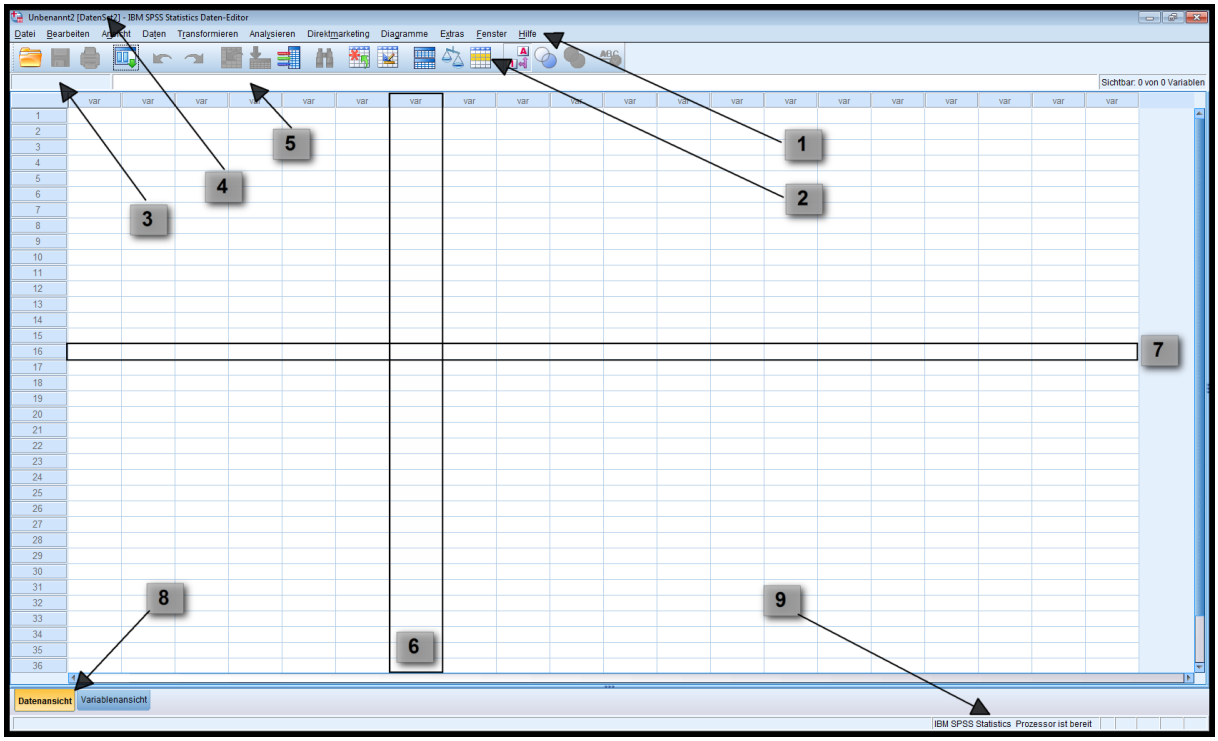
Wenn Sie SPSS das erste Mal starten, öffnet sich jedoch zunächst folgender Dialog (Abbildung 1.2).

Abbildung 1.2: SPSS-Startbildschirm



Neben dem durchaus empfehlenswerten Lernprogramm hat man hier verschiedene Optionen hinsichtlich der Neuerstellung bzw. des Öffnens einer bereits vorliegenden Datenquelle. Sie können Daten aus anderen Programmen importieren, eine leere Datenmaske öffnen oder eine bereits bearbeitete laden. Zunächst beschäftigen wir uns mit der manuellen Dateneingabe und wählen dazu die Option „***Daten eingeben***“. Es öffnet sich ein leerer SPSS-Dateneditor in der Registerkarte „***Datenansicht***“. Deren Aufbau und die verschiedenen Funktionsflächen veranschaulicht Abbildung 1.3.

Abbildung 1.3: Registerkarte ***Datenansicht***

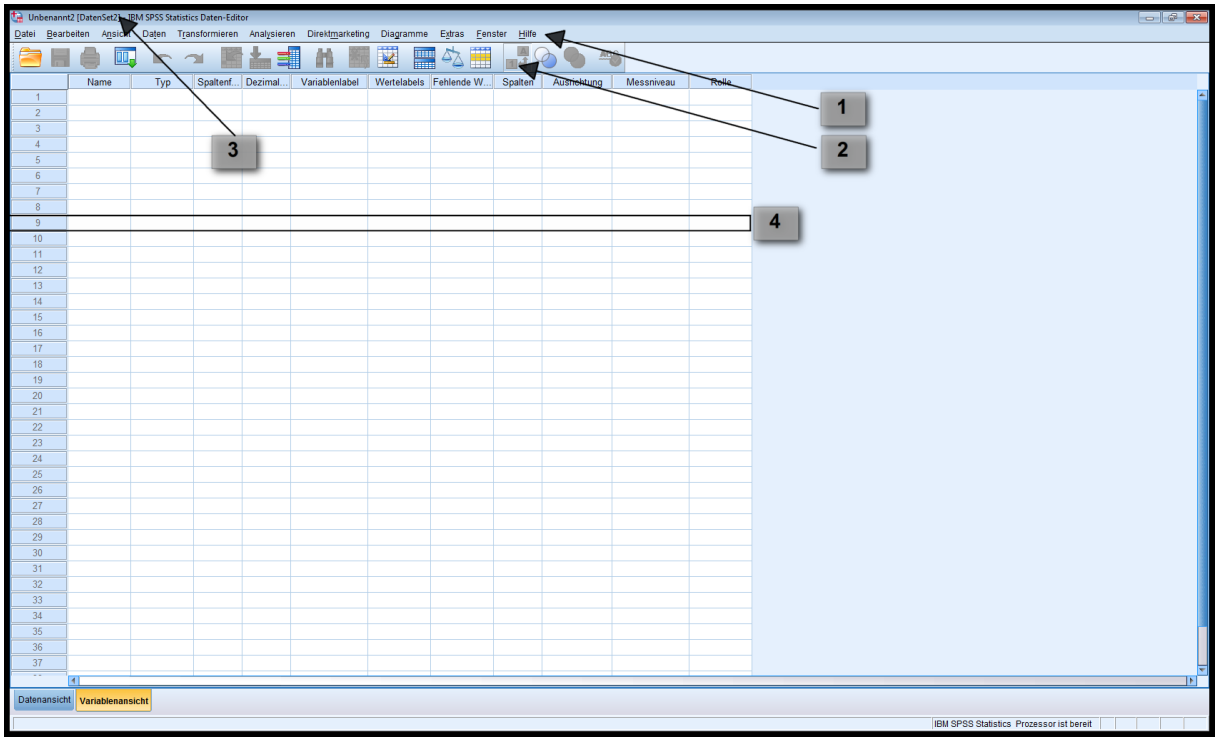


1	Menüleiste	**6**	Variable Nr. 7
2	Symbolleiste	**7**	Zeile oder Fall Nr. 16
3	Adressfeld	**8**	Registerkarten *Datenansicht* und *Variablenansicht*
4	Titelleiste		
5	Eingabezeile	**9**	Statuszeile

In der obigen Registerkarte lassen sich Datendateien erstellen oder öffnen, einsehen und ändern. Die Arbeitsoberfläche von SPSS erinnert zunächst an die gängige Datenansicht in Tabellenkalkulationen wie bspw. Microsoft Excel (Excel ist ein eingetragenes Markenzeichen Microsoft Corporation). Genau wie in Excel setzt sich die Tabellenoberfläche aus Zeilen und Spalten zusammen und wird auch als Spreadsheet bezeichnet. Die einzelnen Zeilen entsprechen den einzelnen Fällen. In einer typischen Befragung enthält Zeile 1 die Antworten von Person 1, Zeile 2 die von Person 2 usw. Die Spalten entsprechen dabei den verschiedenen Variablen bzw. den gestellten Fragen. Die Antworten auf Frage 1 würden also in der ersten Spalte eingetragen werden, die auf Frage 2 entsprechend in Spalte 2. Die einzelnen Zellen enthalten die Werte der jeweiligen Variablen des jeweiligen Falles, jede Zelle entspricht damit einem einzelnen Variablenwert, also der konkreten Merkmalsausprägung.

Neben der Datenansicht verfügt die Daten-Datei über eine weitere Registerkarte namens ***„Variablenansicht“*** am unteren linken Rand des Spreadsheets. Diese dient der Definition von Variablen und wird in Abbildung 1.4 dargestellt und erläutert.

Abbildung 1.4: Registerkarte ***Variablenansicht***



1	Menüleiste	**3**	Titelleiste
2	Symbolleiste	**4**	Eigenschaften von Variable Nr. 9

A VON DER FORSCHUNGSFRAGE ZUM DATENSATZ

2 DATENEINGABE UND DATENMODIFIKATION

Wie formuliert man Forschungsfragen?

Was verbirgt sich hinter dem Begriff Forschungsdesign?

Was muss man bei der Gestaltung eines Fragebogens beachten?

Welche Frage- und Antwortformen stehen zur Verfügung?

Was nützt ein Codeplan?

2.1 FORSCHUNGSFRAGE UND UNTERSUCHUNGSDESIGN

Am Anfang jeder Befragung steht die Formulierung des entsprechenden Fragebogens. Es lohnt sich, etwas Zeit in die Planung dieses Untersuchungsinstruments zu investieren, da man sich auf diese Weise viel Arbeit und Mühe in der Nachbereitung spart.

2.1.1 FORMULIERUNG EINER FORSCHUNGSFRAGE

Die wichtigste Regel ist, dass sich alle Unterfragen an den übergeordneten Forschungsfragen orientieren sollten. Man sollte keine Informationen erfragen, von denen man noch nicht weiß, ob man sie für die eigene Studie wirklich braucht. Die Probanden könnten auch ablehnend reagieren, sollten sie den Eindruck bekommen, man fischt willkürlich nach Daten.

Doch wie formuliert und konkretisiert man seine Forschungsfrage? Zunächst einmal ist es wichtig, sich nicht in der Mehrdimensionalität eines Themas zu verzetteln. Da sich empirisch zumeist nur Ausschnitte eines Themenfeldes erfassen lassen, bietet es sich an, sich auf die Fokussierung einer Teilfrage eines Themas zu beschränken. Je präziser sich die Vorüberlegungen zur zentralen Forschungsfrage gestalten, desto mehr werden Sie während der Bearbeitung des Themas davon profitieren. Im Folgenden werden einige Tipps zur Konkretisierung der Forschungsfrage dargestellt:

Entweder sollte sich Ihre Fragestellung deutlich von den Fragestellungen anderer empirischer Arbeiten unterscheiden oder sich konkret auf die bereits vorliegenden Ergebnisse beziehen, um diese zu überprüfen oder fortzuführen.

Empfehlenswert sind W-Fragen wie etwa wie? Warum? Wozu? Was? (etc.)

Forschungsfragen sollten zeitlich, räumlich und sachlich eingrenzbar sein.

Von widersprüchlichen, mehrdeutigen und missverständlichen Fragen, die evtl. sogar versteckte Behauptungen beinhalten, ist abzuraten.

Die Frageformulierung sollte möglichst präzise ausfallen, um eine konkrete Beantwortung zu ermöglichen.

Vorsicht bei zweifelhaften Vorannahmen, beeinflussenden oder tendenziösen Fragen und vor allem auch vor dem Gebrauch von unklaren Wörtern oder Konzepten.

Es bleibt anzumerken, dass die Ausarbeitung einer Forschungsfrage nicht nur von reinen untersuchungstechnischen Kriterien abhängig gemacht werden sollte. Auch eine ethische Bewertung sollte Beachtung finden.

Haben Sie ihre Forschungsfrage formuliert, so teilen Sie diese in weitere Unterfragen auf, deren Beantwortung die Basis der empirischen Erhebung darstellt. Diese können auch den Ausgangspunkt für eine weiterführende Literaturrecherche bilden.

2.1.2 FORMULIERUNG DER FORSCHUNGSHYPOTHESEN UND AUSWAHL DER STICHPROBE

Hat man sich für eine Forschungsfrage und verschiedene Unterfragen entschieden, sollte man sich der Formulierung von einer oder mehreren Forschungshypothesen widmen, die Sie zu untersuchen beabsichtigen. Hier unterscheidet man zwischen unspezifischen und spezifischen Hypothesen. Während eine unspezifische Hypothese nur behauptet, dass ein „irgendwie“ gearteter Effekt vorliegt und allenfalls noch die Richtung des Effektes angibt, konkretisiert eine spezifische Hypothese auch den Betrag des Effektes bzw. die Effektgröße. Je nach Stand der Forschung und Erkenntnisinteresse kann die Erarbeitung des Forschungsdesigns deshalb unterschiedlich angegangen werden:

Explorativ	Hier geht es um die Erkundung eher unbekannter Themenbereiche in Form von Vorstudien.
Deskriptiv	Hierbei handelt es sich um eine beschreibende Studie. Es geht dabei weniger um Ursachenforschung, sondern um die Schätzung von Merkmalen einer vorher definierten Grundgesamtheit.

Auch die Auswahl der Stichprobe ist von nicht zu unterschätzender Bedeutung. Dabei sollte die gezogene Stichprobe die Grundgesamtheit möglichst genau abbilden, denn je besser die ausgewählte Teilmenge die Grundgesamtheit repräsentiert, desto präzisere Aussagen lassen sich über diese machen.

Nachdem Sie Ihre Forschungsfragen und Hypothesen formuliert sowie sich für ein Stichprobenauswahlverfahren entschieden haben, können Sie mit der konkreten Ausarbeitung des Fragebogens beginnen.

Aufgabe 2.1
Welche Stichprobenarten sind Ihnen bekannt?

Aufgabe 2.2
Wann wird von einer repräsentativen Stichprobe gesprochen? Gibt es 100 % repräsentative Stichproben?

Aufgabe 2.3
Sie werden als Abteilungsleiter Datenerhebung einer Beratungsfirma mit gleich drei neuen Aufträgen betraut. Sie sollen die Qualität von Weinen in Orvieto prüfen, die Elastizität von Nylonstrümpfen in der Produktion testen und das Suchtverhalten Jugendlicher untersuchen. Entscheiden Sie jeweils zwischen einer Stichprobenauswahl und einer Vollerhebung. Begründen Sie Ihre Entscheidung.

2.2 ALLGEMEINE GESTALTUNGSRICHTLINIEN FÜR FRAGEBÖGEN

Ein optisch ansprechender und leicht auszufüllender Fragebogen ist für den Erfolg einer Befragung von großer Bedeutung. Dies gilt insbesondere, wenn der Fragebogen per Post oder E-Mail verschickt werden soll und die Befragten diesen selbstständig auszufüllen haben. Aber auch in der Hand eines Befragers macht ein ansprechender und durchdachter Fragebogen einen professionellen Eindruck auf die Befragten. Es lohnt sich außerdem, den Fragebogen in einzelne Themenbereiche aufzuteilen, die sich auch im Layout voneinander abheben, um mehr Übersicht zu schaffen. Außerdem haben sich folgende Gestaltungsrichtlinien in der Praxis bewährt:

Versehen Sie jeden Fragebogen mit einer Identifikationsnummer auf der ersten Seite. Das kann auch nach der eigentlichen Befragung gemacht werden. Nur so lassen sich evtl. fehlerhafte Eintragungen nachträglich überprüfen.

Man sollte den Fragebogen mit einfachen und interessanten Fragen beginnen. Diese erleichtern den Befragten den Einstieg.

Fragen zu demografischen Einzelheiten oder nach persönlichen Belangen sollten immer zum Ende des Fragebogens gestellt werden, da sich diese negativ auf die Antwortbereitschaft auswirken können.

Halten Sie die Anzahl der Fragen in einem verträglichen Rahmen. Zu viele Fragen können zu Interviewabbrüchen führen und die erhobenen Teildaten vielleicht wertlos machen.

Die einzelnen Fragen sollten nicht zu kompliziert oder zu lang formuliert werden. Kurze und prägnante Formulierungen erhöhen die Antwortrate und stellen sicher, dass die Fragen auch wirklich so verstanden werden, wie es das Erhebungsdesign vorsieht. Eine Höchstzahl von 20-25 Wörtern kann hier als Richtwert dienen.

Fragebögen sollten nicht überladen wirken. Planen sie deshalb freie Flächen im Layout ein.

Bei einem elektronisch oder postalisch verschicken Fragebogen empfiehlt sich ein Format in Form einer Broschüre mit einem etwas kleineren Format als das übliche DIN A4.

Nutzen Sie eine klare und einfache Schriftform mit gut lesbarer Schriftgröße. Dabei sollte man sich auf wenige Schriftarten beschränken, um das Layout nicht zu unruhig erscheinen zu lassen.

Geben Sie klare Anweisungen, was sie von den Befragten erwarten. Sind bspw. Mehrfachantworten erlaubt?

Ankreuzbare Kästchen oder Nummern zum Einkreisen bieten mehr Klarheit als einfache Linien.

Unterscheiden Sie die Fragen typografisch von den Antwortvorgaben. Beispielsweise kann man die Fragen in fetter Schrift und die Antworten in normalen Schrifttyp darstellen. Anweisungen könnten dann wiederum kursiv geschrieben werden. Bei Befragungen (mit zu schulenden!) Befragern sollten die Anweisungen für den Befragenden so einfach und deutlich wie möglich formuliert werden.

Bedrucken Sie den Fragebogen beidseitig.

Versehen Sie das Deckblatt mit einem aussagekräftigen Bild oder einer Grafik.

Man sollte den Fragebogen in thematische Unterabschnitte aufteilen, die mit eigenen Überschriften versehen sind.

Mehr als zwei Farben sollten für die Gestaltung des Fragebogens nicht verwendet werden. Dabei empfiehlt es sich, auf allzu kräftige oder strahlende Farben zu verzichten.

Verwenden Sie viele Fragen mit einer Likert-Skala (bspw. Bewertungsfragen von sehr gut bis mangelhaft), sollten nicht mehr als 10 Fragen dieses Typs untereinander darstellt werden, da ein solcher Überfluss auf Befragte ermüdend wirkt.

Übertragen sie die Codierungen aus dem Codeplan in den Fragebogen, um die manuelle Eingabe der Daten in SPSS zu erleichtern (siehe Kapitel 2.4).

Fragetext und dazugehörige Antwortmöglichkeiten sollten auf derselben Seite platziert werden.

Ordnen Sie die verschiedenen Textelemente so an, dass die Antwortmöglichkeiten aller Fragen in gerader Linie untereinander stehen.

Planen Sie genug Raum zur Beantwortung offener Fragen ein und achten Sie auf den Abstand zwischen evtl. angelegten Linien, um unterschiedlichen Handschriften gerecht zu werden.

Legen Sie dem Fragebogen einen Begleitbrief bei oder fassen Sie die vom Befragenden vorzutragenden Einleitungssätze auf der ersten Seite kurz zusammen.

Danken Sie den Befragten für deren Mitarbeit.

2.3 FRAGEFORMEN UND ANTWORTMÖGLICHKEITEN

Nach diesen eher grundsätzlichen Richtlinien zur Erstellung eines Fragebogens widmen wir uns in diesem Kapitel der Formulierung der einzelnen Fragen. Die zur Auswahl stehenden Fragetypen lassen sich hierbei grundsätzlich in drei Gruppen unterteilen: Einstellungen, Verhaltensweisen und charakteristische Eigenschaften. Einstellungen werden häufig durch Bewertungsfragen untersucht (Schulnoten, Bewertungen auf einer Skala zwischen 1 und 10). Beschreibungen des Verhaltens durch die Erfragung der Häufigkeit bestimmter Tätigkeiten und charakteristischer Eigenschaften umfassen zumeist demografische Angaben wie das Geschlecht, Alter und Einkommen.

Neben den Frageformen sollten auch die verschiedenen Antwortmöglichkeiten in der Untersuchungsplanung Beachtung finden. Man sollte sich hier fragen, welches Datenniveau die betroffenen Variablen aufweisen, ob geschlossene Antwortmöglichkeiten vorgegeben werden sollen oder freie Antworten der Befragten zugelassen sind, ob Mehrfachantworten zulässig sein sollen usw. Die Kategorisierung verschiedener Antwortmöglichkeiten lässt sich auf unterschiedliche Weise durchführen. Im Folgenden werden die wichtigsten Antwortoptionen kurz vorgestellt.

2.3.1 DATENNIVEAU DER VARIABLEN

Bei der manuellen Eingabe der Daten in SPSS ist insbesondere das Skalenniveau der Variable von Bedeutung (Datenniveau). Jede statistische Operation ist an bestimmte Datenniveaus gebunden. Will man die korrekte Auswertung durch SPSS sicherstellen, muss man jeder Variable das jeweilige Datenniveau zuordnen.

NOMINALES DATENNIVEAU

Bei einer Nominalskala stellen die Werte einer Variablen verschiedene Kategorien dar, die nicht in einer Rangreihenfolge geordnet werden können. Fragen nach dem Wohnort, dem Geschlecht oder einfache Ja/Nein-Fragen sammeln nominale Daten. Nominalskalierte Variablen sind in ihren Auswertungsmöglichkeiten deutlich eingeschränkt. Sie eignen sich vor allem zur Auszählung von Häufigkeiten und Darstellung in Kreisdiagrammen. Berechnungen von verbreiteten statistischen Kennzahlen wie dem arithmetischen Mittel oder der Standardabweichung sind hier nicht möglich. Sie eignen sich allerdings gut als Gruppierungsvariablen, indem die Stichprobe anhand der Kategorien dieser Variable in weitere Untergruppen unterteilt werden kann.

Eine spezielle Form der Antwortoptionen auf nominalem Datenniveau sind Variablen, die nur zwei Kategorien vorgeben. Wir bezeichnen sie als dichotome Merkmale. Fragen, die bspw. nur Ja/Nein-Antworten zulassen, weisen eine dichotome Skala auf. Ein Vorteil dieser Variablen liegt darin, dass sie als *erklärende* Variablen bei Regressionsanalysen Anwendung finden können.

ORDINALES DATENNIVEAU

Ordinale Daten werden, wie nominale Daten, häufig als „qualitative Daten“ bezeichnet. Im Unterschied zur Nominalskala lassen sich ordinalskalierte Daten in eine sinnvolle (nicht willkürlich oder beliebig geordnete) Rangreihenfolgen bringen. Beispiele sind kategorisierte Einkommensgruppen, Schuhgrößen oder klassische Bewertungsfragen (Schulnoten). Für ein ordinal skalierbares Merkmal bestehen Rangordnungen der Art „größer“, „kleiner“, „mehr“, „weniger“, „stärker“, „schwächer“ zwischen je zwei unterschiedlichen Merkmalswerten. Über die Abstände zwischen benachbarten Urteilsklassen lässt sich jedoch keine genaue Aussage treffen – dafür sind die Kategorien zu vage. Hier können die Codezahlen eine empirische Bedeutung gewinnen, indem sie eine Ordnungsrelation zulassen. Neben der Häufigkeitsauszählung und der Darstellung in Balken- und Kreisdiagrammen erlauben ordinale Daten die Berechnung gewisser statistischer Kennwerte wie bspw. die des Medians.

Häufig werden ordinalskalierte Antwortkategorien in Form einer Skala zur Einschätzung von Gefühlen und Meinungen verwandt. Eine Skala ist hierbei eine Liste von Antwortmöglichkeiten, die in auf- bzw. absteigender Reihenfolge geordnet sind (Vergabe von Schulnoten, Bewertung zwischen 1 bis 5, s*timme überhaupt nicht zu* bis s*timme voll zu*). Wenn man eine Skala verwenden will, muss man sich entscheiden, wie viele Kategorien die Skala aufweisen soll. Dabei sollten folgende Hinweise bedacht werden:

Je größer die Anzahl der Kategorien, desto eher lassen sich allgemeine lineare Modelle wie bspw. Regressionsanalysen auf diese Daten anwenden. Bei nur drei bis vier Kategorien ist deren Anwendbarkeit eher umstritten.

Haben die Befragten sich nicht bereits mit dem angesprochenen Thema auseinandergesetzt, werden diese keine sehr differenzierten Stellungnahmen abgeben können; entsprechend zurückhaltend sollten die Ergebnisse interpretiert werden.

Hat man sich dazu entschieden, ob die stärkste Antwortmöglichkeit der Kategorien an erster oder an letzter Stelle auf der Skala genannt werden soll, sollte man diese Festlegung über den gesamten Fragebogen hinweg beibehalten.

Die Meinungen über die Verwendung einer mittleren, neutralen Kategorie variieren stark. Wenn man versucht, die Befragten zu einer eindeutig negativen bzw. einer eindeutig positiven Bewertung zu „zwingen“, sollte man auf eine neutrale Alternative verzichten. Allerdings haben empirische Untersuchungen gezeigt, dass die ohne Verwendung einer neutralen Kategorie am häufigsten gewählte Alternative auch dann am häufigsten genannt wurde, wenn eine mittlere Alternative zur Verfügung stand. Das gilt auch für die zweithäufigste Kategorie usw. Häufig scheint es auch sehr plausibel, dass einige Personen bei bestimmten Themen die mittlere Position wählen möchten. Diese Möglichkeit sollte ihnen dann auch eröffnet werden.

INTERVALL- ODER VERHÄLTNISSKALA

Intervallskalen weisen zunächst die Merkmale von Ordinalskalen auf, besitzen jedoch die zusätzliche Eigenschaft, dass die Differenz (des Intervalls) zwischen zwei Werten genau bestimmt werden kann. Die Bearbeitung von intervallskalierten Daten unterliegt keinerlei Ein-

schränkungen. Alle statistischen Kennwerte, Tests und Verfahren können hier Anwendung finden. Manchmal findet man in der Literatur noch eine Unterscheidung zwischen Intervallskala und Verhältnisskala. In der Regel sind intervallskalierte Variablen gleichzeitig auch verhältnisskaliert, wenn diese einen absoluten Nullpunkt aufweisen. Verhältnisskalierte Merkmalsausprägungen lassen sich quantifizieren und im Verhältnis zueinander betrachten. Betrachtet man bspw. das Alter, so lässt sich bei diesem verhältnisskalierten Merkmal sagen, dass eine Person von 60 Jahren doppelt so alt ist wie eine Person im Alter von 30 Jahren. Es bleibt anzumerken, dass SPSS nicht zwischen Intervall- und Verhältnisskala unterscheidet. Deswegen wird im Folgenden auch immer von intervallskalierten oder metrischen Daten die Rede sein.

2.3.2 GESCHLOSSENE/OFFENE FRAGEN

Bei der Konzipierung der einzelnen Fragen muss man sich darüber Gedanken machen, ob man Antwortmöglichkeiten vorgeben will oder ein leeres Feld zur freien, nichtstandardisierten Antwort einbaut. Bei offenen Fragen sollte man allerdings daran denken, dass man bspw. bei postalisch versendeten Fragebögen nicht antizipieren kann, ob die Befragten alle Fragen verstehen und man keine Möglichkeit zur Erläuterung hat. Um große Mengen an unbrauchbaren Daten zu vermeiden, sollten solche Fragekategorien möglichst unkompliziert formuliert werden sollten.

Offenen Fragen sind mit folgenden weiteren Schwierigkeiten verbunden:

Antworten auf offene Fragen sind schwerer zu interpretieren.

Die Codierung und die manuelle Eingabe der Daten nehmen deutlich mehr Zeit in Anspruch.

Die nachträgliche Codierung bedarf einer umfassenden theoretischen Basis und hängt stark von der Beurteilung des Codierenden ab.

Die Bereitschaft der Probanden, offene Fragen zu beantworten, kann stark variieren. Ein hohes Interesse an den Fragestellungen führt hier zu mehr und besseren Antworten. Gering involvierte Befragte neigen eher dazu, keine Antwort zu geben. Sie müssen hier selbst Antworten finden und formulieren und nicht nur aus einem Set fertiger Antworten die zutreffende erkennen.

Trotz dieser Vorsichtsmaßnahmen können offene Fragen unter bestimmten Voraussetzungen sehr hilfreich sein:

Die genaue Ausgestaltung der möglichen Anzahl ist unbekannt. Die Verwendung von vorgefertigten Kategorien würde die tatsächlichen Antworten in diesem Fall zugunsten der ausgewählten Antwortvorgaben verzerren.

Verfolgt man einen eher induktiven Forschungsansatz, lohnt es sich vielleicht die Probanden zu freien Assoziationen zu bewegen, indem sie ihre Ideen, Gedanken oder Auffassungen in Bezug auf bestimmte Fragen äußern dürfen.

Häufig finden sich offene Fragen am Ende eines Fragebogens, um den Befragten die Möglichkeit zu geben, zu Punkten Stellung zu nehmen, die durch die übrigen Fragen nicht abgedeckt werden.

2.3.3 WEITERE ANTWORTMÖGLICHKEITEN (NICHT ZUTREFFEND/WEIß NICHT/KEINE ANGABE)

Eine häufige Kritik an Befragungen ist, dass die Befragten zu Meinungsäußerungen verleitet werden, die sie jenseits der Befragung nicht geäußert hätten. Auch wenn dieser Kritikpunkt nicht ganz von der Hand zu weisen ist, lässt sich das Problem durch den Einbau von Antwortmöglichkeiten für alle Befragten, die keine gültige Antwort geben können oder wollen, abschwächen. Beispielsweise lässt sich aus der häufigen Nennung der „weiß nicht"-Antwortkategorie schließen, dass die Befragten bisher wenig Interesse an bzw. wenig Kontakt mit den für die Forschungsfrage relevanten Sachverhalten hatten.

Besteht die Möglichkeit, dass eine Frage nicht für jeden Befragten relevant ist, sollte man die Kategorie „nicht zutreffend" als zusätzliche Antwortvorgabe erwägen. Sollten Folgefragen auf dieser Frage aufbauen, dann ist es ratsam, eine Verzweigung einzubauen, die den Probanden zur nächsten für ihn relevanten Frage weiterleitet.

Erwartungsgemäß verringert der Einbau der Antwortkategorie „weiß nicht" die Anzahl der gültigen Antworten. Trotzdem ist es ratsam, eine solche Option zu berücksichtigen, vor allem wenn Fragen über zukünftige Absichten sowie Erinnerungsfragen in der Untersuchung enthalten sind.

Um die Befragten nicht zu einer Antwort zu zwingen, die sie nicht zu geben bereit sind, empfiehlt es sich, die Kategorie „keine Angabe (k. A.)" zu ergänzen.

Aufgabe 2.4
Eine Firma interessiert sich im Rahmen der Planung von Parkplätzen und dem Einsatz von firmeneigenen Bussen dafür, in welcher Entfernung ihre Beschäftigten von der Arbeitsstätte wohnen und mit welcher Beförderungsmitteln die Arbeitsstätte überwiegend erreicht wird. Sie greift dazu auf eine Untersuchung zurück, die zur Erfassung der wirtschaftlichen Lage der Mitarbeiterinnen und Mitarbeiter durchgeführt wurde. Bei der Untersuchung wurden an einem Stichtag 50 Beschäftigte ausgewählt und zu folgenden Punkten befragt:
Haushaltsgröße (Anzahl der im Haushalt lebenden Personen); Monatliche Miete; Beförderungsmittel, mit dem die Arbeitsstätte überwiegend erreicht wird; Entfernung zwischen Wohnung und Arbeitsstätte; eigene Einschätzung der wirtschaftlichen Lage mit 1 = sehr gut, …, 5 = sehr schlecht.
(a) Geben Sie die Grundgesamtheit und die Untersuchungseinheiten an.
(b) Welche Ausprägungen besitzen die erhobenen Merkmale und welches Skalenniveau liegt ihnen zugrunde?

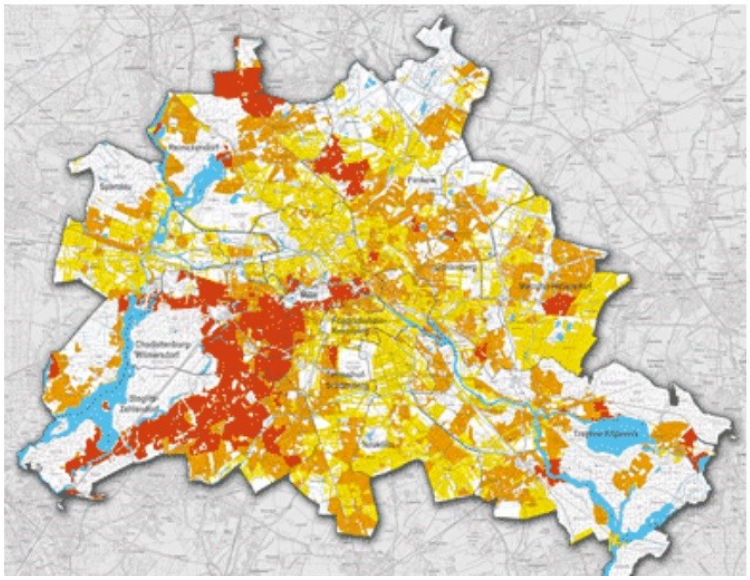
Aufgabe 2.5
Nennen Sie die Skalenniveaus hierarchisch vom informationsärmsten bis zum informationsreichsten.

Aufgabe 2.6
Welches Skalenniveau hat die Variable „Sportlichkeit“, wenn die Probanden gebeten wurden, sich selbst auf einer Skala von 0 % bis 100 % einzuschätzen.

Aufgabe 2.7
Erläutern Sie anhand der Wohnlagenkarte des Berliner Mietspiegels 2013 die folgenden statistischen Grundbegriffe:
Erhebungseinheit und Grundgesamtheit,
Erhebungsmerkmal und Merkmalsausprägungen,
Skalenniveau des Erhebungsmerkmals.

Berliner Mietspiegel 2013

Wohnlagenkarte Berlin



Überwiegend einfache Wohnlage
Überwiegend mittlere Wohnlage
Überwiegend gute Wohnlage
Gebiete ohne betroffenen Wohnraum

Achtung: Die Karte soll lediglich eine erste Orientierung über die mögliche Wohnlagezuordnung geben.

Eine genaue Einordnung des Wohnhauses in die zutreffende Wohnlage ermitteln Sie bitte über

Mietspiegelabfrage
Karte mit Adressuche

Wohnlagenkarte
(pdf; geringe Auflösung; 413 KB)

Wohnlagenkarte
(pdf; hohe Auflösung; 6,9 MB)

Quelle: Berliner Senatsverwaltung für Stadtentwicklung und Umwelt, http://www.stadtentwicklung.berlin.de/wohnen/mietspiegel/de/wohnlagenkarte.shtml (25.9.2014)

Aufgabe 2.8
Welches Skalenniveau liegt den folgenden Aussagen jeweils zugrunde? Begründen Sie kurz Ihre jeweilige Antwort.
1. Eine Merkmalsausprägung ist doppelt so groß wie eine andere.
2. Die Abstände zwischen je zwei Merkmalsausprägungen eines Erhebungsmerkmals lassen sich berechnen bzw. messen und vergleichen.
3. Die Merkmalsausprägungen eines Erhebungsmerkmals sind Rangzahlen.
4. Die Merkmalsausprägungen eines Erhebungsmerkmals können lediglich als gleich- oder verschiedenartig eingeordnet werden.

Aufgabe 2.9
Öffnen Sie die SPSS-Datendatei „Bev_Rhein_Main.sav“. Benennen Sie die Merkmalsträger und die Grundgesamtheit.

2.4 CODIERUNG UND CODEPLAN

Sind alle Fragen formuliert und steht das Layout des Fragebogens fest, sollte man diesen codieren, um die spätere Dateneingabe zu vereinfachen. Der Codeplan ordnet hierbei den einzelnen Fragen des Fragebogens Variablennamen und den möglichen Merkmalsausprägungen Codenummern zu. Bei der Abfrage von metrischen Daten wie Alter, Größe und Gewicht liegen direkt eingebbare Zahlen vor und bei offenen Fragen liegen zumeist ganze Sätze vor. Bei anderen Merkmalen wie dem Geschlecht oder dem Schulabschluss ist jedoch zu überdenken, nach welchen Regeln den „Kreuzen“ oder allgemeiner den Angaben der Probanden verschiedene Zahlen zugewiesen werden sollen. So lassen sich beispielsweise die Variablenausprägungen *„männlich“* bzw. *„weiblich“* für das Merkmal *„Geschlecht“* durch die Zahlen *„1“* für *„weiblich“* und *„2“* für *„weiblich“* codieren. Bei einem solchen Vorgehen erspart man sich zum einen langwierige Texteingaben und zum anderen liegen somit bereits von SPSS interpretierbare numerische Daten vor. Insbesondere bei der Betrachtung vieler sehr ähnlicher Frageformen, wie etwa im Fall einer großen Zahl von Beurteilungsfragen, lohnt sich eine gleiche Codierung gleicher Antwortkategorien. Ein Beispiel einer solchen Codierung, die bereits in einen Fragebogen integriert wurde, veranschaulicht Abbildung 2.1.

Abbildung 2.1 Beispiel eines integrierten Codeplans mit vielen ordinalen Variablen

B Gastronomieangebot

5 Wie zufrieden sind Sie mit dem Gastronomieangebot in Diepholz hinsichtlich der folgenden Aspekte:

Bars/ Kneipen	5.1 Anzahl:	zufrieden □15 □16 □17 □18 unzufrieden	□0 weiß nicht □9 k. A.
	5.2 Abwechslung:	zufrieden □15 □16 □17 □18 unzufrieden	□0 weiß nicht □9 k. A.
	5.3 Öffnungszeiten:	zufrieden □15 □16 □17 □18 unzufrieden	□0 weiß nicht □9 k. A.
Cafés/Eisdielen	5.4 Anzahl:	zufrieden □15 □16 □17 □18 unzufrieden	□0 weiß nicht □9 k. A.
	5.5 Abwechslung:	zufrieden □15 □16 □17 □18 unzufrieden	□0 weiß nicht □9 k. A.
	5.6 Öffnungszeiten:	zufrieden □15 □16 □17 □18 unzufrieden	□0 weiß nicht □9 k. A.
Gaststätten/ Restaurants	5.7 Anzahl:	zufrieden □15 □16 □17 □18 unzufrieden	□0 weiß nicht □9 k. A.
	5.8 Abwechslung:	zufrieden □15 □16 □17 □18 unzufrieden	□0 weiß nicht □9 k. A.
	5.9 Öffnungszeiten:	zufrieden □15 □16 □17 □18 unzufrieden	□0 weiß nicht □9 k. A.
Discotheken/ Tanztreffs	5.10 Anzahl:	zufrieden □15 □16 □17 □18 unzufrieden	□0 weiß nicht □9 k. A.
	5.11 Abwechslung:	zufrieden □15 □16 □17 □18 unzufrieden	□0 weiß nicht □9 k. A.
	5.12 Öffnungszeiten:	zufrieden □15 □16 □17 □18 unzufrieden	□0 weiß nicht □9 k. A.

Dieses Vorgehen erlaubt das Kopieren der Definitionen dieser Variablen und stellt damit eine erheblich Arbeitsersparnis dar (siehe Kapitel 3.2). Abbildung 2.2 zeigt einen weiteren codierten Fragebogenabschnitt, der die Codierung anderer Antwortformen verdeutlicht.

Abbildung 2.2: Beispiel eines integrierten Codeplans mit unterschiedlichen Antwortformen

K Nutzung des Diepholzer Marktplatzes für Veranstaltungen

38 Kennen Sie den so genannten Großmarkt, der jedes Jahr auf dem Diepholzer Marktplatz stattfindet? Falls ja: haben Sie den Großmarkt schon einmal besucht?

□57 nicht bekannt □58 bekannt □59 bekannt und besucht → falls nicht bekannt, weiter mit Frage 43

39 Wie gefällt Ihnen der Diepholzer Großmarkt?

gut □19 □20 □21 □22 schlecht □0 weiß nicht □9 k.A.

40 An wie vielen Tagen haben Sie den Diepholzer Großmarkt im vergangenen Jahr besucht?

_____Tage □0 weiß nicht □9 k.A.

41 Wie gut oder weniger gut ist der Diepholzer Marktplatz Ihrer Meinung nach als Austragungsort für den Großmarkt geeignet? Warum?

gut □19 □20 □21 □22 schlecht □0 weiß nicht □9 k.A.

41.1 Begründung:

Der gesamte Fragebogen sollte vor der eigentlichen Befragung an Kollegen und Freunden, besser noch im Rahmen eines sogenannten Pretests, an einem Teil der Zielgruppe erprobt werden. In den Fragebogen eingebaute Schwächen wie unverständliche Formulierungen u. Ä. können so vor der eigentlichen (Haupt-)Erhebung behoben werden.

Aufgabe 2.10
Entwerfen Sie einen eigenen Fragebogen. Dieser soll fünf sehr unterschiedliche Fragemodelle beinhalten, den allgemeinen Richtlinien zur Erstellung von Fragebögen Genüge tun und mit einer sinnvollen Codierung ausgestattet sein.

3 DATENEINGABE

Wie öffne ich Datendateien aus anderen Programmen als SPSS?

Wie gebe ich Variablen und Daten manuell in SPSS ein?

Wie erleichtere ich mir die Dateneingabe (Daten zusammenführen, Zeilen und Spalten löschen oder kopieren, Deklarationen kopieren etc.)?

3.1 DATENIMPORT AUS FREMDFORMATEN

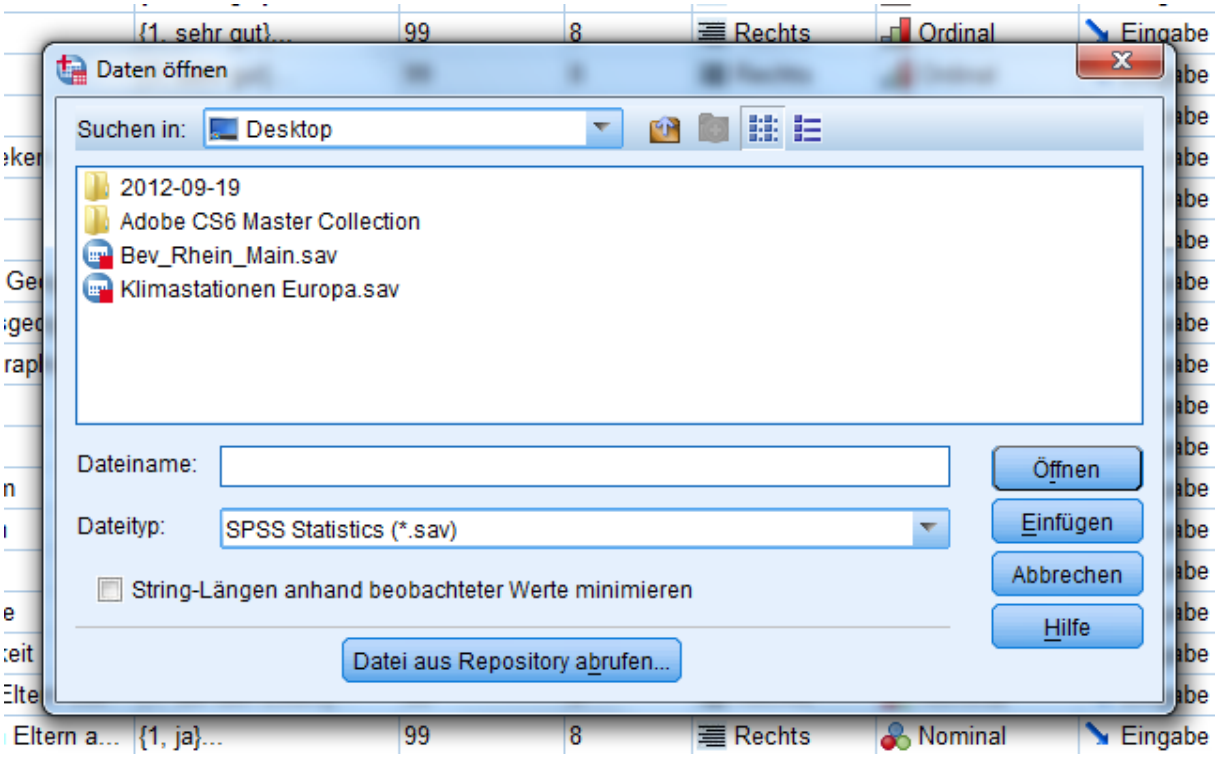
Die Verwendung von SPSS beschränkt sich nicht allein auf die Auswertung einer selbst konzipierten Befragung, sondern kann ebenso gut als Analyseinstrument für bereits vorhandene Daten dienen, beispielsweise bei der Auswertung verschiedener Datenbanken zu demographischen Daten u. Ä. Die zugänglichen Formate solcher Vorlagen sind zumeist nicht im SPSS-Format „sav“ gespeichert, aber eine Umwandlung in lesbare Dateien lässt sich durch wenige Einstellungen leicht vornehmen. Umgekehrt ist SPSS in der Lage, Datendateien zur Weiterverarbeitung in anderen Datendateien zu speichern.

SPSS ist in der Lage, verschiedene Formate aus Datenbanken, Textdateien oder anderen Statistikprogrammen und Tabellenkalkulationen zu öffnen und umzuwandeln. Die Reichweite erstreckt sich von ASCII-Dateien über Statistikprogramme wie SYSTAT, SAS oder STATA bis zu Tabellenkalkulationsprogrammen wie Excel und Datenbankanwendungen wie Access. Dabei unterscheidet sich das Einlesen von Textdateien und Datenbankformaten vom Öffnen anderer Formate.

3.1.1 ÖFFNEN VON DATEIEN AUS ANDEREN STATISTIK- ODER TABELLENKALKULATIONSPROGRAMMEN

Außer in den Sonderfällen wie bei der Benutzung einer Datenbank-Schnittstelle oder der Übernahme von ASCII-Daten geht man zum Öffnen anderer Dateiformate wie folgt vor: ***Datei → Öffnen → Daten***. Es öffnet sich das in Abbildung 3.1 dargestellte Dialogfenster.

Abbildung 3.1: Kontextmenü **Datei öffnen**



Hier wählt man wie von Windows gewohnt das gesuchte Verzeichnis. In diesem werden zunächst nur alle Dateien mit der Endung .sav angezeigt. Diese Auswahl lässt sich jedoch durch einen Klick auf das Dropdown-Menü der Dateitypliste anpassen. Wählen Sie nun das gewünschte Format sowie die gesuchte Datei aus und bestätigen Sie Ihre Auswahl durch einen Klick auf ***Öffnen***. Je nach Format kann es sein, dass SPSS ein weiteres Kontextmenü öffnet und Anweisungen zum Einlesen der Variablennamen und dem relevanten Bereich anfordert.

3.1.2 EINLESEN VON DATENBANKFORMATEN

Datenbankdateien sind ähnlich aufgebaut wie SPSS-Datendateien. Sie eröffnen jedoch verschiedene Komfortfunktionen wie die Anpassung der Bildschirmoberfläche und helfen so, Fehler bei der Dateneingabe zu verringern. SPSS kann verschiedene Datenbankformate einlesen. Unter Umständen muss das dbase-Format für den Transfer gewählt werden, indem in der Datenbankanwendung dieses Format für den Export gewählt wird. Um eine Datenbankdatei zu öffnen, wählen sie unter ***Datei → Öffnen → Daten*** unter ***Dateityp*** z. B. das ***dBase (*.dbf)*** –Format.

3.1.3 EINLESEN VON TEXTDATEIEN (ASCII-DATEIEN)

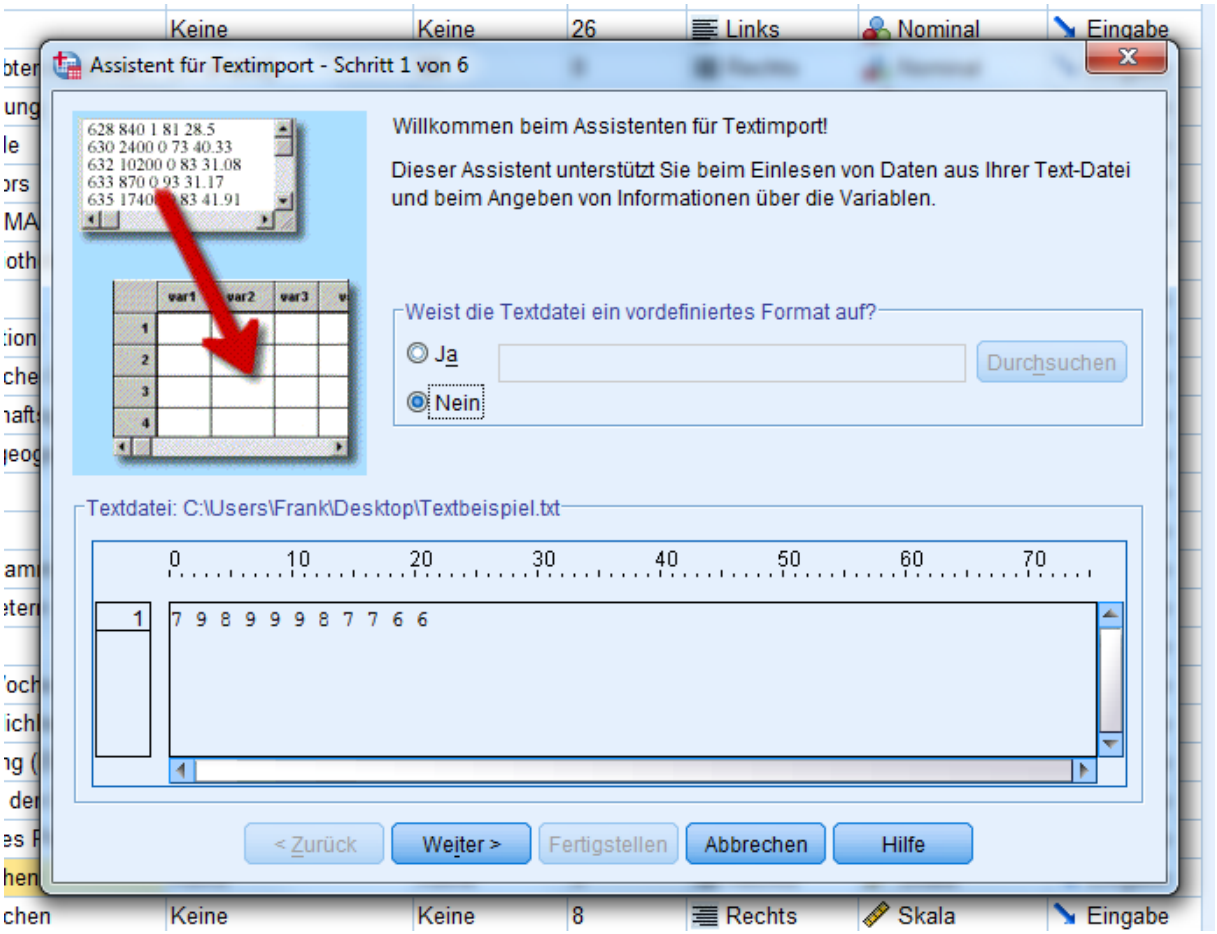
Zwar können empirische Daten manuell im Datenfenster von SPSS Statistics eingegeben werden, dieses Vorgehen ist bei größeren Datensätzen allerdings nicht praktikabel, da der Dateneditor von SPSS recht unkomfortabel ist. Häufig werden die Daten daher mit speziellen Eingabeprogrammen in einfachen Textdateien, sogenannten ASCII-Dateien, abgelegt. Vor allem wenn man bei der Bearbeitung seiner Forschungsfragen auf externe Datenquellen angewiesen ist, kann man es mit ASCII-Dateien oder anderen Textformaten zu tun bekommen. Diese lassen sich jedoch auch in SPSS übertragen. In diesem Fall liegen sie als Zahlenkolonnen vor, die verschieden formatiert sein können. Entweder werden diese durch Trennzeichen strukturiert oder liegen im festen bzw. freien Format vor.

Eingabedaten im Textformat enthalten keine Informationen über die Zuordnung der Spalten der Datenzeilen zu Variablen. Um eine Übertragung in SPSS durchzuführen, müssen diese Daten daher zunächst definiert werden. Dazu sind folgende Informationen notwendig:

In welcher Datei steht der Datensatz?
Wie heißen die Variablen?
Wo stehen welche Variablen im Datensatz?
Wie sind die Variablen im Datensatz getrennt (Leerzeichen, Komma, ...)?

Durch diese Definitionen lassen sich die Zahlenkolonnen inhaltlich füllen. Über die Auswahl von ***Datei → Textdaten lesen*** in der Menüsteuerung gelangt man in den Explorer und wählt zunächst wie gewohnt die zu öffnende Datei aus. Danach ein Klick auf ***Öffnen*** und es öffnet sich das folgende Kontextmenü (Abbildung 3.2).

Abbildung 3.2: Kontextmenü ***Assistent für Textimport***



In sechs Schritten können Sie nun den Import der Daten in SPSS einleiten.

3.2 MANUELLE DATENERFASSUNG IM SPSS DATENEDITOR

Bevor man sich eingehend mit der manuellen Dateneingabe in SPSS beschäftigt, lohnt sich ein Blick auf die zum schnellen Manövrieren im Daten- wie im Variableneditor zur Verfügung stehenden Tastenkombinationen.

Tab oder Pfeil nach rechts	Positioniert die Auswahl eine Zeile nach rechts
Enter oder Pfeil nach unten	Positioniert die Auswahl eine Zeile tiefer
Pfeil nach oben	Positioniert die Auswahl eine Zeile höher
Shift + Tab oder Pfeil nach links	Positioniert die Auswahl eine Zeile nach links
Pos1	Positioniert die Auswahl auf die erste Zelle einer Zeile bzw. eines Fasses
Ende	Positioniert die Auswahl auf die letzte Zelle einer Zeile eines Fasses
Strg + Pfeil nach oben	Positioniert die Auswahl auf den ersten Fall einer Spalte

Strg + Pfeil nach unten	Positioniert die Auswahl auf den letzten Fall einer Spalte
Strg + Pos1	Positioniert die Auswahl auf die erste Zelle des ersten Falles
Strg + Ende	Positioniert die Auswahl auf die letzte Zelle des letzten Falles
Bild rauf	Vollzieht einen Bildlauf nach oben um eine Seite
Bild runter	Vollzieht einen Bildlauf nach unten um eine Seite

Neben dem Manövrieren bieten sich verschiedene Tastenkombinationen zum zeitsparenden Markieren und Editieren von Zeilen und Spalten an:

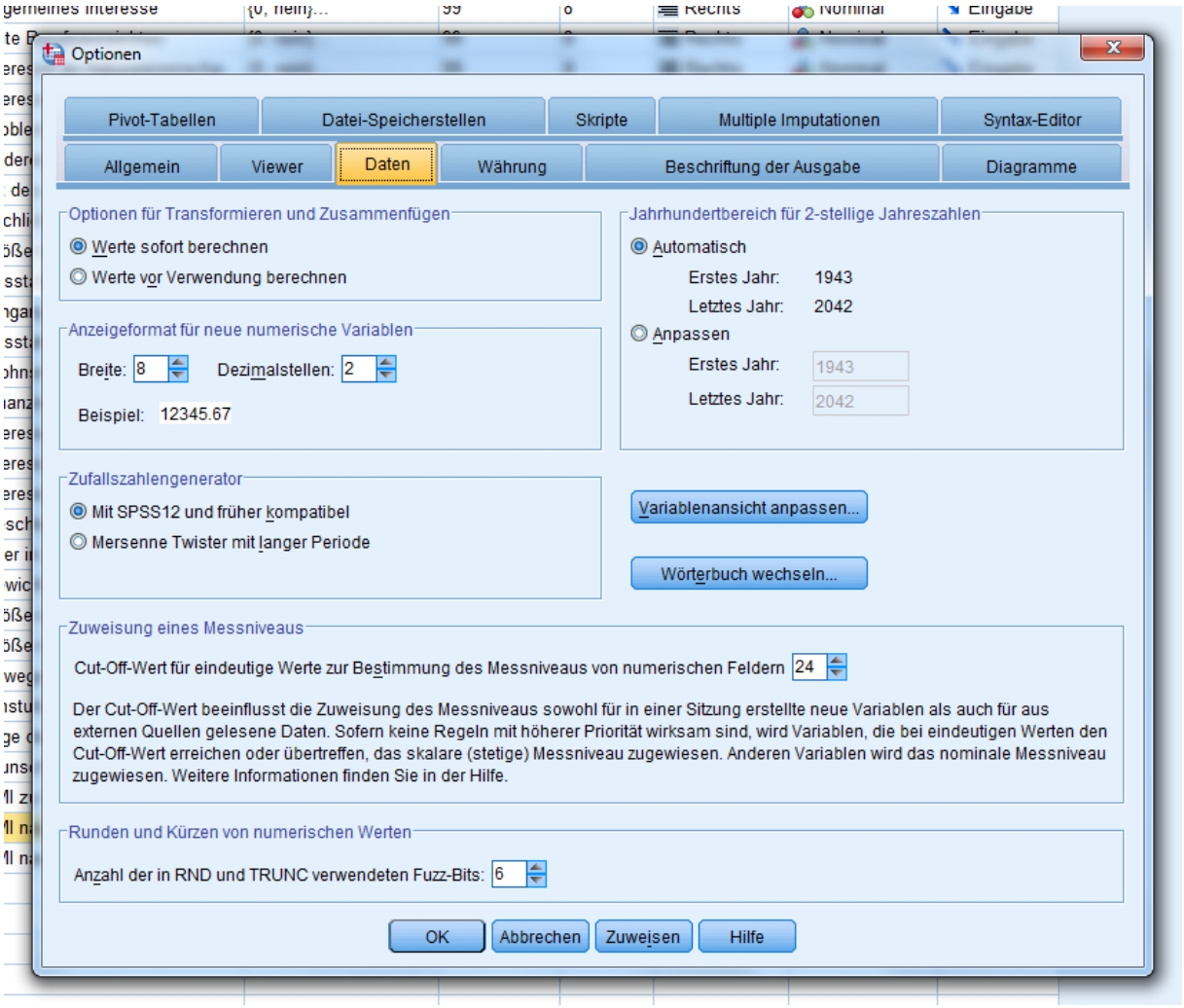
Shift + Leertaste	Markiert die ganze Zeile
Strg + Leertaste	Markiert die ganze Spalte
Shift + Pfeiltaste	Auswahl eines Bereiches von Fällen und Variablen. Alternativ: Klicken und Ziehen der Maus von der oberen linken Ecke bis zur rechten unteren Ecke
F2	Schaltet in den Editiermodus um. Ein erneutes Drücken schaltet diesen wieder aus.

3.2.1 DEFINITION VON VARIABLEN

Nach dem Aufruf von SPSS befindet man sich bei Auswahl einer neuen Datei zunächst in einem leeren Fenster des Dateneditors. Um nun die verschiedenen Variablen in die Datenmaske einzubauen, muss man zunächst die Registerkarte ***Variablenansicht*** (siehe Abbildung 1.4) auswählen. In dieser Ansicht lassen sich die verschiedenen Variablen entsprechend der Angaben aus dem vorher erstellten Codeplan eingeben.

Bei der Eingabe einer neuen Variablen wird diese von SPSS automatisch auf den Variablentyp ***numerisch*** mit einer Maximallänge von acht Zeichen mit zwei Dezimalstellen definiert. Hat man nur mit einer Reihe relativ homogenen Variablendefinitionen zu tun, lassen sich diese Voreinstellungen über ***Bearbeiten*** → ***Optionen*** unter der Registerkarte ***Daten*** verändern. Es erscheint dann folgende Box, in der verschiedene Änderungen vorgenommen werden können (Abbildung 3.3).

Abbildung 3.3: Auswahlmöglichkeiten zur Definition der Voreinstellungen neuer Variablen



VARIABLENNAME

In der in Abbildung 1.4 dargestellten ***Variablenansicht*** geben Sie in das Textfeld ***Name*** den gewünschten Variablennamen ein. Bei der Vergabe von Namen müssen jedoch bestimmte Regeln eingehalten werden.

Variablennamen **müssen:**

mit einem Buchstaben beginnen

Variablennamen **können**:

aus Buchstaben und Ziffern bestehen
aus einer beliebigen Kombination aus Groß- und Kleinschreibung bestehen
die Sonderzeichen _ (Unterstrich), . (Punkt) sowie die Zeichen @, #, $ und § enthalten
Umlaute und ß enthalten (sollten aber vermieden werden)

Nicht erlaubt sind:

Leerzeichen sowie spezifische Zeichen wie !, ?, „ und *
ein . (Punkt) oder _ (Unterstrich) als letztes Zeichen (kann zu Konflikten bei speziellen SPSS-Prozeduren führen)
die doppelte Vergabe eines Variablennamens
die Verwendung reservierter Schlüsselwörter wie: ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO, WITH.

Beispiele für **gültige Variablennamen**:

Fragebogennummer
FBN
Frage_8
Lgehalt

Beispiele für **ungültige Variablennamen:**

1_Geschlecht	**Name beginnt mit einer Ziffer**
Gehalt 2012	**Name enthält ein Leerzeichen**
Park&Ride_jn	**Name enthält unzulässiges Zeichen „&“**

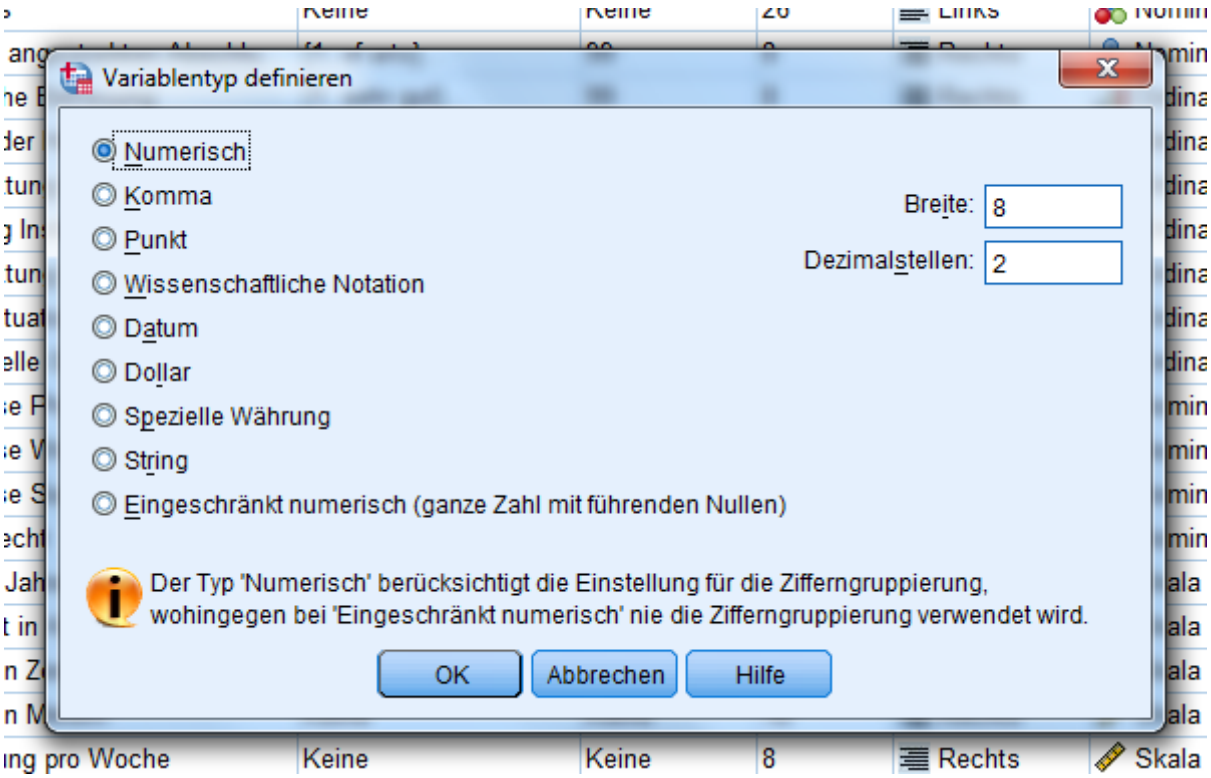
Bei der Anlage eines ungültigen Variablennamens gibt SPSS eine Fehlermeldung aus, sodass die Eingabe von „falschen“ Variablennamen, die weitere Berechnungen unbrauchbar machen würden, automatisch unterbunden wird.

Mit der Tabulator-Taste lässt sich eine Eingabe abschließen und die Auswahl springt in die nächste Spalte.

VARIABLENTYP

Bei der Eingabe einer neuen Variablen wird diese von SPSS automatisch auf den Variablentyp ***Numerisch*** definiert. Um diese Voreinstellung zu ändern, klickt man in der Ansicht auf die Schaltfläche mit den drei Punkten in der zweiten Zelle. Es öffnet sich das Dialogfenster ***Variablentyp definieren***, welches in Abbildung 3.4 dargestellt ist.

Abbildung 3.4: **Auswahlfenster für den Variablentyp**



SPSS bietet eine große Auswahl verschiedener Variablentypen an. Die nachfolgende Tabelle fasst die wichtigsten Spezifika der jeweiligen Ein- und Ausgabemöglichkeiten dieser Variablentypen kurz zusammen.

Numerisch	Ziffern, evtl. vorgestelltes Minuszeichen und ein Dezimaltrennzeichen.
Komma	Ziffern, evtl. vorgestelltes Minuszeichen, einen Punkt als Dezimaltrennzeichen sowie ein oder mehrere Kommas als Tausendertrennzeichen. Diese Kommas werden automatisch eingefügt.
Punkt	Ähnlich wie bei der Auswahl „Komma“. Nur das hier Kommas als Dezimaltrennzeichen verwendet werden und Punkte als Tausendertrennzeichen.
Wissenschaftliche Notation	Alle gültigen numerischen Werte, einschließlich wissenschaftlicher Notation, die durch ein eingebettetes E, D, Plus- oder Minuszeichen gekennzeichnet sind.
Datum	Datums- und Zeitangaben. Es stehen 27 verschiedene Datums- und Zeitformate zur Verfügung.
Dollar	Werte, die ein Dollarzeichen, einen Punkt als Dezimaltrennzeichen und Kommas als Tausendertrennzeichen enthalten. Dollarzeichnen und Kommas werden von SPSS automatisch eingefügt
Spezielle Währung	Möglichkeit der Definition eigener Währungsformate
String	Zeichenkette aus Buchstaben, Ziffern und Sonderzeichen. Stringvariablen dürfen bis zu 255 Zeichen lang sein.

Die für sozialwissenschaftliche Sachverhalte gebräuchlichsten Variablentypen sind numerische und solche im String-Format.

DEZIMALSTELLEN

Um die von SPSS automatisch gesetzte Voreinstellung von zwei ***Dezimalstellen*** zu ändern, klickt man in das entsprechende Feld und passt die Stellen über den am rechten Rand des Feldes angezeigten Regler nach Belieben an. Bei der Festlegung der Dezimalstellen ist jedoch darauf zu achten, dass deren Anzahl kleiner dem Wert in der Zelle Spaltenformat sein sollte. Falls das nicht der Fall ist, gibt das Programm eine Fehlermeldung aus.

VARIABLENLABELS

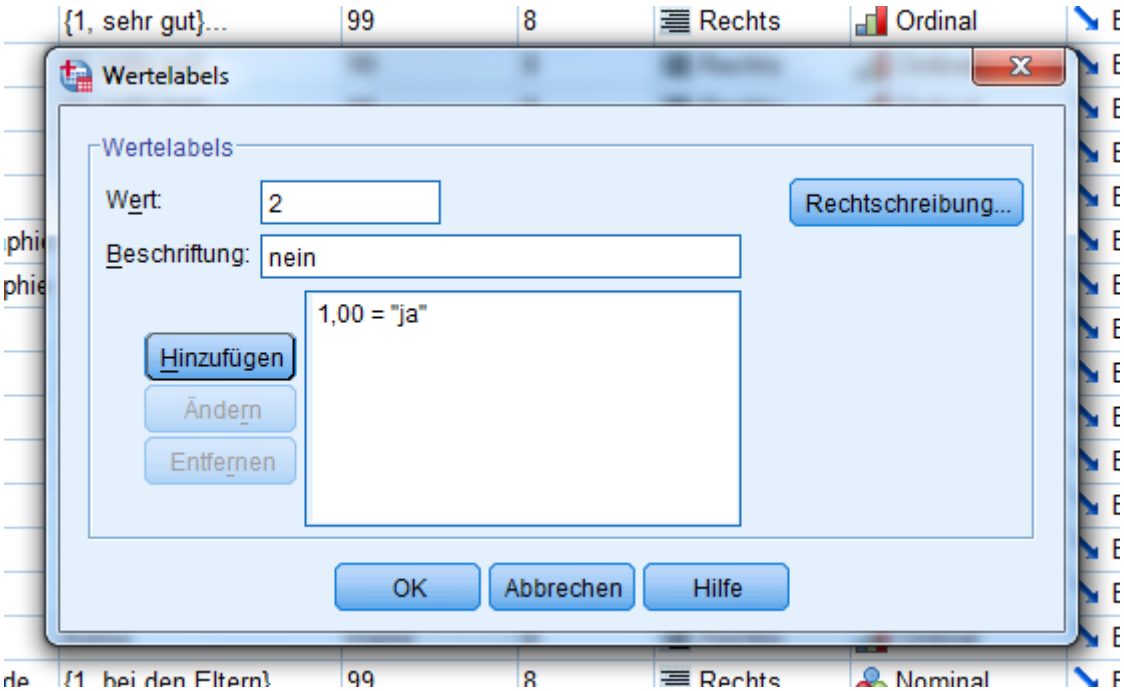
Durch einen Klick auf das Feld unter der Spalte ***Variablenlabel*** lässt sich eine bis zu 256 Zeichen umfassende Beschreibung der Variable bzw. der zugehörigen Frage einfügen. Trotzdem sollte man auf überlange Formulierungen verzichten, da diese Labels bei verschiedenen Ergebnisausgaben über den angeforderten Tabellen und Diagrammen angezeigt werden und die Darstellungen dadurch stark „aufgebläht“ werden. Hier empfiehlt sich eine eher knappe Beschreibung. Sollte man sich trotzdem für einen längeren Text entscheiden, lässt sich dieser durch Anklicken und Ziehen der rechten Begrenzungslinie verbreitern und verschmälern. Korrekturen lassen sich durch einen Doppelklick auf das entsprechende Label durchführen.

WERTELABELS

Im Gegensatz zum Variablenlabel umfasst die auf den ersten Blick sehr ähnlich erscheinende Spalte ***Wertelabel*** keine Beschreibung der Variable, sondern dient der Übertragung der numerischen Codierungen aus dem vorher angelegten Codeplan (siehe Kapitel 2.4). Den verschiedenen Zahlencodierungen lassen sich bis zu 60 Zeichen lange Bezeichnungen zuweisen. Auch hier empfiehlt sich wieder die Beschränkung auf kurze Labels. Um eine Eingabe zu machen, klickt man zunächst auf die entsprechende Zeile und dann auf die angezeigten drei Punkte. Es öffnet sich die Dialogbox, die in Abbildung 3.5 gezeigt wird.

Hier gibt man die einzelnen Werte und die entsprechenden Labels nacheinander ein und bestätigt jede Eingabe einzeln durch einen Klick auf ***Hinzufügen***. Nachträgliche Änderungen oder das Entfernen kompletter Labels nimmt man durch einen Klick auf die entsprechenden Schaltflächen vor.

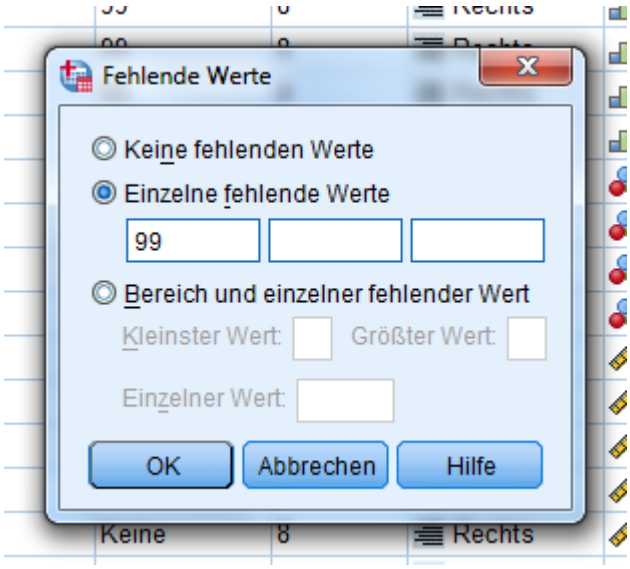
Abbildung 3.5: Dialogbox ***Wertelabels definieren***



FEHLENDE WERTE

SPSS kennt zwei verschiedene Arten von fehlenden Werten, zum einen die *systemdefinierten fehlenden Werte* und zum anderen die *benutzerdefinierten fehlenden Werte*. Werden Datenfelder nicht ausgefüllt, behandelt SPSS sie als systemdefinierte fehlende Werte. Diese werden in der Datenansicht als Komma (,) angezeigt. SPSS bietet jedoch auch die Möglichkeit, eigene Antwortvorgaben als fehlende Werte zu verwenden. Dies ist beispielweise sinnvoll, wenn bestimmte Gründe für das Fehlen der Daten unterschieden werden sollen; etwa eine bewusste Antwortverweigerung aus persönlichen Gründen im Kontrast zu mangelnden Kenntnissen über den Gegenstand der Fragestellung. Hier lassen sich Antwortkategorien wie beispielsweise ***weiß nicht***, ***nicht zutreffend*** oder ***keine Angabe (k. A.)*** einfügen. Abbildung 3.6 veranschaulicht die Dialogbox ***Fehlende Werte definieren***.

Abbildung 3.6: Dialogbox **Fehlende Werte definieren**



Hier lassen sich einzelne fehlende Werte eingeben oder ganze Bereiche für solche abstecken. Haben Sie sich für die Verwendung benutzerdefinierter fehlender Werte entschieden und verwenden Sie Variablen mit Wertelabels, sollten Sie die hier definierten fehlenden Werte auch in die Liste der Wertelabels übertragen.

SPALTEN

Das Spaltenformat einer neuen Variablen wird von SPSS automatisch auf den relativ hohen Wert von 8 gesetzt, sodass im Datenfenster lediglich 10 Variablen bzw. Spalten gleichzeitig auf dem Bildschirm sichtbar sind. Verwendet man hauptsächlich kurze Variablennamen, lässt sich diese Zahl in der Ansicht durch Anpassung der Variablenspalten ändern. Eine weitere Möglichkeit zur Veränderung der Spaltenbreite bietet sich in der Datenansicht. Klickt man auf die rechte Variablenbegrenzung, lässt sich diese durch einfaches Ziehen auf die gewünschte Breite bringen.

AUSRICHTUNG

Standardmäßig werden die Werte der numerischen Merkmale in der Datenansicht rechtsbündig, die von String-Variablen linksbündig ausgerichtet. Diese Voreinstellung lässt sich jedoch durch einen Klick auf das entsprechende Feld angleichen bzw. den eigenen Vorstellungen entsprechend anpassen.

MESSNIVEAU

In der Spalte Messniveau müssen Angaben zum Skalenniveau der Variablen bzw. der Variablenwerte gemacht werden. Die drei möglichen Vorgaben (metrisch – ordinal – nominal) wurden bereits in Kapitel 2.2 beschrieben. Die hier vorgegebene Einordnung dient jedoch

nur der eigenen Übersicht. Sollten Sie es darauf anlegen, berechnet Ihnen SPSS auch das arithmetische Mittel für die Verteilung einer nominalen Variablen.

3.2.2 VARIABLENDEKLARATIONEN KOPIEREN

Die Definition verschiedener Variablenattribute lässt sich auf mehrere gleich aufgebaute Variablen übertragen, sodass man nicht zu einer langwierigen manuellen Eingabe jedes Definitionsschrittes gezwungen ist. Dies stellt insbesondere für die wiederholte Eingabe gleicher Wertelabels eine erhebliche Arbeitserleichterung dar. Das Vorgehen ist hier sehr ähnlich wie in gebräuchlichen Textbearbeitungsprogrammen. Man klickt auf die zu kopierende Zeile und drückt ***Strg + C***, wählt die Zielzeile(n) aus und drückt ***Strg + V***. Ein Kopieren und Einfügen mittels der rechten Maustaste ist leider nicht möglich.

3.2.3 ANZEIGEN UND AUFLISTEN DER VARIABLENATTRIBUTE

Über das Auswählen von ***Extras → Variablen*** kann man sich die Definitionen der verschiedenen Variablenattribute in einem Dialogfenster übersichtlich anzeigen lassen. Auch der Druck der eingegebenen Variablendefinitionen aus der Auflistung im Ausgabefenster ist möglich. Hierzu wählt man ***Extras → Datei-Info***.

3.2.4 SPEICHERN EINER DATENDATEI

Während des Vorgangs der Variablendefinition sollte man seine Eingabefortschritte in gewissen Abständen immer wieder abspeichern. Das Vorgehen orientiert sich auch hier wieder an weitläufig bekannten Office-Programmen. Um zu speichern, klickt man auf ***Datei → Speichern*** bzw. ***Speichern unter*** und wählt einen entsprechenden Namen und Speicherort.

3.3 EIGENTLICHE DATENEINGABE

Nachdem alle Variablen in der Variablenansicht definiert wurden, kann man sich der Eingabe der gesammelten Erhebungsdaten in der Datenansicht widmen. Da die Ergebnisse zumeist personen- bzw. fragebogenweise vorliegen, ist eine zeilenweise manuelle Eingabe vonnöten.

Dazu wählt man zunächst die erste Spalte der ersten Zeile mit einem Linksklick aus und überträgt die erste Merkmalsausprägung in die Datenmaske. Hat man viele geschlossene Fragen verwandt und die Antwortmöglichkeiten mit Wertelabels versehen, kann man sich gut am entsprechenden Codeplan orientieren und einfach den angegebenen Code in die Datenmaske übertragen. Das Eintippen der einzelnen Werte wird nicht (wie in vielen Office-Tabellendarstellungen) durch Enter bestätigt, sondern mit Pfeil nach rechts oder Tab abgeschlossen. Die Auswahl springt dann in die zweite Spalte der ersten Zeile und man kann die Eingabe fortsetzen. Zur Erleichterung der Navigation kann man in der Datenansicht dieselben Tastenfunktionen wie in der Variablenansicht verwenden (siehe oben, Kap. 3.2.2).

LÖSCHEN VON ZEILEN (PERSONEN) UND SPALTEN (VARIABLEN)

Zum Löschen von Zeilen und Spalten wählt man diese mit einem Linksklick aus und drückt die Taste ***Entf***. Will man mehrere Spalten bzw. Zeilen auf einmal bereinigen, wählt man diese durch die *Shift*-Taste zunächst aus und löscht dann durch die *Entf*-Taste. Man sollte jedoch darauf achten, dass beim Löschen von einer oder mehreren Zeilen sich automatisch die Nummerierung der Zellen ändert. Aus diesem Grund lohnt sich das Einfügen einer Identifikationsvariablen, die unabhängig von den vorgegebenen Zeilennummern bestehen bleibt.

EINFÜGEN LEERER ZEILEN (PERSONEN) UND SPALTEN (VARIABLEN)

Zunächst wählt man die Zeile (Spalte) an, vor der man ein weiteres Feld einfügen möchte. Will man eine Zeile einfügen, wählt man die Menüpunkte ***Daten → Fall einfügen*** an. Bei einer Variablen entsprechend ***Daten → Variable einfügen***. Neu erstellte Zeilen bzw. Spalten enthalten nur Kommas, die für systeminterne fehlende Werte stehen. Eine neue Variable erhält automatisch den Titel ***„var00001“***, lässt sich aber über die Variablenansicht genauer definieren (s. o.).

VERSCHIEBEN VON ZEILEN (PERSONEN) UND SPALTEN (VARIABLEN)

Zeilen und Spalten lassen sich durch einen Linksklick auswählen und durch Festhalten der linken Maustaste an eine beliebige Position der Datenmaske verschieben. Außerdem lassen sich ausgewählte Zeilen bzw. Spalten durch die Tastenkombination ***Strg + X*** ausschneiden und in einer vorher angelegten leeren Zeile bzw. Spalte durch Betätigung der Tasten ***Strg + V*** in Position bringen. Auch hier ist wieder auf die evtl. veränderte Nummerierung der Fälle zu achten.

3.4 ZUSAMMENFÜGEN VON DATENDATEIEN

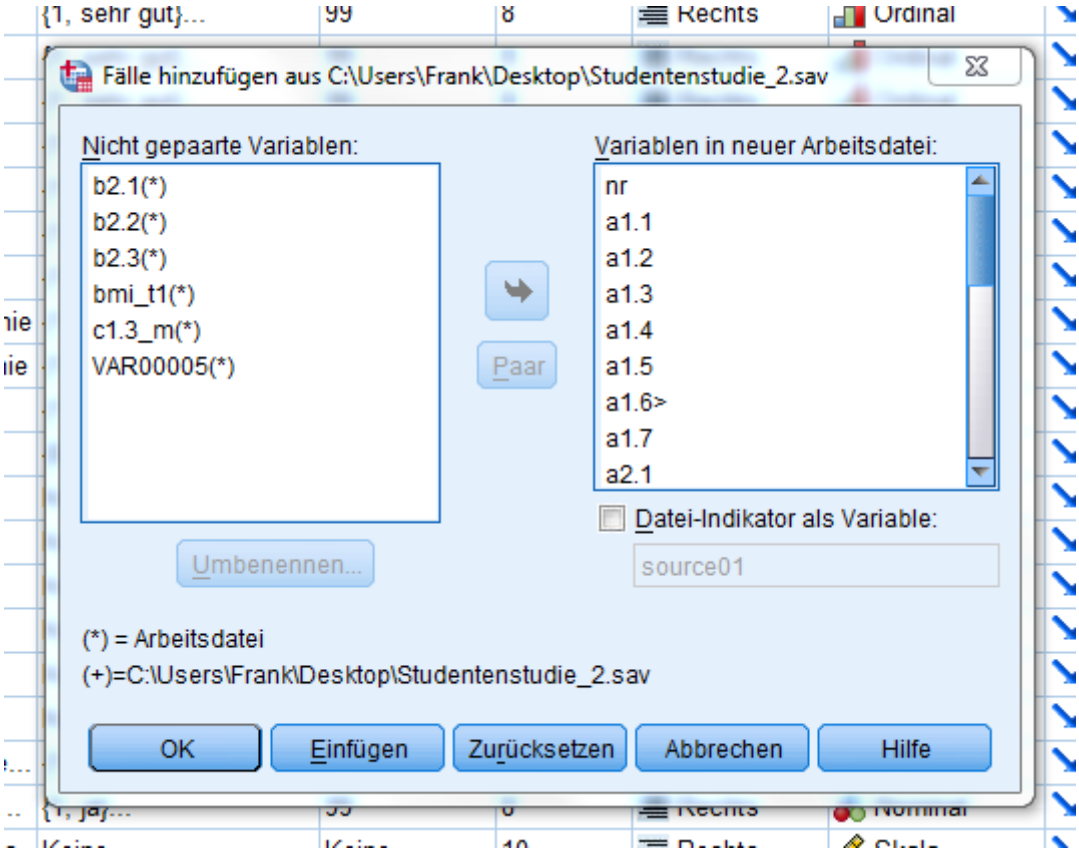
Vor allem, wenn man mit mehreren Personen an der Eingabe von Variablen bzw. Fällen in SPSS arbeitet, ist es hilfreich, sich mit dem Zusammenfügen verschiedener Datendateien auseinanderzusetzen. Wir haben zwei Möglichkeiten der Zusammenführung: Zum einen kann man neue Fälle hinzufügen und zum anderen neue Variablen ergänzen.

Beim Hinzufügen neuer Fälle wurden zwei oder mehr Gruppen von unterschiedlichen Fällen (zumeist Personen) nach einem zumindest ähnlichen Muster befragt. Beim Hinzufügen neuer Variablen werden die gleichen Personen befragt, die Datendatei unterscheidet sich jedoch in der Auswahl der Variablen, wie es bei einer Längsschnittstudie mit Erhebungen zu verschiedenen Zeitpunkten vorkommen kann.

HINZUFÜGEN NEUER FÄLLE

Zum Hinzufügen neuer Fälle geht man auf ***Daten → Dateien zusammenfügen → Fälle hinzufügen***. Nachdem Sie die entsprechende SPSS-Datendatei ausgewählt haben, öffnet sich die dargestellte Dialogbox (Abbildung 3.7).

Abbildung 3.7: Kontextmenü ***Fälle hinzufügen aus …***



Im linken Feld unter ***Nicht gepaarte Variablen*** werden die Variablen angezeigt, die kein Pendant in der zweiten Datei besitzen.

Ein * steht für eine Variable in der geöffneten Arbeitsdatei, die sich nicht in der hinzugefügten Datei befindet.

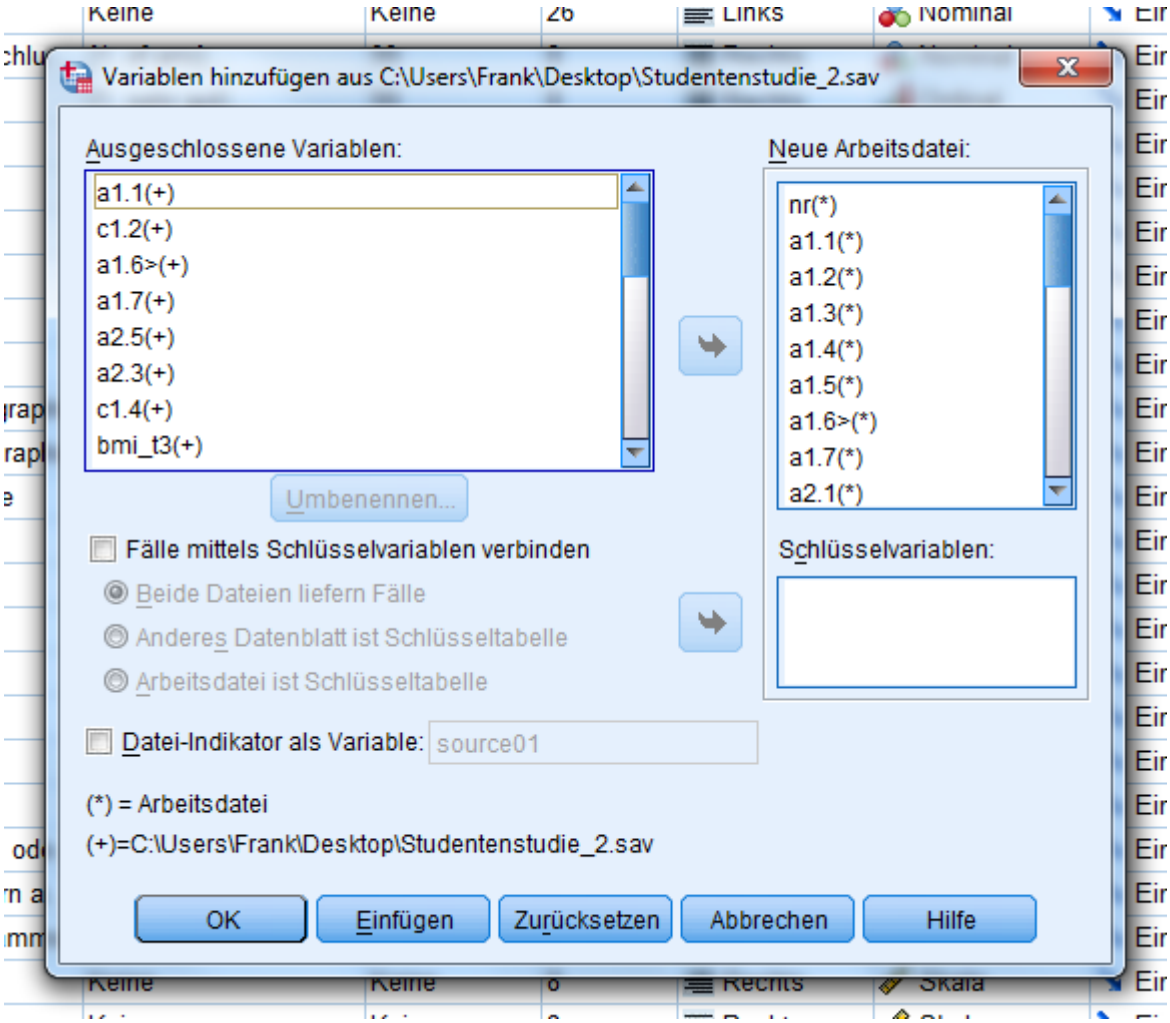
Ein + steht analog dazu für eine Variable in der hinzugefügten Datei, die nicht in der Arbeitsdatei vorliegt.

In der Liste am rechten Rand werden alle gepaarten Variablen aufgeführt. In diesem Auswahlfenster lassen sich Variablen entfernen, hinzufügen oder kombinieren. Auch deren Umbenennung ist möglich. Zum Entfernen wählt man die gewünschte Variable im rechten Fenster aus und verschiebt diese durch einen Klick auf den Pfeil nach rechts in die Liste unter ***Nicht gepaarte Variablen.*** Zum Hinzufügen geht man genau anders herum vor und zieht die Variablen vom linken Feld ins rechte Feld. Will man Variablen zusammenfügen, markiert man zunächst die beiden Variablen und klickt dann auf die Schaltfläche ***Paar***.

HINZUFÜGEN NEUER VARIABLEN

Zum Hinzufügen neuer Variablen wählt man die Befehlsfolge ***Daten → Dateien zusammenfügen → Variablen hinzufügen*** (Abbildung 3.8).

Abbildung 3.8: Kontextmenü ***Variablen hinzufügen aus…***



Grundsätzlich geht man hier ähnlich vor wie beim Zusammenfügen von Fällen. Allerdings muss man beachten, dass nur bei gleich vielen Fällen in beiden Datendateien eine einfache Zusammenführung möglich ist. Wird diese Voraussetzung nicht erfüllt, muss man eine Schlüsselvariable angeben, die den Ablauf der weiteren Operation bestimmt. In den meisten Fällen dient hierzu diejenige Variable, die die Fallnummer enthält. Dazu müssen die Fälle allerdings zuvor nach der Fallnummer sortiert worden sein. Um eine Schlüsselvariable festzulegen, klicken Sie zunächst auf das Kontrollkästchen ***Fälle mittels Schlüsselvariablen verbinden,*** dann zur gleichwertigen Behandlung beider Dateien auf ***Beide Dateien liefern Fälle***. Jetzt markiert man die Schlüsselvariable und überträgt diese durch einen Klick auf den Pfeil in das Feld ***Schlüsselvariablen***. Mit einem Klick auf ***Ok*** bestätigen wir die Auswahl.

Aufgabe 3.1
Sie haben die Aufgabe, den Teildatensatz einer Befragung unter Studenten zu analysieren. Ihr Teil umfasst folgende Variablen mit folgenden Ausprägungen:
G: Geschlecht (1 = weiblich, 2 = männlich)
S: Studiendauer in Semestern
E(S): Engagement im Studium mit 5 Kategorien (1 = sehr engagiert … 5 = gar nicht engagiert)
E(F): Engagement in der Freizeit mit 5 Kategorien (1 = sehr engagiert … 5 = gar nicht engagiert)
A: Ausrichtung der Abschlussarbeit (1 = empirisch, 2 = Literaturarbeit)
N: Note der Abschlussarbeit
Übertragen Sie diese Auflistung in eine SPSS-Dateneingabemaske.

Aufgabe 3.2
Übertragen sie die unten dargestellte Datenmatrix in die unter 3.1 erstellte Datenmaske.

Person (i)	**G**	**S**	**E(S)**	**E(F)**	**A**	**N**
1	1	9	1	2	1	2
2	2	6	3	3	2	2
3	2	12	2	2	1	1
4	1	7	4	1	2	4
5	1	6	2	3	1	3
6	1	10	2	2	1	3

Aufgabe 3.3
Fügen sie die Variable „Engagement im Studium“ (1 = sehr engagiert … 5 = gar nicht engagiert) zur Studentenstudie 2 hinzu. Diese soll außerdem den Namen ***a2.6*** bekommen und an entsprechender Stelle eingefügt werden. Die nach Fällen geordnete Urliste der Ausprägungen lautet wie folgt:
3, 4, 2, 1, 5, 2, 2, 3, 1, 4, 2, 1, 5, 3, 2, 4, 3, 2, 1, 4

4 DATENEXPLORATION

Wie erhält man zusammenfassende Informationen zur Datendatei?

Wie lassen sich doppelte Fälle und andere Fehleinträge ermitteln?

Wie erkennt man Ausreißer und Extremwerte?

Wie wählt man bestimmte Fälle aus?

Nach der Eingabe der Daten in SPSS sollte man sich nicht gleich in die Datenanalyse stürzen. Es empfiehlt sich, die erhobenen Daten zunächst einer ausführlichen Prüfung zu unterziehen. Dabei dient diese Prüfung nicht allein der Aufdeckung von Eingabefehlern, sondern bietet einen wichtigen ersten Eindruck über die Verteilungen der einzelnen Variablen. Auf dieser Grundlage lassen sich weitergehende Analysen aufbauen.

4.1 GRUNDLEGENDE INFORMATIONEN ZUR DATENDATEI

SPSS bietet zunächst einmal verschiedene Abfragen zu grundsätzlichen Informationen über die vorliegende Datendatei. Es lassen sich Variablenlisten und -beschreibungen anzeigen sowie eine Auflistung der Fälle durchführen.

4.1.1 AUFRUFEN VON VARIABLENINFORMATIONEN

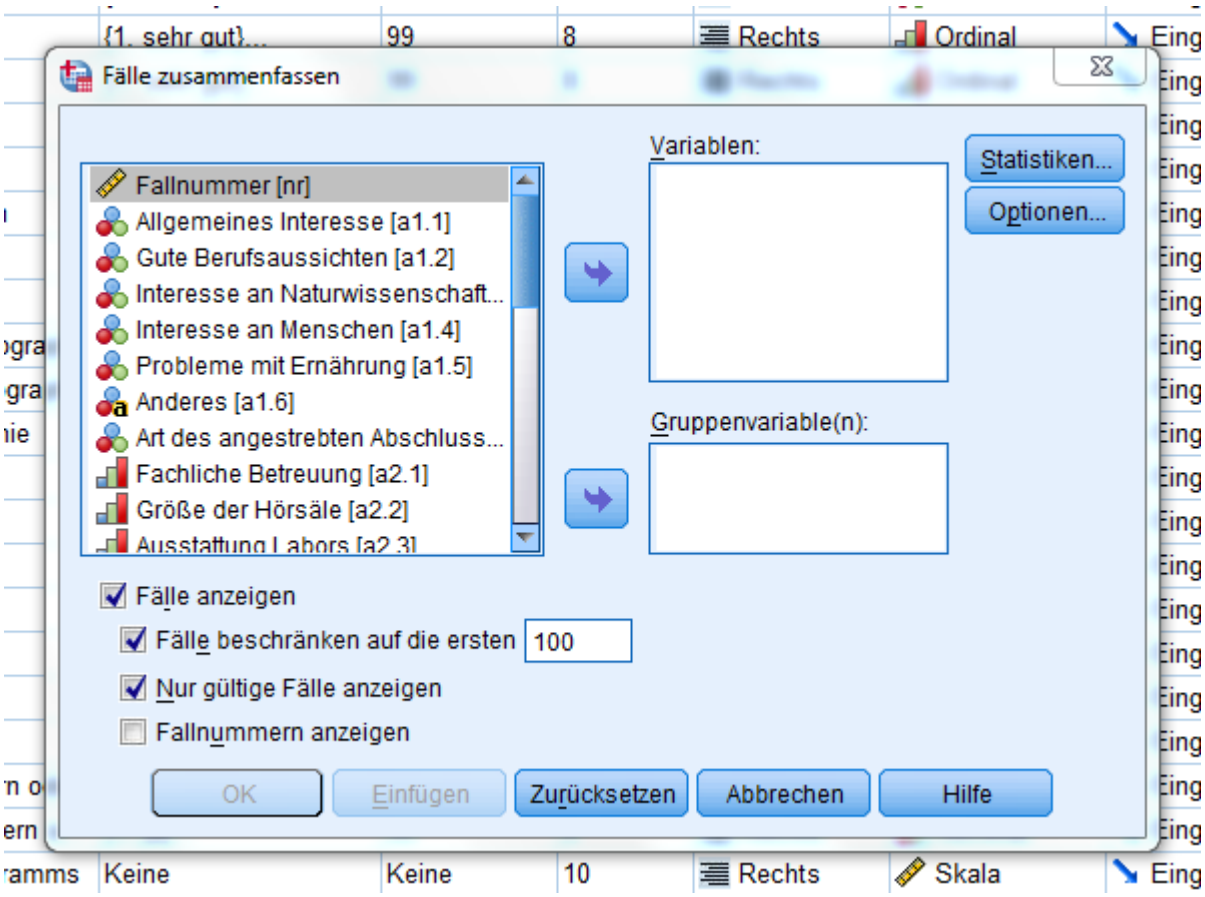
Möchten Sie sich eine einzelne Variable mit ihren Merkmalsausprägungen, ihrem Format und ihrer Etikettierung übersichtlich anzeigen lassen, wählen Sie ***Extras → Variablen***. Durch einen Klick auf ***Gehe zu*** springt man in die Variablenansicht und kann etwaige Änderungen vornehmen. Die wichtigsten Informationen zu den einzelnen Variablen lassen sich auch in verschiedenen anderen Dialogboxen abrufen. Dazu reicht ein ***Rechtsklick auf die entsprechende Variable***, dann wählen Sie ***Variablenbeschreibung***.

Möchten Sie dagegen Informationen über alle Variablen erhalten, beispielsweise für den Fall, dass Sie Ihre SPSS-Datei aus einer externen Quelle erhalten haben und sich zunächst einen Eindruck über deren Inhalt verschaffen wollen, wählen Sie aus dem Menü ***Datei → Datendatei-Informationen anzeigen → Arbeitsdatei***.

4.1.2 ZUSAMMENFASSEN VON FALLINFORMATIONEN

Wenn Sie sich für die Anzeige der Fälle in Bezug auf eine Variable bzw. deren Merkmalsausprägungen interessieren, wählen Sie ***Analysieren → Berichte → Fälle zusammenfassen***. Es öffnet sich folgende Dialogbox (siehe Abbildung 4.1).

Abbildung 4.1: Kontextmenü ***Fälle zusammenfassen***



Auf der linken Seite der Box werden die verschiedenen Variablen aufgelistet, von denen Sie eine oder mehrere anwählen können, um deren Fälle darzustellen. Durch einen Klick auf den Pfeil nach rechts werden die ausgewählten Merkmale in das Dialogfeld ***Variablen*** übertragen. Vor der Bestätigung der Eingabe durch einen Klick auf ***Ok***, kann man sich noch weitere statistische Kennzahlen anzeigen lassen oder Diagramme zur Verteilung anlegen (mehr dazu in den Kapiteln 6, 7 und 9).

Aufgabe 4.1
Lassen Sie sich die Fälle für die Variable „Alter“ in der „Studentenstudie 1“ anzeigen. Welche Fallnummer hat ein Alter von 43 Jahren angegeben?

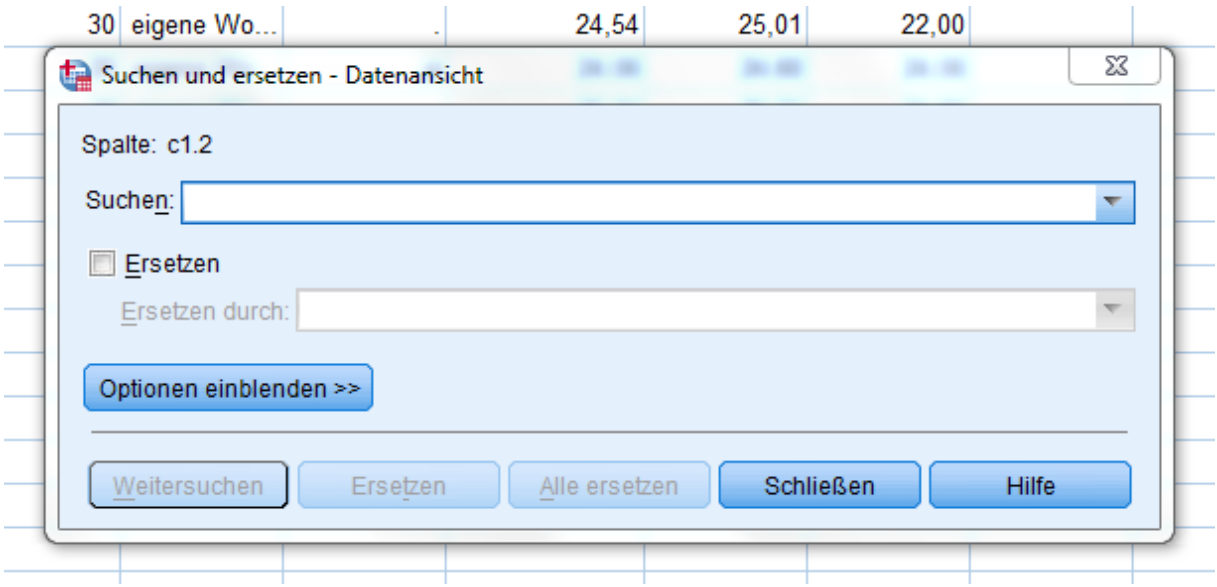
4.1.3 AUFSUCHEN VON FÄLLEN, MERKMALSAUSPRÄGUNGEN UND VARIABLEN

Gerade im Umgang mit besonders umfangreichen Datendateien machen sich die verschiedenen Navigationsbefehle für SPSS bezahlt. Diese bieten verschiedene Funktionen zum schnellen Auffinden von speziellen Fällen, Merkmalsausprägungen oder Variablen und werden im Folgenden kurz vorgestellt.

Geht es darum, bestimmte Merkmalsausprägung einer Variablen zu finden, wählt man in der Datenansicht ***Bearbeiten → Suchen*** oder benutzt wahlweise die Tastenkombination ***Strg + F***.

Es öffnet sich ein Kontextmenü, welches wir durch einen Klick auf ***Optionen einblenden*** erweitern und somit Abbildung 4.2 erhalten.

Abbildung 4.2: Dialogbox **Suchen und ersetzen: Datenansicht**



Durch einen Klick auf die nachgelagerte Datenansicht können wir die uns interessierende Spalte auswählen. Das entsprechende Suchfeld ermöglicht nun die Eingabe des gesuchten Wertes. Weitere Suchoptionen bietet die erweiterte Ansicht und es ist auch möglich, alle den Suchkriterien entsprechenden Merkmalsausprägungen durch neue Eingaben zu ersetzen.

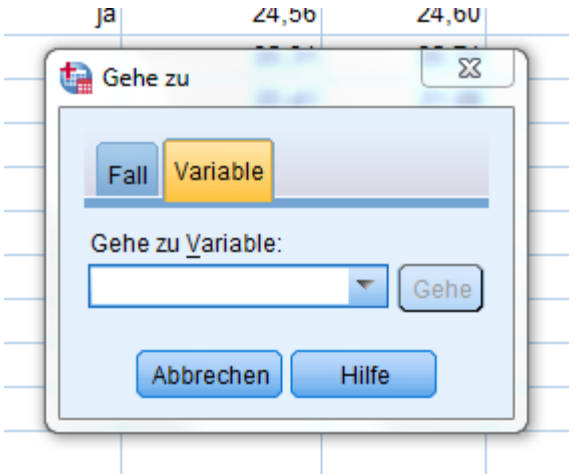
Wollen wir nur einen speziellen Fall aufsuchen, beispielsweise da uns Unregelmäßigkeiten in der Dateneingabe aufgefallen sind, wählen wir ***Bearbeiten → Gehe zu Fall*** und erhalten das in Abbildung 4.3 gezeigte Kontextmenü.

Abbildung 4.3: Dialogbox ***Gehe zu Fall***

Hier geben wir einfach die Nummer des gesuchten Falls ein und die SPSS-Datenansicht springt in die entsprechende Zeile.

Wollen wir eine bestimmte Variable ausfindig machen, können wir entweder die unter Abbildung 4.3 angedeutete Registerkarte Variable auswählen oder über ***Bearbeiten → Gehe zu Variable*** das gezeigte Kontextmenü öffnen (Abbildung 4.4).

Abbildung 4.4: Dialogbox ***Gehe zu Variable***



Hier geben wir die gesuchte Variable ein und die Datenansicht springt nach Bestätigung der Auswahl auf das entsprechende Merkmal.

4.2 DOPPELTE FÄLLE ERMITTELN

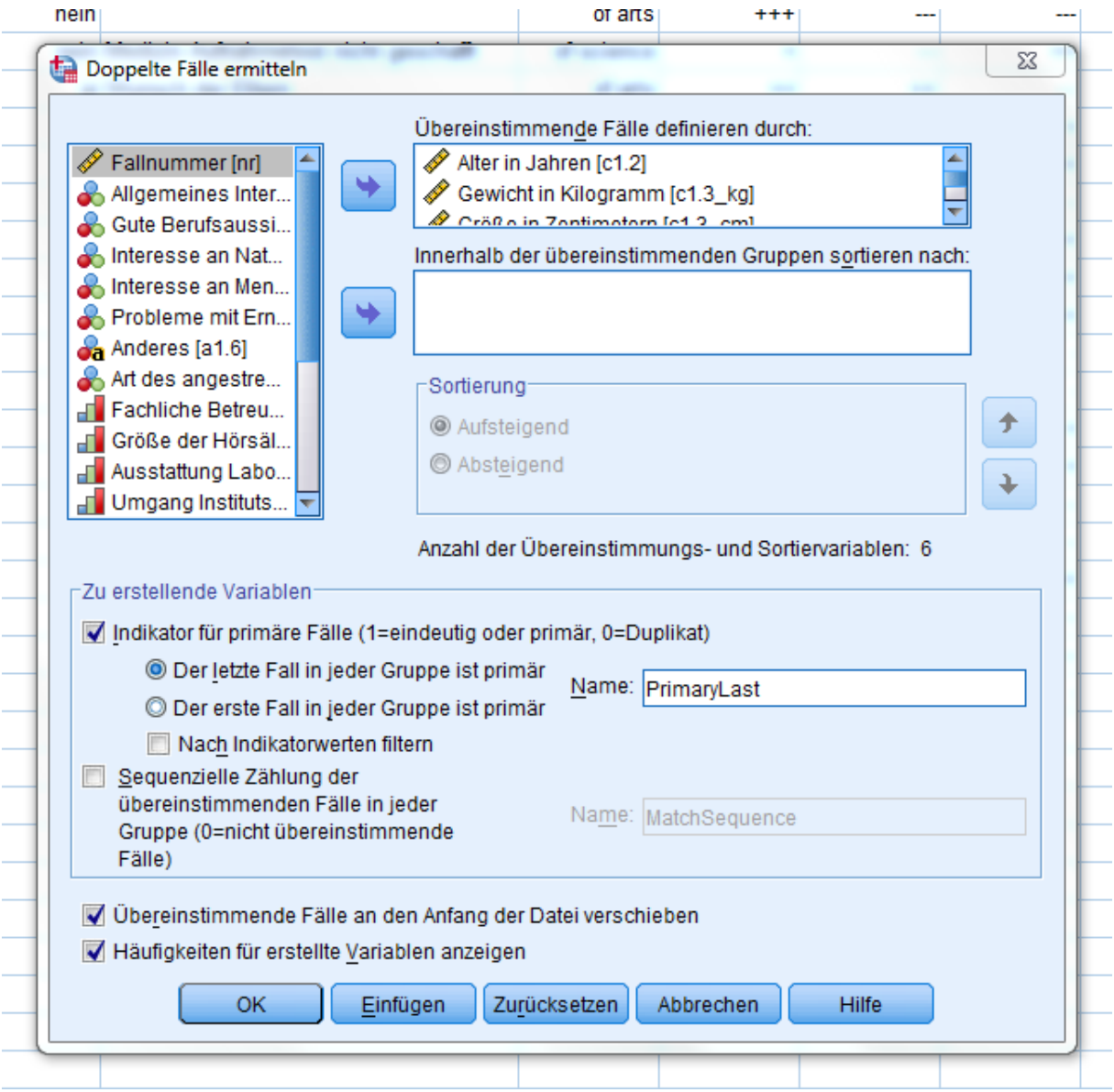
Eine Möglichkeit zur Aufdeckung von Eingabefehlern ist die Ermittlung doppelter Fälle. Dazu wählt man zunächst ***Daten → Doppelte Fälle ermitteln***. Es öffnet sich das in Abbildung 4.5 dargestellte Kontextmenü.

Zunächst müssen nun alle Variablen ausgewählt werden, mit deren Hilfe man Dubletten aufdecken möchte (mit ***Strg + A*** lassen sich alle Variablen auswählen). Wieder klicken wir die relevanten Variablen an und übertragen diese ins rechte Auswahlfenster.

Es bleibt anzumerken, dass man mit der Operation ***Doppelte Fälle ermitteln*** natürlich nicht nur Fehler in der Datendatei ermitteln kann, sondern ebenso Fälle für Matching oder nachträgliches Matching auswählen kann (beispielsweise, um in Bezug auf bestimmte Variablen übereinstimmende Fälle aus verschiedenen Untersuchungsgruppen auszuwählen).

Aus der „Studentenstudie 2“ könnte man beispielsweise die Variablen *„Gewicht“*, *„Alter“*, *„Größe“*, *„Bewegung pro Woche“* und *„BMI nach 8 Wochen“* als Indikatorvariablen wählen. Wir übernehmen die voreingestellten *„Auswahlen für die zu erstellende Variable“*, die *„Ordnung der übereinstimmenden Fälle“* und die *„Anzeige der Häufigkeiten“*. Nun bestätigen wir die getroffene Auswahl mit einem Klick auf ***Ok***.

Abbildung 4.5: Dialogbox ***Doppelte Fälle ermitteln***



Es geschehen drei Dinge: Zunächst einmal öffnet sich das Ausgabefenster, aus dem zu entnehmen ist, wie viele primäre und doppelte Fälle in der Datei enthalten sind. Außerdem ergänzt SPSS eine neue Variable in der Datendatei, die per Voreinstellung den Titel ***PrimaryLast*** erhält. Diese enthält Werte zwischen 1 und 0, wobei 1 für einen primären Fall und 0 für einen doppelten Fall steht. Schließlich werden die Daten in der Datendatei so umsortiert, dass die Fälle mit Dubletten an den Anfang der Datei gestellt werden.

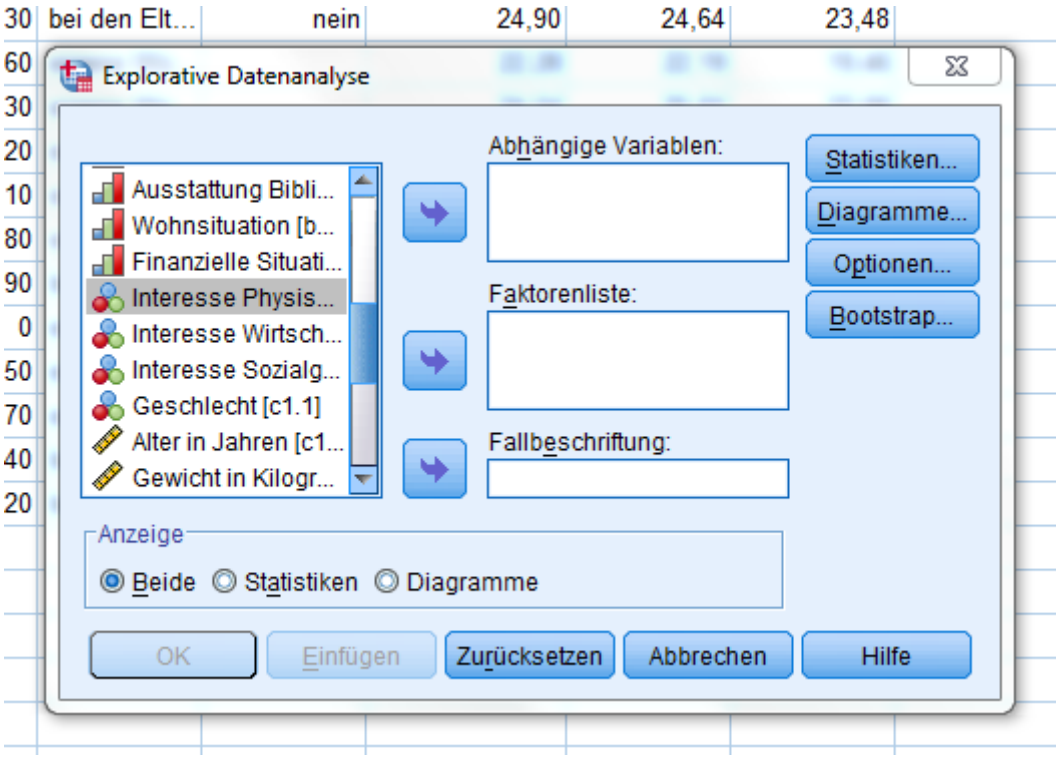
Aufgabe 4.2
Befinden sich Dubletten in der „Studentenstudie 2“? Wenn ja, bereinigen Sie die Datendatei.

4.3 AUSREIßER, EXTREMWERTE UND FEHLEINTRÄGE

Nachdem sichergestellt wurde, dass die Datendatei keine Dubletten enthält, kann man sich einen ersten Überblick über die Verteilungen der einzelnen Variablen verschaffen. Dabei konzentrieren wir uns insbesondere auf die Suche nach Fehleinträgen und der Darstellung von Ausreißern und Extremwerten. Wir wollen diese Operation an der „Studentenstudie 2“ verdeutlichen.

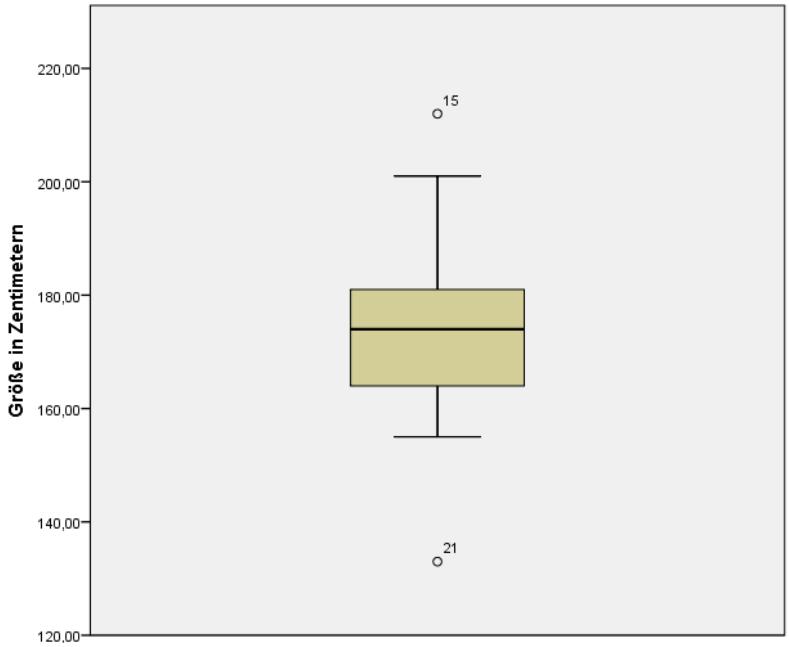
Wir wählen zunächst das Menü ***Analysieren → Deskriptive Datenanalyse → Explorative Datenanalyse*** aus. Es öffnet sich die in Abbildung 4.6 dargestellte Dialogbox.

Abbildung 4.6: Dialogbox ***Explorative Datenanalyse***



In dieser Ansicht wird zwischen einer abhängigen Variablen und Faktoren unterschieden. Diese Unterscheidung dient der getrennten Analyse für Gruppen und für Fälle, die uns allerdings im Moment nicht interessieren. Also ignorieren wir die Faktorenliste und wählen die Variablen *„Größe“* aus der *„Studentenstudie 2“* als zu untersuchende abhängige Variable aus. Dazu wählen wir die Variable an und fügen diese über den Pfeil nach rechts in das Auswahlfenster ein. Bevor wir die Berechnung durch einen Klick auf ***Ok*** starten, wählen wir unter ***Diagramme*** noch die Option, ein Histogramm anzeigen zu lassen. Nun müsste sich abermals das Ausgabefenster von SPSS öffnen. Hier finden wir verschiedene tabellarische Zusammenfassungen zu Anzahl, Gültigkeit und statistischen Kennzahlen der ausgewählten Variablen (siehe Kapitel 9) sowie das ausgewählte Histogramm und einen Boxplot (siehe Kapitel 7). Zum Auffinden von Ausreißern und Extremwerten eignet sich insbesondere der in Abbildung 4.7 dargestellte Boxplot.

Abbildung 4.7: Boxplot der Variable *„Größe“*



Eine genauere Erläuterung zur Erstellung und Interpretation des Boxplot-Diagramms findet sich in Kapitel 7.3. Einiges muss an dieser Stelle jedoch vorweggenommen werden. Die etwas dickere schwarze Linie stellt die Grenze zwischen den oberen und den unteren 50 % der Werte dar (den Median). Das braune Rechteck gibt den Bereich zwischen dem 25 %-Perzentil und dem 75 %-Perzentil an. Somit entspricht dieses Rechteck 50 % der Verteilung. Wir können am dargestellten Boxplot außerdem erkennen, dass zwei Werte deutlich außerhalb des eigentlichen Boxplots liegen. Hier haben wir es mit Ausreißern zu tun. Würden bei der zu untersuchenden Verteilung Extremwerte vorliegen, wären diese durch Sternchen gekennzeichnet werden. Diese Ausreißer bzw. Extremwerte könnten Fehleinträge darstellen, vielleicht sind sie jedoch auch nur ein Teil der erhobenen Verteilung. Auf jeden Fall lohnt sich hier eine genauere Betrachtung der Fälle 15 und 21.

Aufgabe 4.3
Untersuchen Sie die Variable „Gewicht in kg“ in der „Studentenstudie 2“ auf Ausreißer, Extremwerte und Fehleinträge. Müssen Datenmodifikationen zur Homogenisierung der Verteilung vorgenommen werden?

Eine Möglichkeit, Fehleinträge insbesondere bei der Untersuchung gelabelter Variablen aufzudecken, bietet sich über die Anzeige der Häufigkeitsverteilung. Hierzu gehen Sie bitte auf ***Analysieren → Deskriptive Statistiken → Häufigkeiten***. Wir befinden uns im Menü Häufigkeiten wie in Abbildung 4.8 dargestellt.

Abbildung 4.8: Kontextmenü ***Häufigkeiten***

Wieder wählen wir die zu untersuchende Variable aus. In diesem Fall die Variable *„Finanzielle Situation“* aus der Studentenstudie 2 und klicken danach auf ***Ok***. Abermals öffnet sich das Ausgabefenster und es erscheint die in Abbildung 4.9 gezeigte Tabelle.

Abbildung 4.9: Darstellung der Häufigkeiten für die Variable *Finanzielle Situation*

Finanzielle Situation

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	+++	3	15,0	15,8	15,8
	++	1	5,0	5,3	21,1
	+	3	15,0	15,8	36,8
	-	3	15,0	15,8	52,6
	--	4	20,0	21,1	73,7
	---	4	20,0	21,1	94,7
	7,00	1	5,0	5,3	100,0
	Gesamt	19	95,0	100,0	
Fehlend	99,00	1	5,0		
Gesamt		20	100,0		

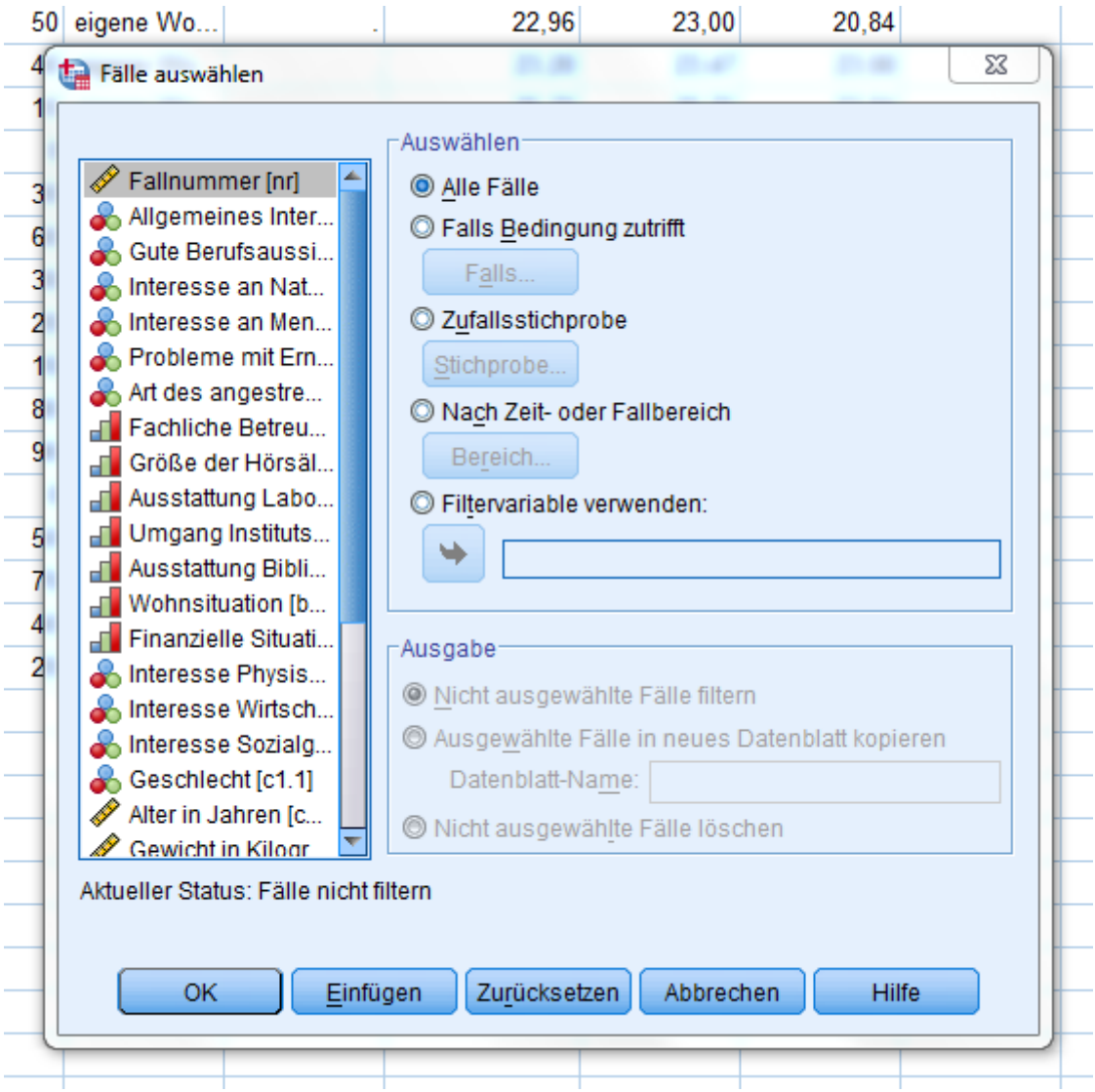
In dieser Tabelle kann man erkennen, dass eine Person die Frage nicht beantworten konnte (oder wollte) und damit als fehlender Wert gezählt wird. Da die Variable neben den sechs Kategorien und einem fehlenden Wert eine weitere Kategorie beinhaltet, haben wir hier eine unzulässige Codierung entdeckt: Bei der Ausprägung ***„7“*** handelt es sich offensichtlich um einen Fehleintrag.

Aufgabe 4.4
Finden Sie den oben dargestellten Fehleintrag in der Studentenstudie 2 und ersetzten Sie ihn durch die Codierung für einen fehlenden Wert.

4.4 FÄLLE AUSWÄHLEN UND INKONSISTENTE ANTWORTEN FINDEN

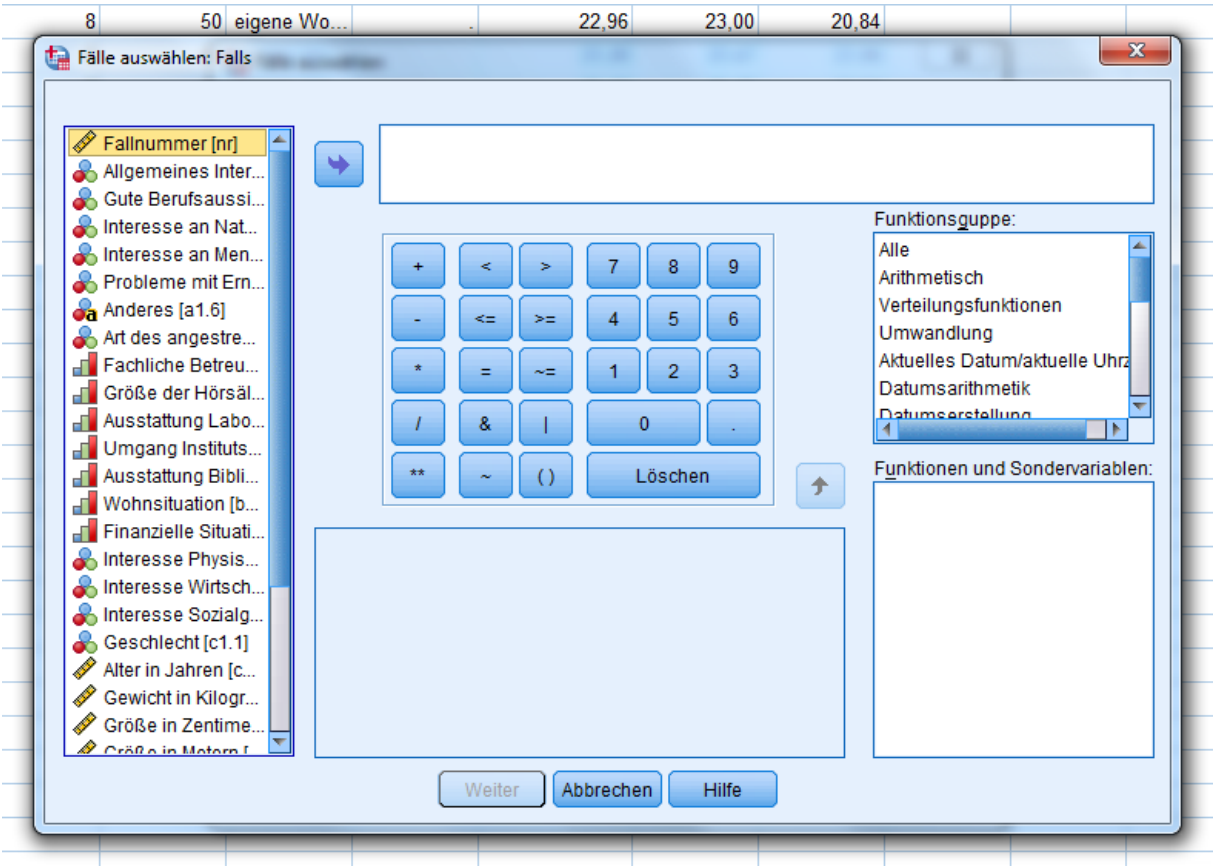
Viele Fehler lassen sich durch die bereits beschriebenen Methoden der Explorativen Datenanalyse und der Darstellung der Häufigkeiten ausfindig machen. Doch wie geht man bei inkonsistenten Antworten vor? Eine inkonsistente Antwort liegt dann vor, wenn eine Person eine Angabe zu einer Frage gemacht hat, die ihn aus inhaltlichen Gründen gar nicht betrifft. Dieser Fall tritt häufig auf, wenn viele Sprunganweisungen in einem Fragebogen verbaut haben. Um in SPSS nach solchen Fällen zu suchen, kann man entweder die Prozedur ***Fälle auswählen*** oder ***Kreuztabellen*** verwenden. Da Kreuztabellen hier erst in Kapitel 6 näher beschrieben werden, wählen wir in diesem Fall ***Daten → Fälle auswählen***. Es öffnet sich folgendes Dialogfenster (siehe Abbildung 4.10).

Abbildung 4.10: Dialogfenster ***Fälle auswählen***



Sie müssen nun zunächst eine Bedingung formulieren, durch die inkonsistente Antworten aufgedeckt werden können. Nehmen wir abermals die Studentenstudie 2 zur Hand, fällt bei der Betrachtung der Variablen auf, dass eigentlich nur die Personen, die bei Frage d1.1 angegeben haben, dass sie noch zu Hause wohnen (Ausprägung „1“), auf Frage d1.2 (Wunsch bald bei den Eltern auszuziehen) eine Antwort geben dürften. Überprüfen wir nun, ob dem auch so ist. Dazu klicken wir zunächst auf die Schaltfläche ***Falls Bedingung zutrifft*** und danach auf ***Falls***. Dann öffnet sich folgende Dialogbox (siehe Abbildung 4.11)

Abbildung 4.11: Dialogbox ***Fälle Auswählen: Falls***



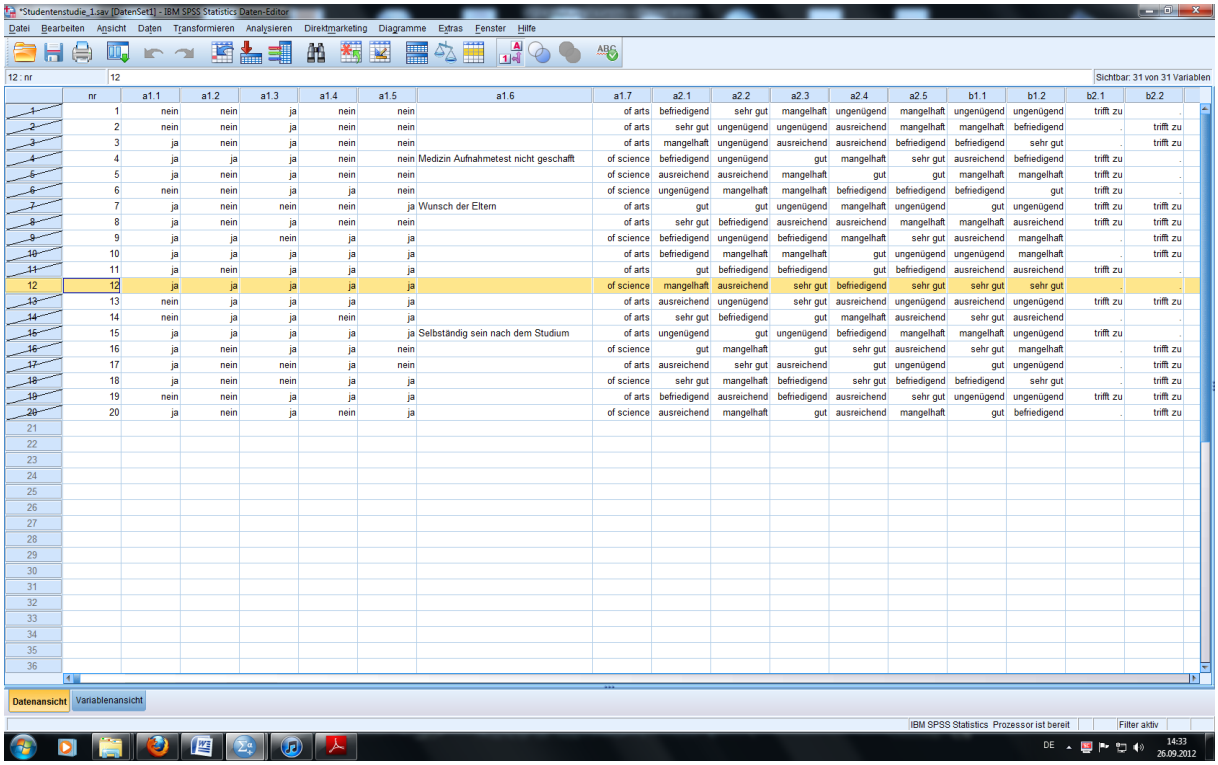
In diesem Optionsmenü können wir nun verschiedene Bedingungen der Variablenauswahl mathematisch beschreiben. Eine detaillierte Beschreibung der hier verwendeten Operatoren finden Sie in Kapitel 5.2.3. Wir beschränken uns hier beispielhaft auf die Bedingung

d1.1 = 2 AND (d1.2 = 1 OR d1.2 = 2)

die sie entweder per Hand oder unter Zuhilfenahme der Variablenauswahl und des vorgegebenen Tastenfeldes eingeben können. Im Einzelnen bedeutet diese Bedingung: Wähle alle Fälle aus, in denen der Proband die Frage nach dem „Wunsch bald bei den Eltern auszuziehen“ mit „Ja“ oder „Nein“ beantwortet, obwohl er die Frage nach dem Wohnort mit 2 für „eigene Wohnung“ beantwortete hat.

Wir bestätigen die Auswahl und gehen zurück in die Datenansicht. Hier können Sie auf den ersten Blick schnell erkennen, welche Fälle beide Bedingungen erfüllen und damit ungültig sind (Abbildung 4.12).

Abbildung 4.12: Datenansicht nach Auswahl der Fälle



	nr	a1.1	a1.2	a1.3	a1.4	a1.5	a1.6	a1.7	a2.1	a2.2	a2.3	a2.4	a2.5	b1.1	b1.2	b2.1	b2.2
1	1	nein	nein	ja	nein	nein		of arts	befriedigend	sehr gut	mangelhaft	ungenügend	mangelhaft	ungenügend	ungenügend	trifft zu	.
2	2	nein	nein	ja	nein	nein		of arts	sehr gut	ungenügend	ungenügend	ausreichend	mangelhaft	mangelhaft	befriedigend	.	trifft zu
3	3	ja	nein	ja	nein	nein		of arts	mangelhaft	ungenügend	ausreichend	ausreichend	befriedigend	befriedigend	sehr gut	.	trifft zu
4	4	ja	ja	ja	nein	nein	Medizin Aufnahmetest nicht geschafft	of science	befriedigend	ungenügend	gut	mangelhaft	sehr gut	ausreichend	befriedigend	trifft zu	.
5	5	ja	nein	ja	nein	nein		of science	ausreichend	ausreichend	mangelhaft	gut	gut	mangelhaft	mangelhaft	trifft zu	.
6	6	nein	nein	ja	ja	nein		of science	ungenügend	mangelhaft	mangelhaft	befriedigend	befriedigend	befriedigend	gut	trifft zu	.
7	7	ja	nein	nein	nein	ja	Wunsch der Eltern	of arts	gut	gut	ungenügend	mangelhaft	ungenügend	gut	ungenügend	trifft zu	trifft zu
8	8	ja	nein	ja	nein	nein		of arts	sehr gut	befriedigend	ausreichend	ausreichend	mangelhaft	mangelhaft	ausreichend	trifft zu	trifft zu
9	9	ja	ja	nein	ja	ja		of science	befriedigend	ungenügend	befriedigend	mangelhaft	sehr gut	ausreichend	mangelhaft	.	trifft zu
10	10	ja	ja	ja	ja	ja		of arts	befriedigend	mangelhaft	mangelhaft	gut	ungenügend	ungenügend	mangelhaft	.	trifft zu
11	11	ja	nein	ja	ja	ja		of arts	gut	befriedigend	befriedigend	gut	befriedigend	ausreichend	ausreichend	trifft zu	.
12	12	ja	ja	ja	ja	ja		of science	mangelhaft	ausreichend	sehr gut	befriedigend	sehr gut	sehr gut	sehr gut	.	.
13	13	nein	ja	ja	ja	ja		of arts	ausreichend	ungenügend	sehr gut	ausreichend	ungenügend	ausreichend	ungenügend	trifft zu	trifft zu
14	14	nein	ja	ja	nein	ja		of arts	sehr gut	befriedigend	gut	mangelhaft	ausreichend	sehr gut	ausreichend	.	.
15	15	ja	ja	ja	ja	ja	Selbständig sein nach dem Studium	of arts	ungenügend	gut	ungenügend	befriedigend	mangelhaft	mangelhaft	ungenügend	trifft zu	.
16	16	ja	nein	ja	ja	nein		of science	gut	mangelhaft	gut	sehr gut	ausreichend	sehr gut	mangelhaft	.	trifft zu
17	17	ja	nein	nein	ja	nein		of arts	ausreichend	sehr gut	ausreichend	gut	ungenügend	gut	ungenügend	.	trifft zu
18	18	ja	nein	nein	ja	ja		of science	sehr gut	mangelhaft	befriedigend	sehr gut	befriedigend	befriedigend	sehr gut	.	trifft zu
19	19	nein	nein	ja	ja	ja		of arts	befriedigend	ausreichend	befriedigend	ausreichend	sehr gut	ungenügend	ungenügend	trifft zu	trifft zu
20	20	ja	nein	ja	nein	ja		of science	ausreichend	mangelhaft	gut	ausreichend	mangelhaft	gut	befriedigend	.	trifft zu

Abbildung 4.12 zeigt, dass bis auf Zeile Nr. 12 alle Fälle durchgestrichen wurden und damit Zeile 12 – der selektierte Fall – einer näheren Überprüfung bedarf.

Die Auswahl von Fällen nützt natürlich nicht nur bei der Fehlersuche, sondern kann auch zur Analyse einer bestimmten Untergruppe der Befragten verwandt werden. Beispielsweise könnte es sein, dass man nur die Antworten der Männer betrachten möchte und die Frauen aus den folgenden Berechnungen außen vor bleiben sollen. Hier geht man wie oben beschrieben vor. Es ist jedoch stets darauf zu achten, dass man die getätigte Auswahl vor den weiteren Berechnungen für die gesamte Stichprobe wieder entfernt. Dazu wählt man unter ***Daten → Fälle auswählen*** die Schaltfläche ***Alle Fälle*** an.

Aufgabe 4.5
Gibt es weitere inkonsistente Antworten in der Studentenstudie 2? Wenn ja, finden Sie diese und entscheiden Sie über deren Verbleib in der Datendatei.

5 WEITERE DATENMODIFIKATIONEN

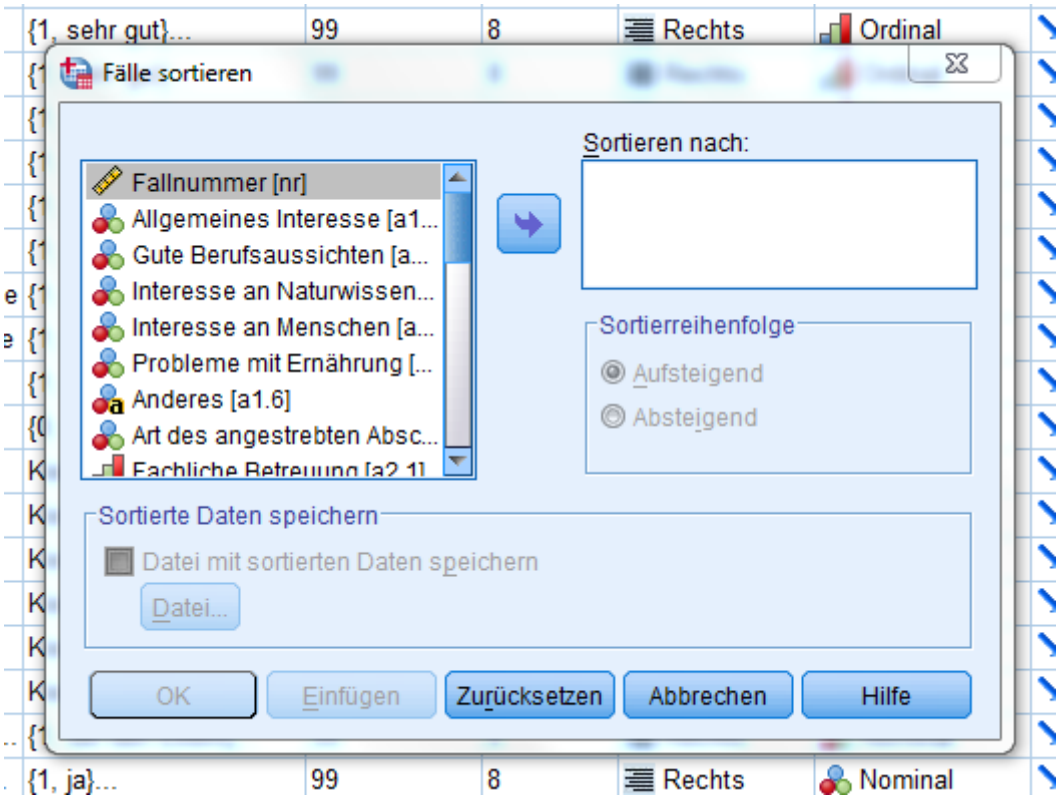
Wie kann man Daten sortieren, umstrukturieren, klassieren und transformieren?

SPSS bietet eine Reihe von Möglichkeiten, Daten zu modifizieren. Diese lassen sich sortieren, gewichten oder neu berechnen.

5.1 DATEN SORTIEREN

Für verschiedene Zwecke ist es nützlich, die vorliegenden Daten in einer bestimmten Reihenfolge zu sortieren. Hat man beispielsweise einen Fehler in der Datenmaske entdeckt, lässt sich dieser nach einer Sortierung nach der Fallnummer viel einfacher aufdecken. Um eine Sortierung vorzunehmen, klickt man auf ***Daten → Fälle sortieren***. Jetzt öffnet sich das in Abbildung 5.1 dargestellte Kontextmenü.

Abbildung 5.1: Dialogbox ***Fälle sortieren***



Wir wählen nun die Sortiervariable aus der Quellvariablenliste aus. Es lassen sich auch mehrere Variablen auswählen und in eine Sortierreihenfolge bringen. Wird nach mehr als einer Variablen sortiert, erfolgt die Sortierung der folgenden jeweils innerhalb der vorausgehenden Variablen. Durch Abspeichern kann die Datei dauerhaft in die sortierte Form gebracht werden.

Aufgabe 5.1
Sortieren Sie die Daten in der „Studentenstudie 2“ sowohl nach dem Geschlecht als auch nach dem Alter. Dabei sollte die Sortierreihenfolge auf „Absteigend“ eingestellt werden und die Sortierung in einer neuen Datei gespeichert werden. Welche Fallnummer steht in der ersten Zeile der SPSS-Datenansicht?

5.2 DATEN UMSTRUKTURIEREN

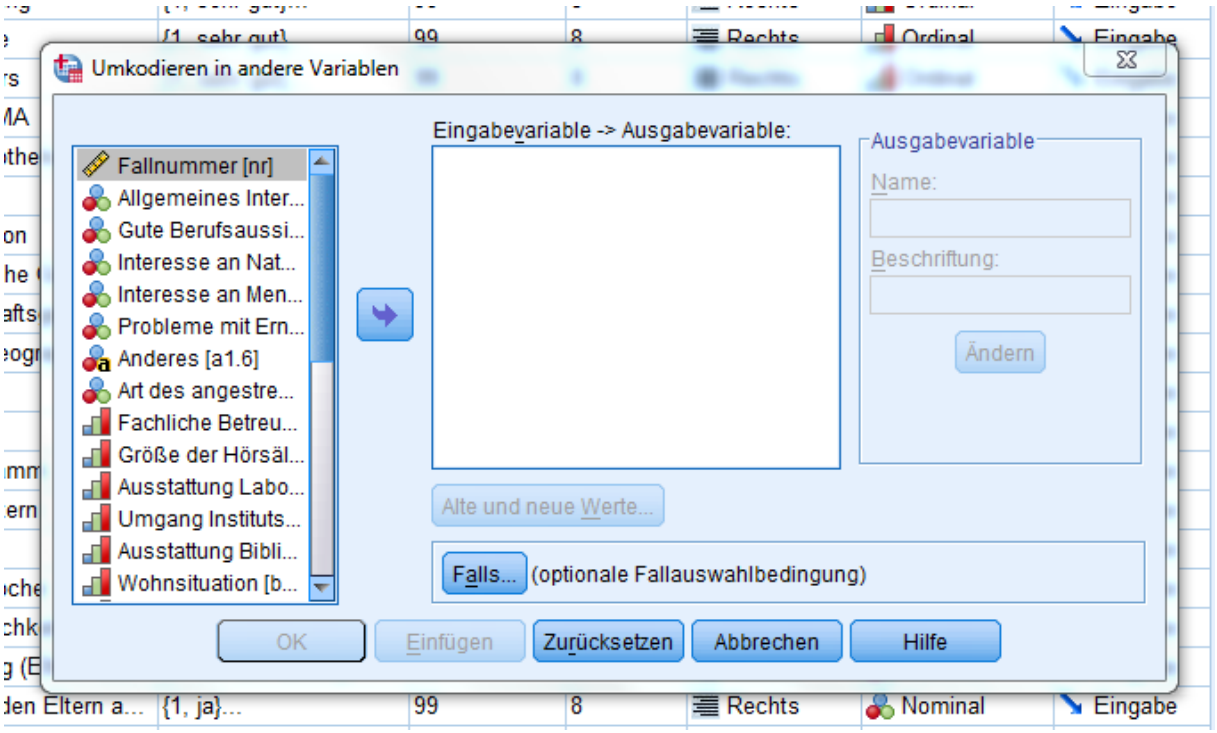
5.2.1 UMCODIEREN VON VARIABLEN

Eine der am häufigsten verwendeten Datenmodifikation beschäftigt sich mit der Umcodierung von Variablen. Diese Funktion lässt sich beispielsweise zur Zusammenfassung von Kategorien bzw. der Bildung von Klassen bei der Analyse metrischer Daten nutzen. Das primäre Ziel einer solchen Umcodierung sollte die möglichst genaue Übertragung der Verteilung der (ursprünglichen) Einzelwerte in die Verteilung der Klassenwerte sein. Hierzu gilt es, sechs Faustregeln zu beachten:

1. Verschiedene Klassen dürfen sich nicht überdecken und es dürfen keine Lücken entstehen
2. Alle Werte müssen abgedeckt sein
3. Klassengrenzen sollten Bereiche ähnlicher Werte nicht trennen
4. Alle Klassen, insbesondere die mittleren, sollten besetzt sein
5. Die Klassengrenzen sollten möglichst einfache Zahlen sein
6. Die Größen der Klassen sollten nicht übergebührlich voneinander abweichen

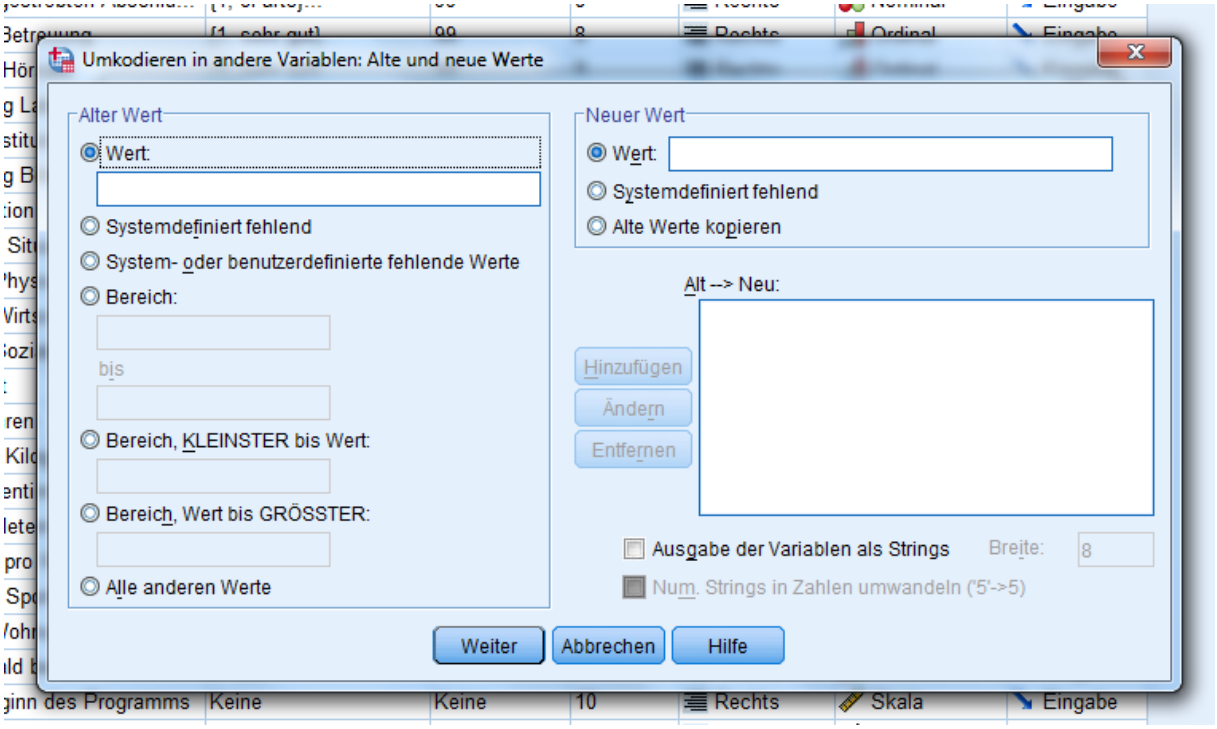
Zur Bildung von Klassen bietet SPSS zwei verschiedene Prozeduren an. Zum einen das Umcodieren in einer bereits bestehenden Variablen und zum anderen das Umcodieren in eine andere Variable. Im Zweifel sollte man sich immer auf die Transformation in eine neue Variable beschränken, da man auf diese Weise Datenverluste vermeidet. Um eine solche Umcodierung durchzuführen, wählen Sie den Menübefehl: ***Transformieren → Umcodieren in andere Variable*** (Abbildung 5.2).

Abbildung 5.2: Kontextmenü **Umcodieren in eine andere Variable**



Wir wählen zunächst die Variable, die transformiert werden soll. Dann muss man den Namen der neuen Variablen festlegen und durch einen Klick auf ***Alte und neue Werte*** eben diese definieren (Abbildung 5.3).

Abbildung 5.3: Kontextmenü **Umcodieren in eine andere Variable: Alte und neue Werte**



Hier lassen sich Einzelwerte, Reichweiten, fehlende Werte und die dazu passenden neuen Werte definieren. Im linken Teil der Abbildung sieht man die Angaben für die alten Werte bzw. Wertebereiche, im rechten die neu zu definierenden Einzelwerte (Reichweitenangaben sind hier nicht möglich). Hat man eine Umkodierungsvorschrift definiert, klickt man auf ***Hinzufügen*** und die Bedingung wird unter dem Feld ***Alt → Neu*** angezeigt. Diesen Vorgang wiederholt man so lange, bis alle Vorschriften beschrieben wurden. Auch die Umcodierung einer bestimmten Auswahl der Fälle ist durch einen Klick auf die Schaltfläche ***Falls*** möglich. Die hierfür zuständige Dialogbox hat denselben Aufbau wie Abbildung 4.11 und wurde in Kapitel 4.4 näher erläutert. Zusätzlich lässt sich unter dem Feld ***Label*** ein Variablen-Label für die neue Variable vergeben. Im rechten unteren Rand lassen sich die die Breite und der Variablentyp anpassen. Durch einen Klick auf ***Weiter*** kommen wir in das in Abbildung 5.2 gezeigte Menü zurück. Hier bestätigen wir durch einen Klick auf ***OK***. Nun setzt SPSS die gegebene Definition um und erstellt eine neue Variable, die an das Ende der Variablenliste angefügt wird.

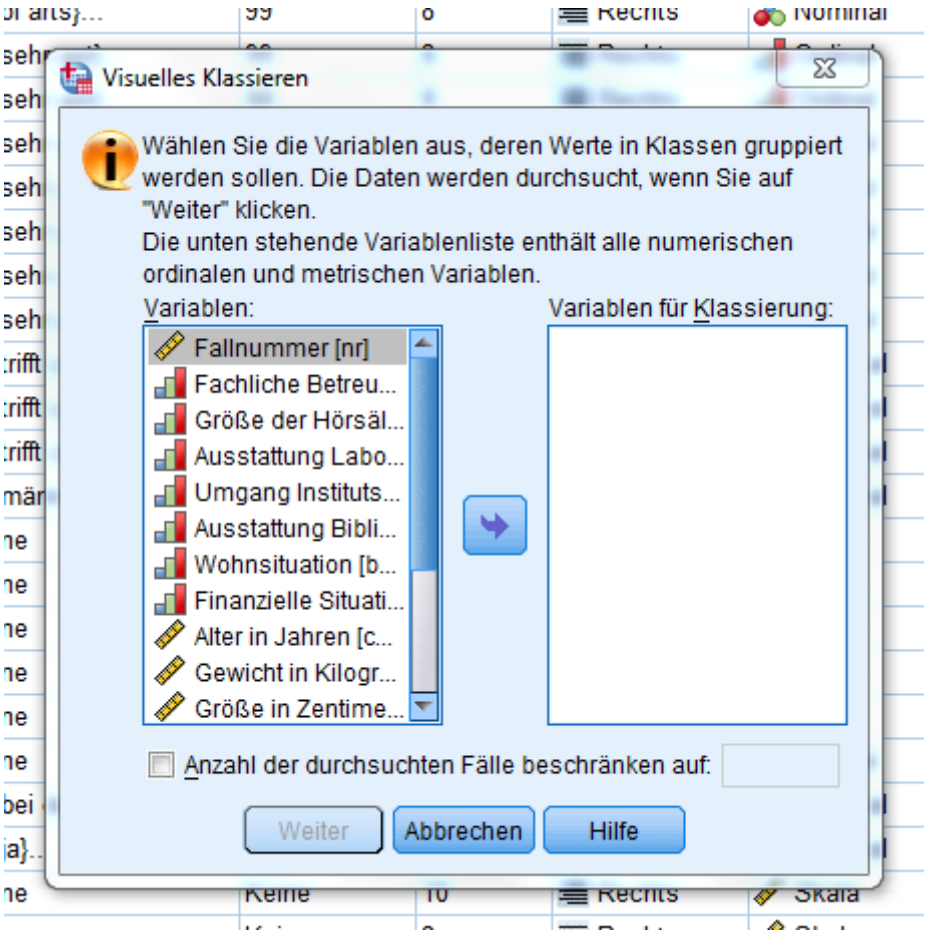
Aufgabe 5.2
Welche Variablen der „Studentenstudie 1“ bieten sich für die Bildung von Klassen an? Führen Sie eine solche Umkodierung an einer dieser Variablen durch. Welche Regeln der Gruppenbildung haben Sie beachtet und welche nicht? Begründen Sie Ihr Vorgehen.

Aufgabe 5.3
Klassieren Sie den „Mittleren Niederschlag (Jahr) in mm“ aus der SPSS-Datendatei „Klimastationen Europa.sav“ derart, dass diese gemäß dem Klassierungsprinzip „von … bis unter …“ in sechs äquidistante Klassen gegliedert werden. Geben Sie die prozentuale relative Häufigkeitsverteilung der klassierten Niederschlagswerte an und charakterisieren Sie die Häufigkeitsverteilung.

5.2.2 VISUELLES KLASSIEREN

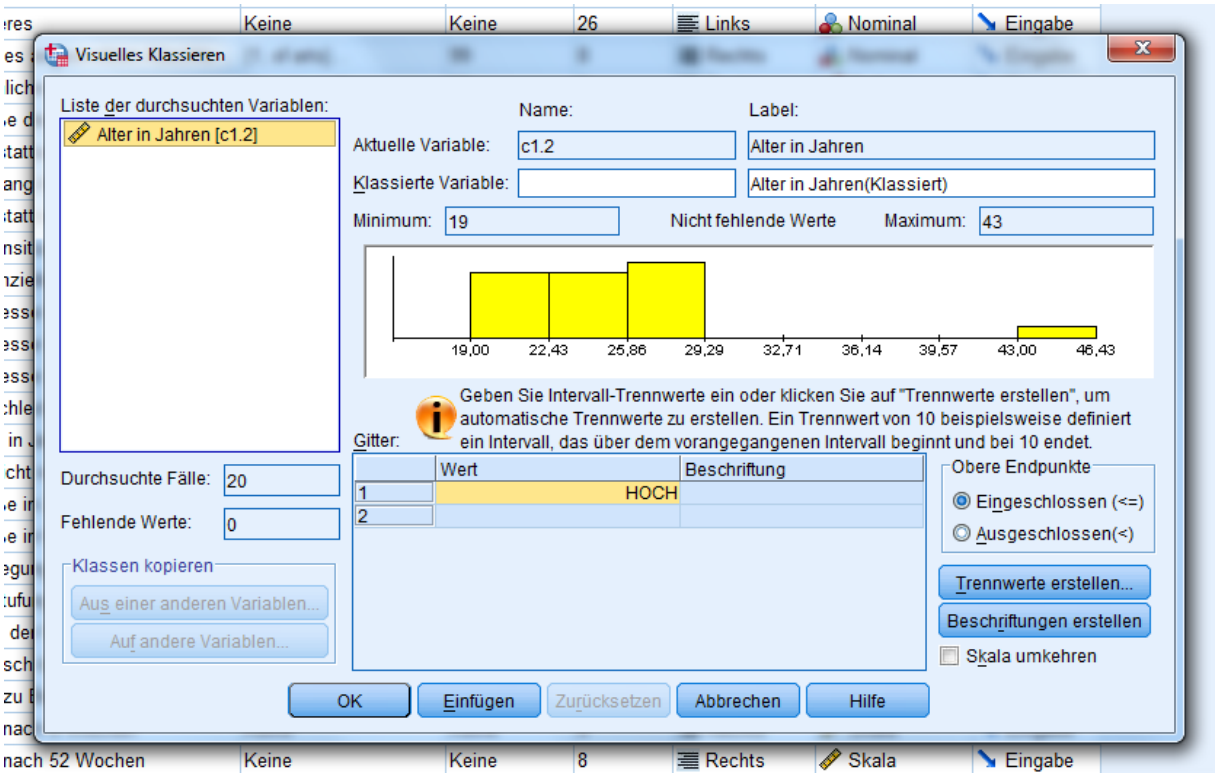
Die Option „Visuelles Klassieren“ bietet eine weitere Möglichkeit des Umcodierens von Variablen. Der Vorteil dieses Vorgehens liegt in der besonderen Übersichtlichkeit bei der Zusammenfassung von metrischen bzw. ordinalen Variablen zu einer kleineren Zahl an Klassen bzw. Kategorien. Insbesondere beim Umcodieren von metrischen Variablen mit sehr vielen Ausprägungen bietet das visuelle Klassieren deutliche Vorteile gegenüber der Umcodierung in neue Variablen. Wir wählen also ***Transformieren → Visuelles Klassieren***. Es öffnet sich das in Abbildung 5.4 dargestellte Kontextmenü.

Abbildung 5.4: Kontextmenü ***Visuelles Klassieren 1***



Wieder übertragen wir die zu codierenden Variablen in das Auswahlfenster, in diesem Fall unter ***Variablen für Klassierung***. Mit einem Klick auf ***Weiter*** öffnet sich eine weitere Dialog-box, die ebenfalls mit ***Visuelles Klassieren*** überschrieben ist.

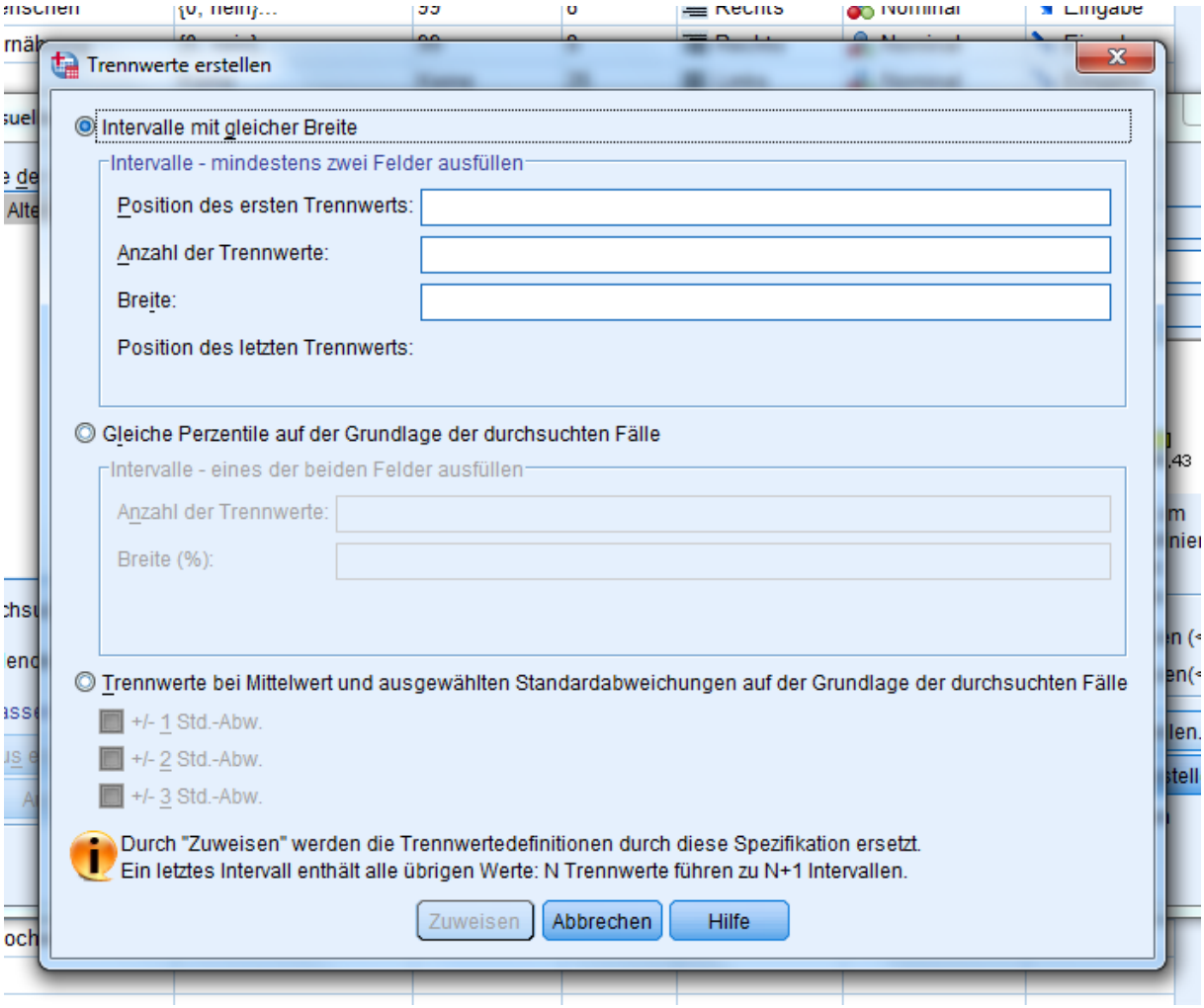
Abbildung 5.5: Kontextmenü ***Visuelles Klassieren 2***



Abermals wählen wir die zu analysierende Variable aus dem Feld „Variablen“. Im Feld „Nicht fehlende Werte“ erscheint nun ein Histogramm der Verteilung dieser Variablen. Außerdem werden Minimum, Maximum, die Zahl der Fälle und die fehlenden Werte mit angegeben.

Zunächst sollte man neben „Klassierte Variable“ einen Namen für die neue Variable eingeben (dabei bleibt die alte Variable natürlich erhalten) und ggf. die Beschreibung ändern. Danach kann man sich an die eigentliche Umcodierung machen. Dazu klicken wir auf die Schaltfläche ***Trennwerte erstellen…*** Es öffnet sich folgende Dialogbox (Abbildung 5.6).

Abbildung 5.6: Kontextmenü ***Trennwerte erstellen***



Für die Klassenbildung stehen hier drei Möglichkeiten zur Verfügung:

Intervalle mit gleicher Breite: Breite und Anzahl der Trennwerte lassen sich frei bestimmen.

Gleiche Perzentile auf der Grundlage der durchsuchten Fälle: Es werden Klassen mit gleicher Fallzahl (nicht gleicher Breite) gebildet. Hier lässt sich entweder Breite oder Anzahl der Trennwerte bestimmen, das zweite Auswahlfenster wird dann automatisch ausgefüllt.

Trennwerte bei Mittelwert und ausgewählten Standardabweichungen auf Grundlage der durchsuchten Fälle: Ein Trennwert liegt hier beim Mittelwert und weitere können zwischen dem Mittelwert und bis zu +/- zwei Standardabweichungen gesetzt werden.

Die Umcodierung startet man dann durch einen Klick auf ***Zuweisen.*** Es erscheint wieder die Dialogbox ***Visuelles Klassieren 1.*** Die gesetzten Trennwerte erscheinen nun als Linien im Histogramm und als Obergrenzen der einzelnen Klassen unter dem Feld „Wert“. Diesen Werten kann man nun durch einen Klick auf ***Beschriftungen erstellen*** automatisch Label zuteilen. Sowohl die Trennwerte als auch die Labels lassen sich durch Anklicken im Histogramm bzw. durch Umstellen der Tabelle anpassen. Dabei lassen sich auch Klassen mit unterschiedlicher

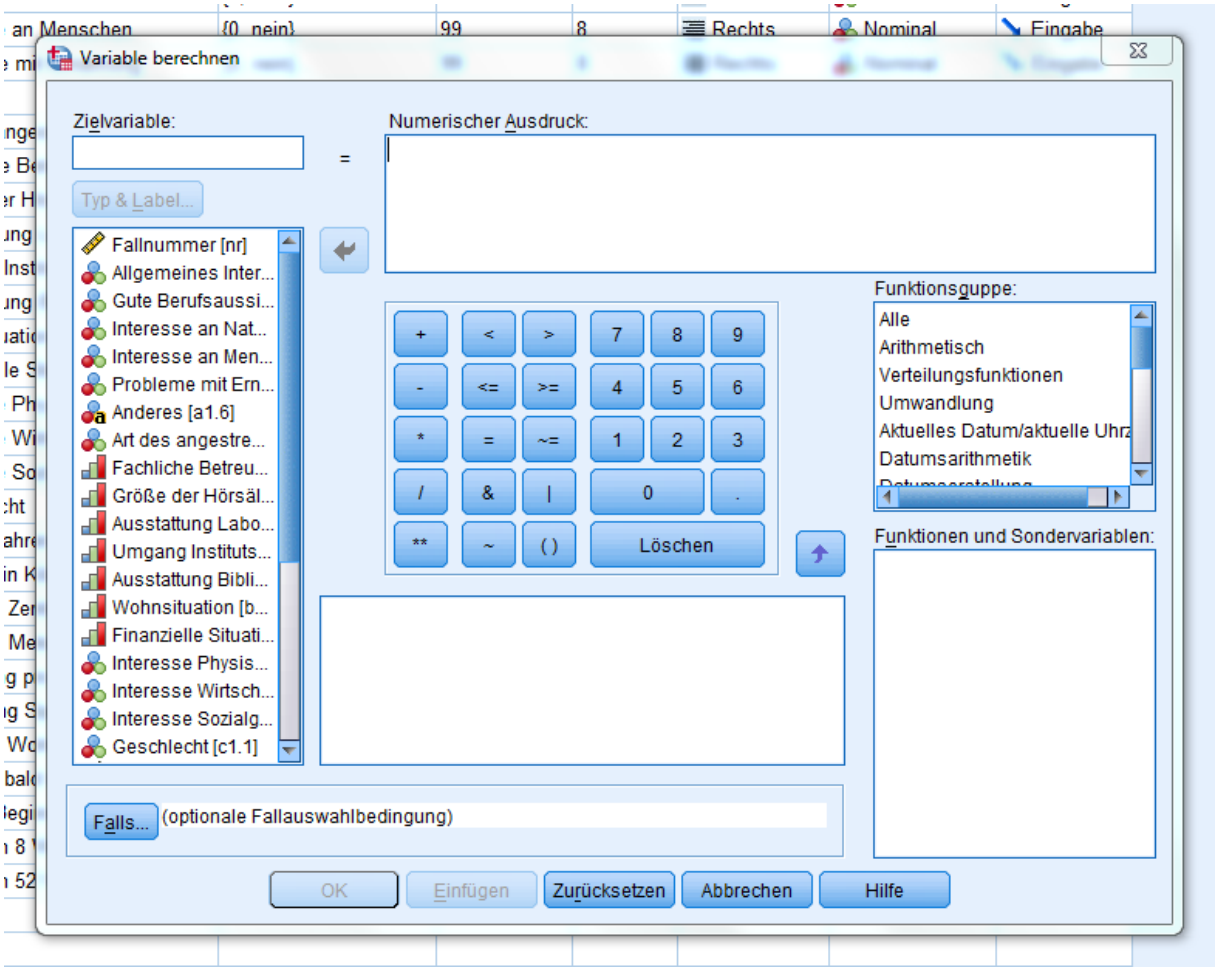
Breite anlegen. Mit ***Ok*** schließt man die Umcodierung und SPSS fügt die neu erstellten Variablen zur Variablenliste hinzu.

Aufgabe 5.4
Bilden Sie mithilfe der Operation ***Visuelles Klassieren*** abermals Gruppen für die in Aufgabe 5.1 ausgewählte Variable. Haben Sie sich hinsichtlich der Klassengrenzen anders entschieden als in Aufgabe 5.1? Welches Verfahren erscheint Ihnen praktischer?

5.2.3 TRANSFORMATIONSBEFEHLE: SKALENERSTELLUNG, GRUPPENBILDUNG

Variablen lassen sich in SPSS nicht nur Umcodieren, es besteht außerdem die Möglichkeit aus den Werten verschiedener Variablen neue Ergebnisvariablen zu berechnen oder bereits bestehende umzurechnen. Beispielsweise könnte man aus zwei Variablen für Körpergröße und Körpergewicht den Body Maß Index berechnen oder die in cm gemessene Körpergröße in Meter umrechnen. SPSS bietet dazu verschiedene Transformationsbefehle an, so u. a. einfache Summenbildung, z-Transformation oder Logarithmieren. Für eine Berechnung wählt man: ***Transformieren → Berechnen***. Es öffnet sich folgende Dialogbox (siehe Abbildung 5.7).

Abbildung 5.7: Kontextmenü ***Variable berechnen***



Zwar lassen sich Neuberechnungen wieder in einer bereits existierenden Variablen durchführen und damit die empirischen Daten überschreiben, es empfiehlt sich jedoch, abermals eine neue Variable zu erstellen. Im Eingabefeld gibt man deshalb unter ***Zielvariable*** einen neuen Variablennamen ein. Danach lässt sich unter ***Numerischer Ausdruck*** die gewünschte Berechnungsformel eingeben. Wir wählen die gewünschten Variablen wie gewohnt aus dem Variablenfenster und entscheiden uns für die passenden Operatoren und Funktionsgruppen. An dieser Stelle werden die Operatoren beschrieben, die in der „Rechnertastatur" der Dialogbox enthalten sind:

Die arithmetischen Operatoren Addition (+), Subtraktion (-), Multiplikation (*), Division (/) und Potenzieren (**) werden nach den üblichen Regeln abgearbeitet („Punkt vor Strich"). Dabei kann man auch hier durch Klammersetzungen diese Regeln beeinflussen.

Die relationalen Operatoren (Vergleichsoperatoren) vergleichen verschiedene Werte miteinander. Vor allem bei der Auswahl spezifischer Merkmalsausprägungen können diese hilfreich sein. Hier lassen sich beispielsweise den einzelnen Fällen durch die Abhängigkeit von bestimmten Bedingungen unterschiedliche Werte zuweisen. Zur Auswahl stehen hier: < (kleiner), <= (kleiner oder gleich), > (größer), >= (größer oder gleich), = (gleich), ~= (ungleich).

Die logischen Operatoren können durch die Verbindung zweier Bedingungen die Auswahl bestimmter Merkmalsausprägungen weiter präzisieren oder den Wahrheitswert eines Bedingungsausdrucks umkehren. Zu diesen zählen & (beide Ausdrücke müssen wahr sein), | (einer der beiden Ausdrücke muss wahr sein), ~ (kehrt den Wahrheitswert des Ausdrucks um).

Auch die in vielen Befragungen behandelten Bewertungsfragen lassen sich durch eine Skalierung zusammenfassen. Dabei werden häufig mehrere Fragen verwendet, die die gleiche Problematik aus unterschiedlichen Blickwinkeln betrachten. In Abbildung 19 wurden die Probanden beispielsweise darum gebeten, das Gastronomieangebot einer Stadt zu beurteilen. Dabei wurde nach der Beurteilung der Eisdielen, Bars, Restaurants und Diskotheken gefragt. Jede dieser Fragen könnte eigenständig bearbeitet werden, jedoch lassen sich diese Einzelfragen auch zu einer Gesamtbewertung zusammenfassen. Wurde bei allen Fragen die gleiche Skala verwandt, lassen sich die verschiedenen Werte über die Variablentransformation beispielsweise addieren und daraus ein Gesamturteil bestimmen.

Aufgabe 5.5
Berechnen sie den sogenannten Broca-Index (bi = 100 x Gewicht / Größe – 100) für alle Probanden der „Studentenstudie 1".

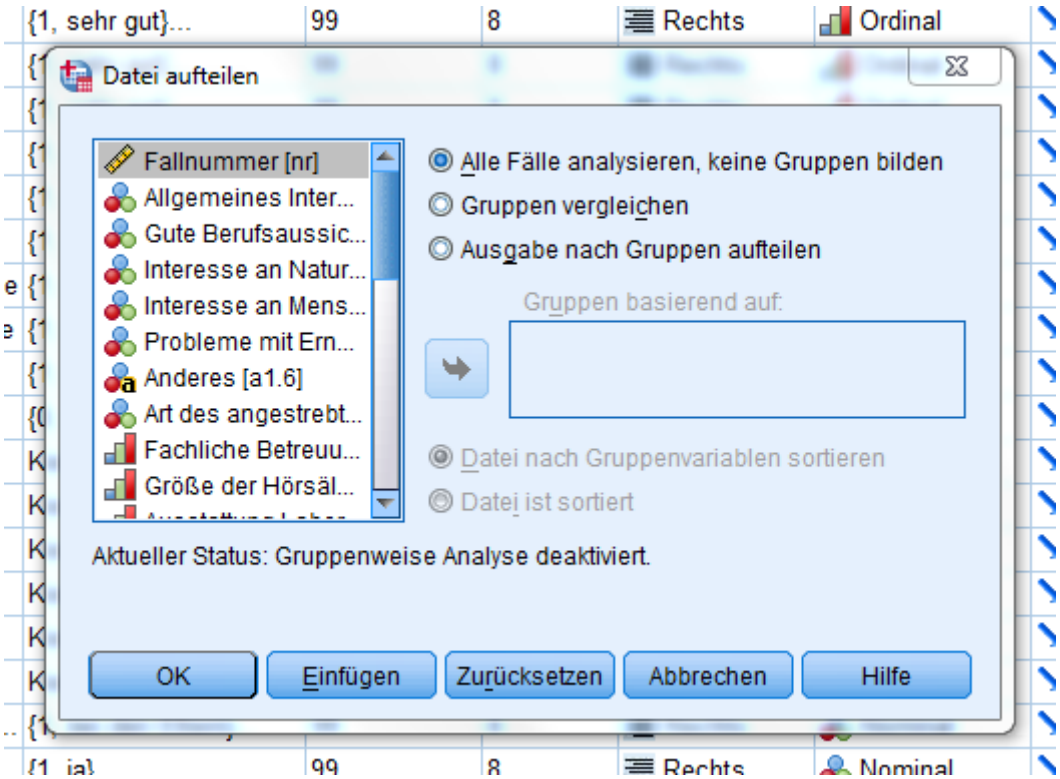
Aufgabe 5.6
Nach dem Broca-Index gelten alle Menschen als übergewichtig, deren Broca-Wert über der eigenen Körpergröße in cm – 100 liegt. Wie viele Probanden aus der „Studentenstudie 1" sind übergewichtig?

Aufgabe 5.7
Wie viele Studenten aus der Studentenstudie sind zwischen 1,80 m und 2,10 m groß?

5.3 AUFTEILEN VON DATEN IN GRUPPEN

In Studien, in denen es um die separate Analyse verschiedener Gruppen von Befragten geht, kann es von Nutzen sein, den Datensatz in verschiedene Untergruppen aufzuteilen (beispielsweise Männer / Frauen). Diese stehen dann separat zur Analyse zur Verfügung. Voraussetzung für eine Gruppenverarbeitung ist das Vorhandensein einer Gruppierungsvariable. Im vorliegenden Beispiel dient dazu die Variable Sex. Alle Befragten mit der Merkmalsausprägung 1 (weiblich) bilden die erste Gruppe, Probanden mit 2 (männlich) die zweite. Für jeden Datensatz können im Folgenden getrennte Auswertungen durchgeführt werden. Wir gehen wie folgt vor: ***Daten → Datei aufteilen***. Es öffnet sich das dargestellte Menü (Abbildung 5.8).

Abbildung 5.8: Kontextmenü ***Datei aufteilen***



Nach den von SPSS getroffenen Voreinstellungen wird keine Aufteilung in Gruppen vorgenommen. Deswegen wählen wir zunächst die Schaltfläche ***Ausgabe nach Gruppen aufteilen***. Jetzt muss man eine Variable in das Feld unter ***Gruppen basierend auf*** übertragen. Eine Voraussetzung für die Aufteilung in Gruppen ist die vorhergehende Sortierung der Datendatei anhand der entsprechenden Gruppenvariable. Die Option ***Datei nach Gruppenvariable sortieren*** übernimmt diesen Arbeitsschritt jedoch auch nachträglich. Bestätigen Sie die Anweisungen mit einem Klick auf ***OK***.

Auch hier ist darauf zu achten, dass (ähnlich wie bei der Anweisung zum Auswählen von Fällen) die getätigten Einstellungen so lange bestehen bleiben, bis man die Option ***Alle Fälle analysieren*** in der Dialogbox ***Datei aufteilen*** wieder angewählt hat.

Aufgabe 5.8
Vergleichen Sie die Verteilung der Variable „Gewicht“ getrennt nach Männern und Frauen in der „Studentenstudie 1“.

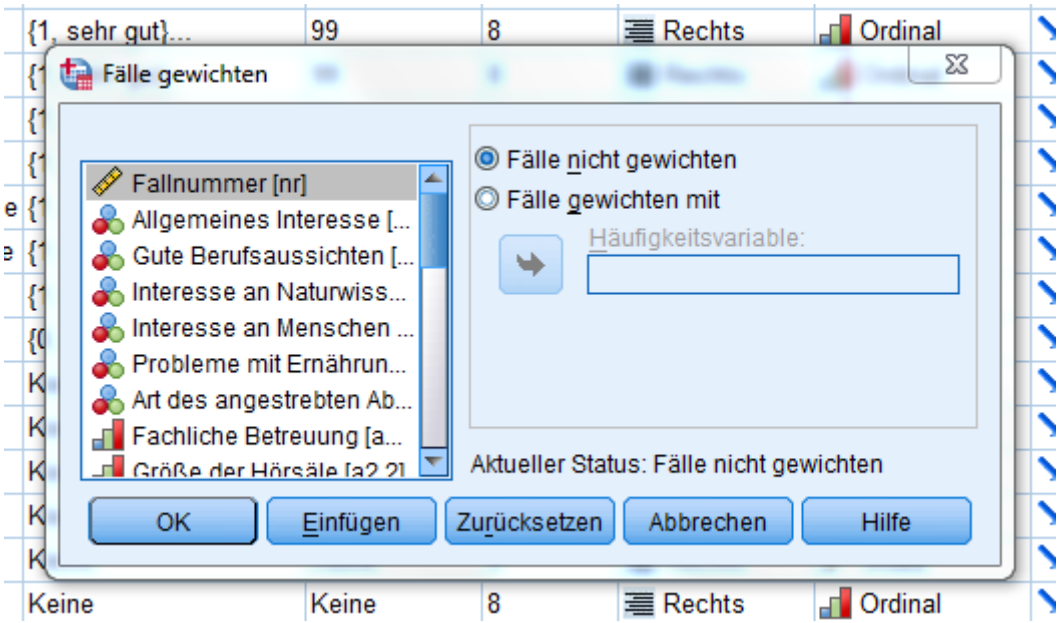
5.4 GEWICHTEN VON FÄLLEN

SPSS bietet auch die Möglichkeit, Fälle zu gewichten. In diesem Prozess werden die Daten fallweise über Gewichtungsvariablen bewertet. Mit der Gewichtung einer Stichprobe kann beispielsweise erreicht werden, dass das Stichprobenprofil der Untersuchung einem gewünschten Profil wie beispielsweise der zugrunde liegenden Grundgesamtheit angenähert wird. Vielleicht haben Sie eine Stichprobe erhoben, in der der Anteil an Beamten 10,5 % beträgt, während dieser Anteil an der Gesamtbevölkerung lediglich 8 % beträgt. Die Stichprobe ist damit nicht repräsentativ, d.h. die Häufigkeitsverhältnisse der relevanten Variablen entsprechen nicht denen in der Grundgesamtheit.

In der „Studentenstudie 1“ haben wir zehn Männer und zehn Frauen befragt. Der Anteil der Frauen an der Studentenschaft der untersuchten Beispieluniversität liegt jedoch nicht bei 50 % sondern bei 47,5 %. Um eine bessere Abbildung der Grundgesamtheit aller Studenten an der untersuchten Universität sicher zu stellen, wollen wir eine Gewichtung durchführen. Dazu müssen wir zunächst eine Gewichtungsvariable erstellen, mit deren Hilfe die Fälle mit einem bestimmten Faktor < 1 oder > 1 multipliziert werden müssen. Der Gewichtungsfaktor bestimmt sich dadurch, dass für jede Ausprägung der betreffenden Variable (hier die Variable Geschlecht mit ihren Ausprägungen „männlich“ und „weiblich“) ein Verhältnis zwischen „Sollzustand“ und „Istzustand“ gebildet wird: ***Gewichtungsfaktor = Soll/Ist***. Legen wir einen Anteil von 47,5 % Frauen in der Studentenschaft zugrunde, bestimmt sich der Gewichtungsfaktor für diese Gruppe mit: 50/47,5= 1,05. Analog ergibt sich für die Gruppe der Männer ein Gewichtungsfaktor von 50/52,5 = 0,95. Die Gewichte können nun einzeln eingetippt werden, es lohnt sich jedoch hier, auf eine Datentransformation zurückzugreifen (siehe Kapitel 5.2.3).

Um die vorliegenden Fälle zu gewichten, betätigen wir die Befehlsfolge ***Daten → Fälle gewichten*** und es erscheint Abbildung 5.9.

Abbildung 5.9: Kontextmenü ***Fälle gewichten***



In diesem Menü klickt man auf ***Fälle gewichten mit*** und wählt die neu erstellte Gewichtungsvariable aus. Diese übertragen wir in das Eingabefeld ***Häufigkeitsvariable*** und bestätigen mit einem Klick auf ***OK***.

Aufgabe 5.9
In der „Studentenstudie 1“ haben wir mit Variable „a 1.7“ nach der Art des angestrebten Abschlusses gefragt. Aus der Statistik der Beispieluniversität wissen wir, dass durchschnittlich 50 % der Studenten einen Abschluss „of arts“ und 50 % einen Abschluss „of science“ machen. Führen Sie eine Gewichtung der Fälle durch, um die empirischen Daten diesem Verhältnis anzupassen.

B DESKRIPTIVE STATISTIK

Nach einer umfangreichen und gewissenhaften Datenprüfung kann man sich mit der eigentlichen Datenanalyse auseinandersetzen. Dabei geht es zunächst darum, wesentliche Informationen aus dem Datensatz herauszufiltern und erste Erkenntnisse festzuhalten. Hier greift man auf die Methoden der sogenannten Deskriptiven Statistik zurück. Diese liefert eine Visualisierung und Beschreibung der vorliegenden Daten in Form von Tabellen, Diagrammen und verschiedenen Kennwerten. Welche Analysemethode verwendet werden kann, hängt dabei vor allem vom Skalenniveau der zu untersuchenden Variablen ab.

6 TABELLARISCHE DARSTELLUNG VON DATEN

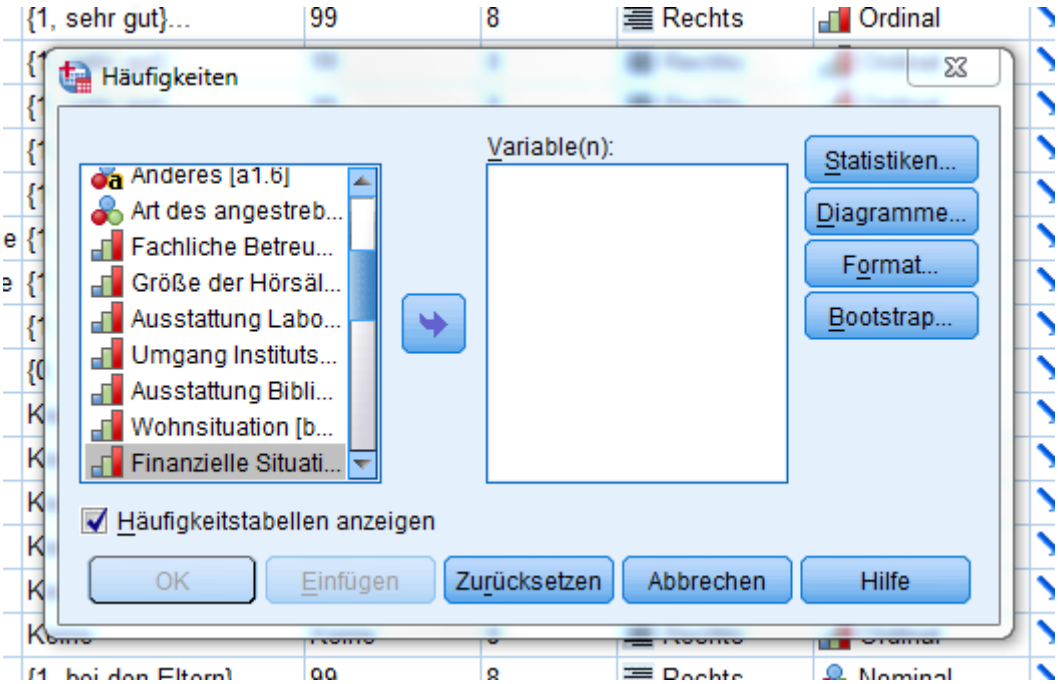
Wie stellt man Daten tabellarisch dar?

Wie kann man Tabellen bearbeiten?

6.1 HÄUFIGKEITSTABELLEN

Das Vorgehen zur Erstellung einer Häufigkeitstabelle wurde bereits angesprochen, soll hier jedoch um einige Aspekte erweitert werden. Häufigkeitstabellen fassen die abgezählten Messwerte der Versuchspersonen in einer gesammelten Darstellung zusammen. Dabei werden absolute, relative (Prozentangaben) und kumulierte Häufigkeiten (kumulieren = anhäufen) gemeinsam tabellarische dargestellt. Zur Erstellung einer Häufigkeitstabelle wählt man den Menüpunkt ***Analysieren → Deskriptive Statistik → Häufigkeiten***.

Abbildung 6.1: Kontextmenü ***Häufigkeiten***



Hier können Sie wieder eine oder mehrere Variablen, die analysiert werden sollen, bestimmen. Interessant sind hier die Einstellungen unter den Schaltflächen ***Statistiken***, ***Diagramme*** und ***Format.*** Unter ***Statistiken*** lassen sich verschiedene statistische Kennwerte zur Beschreibung der Verteilung der Variable anwählen, die in Kapitel 9 näher beschrieben werden. Unter der Schaltfläche ***Diagramme*** kann man sich ergänzend zur Häufigkeitstabelle ein Balkendiagramm, ein Kreisdiagramm oder ein Histogramm anzeigen lassen (mehr zu diesen Diagrammtypen in Kapitel 7). Die Auswahl unter ***Format*** ermöglicht abschließend einen Vergleich zwischen zwei Variablen und eine individuelle Sortierreihenfolge. Mit einem Klick auf ***Ok*** führt SPSS die Anweisungen aus. Abbildung 6.2 zeigt ein Beispiel für die Ausgabe einer Häufigkeitstabelle für die Variable *„Mittl. Temperatur (Jahr) in °C gruppiert“* aus der SPSS-Datendatei *„Klimastationen Europa“*.

Abbildung 6.2: Beispielausgabe für „Mittl. Temperatur (Jahr) in °C gruppiert“

Statistiken

Mittl. Temperatur (Jahr) in °C gruppiert

N	Gültig	397
	Fehlend	0

Mittl. Temperatur (Jahr) in °C gruppiert

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	bis 5,00°	57	14,4	14,4	14,4
	5,01 - 10,00°	179	45,1	45,1	59,4
	10,01-15,00°	99	24,9	24,9	84,4
	über 15,00°	62	15,6	15,6	100,0
	Gesamt	397	100,0	100,0	

Die erste Tabelle enthält Informationen über die zugrunde liegende Fallzahl und die fehlenden Werte. In der zweiten finden wir die gewünschten Häufigkeiten und verschiedene Prozentangaben.

Aufgabe 6.1
Wofür stehen die „gültigen Prozentwerte“ in der Darstellung einer Häufigkeitstabelle?

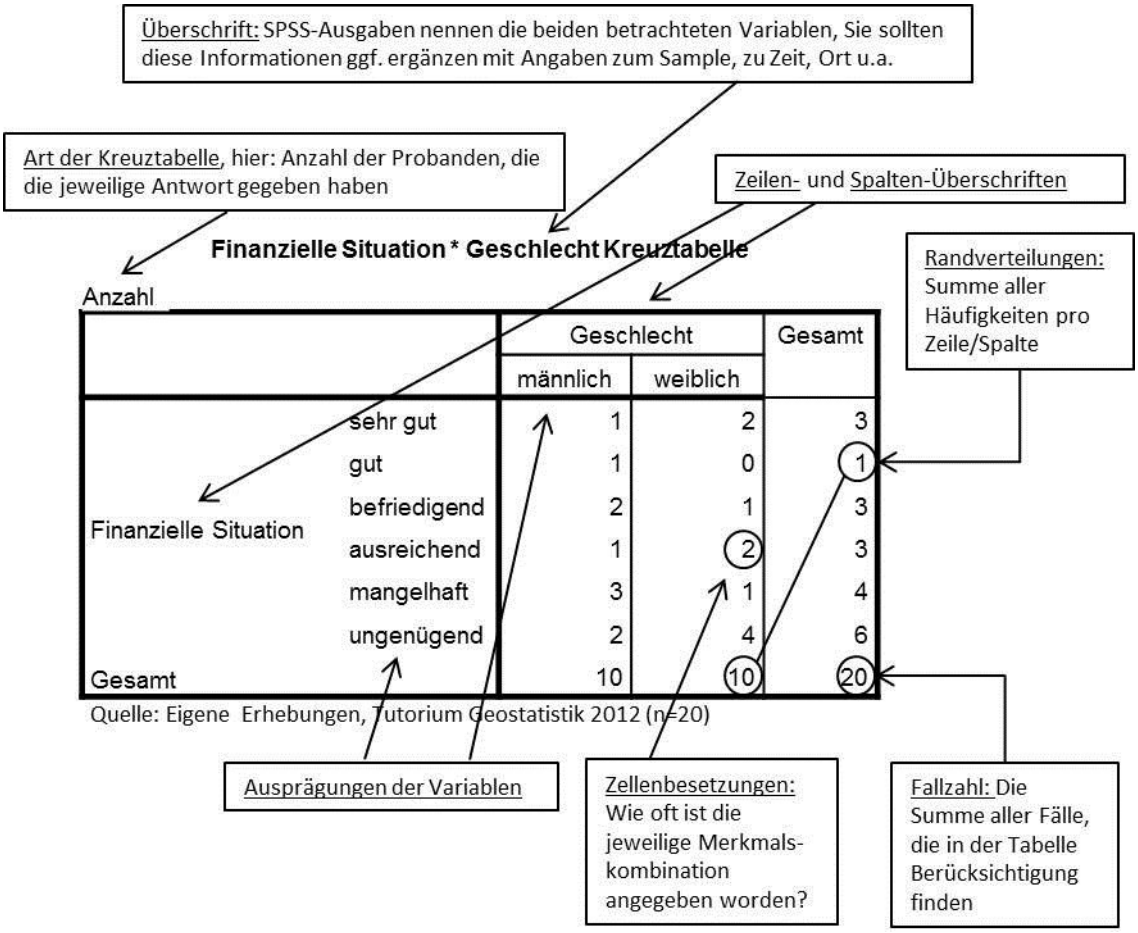
Aufgabe 6.2
Wie lassen sich die kumulierten Prozentwerte interpretieren?

Aufgabe 6.3
Wie viele Gemeinden gehören zur Bevölkerungsgrößenklasse „20.001 bis 30.000“ in der SPSS-Datendatei „Bev_Rhein_Main.sav“?

6.2 KREUZTABELLEN

Da Befragungsdaten zumeist vor allem aus nominalen bzw. ordinalen Daten bestehen, stellen Kreuztabellen oder auch Kontingenztafeln oftmals das Herzstück der anschließenden Analyse dar. Diese stellen die Beziehung der Häufigkeitsverteilungen zweier (oder mehrerer) Merkmale untereinander dar. Einen Überblick über den Aufbau einer Kreuztabelle vermittelt Abbildung 6.3.

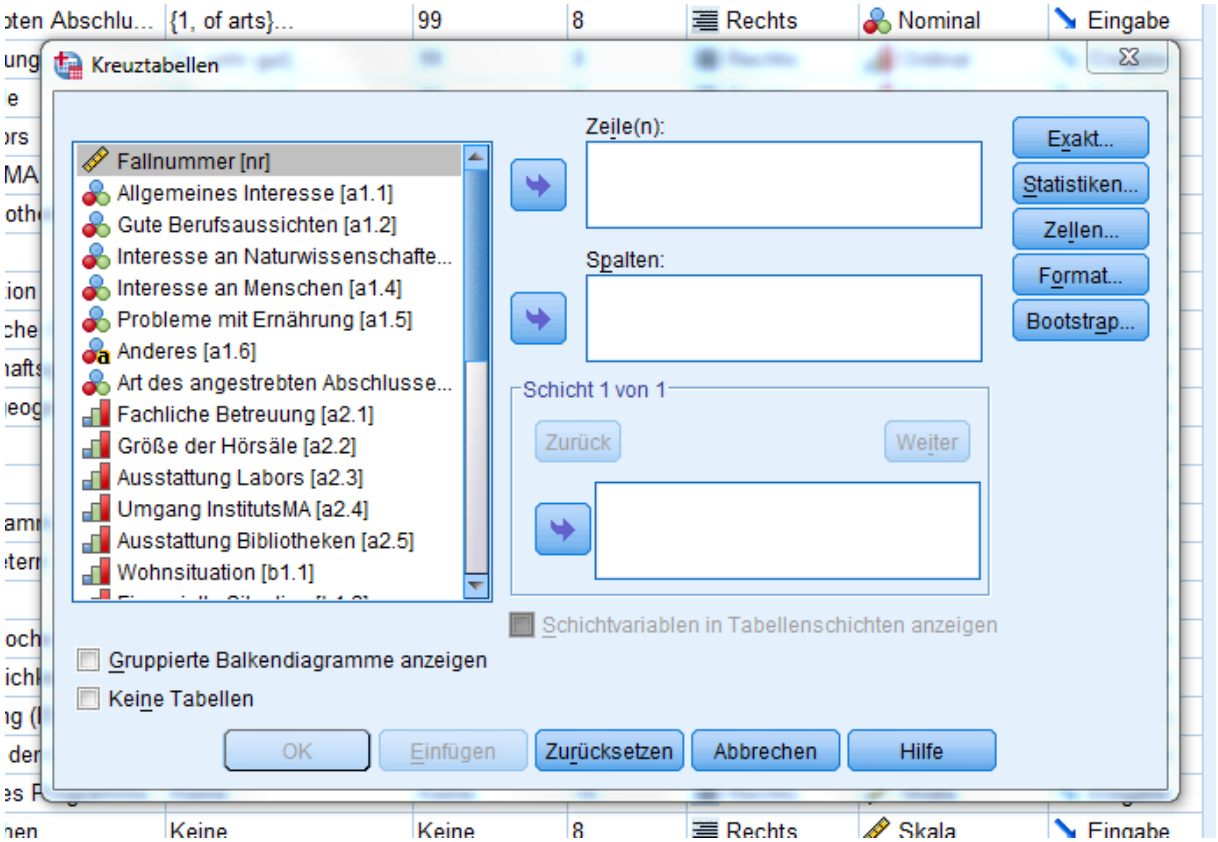
Abbildung 6.3: Aufbau einer Kreuztabelle (hier mit Angabe absoluter Häufigkeiten)



Quelle: Eigene Darstellung

Bei ordinal- oder nominalskalierten Variablen mit mehreren Kategorien können Kreuztabellen einen guten Eindruck über den Zusammenhang zwischen den beiden Merkmalen vermitteln. Dieser Zusammenhang lässt sich auch mit einem statistischen Kennwert bemessen. Dieser Indikator wird auch als Chi^2-Wert bezeichnet. Mit diesem Kennwert wird überprüft, ob statistisch auffällige Kombinationen von Kategorien vorliegen. Die Berechnung und Interpretation des Chi^2-Wertes wird im Kapitel 11 weiter behandelt. Um eine Kreuztabelle zu erstellen, öffnet man das Dialogmenü unter ***Analysieren → Deskriptive Statistik → Kreuztabellen*** (Abbildung 6.4).

Abbildung 6.4: Kontextmenü ***Kreuztabellen***



Wir wählen die darzustellenden Variablen für die gewünschten Spalten und Zeilen. Werden nur zwei Variablen betrachtet, werden die Ergebnisse einer der beiden in den Zeilen und die der anderen in den Spalten abgetragen. In den Schnittpunkten zwischen Spalten und Zeilen werden die Häufigkeiten und andere Maßzahlen für eine Kombination der entsprechenden Kategorien angegeben. Bei der Analyse von mehr als zwei Variablen erstellt SPSS jeweils eine neue Tabelle.

Neben verschiedenen statistischen Maßzahlen (siehe Kapitel 9), die sich unter der Schaltfläche ***Statistik*** verbergen, bietet ***Zellen auswählen*** diverse Auswahlen zu den Angaben in den Feldern der Kreuztabelle (beobachtete und erwartet Werte, Prozentwerte, Residuen und nichtganzzahlige Gewichtungen). ***Format*** legt die Anordnung der dargestellten Werte fest.

Abermals bestätigt man die getroffene Auswahl mit einem Klick auf ***OK***.

Aufgabe 6.4
Wie viele Befragte sind „männlich“ und haben „Interesse an Menschen“ als Grund für die Studienfachwahl in der „Studentenstudie 1“ angegeben?

Aufgabe 6.5
Welchen Eindruck auf einen evtl. bestehenden Zusammenhang zwischen den Variablen macht die bei Aufgabe 6.4 erstellte Kreuztabelle auf Sie? Warum?

Aufgabe 6.6
Erstellen Sie eine Kreuztabelle für die Darstellung der Variablen „Umgang InstitutsMA“ und „Ausstattung Bibliotheken“. Verwenden sie die Variable „Geschlecht“ als Schichtvariable.

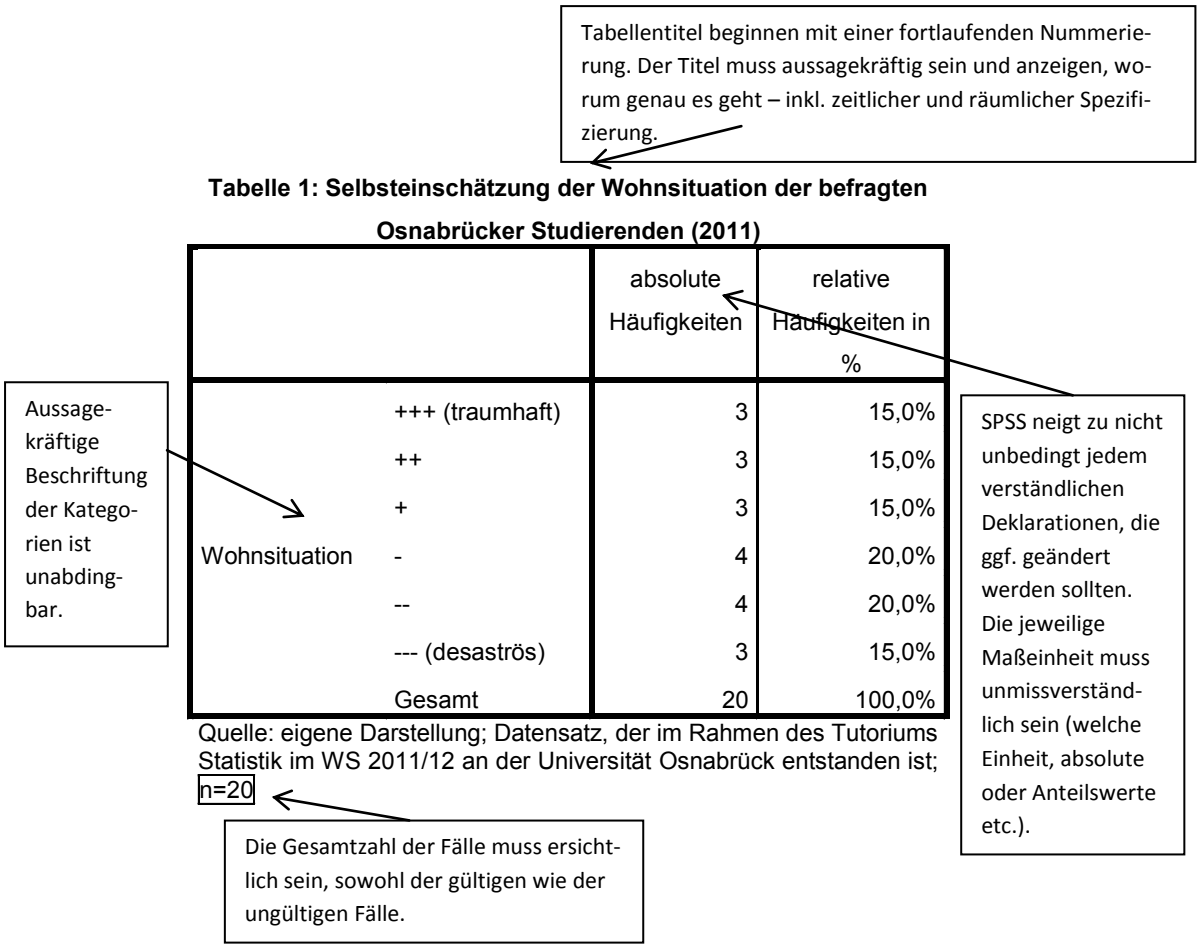
6.3 TABELLEN BEARBEITEN

Die in Tabellen dargestellten Ergebnisse lassen sich in ihrem Aufbau in Spalten, Zeilen und Schichten verändern. Man spricht hier vom Pivotieren einer Tabelle. Hier bieten sich dem Anwender viele verschiedene Möglichkeiten der individuellen Gestaltung wie etwa:

Transponieren von Zeilen und Spalten
Verschieben von Zeilen und Spalten
Erstellen von mehrdimensionalen Schichten
Anlegen und Aufheben von Gruppierungen für Zeilen und Spalten
Anzeigen und Ausblenden von Zeilen, Spalten und anderen Informationen
Drehen von Zeilen- und Spaltenbeschriftungen
Anzeigen von Definitionen für Terme
Erklärungen hinzufügen
Vorlagen auswählen
Schriftart/Größe/Form ändern
Zeilen- und Spaltenbeschriftungen vergrößern/verkleinern
Rahmen ändern/einfügen
Fußnoten einfügen
usw.

Zur Bearbeitung einer Tabelle wählt man diese zunächst in der Ausgabe-Ansicht mit einem Doppelklick aus. Woraufhin sich der Diagramm-Editor öffnet. Hier stehen dem Anwender verschiedene Möglichkeiten der Bearbeitung zur Verfügung, beispielsweise über die verschiedenen Menüs des Diagrammeditors oder über einen Rechtsklick auf das zu bearbeitende Element und dem darauffolgenden Öffnen des Dialogfeldes ***Eigenschaften***. Dieses enthält Registerkarten, mit deren Hilfe man verschiedene Optionen festlegen kann. Da sich die Bearbeitung von Tabellen und Grafiken in SPSS ähnelt, finden Sie weitere Details zur individuellen Gestaltung von Ausgabedateien in Kapitel 7.6. Eine Übersicht wichtiger Anforderungen an eine Tabelle findet sich in Abbildung 6.5.

Abbildung 6.5: Übersicht über wichtige Anforderungen an Tabellendarstellungen



Quelle: Eigene Darstellung

7 GRAFISCHE DARSTELLUNG VON DATEN

Wie stellt man welche Daten grafisch dar?

Wie kann man Grafiken bearbeiten?

Daten und ihre Verteilungen können durch tabellarische Darstellungen veranschaulicht werden. Zur Visualisierung können aber auch Grafiken angefertigt werden, die die gewünschten Informationen quasi ‚auf einen Blick' liefern. Grafiken helfen dabei, Verhältnisse und absolute Werte zu verdeutlichen und ggf. interessante Zusammenhänge im Datenmaterial aufzudecken. Man sollte bei der Erstellung und Interpretation von Grafiken jedoch immer mit einer gewissen Vorsicht vorgehen, da diese leicht fehlinterpretiert werden können. Dem kann auf vielfältige Weise Vorschub geleistet werden, etwa durch verwirrende Achsenskalierungen, zweideutige Beschriftungen usw.

Mit den Prozeduren des Grafik-Menüs von SPSS können verschiedene Diagrammtypen erzeugt werden. Das folgende Kapitel stellt die Vorgehensweise für einige Diagrammtypen dar und beschreibt, wie deren Layout nachträglich verändert werden kann. Bei der Erstellung von Diagrammen ist zu beachten, dass es mehrere Menübefehle gibt, die aber alle zum gleichen Ziel führen. Grundsätzlich finden sich alle Optionen unter dem Menü ***Diagramme***. Hier lassen sich nun Darstellungen über die ***Diagrammerstellung*** anlegen, über die ***Grafiktafel-Vorlagenauswahl*** vorsortieren oder über die sogenannten ***Veralteten Dialogfelder*** erstellen. Im Folgenden werden die drei verschiedenen Wege anhand der unterschiedlichen Diagrammtypen dargestellt.

7.1 BALKENDIAGRAMM

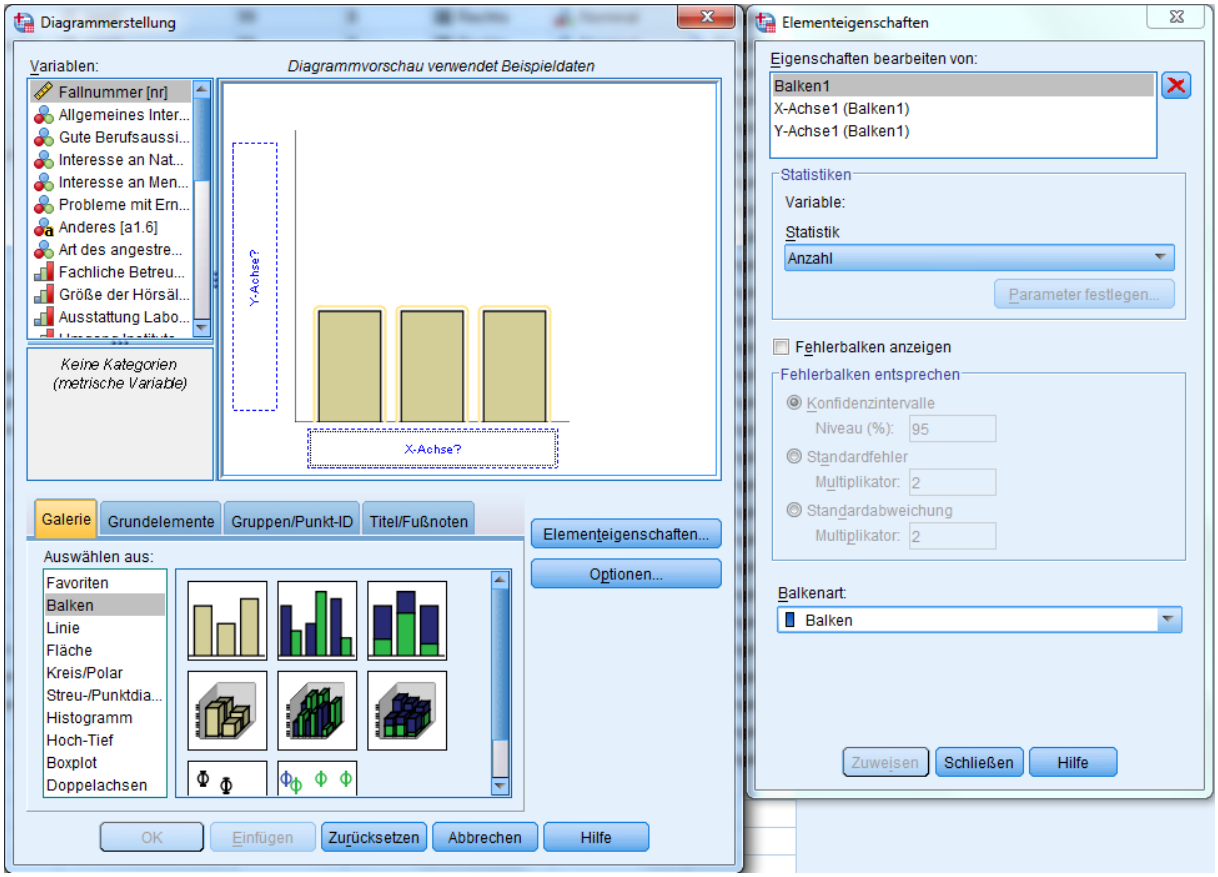
Balkendiagramme werden bevorzugt zur Darstellung von Häufigkeiten nominal- und ordinalskalierter Variablen verwendet. Jeder Merkmalsausprägung wird dabei ein Balken zugeordnet, dessen Länge deren absoluter bzw. relativer Häufigkeit entspricht. Bei der Darstellung von nominal skalierten Merkmalen ist jedoch zu beachten, dass die Anordnung auf der x-Achse willkürlich ist, da nominale Variablen nicht der Größe nach zu ordnen sind. Bei ordinalskalierten Variablen liegt zwar eine (wie auch immer ‚natürliche') Ordnung der Daten vor, aber die Abstände lassen sich nicht sinnvoll interpretieren. Man sollte darauf achten, dass Balkendiagramme am Nullpunkt starten, um dem Gesetz der Flächentreue in der Darstellung zu entsprechen.

Um ein einfaches Balkendiagramm zu erstellen, öffnet man die Diagrammerstellung über den Menübefehl ***Diagramme → Diagrammerstellung*** (Abbildung 7.1).

Nach dem Anklicken gibt SPSS einen Hinweis auf die Festlegung des Messniveaus jeder Variable. Hat man diese noch nicht für alle Variablen festgelegt, sollte man dies nun nachholen. Ansonsten kann man mit einem Klick auf ***OK*** diese Frage bestätigen.

Im dargestellten Auswahlfenster (Abbildung 7.1) wurde im unteren Feld (***Galerie***) bereits das Symbol für einfache Balkendarstellung (einfarbig beige) angeklickt und per „drag and drop" in das obere Fenster Diagrammvorschau gezogen. Nun lassen sich die zu verwendenden Variablen links oben auswählen und durch Ziehen mit der Maus der gewünschten Achse zuweisen. Außerdem bieten sich dem Nutzer weitere Feineinstellungen unter den Menüpunkten ***Elementeigenschaften*** und ***Optionen***.

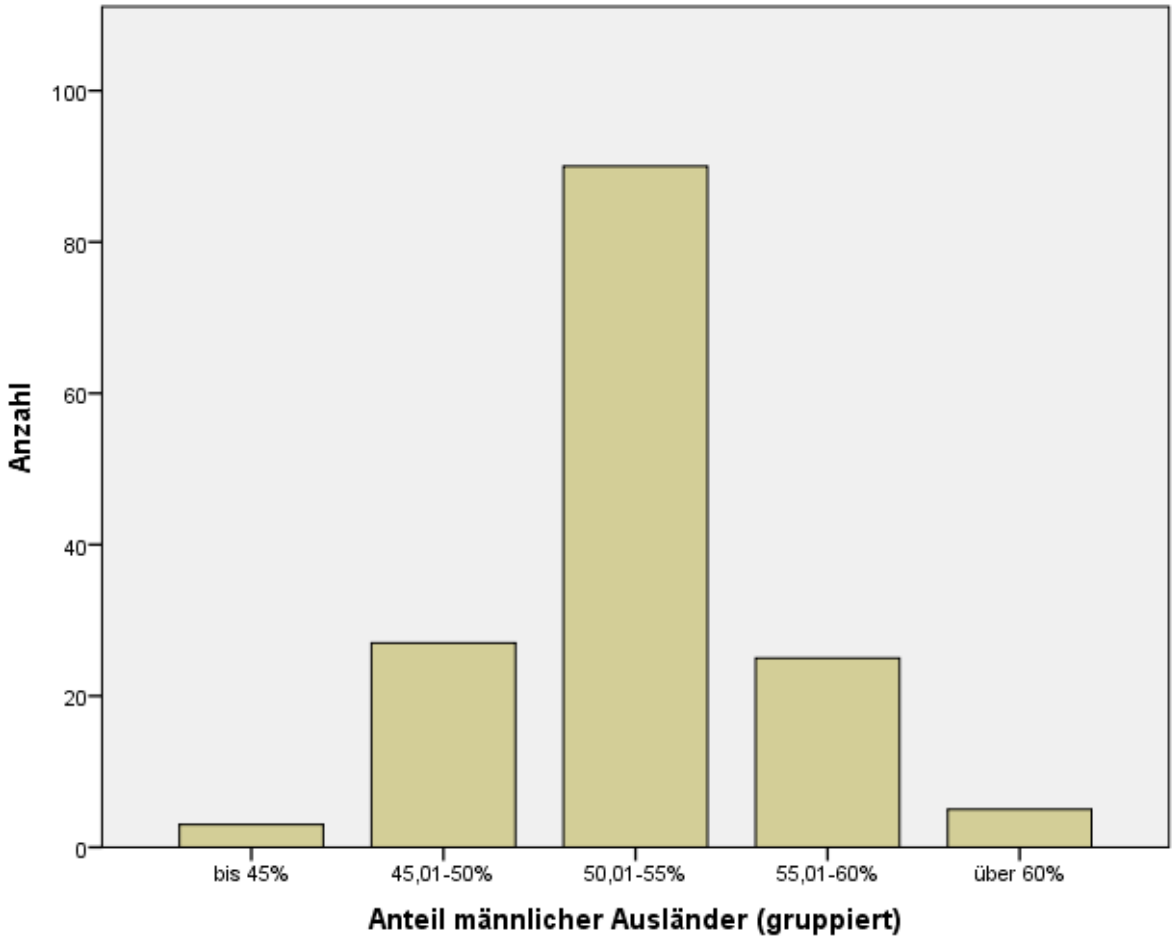
Abbildung 7.1: Kontextmenü ***Diagrammerstellung***



Zur Veranschaulichung ziehen wir nun das Balkendiagramm in das Diagrammvorschaufenster und legen die Variable *„Anteil männlicher Ausländer (gruppiert)“* aus der SPSS-Datendatei Bev_Rhein_Main.sav auf die x-Achse. SPSS bestimmt jetzt automatisch, dass auf der y-Achse die entsprechende Anzahl der Nennungen abgetragen wird. Wir wollen jedoch die Prozentzahlen darstellen und klicken unter den sich selbst öffnenden Menüpunkt Elementeigenschaften auf die Schaltfläche unter ***Statistik***. Im folgenden Menü wählen wir „Prozentsatz“ und klicken danach auf ***Zuweisen***. Hier finden sich auch weitere Möglichkeiten wie die Darstellung kumulierter Prozente, Werte, Mittelwerte und weitere. Zur Bestätigung der Auswahl bedarf es eines Klicks auf die Schaltfläche ***OK***.

Es öffnet sich das Ausgabefenster und es erscheint folgendes Ergebnis (Abbildung 7.2).

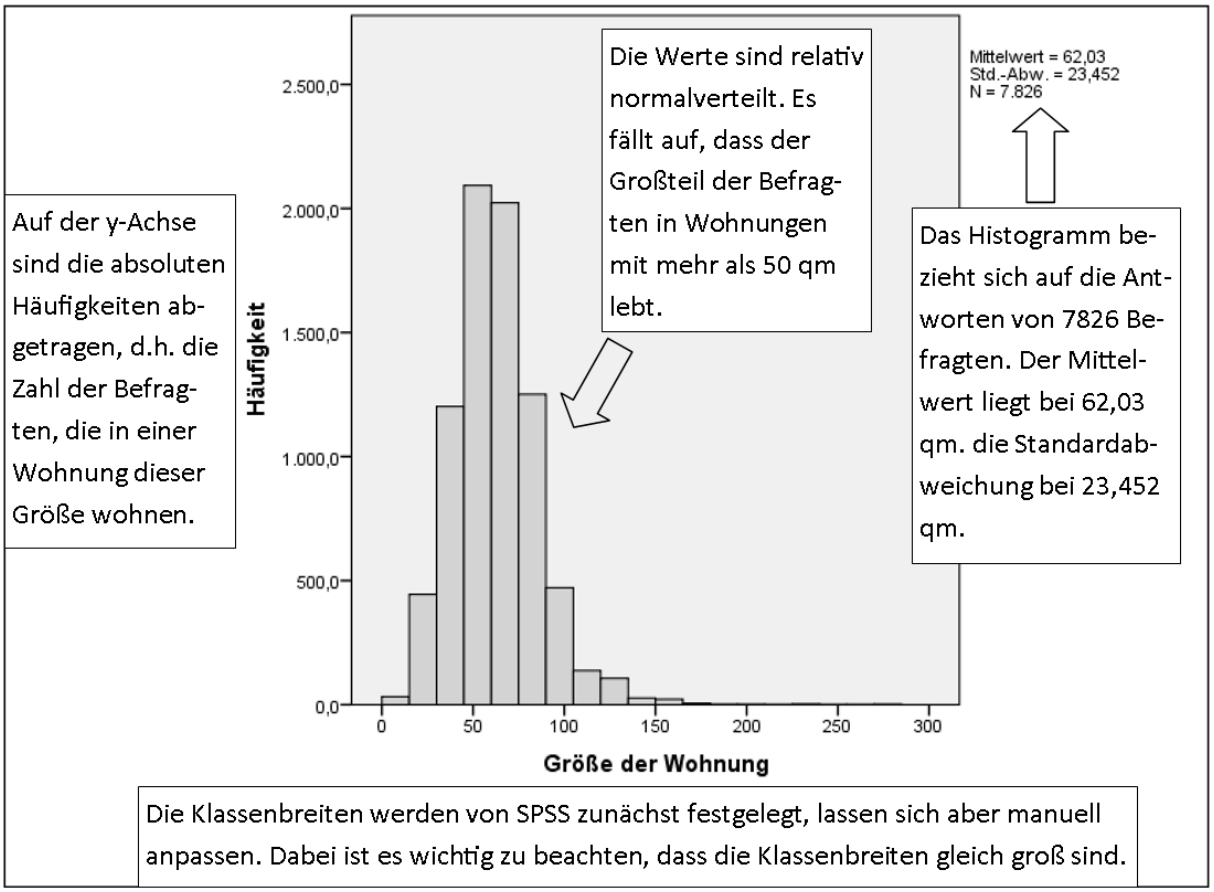
Abbildung 7.2: Balkendiagramme zur Variable *„Anteil männlicher Ausländer (gruppiert)“*



7.2 HISTOGRAMM

Ein Histogramm stellt die Häufigkeitsverteilung von intervallskalierten Variablen dar und eignet sich vor allem für klassierte Daten. Die Flächen des Diagramms sind flächenproportional zur (meist: relativen) Häufigkeit der entsprechenden Merkmalsausprägung. Anders als beim Balkendiagramm berühren sich die einzelnen Teilflächen, die in Abb. 7.2 erkennbaren deutlichen Lücken zwischen den dortigen Balken fehlen in Histogrammen. Gerade wenn besonders viele unterschiedliche Merkmalsausprägungen vorliegen eignet sich die Darstellung in einem Balkendiagramm nicht mehr und man sollte auf ein Histogramm zurückgreifen. Auch bietet SPSS die Möglichkeit der Anzeige einer Normalverteilungskurve, mit der die empirische Verteilung verglichen werden kann. Abbildung 7.3 erläutert den Aufbau eines Histogramms.

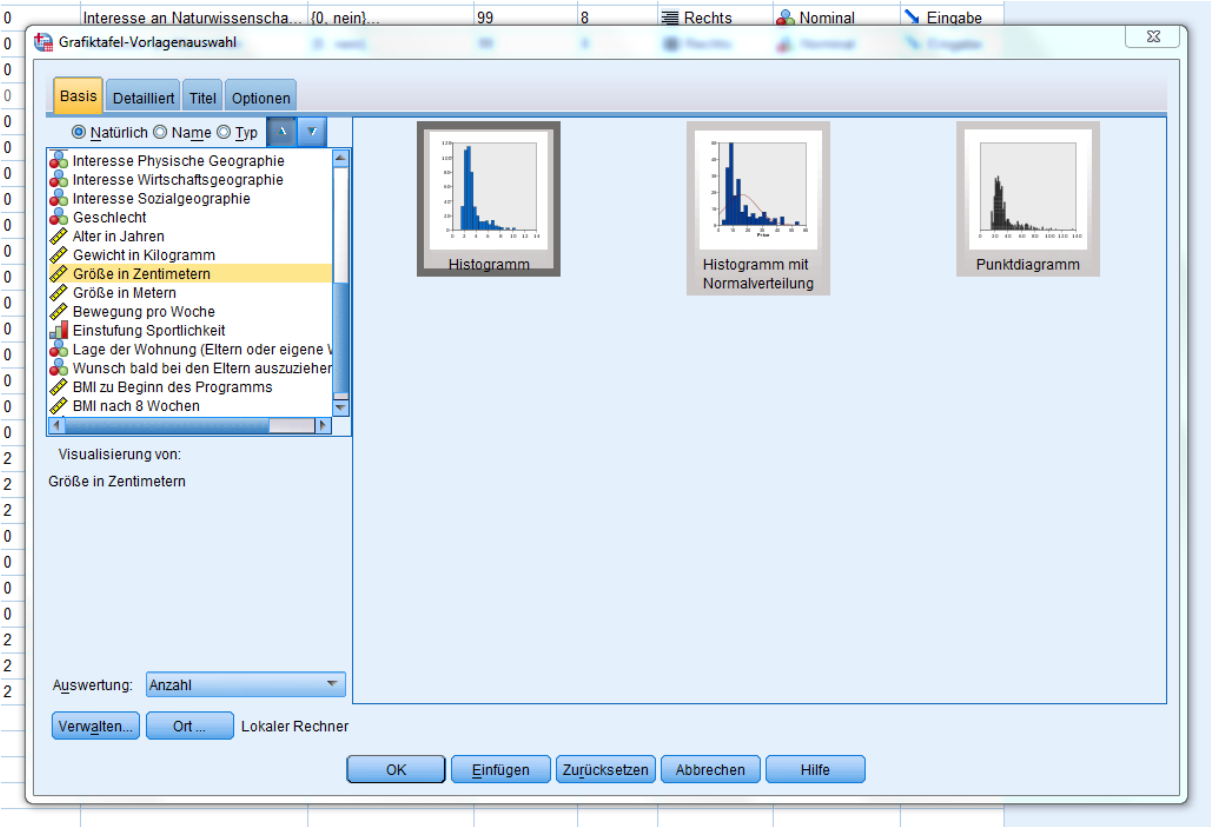
Abbildung 7.3: Aufbau eines Histogramms



Quelle: Modifiziert nach Baur und Fromm 2008, 234

In diesem Fall wollen wir die Grafik unter dem Menüpunkt ***Diagramme → Grafiktafel-Vorlagenauswahl*** erstellen. Es öffnet sich folgende Dialogbox (Abbildung 7.4).

Abbildung 7.4: Kontextmenü ***Grafiktafel-Vorlagenauswahl***



Wir können die am linken oberen Rand dargestellte Variablenliste nach verschiedenen Kriterien wie Name, Typ oder natürlicher Reihenfolge anordnen. Die Auswahl einer Variablen erwirkt, dass SPSS dem Nutzer mehrere zum Skalenniveau passende Diagrammtypen zur Wahl stellt. Unter den Schaltflächen ***Detailliert***, ***Titel*** und ***Optionen*** lassen sich weitere Feineinstellungen für die Darstellung des gewählten Diagramms vornehmen. Hier finden sich Einstellungen für individuelle Titel und Untertitel, die Farbgebung, Reihenfolgen von Kategorien, Diagrammvorlagen, Achsen- sowie Skalenbereiche.

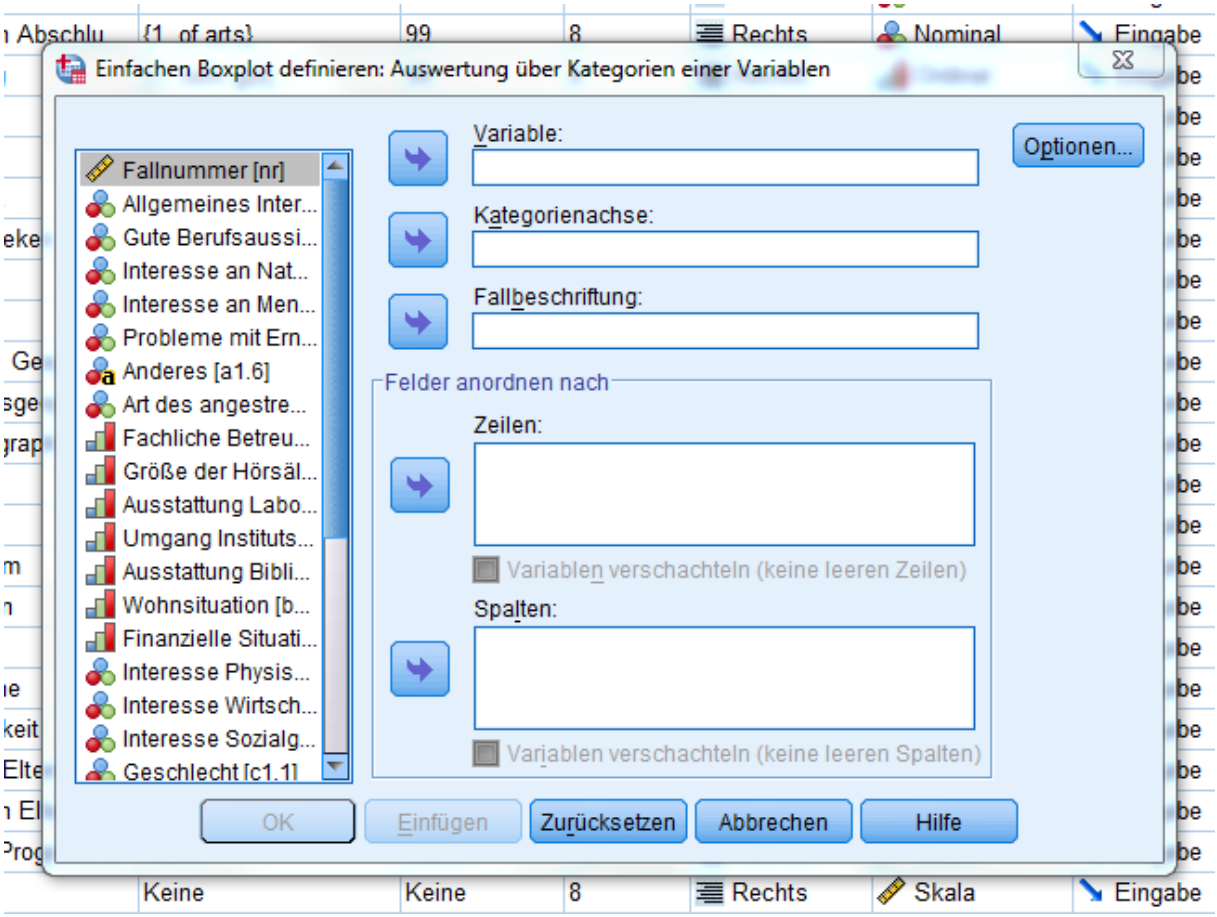
7.3 BOXPLOT

Sowohl der Median als auch die beiden Quartile bzw. der empirische Quartilsabstand von ordinalen oder metrischen Variablen lassen sich trefflich in einem Boxplot (auch: Kastendiagramm) darstellen; sie sind graphisch als Trennlinie in der Box (Median) und als deren obere (oberes Quartil) und untere Begrenzung wiedergegeben (siehe Abbildung 4.7). Die „Antennen“ (Whisker) an beiden äußeren Enden der Darstellung repräsentieren jene Hälfte aller Beobachtungen, die außerhalb des Quartilsabstands liegen, also die 25 % niedrigsten und die 25 % höchsten/größten Merkmalsausprägungen. Außerdem lassen sich ggf. vorhandene Ausreißer und Extremwerte aus der SPSS Darstellung in einem Boxplot ablesen. Extremwerte liegen mehr als 3 Box-Längen vom unteren bzw. oberen Rand der Box entfernt und werden durch „*“ gekennzeichnet. Die Ausreißer liegen zwischen 1,5 und 3 Box-Längen vom unte-

ren bzw. oberen Rand der Box entfernt und werden durch einen „◦“ wiedergegeben. Hier können ggf. Label für definiert werden (üblicherweise die Fallzahl).

Wir stellen die Diagrammerstellung jetzt unter dem Menüpunkt ***Veraltete Dialogfelder*** dar. Dazu wählt man ***Diagramme → Veraltete Dialogfelder → Boxplot***. Im Anschluss kann man sich zwischen einfachen und gruppierten Boxplots entscheiden. Wir wählen einen einfachen Boxplot und belassen die Voreinstellung bei ***Auswertung über Kategorien einer Variable***. Es öffnet sich folgende Dialogbox (Abbildung 7.5). SPSS geht davon aus, dass auch einfache Boxplots für Kategorien ausgegeben werden sollen (etwa getrennt für Frauen und Männer), so dass die Ausgabe mindestens paarweise erfolgt.

Abbildung 7.5: Kontextmenü **Einfachen Boxplot definieren: Auswertungen über Kategorien einer Variable**



Aufgabe 7.1
Öffnen sie die „Studentenstudie 1“ und erstellen sie eine Grafik, die zwei Boxplots enthält. Einen für die Verteilung der Variable „Gewicht in Kilogramm“ und einen für „Größe in cm“. Vergleichen sie die Verteilungen miteinander.

7.4 STREUDIAGRAMM

In einem Streudiagramm wird der Zusammenhang zwischen zwei Variablen dargestellt. Dazu wird das Koordinatensystem mit der x-Achse (Abzisse) und der y-Achse (Ordinate) benutzt. Jeder Punkt im Koordinatensystem kann dann einem Wertepaar zugeordnet werden. Die Erstellung eines Streudiagrammes erfolgt analog zu einer der drei oben dargestellten Möglichkeiten.

7.5 KREISDIAGRAMM

Kreisdiagramme eignen sich zur Darstellung von Häufigkeiten qualitativer, diskreter oder klassierter Merkmale (vor allem auf nominalem Datenniveau). Die Aufteilung in einzelne Sektoren („Kuchenstücke“) ist hierbei proportional zur absoluten bzw. relativen Häufigkeit. Kreisdiagramme sind jedoch nur bedingt dazu in der Lage, Aufschluss über die Größenordnungen in der Datenverteilung zu vermitteln. Deshalb eignet sich diese Darstellungsform weniger für die Bearbeitung von ordinalen oder intervallskalierten Daten.

Insbesondere bei der Erstellung von Kreisdiagrammen sollte auf die 3-dimensionale Darstellung des Diagrammes verzichtet werden. In einer solchen Abbildung entspricht die zugewiesene Fläche nicht mehr den relativen Häufigkeiten, wodurch das Gesetz der Flächentreue verletzt wird. Außerdem wirken die Kuchensegmente im vorderen Bereich aufgrund der perspektivischen Darstellung generell größer. Dieser Effekt wird durch die auf den vorderen Bereich beschränkte Abbildung des „Kuchenrands“ noch verstärkt. Auch Kreisdiagramme lassen sich über einen der drei weiter oben beschriebenen Menüpfade erstellen.

Aufgabe 7.2
Sie werden aufgefordert, die Häufigkeitsverteilung des Erhebungsmerkmale „Zufriedenheit mit dem bisherigen Studium (auf einer Skala von 1 bis 10)“, „Ausgaben für Lebensmittel in Prozent des Gesamteinkommens“ und „Jahreseinkommen in Euro“ und „Heimatbundesland“ grafisch zu präsentieren. Welche Formen der grafischen Darstellung würden Sie für die einzelnen Variablen wählen? Warum?

7.6 GESTALTUNG DER GRAFISCHEN AUSGABE

Die Nachbereitung von Grafiken dient oftmals nicht nur kosmetischen Verfeinerungen, sondern ist in vielen Fällen unumgänglich. Häufig sollten die von SPSS automatisch erstellten Grafiken für eine spätere Präsentation etwas ansprechender gestaltet werden. Zu den wichtigsten Anforderungen an die Gestaltung von Grafiken gehören:

- Der Titel der Abbildung muss aussagekräftig sein. Vergessen Sie auch die Nummerierung der Abbildungen nicht.
- Die Fallzahl muss angegeben werden (n= …)
- Die Datenquelle muss eindeutig erkennbar sein. Nur wenn Sie stets denselben Datensatz verwenden und dieser zu Beginn der Ausarbeitung explizit genannt wird kann

u.U. darauf verzichtet werden. Wichtig ist auch, dass Ihre Eigenleistung markiert wird, z.B. durch: eigener Entwurf, Datengrundlage …

- Achsen bedürfen der Beschriftung, ggf. ist auch die Einheit anzugeben (Bsp.: „Temperatur [in °C]“ oder „Wohndauer (in Jahren)“ oder Bruttoeinkommen [in k€]
- Bedenken Sie bei der Skalierung der Achsen, welchen sachlich-logischen Gesichtspunkten Sie Beachtung schenken müssen. So empfiehlt sich die Abbildung der vollen Skala von 1 bis 100 wenn Sie Angaben in Prozent darstellen – es sei denn, die Balken werden dann so klein, dass kaum mehr etwas erkennbar ist.
- Wenn Sie mehrere Grafiken anfertigen, die vergleichbar sein sollen, dann sollten auch die Skalierungen identisch sein.
- Gehen Sie sparsam mit Spezialeffekten um, es könnte wie Effekthascherei wirken.
- Stets sollte eine Abbildung ‚für sich‘ verständlich, also selbsterklärend sein; sie muss verstanden werden können ohne die Lektüre des Textes. Notfalls müssen Sie mit Fußnoten/Verweisen arbeiten.
- Vergessen Sie niemals, explizit aus dem Text auf die Grafik/Abbildung zu verweisen. Sie wäre sonst verzichtbar (und wird von manchen Gutachtern deshalb auch ignoriert).

Zur Überarbeitung einer Grafik ruft man zunächst den Diagramm-Editor auf. Um diesen zu öffnen, ruft man das ***SPSS-Ausgabefenster*** auf, in dem die Grafik dargestellt wird, die verändert werden soll. Nach einem Doppelklick auf das Diagramm (SPSS erzeugt dann einen Rahmen um die Abbildung) öffnet sich der Diagramm-Editor. Hier kann man nun verschiedene Layout Einstellungen vornehmen, die, nachdem man das Fenster wieder geschlossen hat, im Ausgabefenster übernommen werden.

Grundsätzlich besteht die Layout-Gestaltung von Grafiken in SPSS aus drei Schritten:

1. Man markiert das zu bearbeitende Grafikelement (Überschrift, Balken, Legende, Achsenbeschriftung) mit einem Linksklick und wählt dieses damit aus.

2. Nun führen wir einen Rechtsklick auswählen in der Dialogbox das *„Eigenschaftsfenster“* aus. Die Registerkarten dieses Kontextmenüs beinhalten verschiedene Optionen zur Veränderung des jeweiligen Grafikelements. Alternativ öffnet ein Doppelklick auf das jeweilige Grafikelement ebenfalls das *„Eigenschaftsfenster“*.

3. Nachdem Änderungen vorgenommen worden sind, überträgt SPSS mit einem Klick auf „Zuweisen“ diese auf die jeweilige Grafik.

Die Vielzahl sämtlicher verfügbaren Optionen kann an dieser Stelle nicht umfassend behandelt werden, die Gestaltung des Layouts einer Grafik kann u.a. über folgende Elemente erfolgen: Textfelder mit Titel/Legende/Fußnote, Achsentitel und -beschriftung, Datenelemente des Diagramms, verschiedene Rahmen (innerer und äußerer, Daten-) usw. – hier hilft nur ‚training on the job‘, also Experimentieren bis das Ergebnis den eigenen Wünschen genügt.

Neben der Veränderung der ursprünglich definierten Grafikelemente kann man auch weitere Elemente hinzufügen (Datenbeschriftungen, Anmerkungen, Gitterlinien etc.). Dazu wählt man im Diagrammeditor die Schaltfläche ***Optionen*** bzw. ***Elemente*** und nimmt die gewünschte Einstellung vor.

Auf eine wichtige Funktion soll an dieser Stelle noch hingewiesen werden: Den Datenbeschriftungsmodus. Durch einen Klick auf die an ein Fadenkreuz erinnernde Schaltfläche ***Markierung von Datenwerten*** schaltet SPSS in den Datenbeschriftungsmodus, in dem man Datenpunkte, Balken, Kreissegmente und ähnliches anklicken und damit den dargestellten Datenwert hervorheben kann. Ein abermaliger Klick auf das Fadenkreuz beendet diesen Modus wieder. Alternativ kann man auch ***Elemente → Datenbeschriftungsmodus*** anwählen. Hier lässt sich die Option ***Datenbeschriftungen einblenden*** anklicken, die sofort sämtliche Beschriftungen einfügt.

Aufgabe 7.3
Wir wollen einen Zusammenhang zwischen Größe und Gewicht aus der „Studentenstudie 1“ darstellen. Die Darstellung soll weiterhin nach dem Geschlecht gruppiert werden.

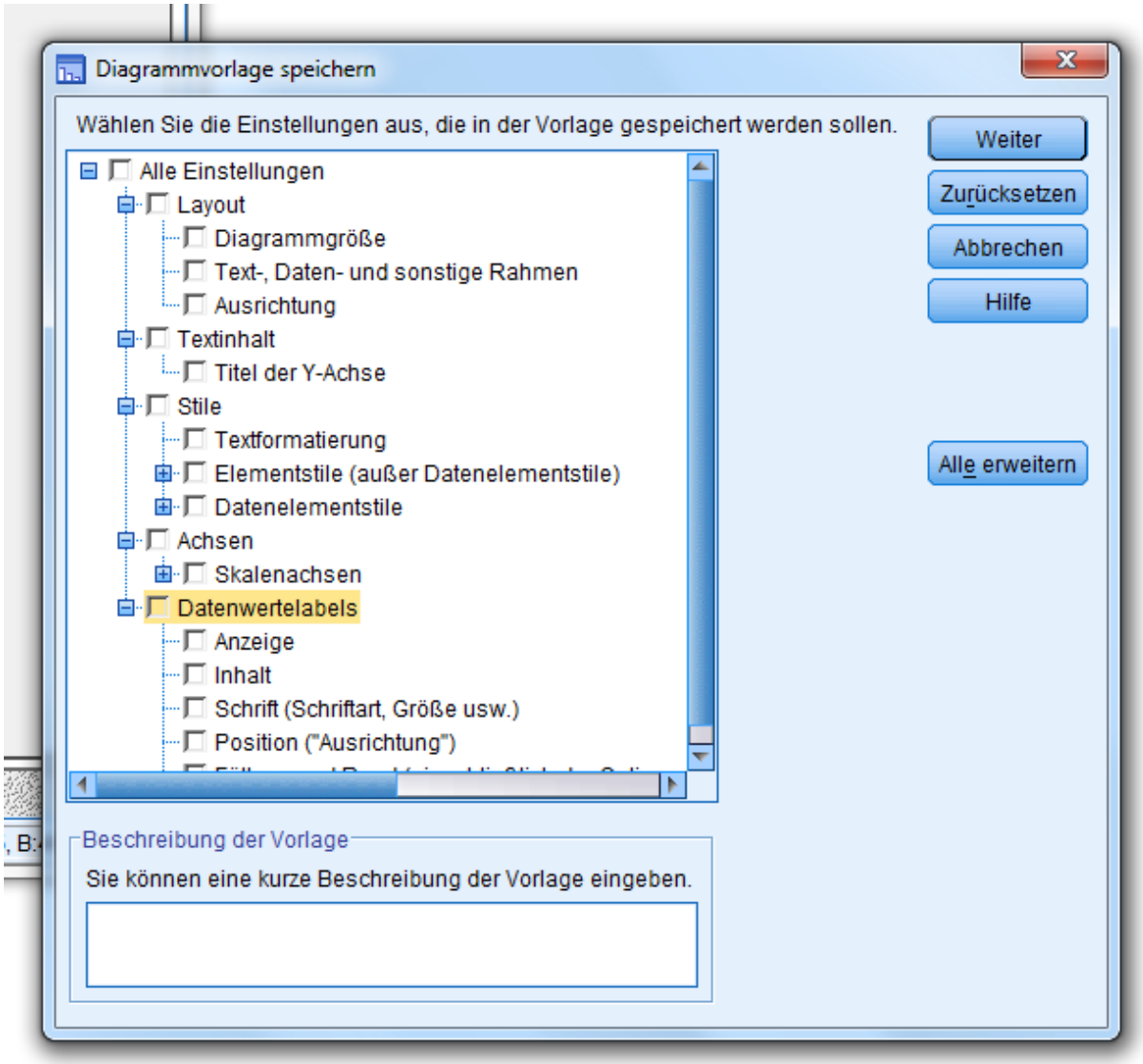
7.6.1 DIAGRAMME ALS VORLAGE SPEICHERN

Ist man mit der Bearbeitung eines Diagramms zufrieden und möchte aus Gründen der einheitlichen Darstellung auch alle weiteren Diagramme seines Auswertungsberichtes mit dem gleichen Layout gestalten, sollte man die Möglichkeit der Vorlagenerstellung in SPSS nutzten.

Im Diagrammeditor kann man unter ***Daten → Diagrammvorlage speichern*** verschiedene Layout-Einstellungen in die neue Vorlage übernehmen (siehe Abbildung 7.6).

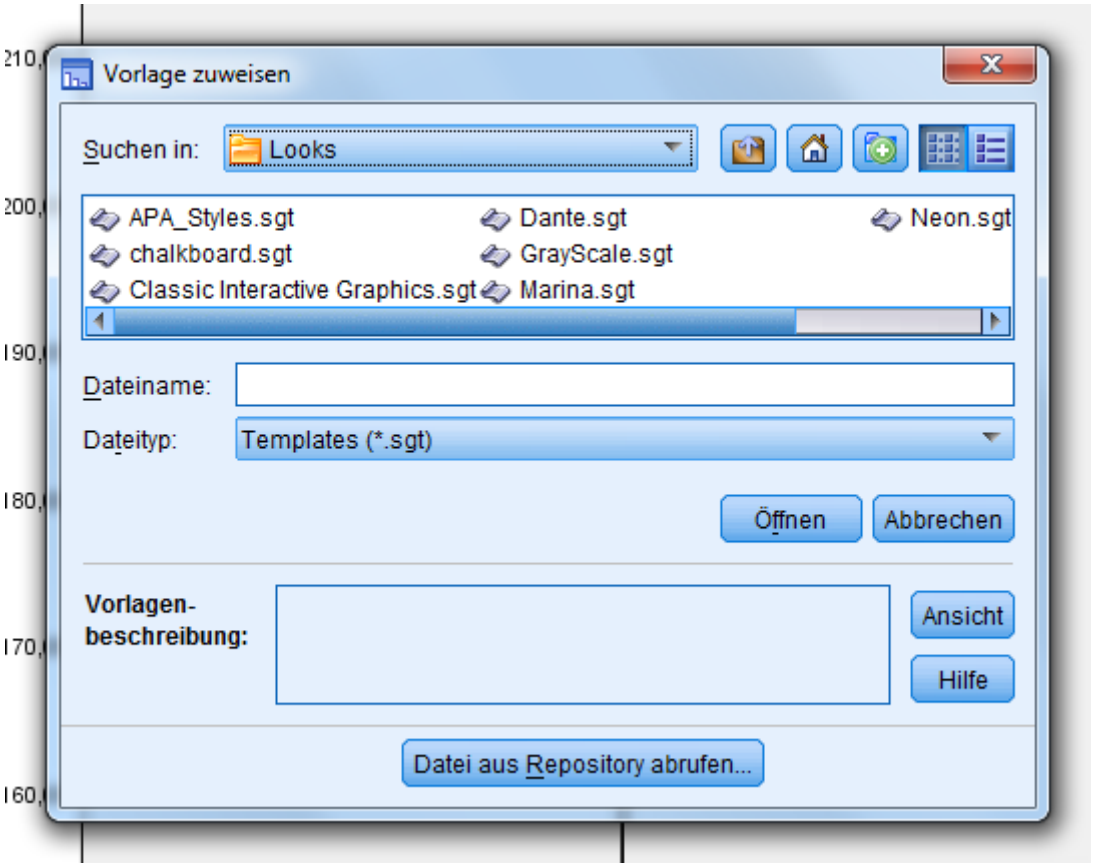
Nach einem Klick auf ***Weiter*** bittet uns SPSS, diese neue Vorlage zu betiteln, um ein späteres Aufrufen zu erleichtern.

Abbildung 7.6: Dialogbox ***Diagrammvorlage speichern***



Will man eine bereits erstellte (oder vorgefertigte) Vorlage verwenden, wählt man im Diagrammeditor ***Daten → Diagrammvorlage*** zuweisen und wählt in folgendem Kontextmenü die entsprechende Datei (Abbildung 7.7).

Abbildung 7.7: Dialogbox ***Diagrammvorlage zuweisen***



Ein Klick auf ***Öffnen*** überträgt die Layouteinstellungen auf die aufgerufene Grafik.

7.6.2 TRANSFER VON GRAFIKEN UND TABELLEN

Oftmals will man die in SPSS erstellten Diagramme in ein anderes Programm übertragen. Beispielsweise, um die mit SPSS erstellten Grafiken in einen textbasierten Bericht zu übernehmen. Oftmals reicht dazu ein einfaches ***Kopieren*** und ***Einfügen***. In manchen Fällen kann es jedoch von Vorteil sein, die erzielten Ergebnisse als eigenständige Dateien zu speichern. Dazu wählt man über ***Bearbeiten → Kopieren Spezial*** das Format des kopierten Objektes (siehe Abbildung 7.8).

Des Weiteren bietet SPSS die Möglichkeit, Elemente aus dem Ausgabefenster als Dateien zu exportieren. Dazu markiert man zunächst das gewünschte Objekt und öffnet dann mit der rechten Maustaste eine Auswahlliste, aus der man den Menüschritt ***Exportieren*** wählt (Abbildung 7.9).

Unter ***Typ*** lassen sich nun die Art der Ausgabe festlegen (Excel, HTML, PowerPoint, Word, …). Unter ***Optionen Ändern*** kann man die Übertragung in Word oder RTF-Format weiter optimieren.

Abbildung 7.8: Kontextmenü ***Kopieren Spezial***

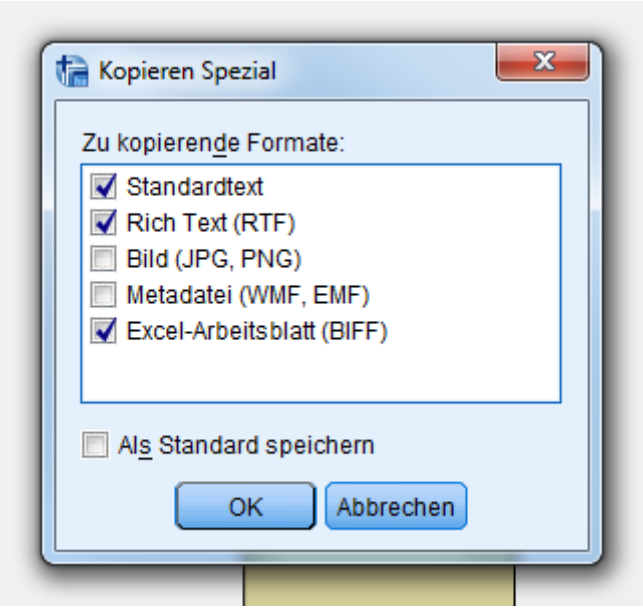
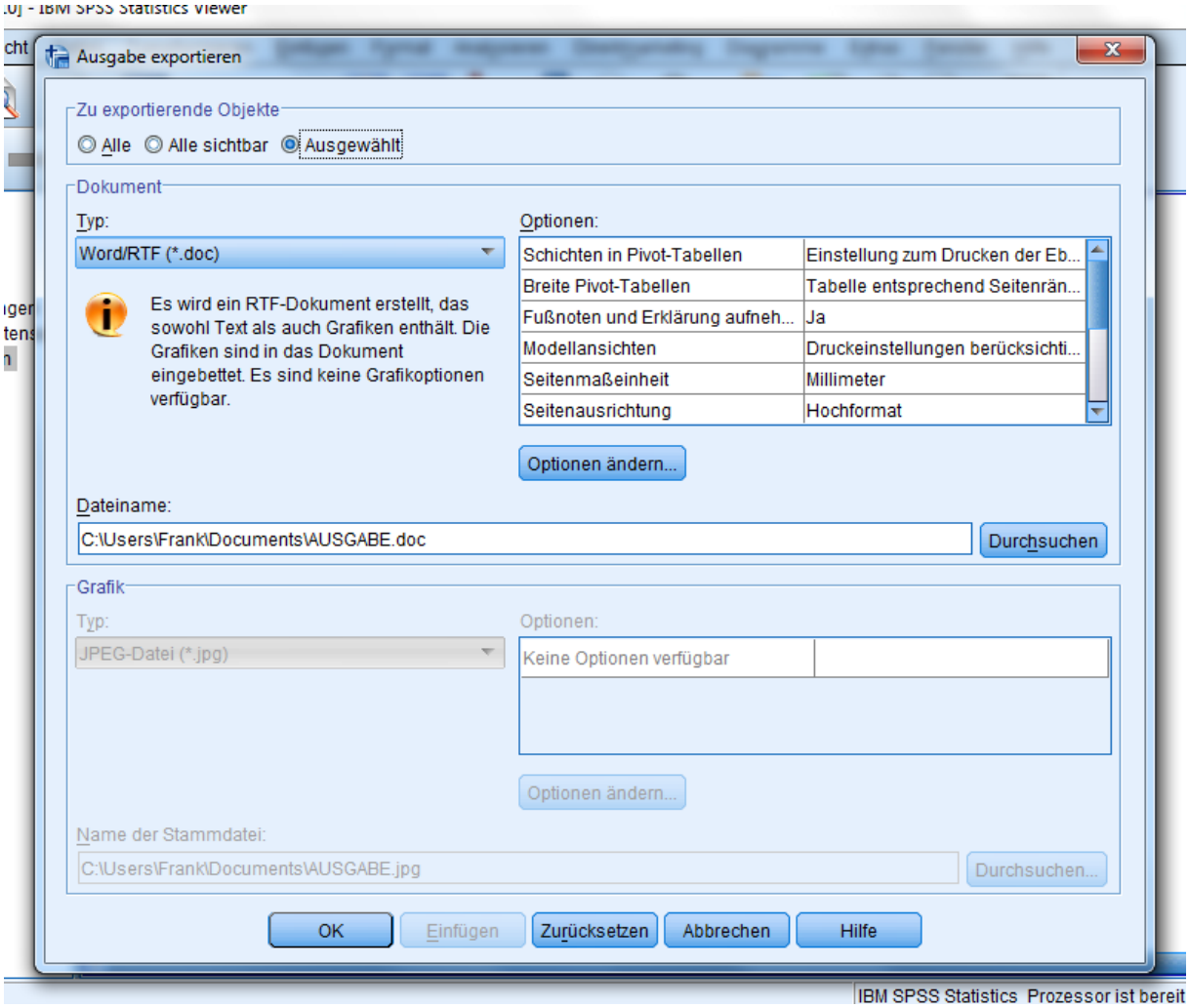


Abbildung 7.9: Kontextmenü ***Ausgabe Exportieren***



8 DER UMGANG MIT MEHRFACHANTWORTEN

Wie erstellt man Fragen, auf die mehrere verschiedene Antworten gegeben werden dürfen?

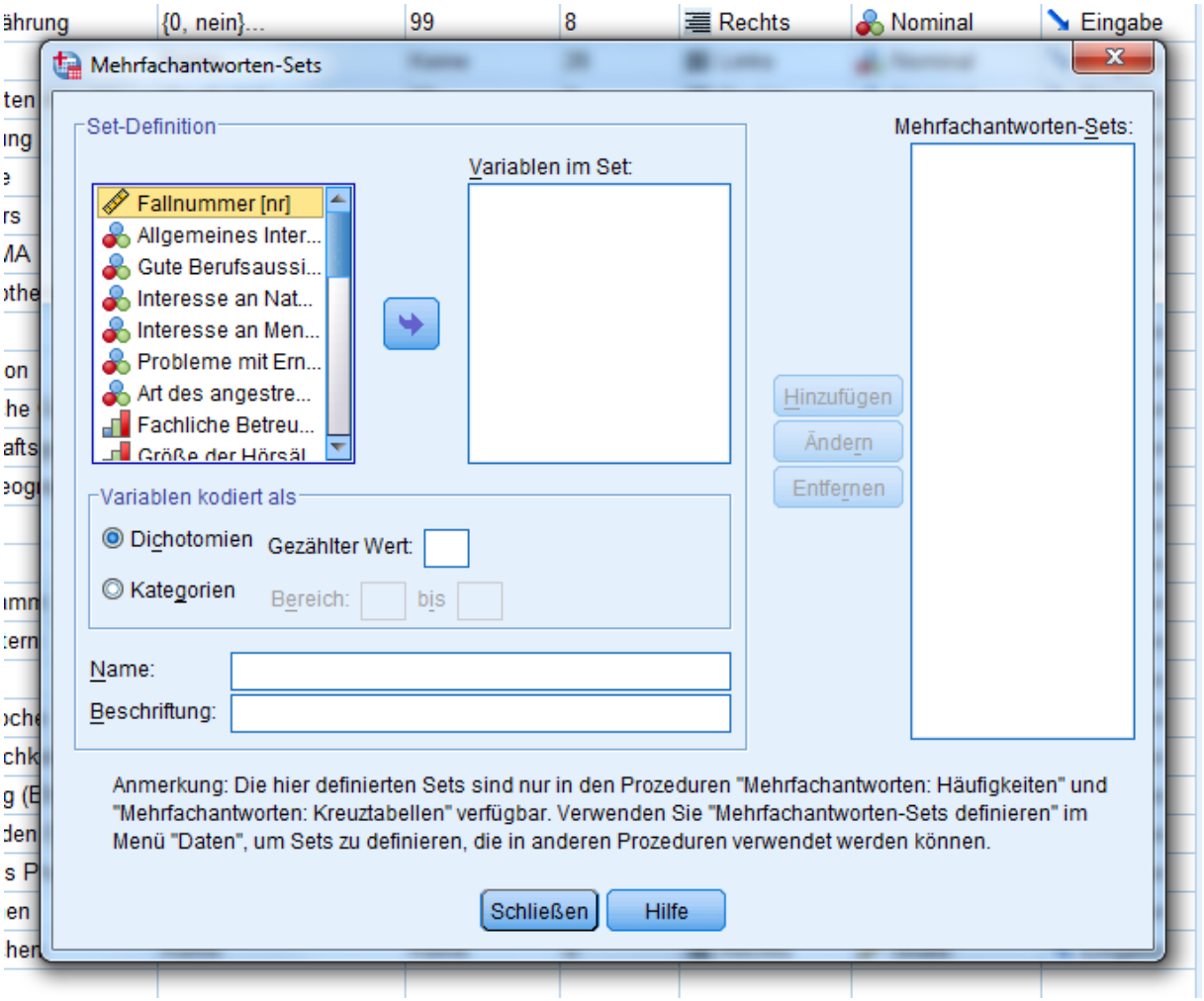
Wie wertet man Mehrfachantwortensets aus?

In der Regel strebt man eindimensionale Messergebnisse an, bei denen jeweils nur eine einzige Antwortmöglichkeit zutrifft. Bei Befragungen kann es jedoch sinnvoll sein, den Probanden mehrere zutreffende Antwortmöglichkeiten offen zu halten.

SPSS stellt zur Handhabung solcher Mehrfachmessungen verschiedene Operation bereit. Wir beschränken uns im Folgenden auf die Einstellungen im Menü „Mehrfachantwort“ und der so genannten *Multiple Dichotomien-Methode.* Bei dieser Methode wird für jeden Wert, den eine Variable erreichen kann, eine eigene Variable in SPSS angelegt. Beispielsweise haben wir in der „Studentenstudie 1“ die Probanden nach ihrem Hauptinteresse am Geographiestudium befragt. Dabei war eine Mehrfachbeantwortung mit den folgenden Antwortvorgaben erlaubt: Physische Geographie, Wirtschaftsgeographie und Sozialgeographie. Für jede dieser drei Kategorien wurde eine eigene Variable angelegt (b2.1 bis b2.3). Für diese Variable wurden wiederum Wertelabels nach folgendem Schema erstellt: 1 = trifft zu, 2 = trifft nicht zu, 3 = keine Angabe.

Wenn wir nun eine Häufigkeitstabelle für dieses Mehrfachantwortenset erstellen möchten, müssen wir eine zusammenfassende Variable erstellen. Diese Variable definiert man unter ***Analysieren → Mehrfachantworten → Sets definieren…*** wie unter Abbildung 8.1 gezeigt.

Abbildung 8.1: Dialogbox Mehrfachantworten-Sets



Nun wählen wir die Variablen aus, die zu einem Set zusammengefasst werden sollen und wählen unter ***Variable kodiert als*** die Option ***Dichotomien***. Unter ***Gezählter Wert*** geben wir die Zahl 1 für ***trifft zu*** an. Im Feld ***Name*** lässt sich die zusammenfassende Variable benennen. Wir beenden die Auswahl mit einem Klick auf ***Schließen***. Die so definierte Variable wird dann im Untermenü ***Mehrfachantworten*** als Variable aufgeführt.

Wollen wir nun eine Häufigkeits- oder Kreuztabelle für die Zusammenfassungsvariablen erstellen, wählen wir ***Analysieren → Mehrfachantworten → Häufigkeitstabellen*** oder ***Analysieren → Mehrfachantworten → Kreuztabellen***. Wie gewohnt wählen wir hier eine oder mehrere Variablen aus und klicken auf ***Ok***. Unter dem Menü Kreuztabellen lassen sich jetzt auch Kreuztabellen zwischen Zusammenfassungsvariablen und „normalen“ Variablen erstellen.

9 STATISTISCHE KENNWERTE

Welche statistischen Kennwerte lassen sich wie in SPSS berechnen?

Wann bietet sich welcher Kennwert an?

Mit der Bestimmung eines statistischen Kennwertes lässt sich eine summarische Auskunft über spezielle Eigenschaften von Verteilungen geben. Im Unterschied zu den vorhergehenden Kapiteln sollen hier zunächst alle statistischen Kennwerte vorgestellt werden, um danach auf die Vorgehensweise in der Berechnung einzugehen und eine Sammlung an Aufgaben zusammenzufassen.

9.1 MAßE DER ZENTRALTENDENZ

Die Maße der Zentraltendenz ermöglichen es, die Verteilung bezüglich ihrer zentralen Lage durch einen einzelnen Kennwert zu charakterisieren. Sie geben sozusagen an, welcher einzelne Wert eine Reihe von Daten am besten repräsentiert.

9.1.1 DAS ARITHMETISCHE MITTEL

Das Arithmetische Mittel $\bar{x}$ ist das bekannteste Lagemaß und wird alltagssprachlich als Durchschnittswert bezeichnet. Dieser Mittelwert ist ein passender Kennwert für intervallskalierte und annähernd normalverteilte Variablen. Eine Eigenschaft des arithmetischen Mittels ist, dass alle Abweichungen der Messwerte über und unter dem Mittelwert zusammengezählt Null ergeben. Zur Veranschaulichung der inhaltlichen Bedeutung des arithmetischen Mittels kann man sich einen Balken vorstellen, der steif und völlig gewichtslos ist. Ergänzen wir nun verschiedene gleich schwere Gewichte, die für die Messwerte stehen. Das arithmetische Mittel ist dann der Wert, der den Balken ausbalanciert und damit im Gleichgewicht hält. Die Angabe eines Mittelwertes zur Charakterisierung einer Verteilung ist jedoch nicht völlig unproblematisch. Die Beschränkung auf einen einzelnen Kennwert kann zu Informationsverlusten führen oder bei der Interpretation bestimmte Effekte verschleiern.

9.1.2 DER MEDIAN

Für ordinal- und für intervallskalierte Daten kann auch noch ein anderer Mittelwert angegeben werden: der Median. Der Median ($\tilde{x}$ oder *Me*) repräsentiert jenen Messwert, der exakt in der Mitte der geordneten Datenreihe erhoben wurde. Damit liegt die Hälfte aller Merkmalsträger der Messreihe unterhalb des Medians, die andere Hälfte oberhalb. Der besondere Vorteil des Medians liegt in seiner Unempfindlichkeit gegenüber Ausreißern. Bei nominalskalierten Variablen ist eine Berechnung jedoch sinnlos, weil keine irgendwie ‚natürliche' Ordnung der Merkmale vorliegt, die Beobachtungen also auch ganz anders sortiert/geordnet werden können.

9.1.3 DER MODUS

Vor allem bei der Untersuchung nominalskalierter Daten bietet sich der Modus ($\hat{x}$ oder *Md*) als passender statistischer Kennwert zur Beschreibung der zentralen Lage der Verteilung an. Der Modalwert stellt hierbei den am häufigsten auftretenden Wert einer Stichprobe dar. Wenn sich mehrere Werte durch dieselbe maximale Häufigkeit auszeichnen, wird SPSS nur den kleinsten davon anzeigen.

9.2 STREUUNGSMAßE

Die so genannten Streuungs- oder Dispersionsmaße geben Auskunft über die Streuung bzw. die Breite einer Verteilung. Diese Angaben sind von besonderer Bedeutung, da eine Verteilung durch die Angabe von einem oder mehreren Mittelwerten allein nur unzureichend beschrieben werden kann. Die Streuungsmaße ergänzen Details zur Abweichung der einzelnen Werte von der Mitte.

9.2.1 VARIANZ

Die Varianz ($\boldsymbol{s^2}$) beschreibt die durchschnittliche quadrierte Abweichung vom Mittelwert. Sie wird umso größer, je stärker die Messwerte vom Mittelwert abweichen. Genau wie das arithmetische Mittel lässt sich auch die Varianz nur für intervallskalierte und annähernd normalverteilte Variablen berechnen. In verschiedenen Lehrbüchern finden sich zwei geringfügig unterschiedliche Formeln zur Berechnung der Varianz: Mal wird die Summe der quadrierten Abweichungen durch ***n*** geteilt, mal durch ***n-1***. Formal betrachtet sollte man durch ***n-1*** teilen, wenn man es mit einer Stichprobe zu tun hat und ***n*** verwenden, wenn man die Varianz einer Grundgesamtheit betrachtet. Es sei darauf hingewiesen, dass SPSS immer ***n*** verwendet, egal ob Grundgesamtheiten oder Stichproben betrachtet werden. Die Differenz ist bei größeren Stichproben zu vernachlässigen.

9.2.2 STANDARDABWEICHUNG

Da die Varianz inhaltlich schwer interpretierbar erscheint, bedient man sich deutlich häufiger der Standardabweichung (***s***, die Wurzel der Varianz). Bei normalverteilten Werten liegen 68,26 % aller Fälle im Bereich von $\bar{x} + s$ bis $\bar{x} - s$, im Bereich $\bar{x} + 2s$ bis $\bar{x} - 2s$ sind es schon 95,44 %. Im Allgemeinen sprechen kleine Standardabweichungen für eine gewisse Nähe der einzelnen Messwerte zum Arithmetischen Mittel, hohe Standardabweichungen für größere Abstände zwischen Einzelwerten und Mittelwert.

9.2.3 QUARTILSABSTAND

Wenn die Voraussetzungen zu Berechnung des arithmetischen Mittels nicht gegeben sind oder ein ordinales Datenniveau vorliegt, kann man den empirischen Quartilsabstand als ergänzendes Streuungsmaß zum Median verwenden. Dieser Kennwert macht eine Aussage

darüber, in welchem Bereich die mittleren 50 % einer Reihe von Messwerten liegen. Das erste Quartil schneidet die unteren 25 % einer Verteilung ab, das zweite die unteren 50 % (damit ist es identisch mit dem Median) und das dritte trennt die unteren 75 % ab. Die Quartilsabstand wird als Differenz zwischen Q3 (75 %) und Q1 (25 %) definiert.

9.2.4 SPANNWEITE UND PERZENTILWERTE

Die Spannweite gibt die Ausdehnung zwischen dem Maximum (höchster Messwert) und dem Minimum (niedrigster Messwert) an. Zur Berechnung der Spannweite bildet man einfach die Differenz aus dem größten und dem kleinsten Messwert. Die Aussagekraft der Spannweite ist allerdings sehr begrenzt, da keinerlei Angaben über die dazwischenliegenden Werte gemacht werden. Extremwerte und Ausreißer haben einen großen Einfluss auf das Ergebnis der Spannweite und können zu Interpretationsfehlern führen. In einem solchen Fall bietet sich die Betrachtung eines eingeschränkten Bereiches der Streuung an, wie zum Beispiel nur die der mittleren 80 % der Werte. Die Berechnung verläuft analog zum Quartilsabstand, nur das die Perzentile vorher erst festgelegt werden müssen.

9.2.5 VARIATIONSKOEFFIZIENTEN

Beim Variationskoeffizienten *v* oder *Vk* handelt es sich um ein relatives Streuungsmaß, das anders als Varianz und Standardabweichung einen Vergleich zwischen der Streuung zweier metrischer Datenverteilungen erlaubt. Da die Varianz und die daraus abgeleitete Standardabweichung nicht normiert sind, kann im Allgemeinen nicht beurteilt werden, ob eine Varianz groß oder klein ist. So schwanken beispielsweise die Preise für ein Pfund Salz, das im Durchschnitt wohl etwa 50 Cent kostet, im Cent-Bereich, während Preise für ein Auto, das im Mittel beispielsweise 20.000 Euro kostet, im 1000-Euro-Bereich variieren. Der Variationskoeffizient erlaubt sozusagen den Vergleich von Äpfeln mit Birnen.

Der Variationskoeffizient ist ein durch das arithmetische Mittel gewichtetes (bereinigtes) Streuungsmaß. Ist die Standardabweichung größer als der Mittelwert, so ist der Variationskoeffizient größer 1.

Leider bietet SPSS kein Verfahren zur Berechnung des Variationskoeffizienten für eine einzelne Variable. Allerdings kann man bei Vorliegen des arithmetischen Mittels ($\bar{x}$) und der Standardabweichung (s) diesen leicht manuell berechnen. Der Variationskoeffizient entspricht dabei der Formel:

$$v = \frac{s}{|\bar{x}|} \times 100$$

9.3 BEISPIELAUSWERTUNGEN UND ÜBUNGSAUFGABEN

9.3.1 BERECHNUNG DER VORGESTELLTEN KENNWERTE IN SPSS

Zur Berechnung dieser Kennwerte stehen verschiedene Menüwahlen zur Verfügung:

Analysieren → Deskriptive Statistiken → Deskriptive Statistiken

Analysieren → Deskriptive Statistiken → Häufigkeiten

Analysieren → Deskriptive Statistiken → Explorative Datenanalyse

Analysieren → Berichte → Fälle zusammenfassen

Eine Übersicht über die Vielzahl der mit SPSS berechenbaren Kennzahlen und die entsprechenden Menüs bietet Abbildung 9.3.

Abbildung 9.1: Übersicht über die Möglichkeiten zur Berechnung statistischer Kennzahlen in SPSS

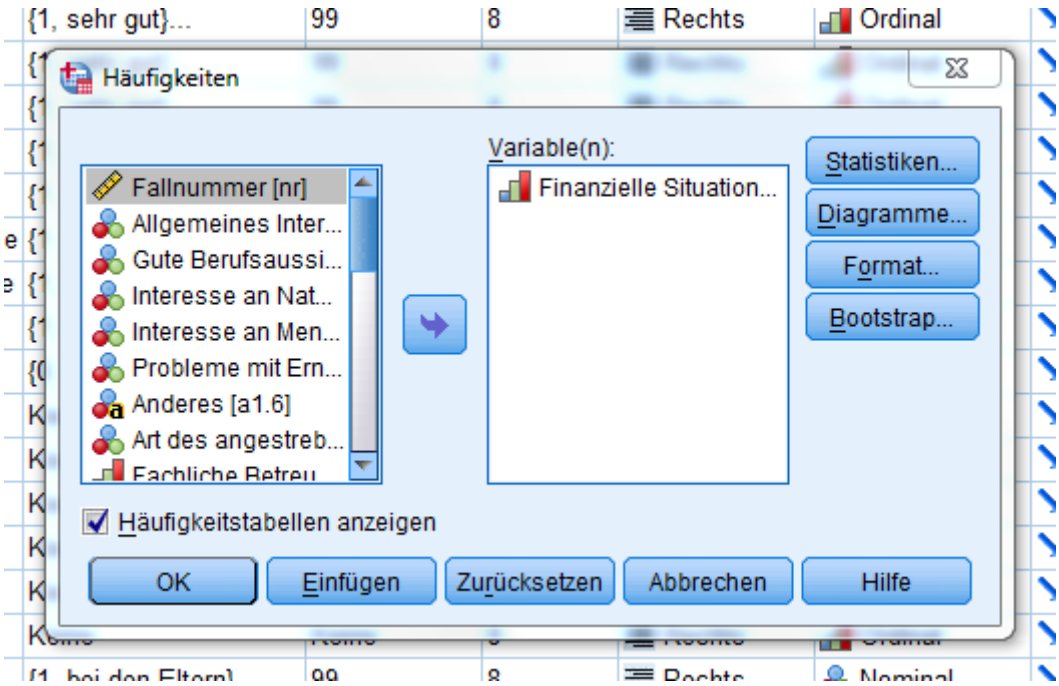
Kennwerte	Deskriptive Statistiken	Häufigkeiten	Explorative Datenanalyse	Fälle zusammenfassen
Mittelwert	X	X	X	X
Summe	X	X		X
Median		X	X	X
Gruppierter Median		X		X
Quartile		X		
Perzentile		X	X	
Modus		X		
Standardabweichung	X	X	X	X
Standardfehler	X	X	X	X
Varianz	X	X	X	X
Minimum	X	X	X	X
Maximum	X	X	X	X
Spannweite	X	X	X	X
Quartilsabstand			X	
Kurtosis (Exzess)	X	X	X	X
Schiefe	X	X	X	X
Standardfehler des Exzess	X	X	X	X
Standardfehler der Schiefe	X	X	X	X
Konfidenzintervall			X	
Harmonisches Mittel				X
Geometrisches Mittel				X
M-Schätzer			X	
Ausreißer			X	
Gestutzter Mittelwert			X	

Diese Übersicht verdeutlicht einerseits das weite Spektrum von Möglichkeiten, die SPSS für die Berechnung statistischer Kennwerte bereithält. Zugleich wird auch veranschaulicht, dass häufig mehrere Wege durch die Menüstruktur des Programms zur gewünschten Kennzahl

führen. In dieser Anleitung beschränken wir uns allerdings auf eine kleine Auswahl und hier insbesondere auf die häufig verwendeten Maßzahlen für die Zentraltendenz und die Streuungsparameter.

Beginnen wir die Berechnung statistischer Kennwerte zunächst über das Menü ***Häufigkeiten***. Dazu wählt man ***Analysieren → Deskriptive Statistiken → Häufigkeiten*** an. Es öffnet sich folgendes Menü (Abbildung 9.2):

Abbildung 9.2: Kontextmenü ***Häufigkeiten***



Zunächst bringen Sie die Variable Ihrer Wahl, wie in der Darstellung angedeutet, in das Testvariablenfeld. Danach ein Klick auf ***Statistiken*** und es öffnet sich das in Abbildung 9.3 wiedergegebene Menü.

Unter dem Menüpunkt ***Statistiken*** finden wir alle oben beschriebenen Kennwerte sowie verschiedene weitere Auswahloptionen wie die Berechnung der Kurtosis und der Schiefe. Wir beschränken uns jedoch zunächst auf die Berechnung des arithmetischen Mittels, des Medians und der Standardabweichung (s. Abbildung 9.3).

Da wir uns im Menü ***Häufigkeiten*** befinden, haben wir hier auch die Möglichkeit, verschiedene Diagrammdarstellungen anzuwählen und können bei der Auswahl mehrerer Variablen unter der Schaltfläche ***Format*** deren Anordnung ändern (siehe Abbildungen 9.4 und 9.5).

Abbildung 9.3: Dialogbox ***Statistiken***

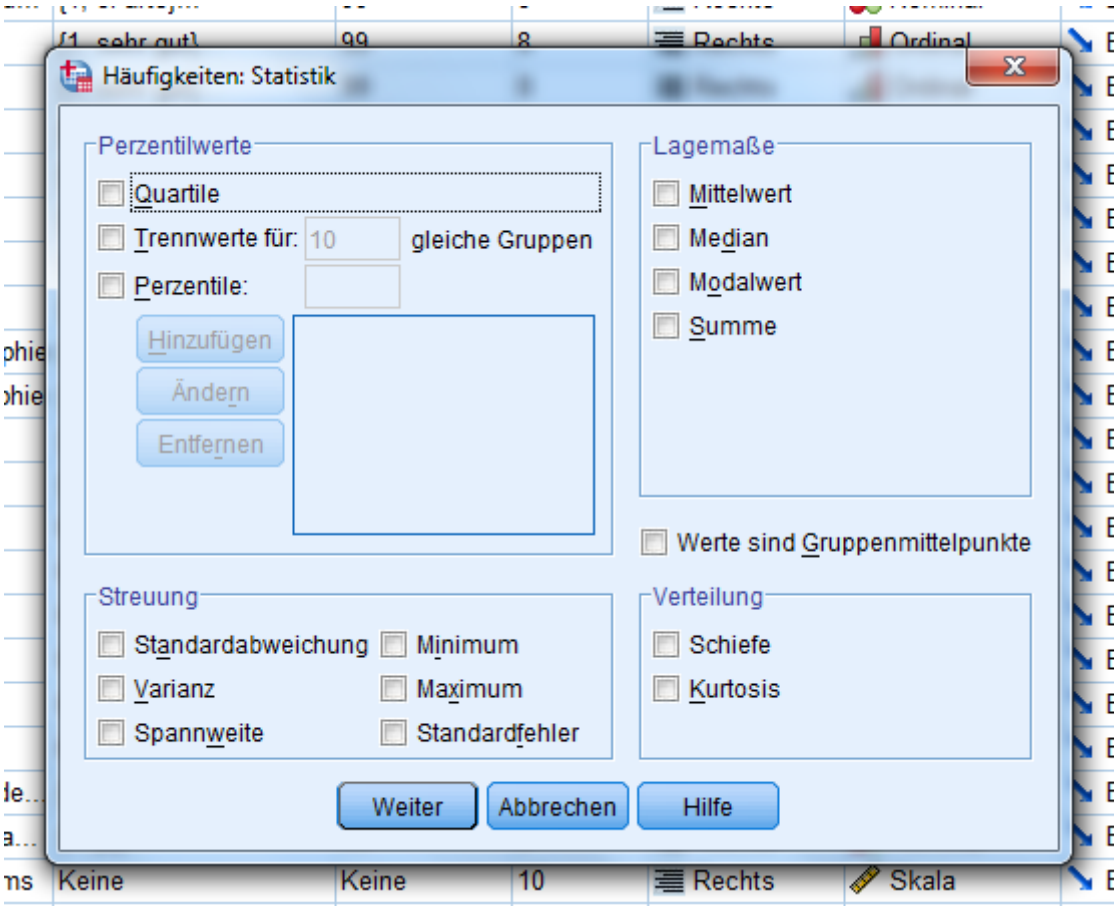


Abbildung 9.4: Kontextmenü ***Diagramme***

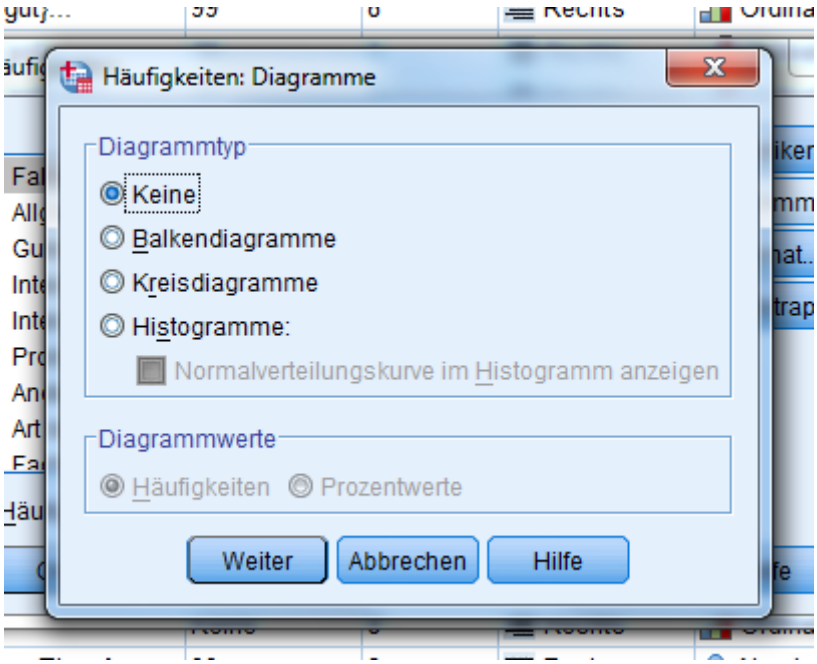
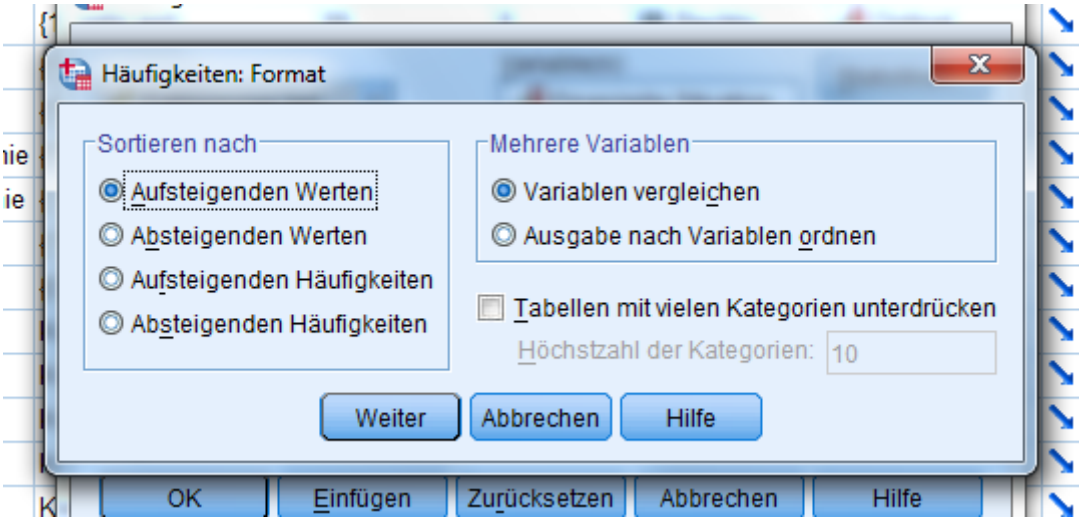


Abbildung 9.5: Dialogbox ***Format***



Wir beschränken uns jedoch zunächst auf eine einfache Berechnung der oben genannten Kennwerte. Deshalb gehen Sie zurück ins Hauptmenü ***Häufigkeiten*** und auf ***Ok***. Nun sollte sich das SPSS-Ausgabefenster öffnen und die Ergebnisse in tabellarischer Form darstellen.

9.3.2 ÜBUNGSAUFGABEN

Aufgabe 9.1
Wir befragen vier Angestellte eines Betriebs nach ihrem Bruttoeinkommen:
Person 1: 2500 €
Person 2: 3000 €
Person 3: 3200 €
Person 4: 25000 €
Berechnen Sie das arithmetische Mittel und interpretieren Sie das Ergebnis. Welche Voraussetzung für die Berechnung des arithmetischen Mittels wurde nicht beachtet?

Aufgabe 9.2
Kind A erzielt bei einem Rechentest von 12 zu erreichenden Punkten zuerst 11, dann 12, dann 4 und dann 0.
Kind B erreichte im selben Rechentest zuerst 0, dann 4, dann 12 und dann 11 Punkte.
Berechnen Sie das arithmetische Mittel der Testergebnisse für beide Kinder und vergleichen Sie die Ergebnisse. Welches Problem ergibt sich für die Interpretation?

Aufgabe 9.3
Bestimmen Sie für die „Studentenstudie 1“ die Variable mit der höchsten Varianz.

Aufgabe 9.4
Kreuzen sie den (die) passenden Mittelwerte für die Beschreibung der verschiedenen Verteilungen an.

	Modus	Median	Arithmetisches Mittel
Mehrgipflige Verteilung			
Asymmetrische Verteilung			
Metrische Daten			
An den Enden offene Klassen			
Extrem kleine Stichprobe			
Nominaldaten			

Aufgabe 9.5
Inwiefern kann die Beschränkung auf einen Mittelwert bei der Beschreibung einer Verteilung zu Problemen führen?

Aufgabe 9.6
Vergleichen Sie die Variablen „Größe“ und „Körpergewicht“ der „Studentenstudie 1“ hinsichtlich ihrer Standardabweichung. Dabei sollen jedoch nur Personen über 25 Jahre untersucht werden.

Aufgabe 9.7
Um die Entwicklung der Telefonkosten des letzten Jahres zu analysieren wird Tochter Bärbel von Ihrem Vater beauftragt, die mittleren Telefonkosten und deren Streuung zu berechnen. Die Rechnung betrug jeweils in Euro:

Jan. 35,46	Feb. 33,60	Mrz. 40,44	Apr. 34,20	Mai. 36,18	Jun. 36,84
Juli. 31,44	Aug. 30,18	Sep. 41,04	Okt. 33,60	Nov. 38,16	Dez. 132,30

(a) Berechnen Sie das arithmetische Mittel und die Standardabweichung der monatlichen Telefonkosten.
(b) Bärbel hat, auf Anraten ihrer Freundin, im Dezember häufig bei teuren 0190-Talklines angerufen. Sie ist nun entsetzt über den hohen Mittelwert und befürchtet Taschengeldentzug. Wie kann man Bärbel aus der Patsche helfen?

Aufgabe 9.8
Führen Sie die unter Kapitel 9.3.1 beschriebene Auswertung über das Menü ***„Deskriptive Statistiken“*** durch. Wählen Sie dabei auch die Option ***„standardisierte Werte in einer Variablen speichern“*** in der Hauptdialogbox an. Wie lässt sich die von SPSS neu angelegte Variable ***„ZAlter“*** interpretieren? Welchen Nutzen bringt diese Standardisierung?

Aufgabe 9.9
Berechnen Sie die Varianz für die Variable „Alter“ aus der Studentenstudie. Welche Wirkung hat die Quadrierung der Abweichungen vom Mittelwert auf die Berechnung der Varianz? Wie lässt sich die Einheit inhaltlich interpretieren?

Aufgabe 9.10
Teilen Sie die Verteilung der Variable „Körpergröße in cm“ in 10 Dezile auf und berechnen sie die Spannweite für die mittleren 80 % der Werte.

Aufgabe 9.11
Berechnen Sie das 5 % getrimmte Mittel für die Variable „Gewicht in Kilogramm“ aus der „Studentenstudie 1“ und interpretieren Sie das Ergebnis im Vergleich zum arithmetischen Mittel.

Aufgabe 9.12
Berechnen Sie das arithmetische Mittel der Körpergröße der Studenten aus der „Studentenstudie 1“. Inwieweit unterscheiden sich diese bezogen auf die getrennte Betrachtung von Männern und der Frauen?

Aufgabe 9.13
Wie groß ist der prozentuale Anteil der Gemeinden aus der SPSS-Datendatei „Bev_Rhein_Main“, die sich hinsichtlich ihrer Bevölkerungszahl im Intervall von arithmetischem Mittel plus/minus eine Standardabweichung befinden?

Aufgabe 9.14
Fügen Sie in der SPSS-Datendatei „Bev_Rhein_Main“ eine Variable hinzu, welche die standardisierten Bevölkerungszahlen zum Inhalt hat. Bestimmen Sie für die standardisierten Bevölkerungszahlen das arithmetische Mittel und die Standardabweichung und interpretieren Sie jeweils den Wert.

10 KORRELATIONEN UND LINEARE REGRESSION

Wie berechnet man ein Maß für den Zusammenhang zwischen zwei Variablen?

Wie interpretiert man diesen Kennwert?

Was ist eine lineare Regression und kann man eine solche darstellen?

In vielen Untersuchungen stellt man sich die Frage nach einem Zusammenhang zwischen zwei Variablen: Hängen das Geschlecht und die Studienfachwahl zusammen? Gibt es einen Zusammenhang zwischen Temperatur und Niederschlag? Wie steht es mit Kreativität und Intelligenz? Sind die Variablen voneinander abhängig?

Hängen zwei Variablen zusammen, dann hat die Ausprägung von Variable 1 einen Einfluss auf Variable 2. Einen Messwert, der diesen Zusammenhang auszudrücken vermag nennt man Korrelationsmaß. Je nach Skalenniveau und Verteilung der vorliegenden Daten verwendet man unterschiedliche Korrelationsmaße:

Nominales Datenniveau	Phi-Koeffizient Kontingenzkoeffizient Cramer‘sches Assoziationsmaß V („Cramers V“)
Ordinales Datenniveau	Rangkorrelation nach Spearman Kendalls τ

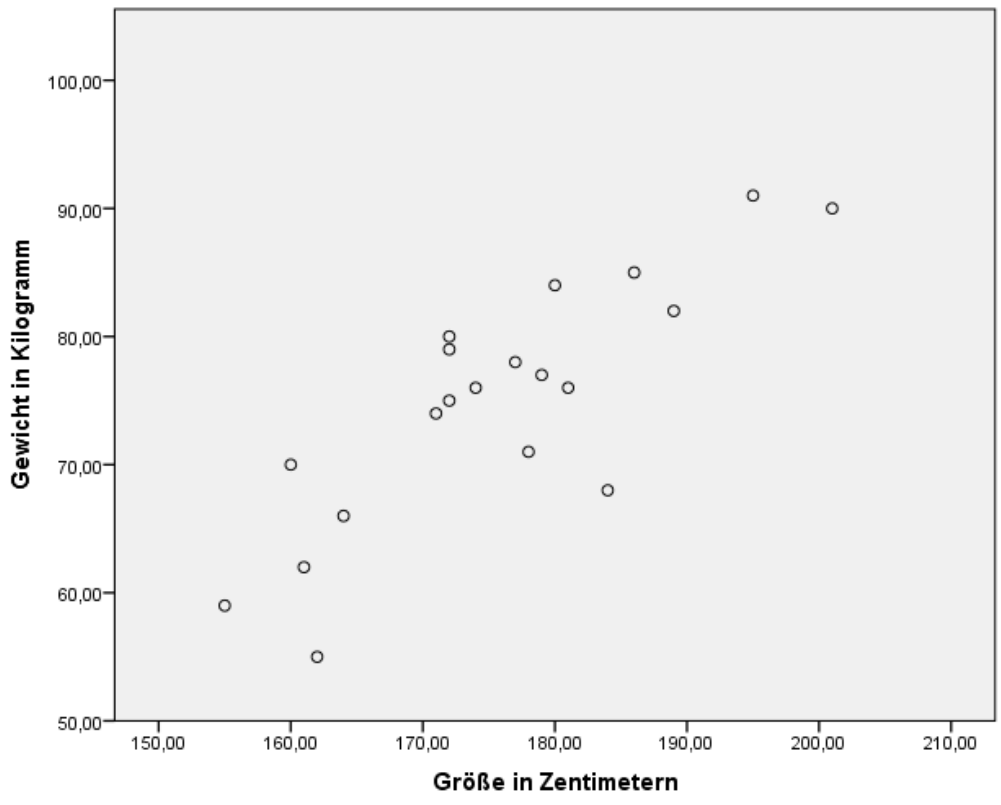
Metrisches Datenniveau	**Produkt-Moment-Korrelation nach Pearson** **Rangkorrelation nach Spearman** **Kendalls τ**
Kombination aus metrischer (intervallskalierter) und nominalskalierter Variable	Biseriale Korrelation

Neben der Berechnung des eigentlichen Zusammenhangsmaßes geben alle Korrelationsausgaben in SPSS noch den Signifikanzwert mit an. Dieser wird unter der Spalte ***Näherungsweise Signifikanz*** angegeben. Interpretieren lässt sich dieser Zahlenwert als Prozentzahl, die die Wahrscheinlichkeit eines statistischen Irrtums angibt. Wird beispielsweise ein Signifikanz von 0,097 ausgegeben, heißt das: Mit einer Wahrscheinlichkeit von 9,7 %, ist das berechnete Zusammenhangsmaß nur per Zufall zustande gekommen. Nach der üblichen Konvention, eine maximale Irrtumswahrscheinlichkeit von 5 % anzunehmen, wäre der Zusammenhang also nicht signifikant. Der Zusammenhang ist somit statistisch nicht abgesichert (mehr zum Thema Signifikanz in Kapitel 11).

10.1 METRISCHES DATENNIVEAU

Zur Veranschaulichung des Prinzips einer Korrelation bietet sich die Betrachtung des Zusammenhangs zwischen zwei metrischen Variablen in Form eines Streudiagramms an. Abbildung 10.1 zeigt ein Beispiel für den Zusammenhang zwischen „Größe“ und „Gewicht“ der Probanden aus der Studentenstudie.

Abbildung 10.1: Streudiagramm für ***„Größe“*** **und** ***„Gewicht“***



Fährt man die x-Achse entlang, bemerkt man, dass die y-Werte in ähnlichem Maße ansteigen. Diese Beobachtung spricht für eine positive Korrelation. Bei einer negativen Korrelation würden z.B. höhere Werte auf der x-Achse mit niedrigeren Werten auf y-Achse einhergehen.

10.1.1 DER PRODUKT-MOMENT-KORRELATIONS-KOEFFIZIENT NACH PEARSON

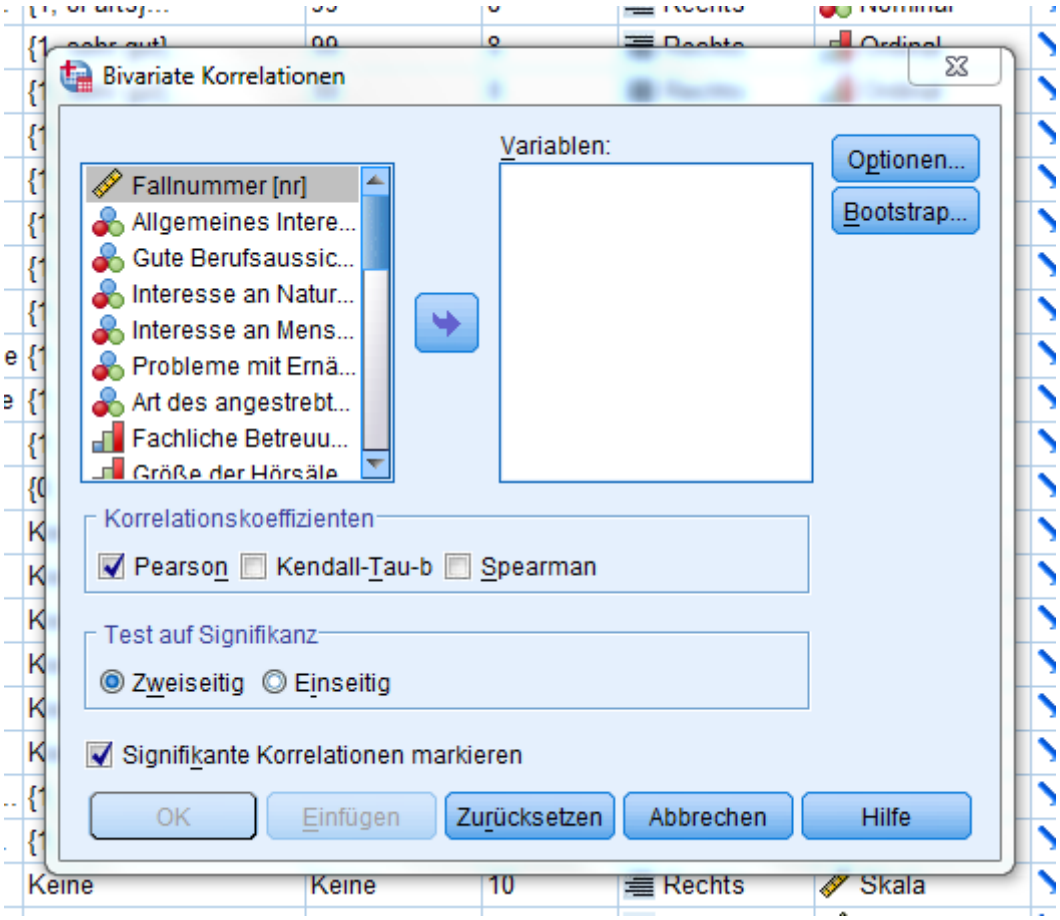
Der Produkt-Moment-Korrelations-Koeffizient würde im Fall einer positiven Korrelation gegen 1 gehen, im Fall einer negativen gegen -1. Liegt keine lineare Korrelation vor, wäre der Wert 0. Hier erkennt man schon, dass der Produkt-Moment-Korrelations-Koeffizient (r) nicht nur ein Maß für die Stärke des Zusammenhangs ist, sondern auch die Richtung einer linearen Beziehung zwischen zwei Variablen auf metrischem Datenniveau anzugeben vermag.

Der erste Schritt zur Berechnung des Korrelationskoeffizienten nach Pearson ist die Ermittlung der Kovarianz. Hierzu bestimmt man zunächst die Abweichung eines jeden Punktepaares vom sogenannten bivariaten Schwerpunkt eines Streudiagramms. Teilt man die Summe nun durch die Anzahl der Beobachtungen, so erhält man die durchschnittliche Abweichung vom bivariaten Schwerpunkt und dadurch die sogenannte Kovarianz. Nach der Division durch die Standardabweichungen ergibt sich daraus der Korrelationskoeffizient nach Pearson. Hierbei ist besonders zu beachten, dass hohe Standardabweichungen den Wert des Koeffizienten über die Maßen erhöhen können.

Die Berechnung des Produkt-Moment-Korrelations-Koeffizienten kann einmal über die Menüfolge ***Analysieren → Korrelationen → Bivariat*** geschehen oder wahlweise über ***Analysieren → Deskriptive Statistiken → Kreuztabellen***. Da das Vorgehen über das Kontextmenü Kreuztabellen in Kapitel 10.2 und 10.3 näher beschrieben wird, beschränken wir uns hier auf die Erläuterung der Dialogbox ***Korrelationen*** (Abbildung 10.2).

Hier wählen wir wieder die gewünschten Variablen und den oder die zu berechnenden Korrelationsmaße aus. Außerdem hat man die Wahl zwischen einem einseitigen bzw. zweiseitigen Signifikanztest (mehr zu Signifikanztests in Kapitel 11.2). Unter ***Optionen*** kann man sich weitere statistische Kennzahlen zu den Verteilungen der ausgewählten Variablen anzeigen lassen. Mit einem Klick auf ***Ok*** beenden Sie die Eingabe und starten die Berechnung.

Abbildung 10.2: Dialogbox **„Bivariate Korrelationen“**



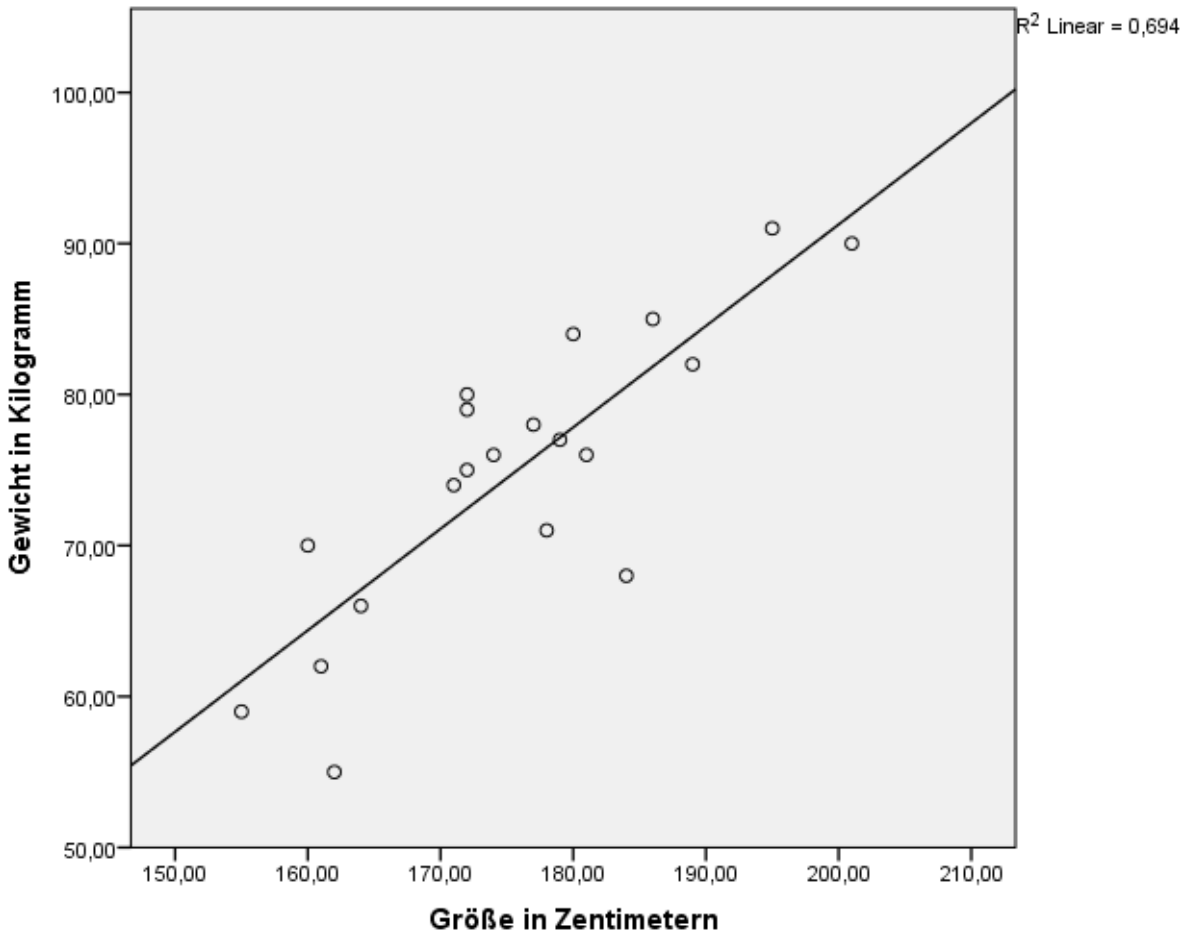
Aufgabe 10.1
Berechnen Sie ein Zusammenhangsmaß für ***„Tage mit Niederschlag (Jahr) über 0,1mm“*** und der ***„Geograph. Breite“*** aus der SPSS-Datendatei ***„Klimastationen Europa“***.

10.1.2 EINFACHE LINEARE REGRESSION

Ziel der Regressionsanalyse ist es, Beziehungen zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen festzustellen. Während der Korrelationskoeffizient nach Pearson nur die Stärke und die Richtung eines Zusammenhanges zwischen zwei metrischen Variablen ermittelt, ermöglicht die Regressionsanalyse neben der quantitativen Beschreibung eines Zusammenhangs auch eine Prognose der Werte der abhängigen Variable.

Zur Verdeutlichung dieses Prinzips dient das folgende Beispiel. Wir greifen abermals auf ein Streudiagramm für die Variablen „Größe in cm“ und „Körpergewicht in Kg“ aus der „Studentenstudie 1“ zurück. Dieses Mal ergänzen wir jedoch eine Regressionsgerade, die den Trend des dargestellten Zusammenhanges symbolisiert (Abbildung 10.3).

Abbildung 10.3: Streudiagramm für ***„Größe“*** und ***„Gewicht“*** mit Trendlinie



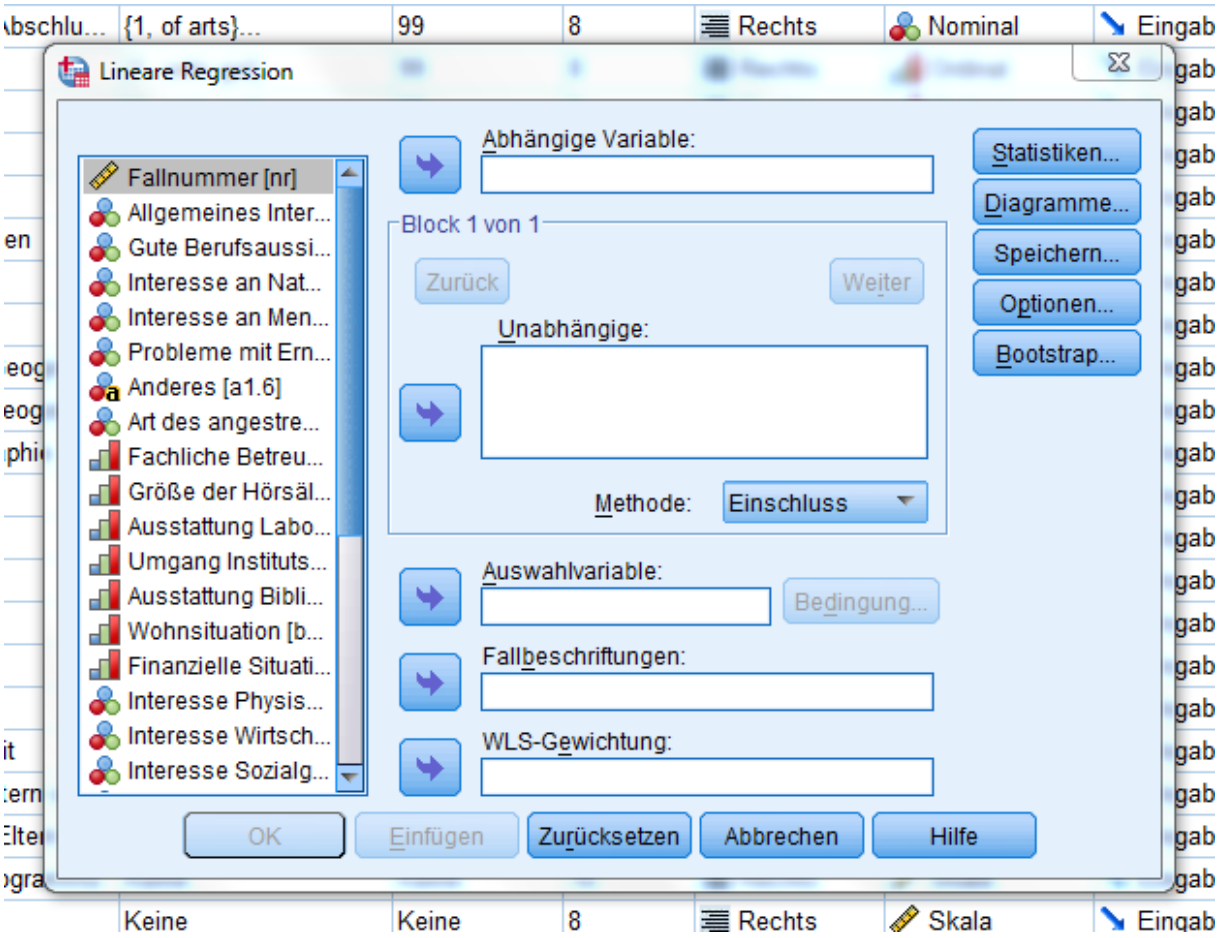
Wir erkennen einen positiven Zusammenhang: Beide Variablen entwickeln sich, von ein paar Ausnahmen abgesehen, in die gleiche Richtung. Die Punktwolke schmiegt sich an die dargestellte Gerade. Diese Trendlinie stellt die zu ermittelnde Regressionsgerade dar. Diese Gerade lässt sich mit folgender Gleichung abbilden:

$$y = b * x + a$$

Hierbei nennt man b den Regressionskoeffizienten und a die Regressionskonstante, also jenen Punkt, an dem die Gerade die y-Achse schneidet. Aufgabe der linearen Regression ist es, diese beiden Parameter abzuschätzen. Dabei gilt jene Gerade als optimal angepasst, bei der die Summe der quadrierten Abstände der Wertepaare von der Geraden minimal ist.

Um eine einfache lineare Regressionsanalyse durchzuführen wählen Sie ***Analysieren → Regression → Linear*** und es öffnet sich folgendes Kontextmenü (Abbildung 10.4).

Abbildung 10.4: Dialogbox ***Regression***



Hier bestimmen wir zunächst die abhängige und die unabhängige Variable. Außerdem kann man unter ***Statistiken***, ***Diagramme*** oder ***Optionen*** ergänzende Ausgaben anfordern sowie unter ***Speichern*** zahlreiche Werte, die im Zusammenhang mit der Regressionsgleichung berechnet werden, dem Datensatz hinzufügen. Wie gewohnt starten wir die Berechnung mit einem Klick auf ***Ok***.

Führt man diese Operation für das oben dargestellte Beispiel aus „Größe“ als unabhängige Variable und „Körpergewicht“ als abhängige Variable aus, erhält man folgendes Ergebnis (Abbildung 10.5).

Aus der untersten Tabelle in Abbildung 10.5 lassen sich der Regressionskoeffizient b (in diesem Fall 0,672) und der Schnittpunkt mit der Ordinate (hier als Konstante bezeichnet und mit einem Wert von -43,125 charakterisiert) ablesen. Daraus ergibt sich die Regressionsgleichung: Körpergewicht = 0,672 * Größe – 43,125. Auch hier wird ein Signifikanzniveau mit angegeben. Im oben dargestellten Fall beträgt die Irrtumswahrscheinlichkeit 3,2 % (s. Abbildung 10.5).

Abbildung 10.5: Ergebnisdarstellung einer Regressionsanalyse

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,833[a]	,694	,677	5,47730

a. Einflußvariablen : (Konstante), Größe in Zentimetern

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	1223,786	1	1223,786	40,792	,000[b]
	Nicht standardisierte Residuen	540,014	18	30,001		
	Gesamt	1763,800	19			

a. Abhängige Variable: Gewicht in Kilogramm

b. Einflußvariablen : (Konstante), Größe in Zentimetern

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		RegressionskoeffizientB	Standardfehler	Beta		
1	(Konstante)	-43,125	18,520		-2,329	,032
	Größe in Zentimetern	,672	,105	,833	6,387	,000

a. Abhängige Variable: Gewicht in Kilogramm

Die mittlere Tabelle gibt unter Quadratsumme den Anteil der Varianz aus, der durch die Regressionsgleichung erklärt wird (Regression) bzw. nicht erklärt werden kann (die nicht standardisierte Residuen). Die Wurzel des Quotienten aus dem erklärten Teil der Varianz und der Gesamtvarianz wird als R-Quadrat in der obersten Tabelle aufgeführt und ist ein Maß für die Qualität der Regressionsgerade (entspricht dem andernorts als B bezeichnetem Bestimmtheitsmaß). Es kann Werte zwischen -1 und 1 erreichen. Bei der hier dargestellten linearen Regression entspricht dieser Wert zugleich dem Quadrat des Korrelationskoeffizienten nach Pearson.

Neben der Regressionsberechnung in der Dialogbox ***Regression*** kann man bei der Bearbeitung eines Streudiagramms im Diagrammeditor diesem unter ***Elemente → Anpassungslinie bei Gesamtwert*** eine Regressionsgerade hinzufügen.

Aufgabe 10.2
Berechnen Sie die Gleichung für die Regressionsgerade zwischen den Variablen „Bevölkerung insgesamt“ und „Ausländeranteil“ aus der SPSS-Datendatei „Bev_Rhein_Main.sav“. Stellen Sie diese auch grafisch dar.

10.1.3 BERECHNUNG VON KORRELATIONEN BEI NICHT NORMALVERTEILTEN DATEN

Korrekterweise sollte man den Korrelationskoeffizienten nach Pearson nur bei normalverteilten metrischen Daten verwenden. Liegen nicht normalverteilte Daten vor (oder wenn dies noch nicht geprüft worden ist), sollte man die Verwendung von Koeffizienten in Erwägung ziehen, die originär für das ordinale Skalenniveau entwickelt wurden. Die im Folgenden beschriebenen Koeffizienten lassen sich problemlos auch für metrische Daten ermitteln. SPSS bietet außerdem verschiedene Operationen zur Bestimmung, ob die Werte einer Variable tatsächlich normalverteilt sind. Diese werden in Kapitel 12.1 näher beschrieben.

10.2 ORDINALES SKALENNIVEAU

Da ordinale Daten einer Ordnung entsprechend einer Rangfolge ermöglichen, lässt sich auch auf diesem Skalenniveau eine Auskunft über die Richtung des ermittelten Zusammenhanges geben. Somit sind Aussagen über positive korrelierende Variablen bzw. negativ korrelierende Variablen zulässig.

10.2.1 SPEARMANS RANGKORRELATIONSKOEFFIZIENT

Prinzipiell basiert die Berechnung des Spearmann‘schen Rangkorrelationskoeffizienten (auch Spearman‘s Rho) auf dem bereits besprochenen Produkt-Moment-Korrelationskoeffizienten nach Pearson. Das Problem, dass man für diese Berechnung eigentlich intervallskalierte Daten benötigt, umgeht der Spearmann‘sche Rangkorrelationskoeffizienten, indem er statt der Variablenwerte die Rangplätze der Fälle verwendet. Die Ausprägungen aus der Urliste werden zuerst Rangzahlen zugeordnet, d.h. die kleinste Ausprägung erhält den Rang 1, die größte Ausprägung den Rang N. Weisen mehrere Erhebungseinheiten die gleiche Ausprägung auf, so werden Durchschnittsränge vergeben, die als Mittel der in Frage kommenden Ränge berechnet werden. Nachdem die Rangplätze vergeben wurden, verwendet man eine für die Verwendung von Rangplätzen angepasste Formel für den Pearson‘schen Produkt-Moment-Korrelationskoeffizienten.

Ein häufiger Kritikpunkt an Spearmann‘s Rho ist die Annahme der äquidistanten Position der einzelnen Ränge. Bei ordinalen Variablen ist diese Annahme allerdings schwer zu begründen. Eine gute Alternative zum Rangkorrelationskoeffizienten nach Spearman ist Kendalls Tau, welches keine Rangkorrelation darstellt, sondern einen paarweisen Vergleich der Maßzahlen durchführt.

10.2.2 KENDALLS TAU

Das Grundprinzip der Berechnung von Kendalls Tau (auch Kendalls τ) ist der paarweise Vergleich aller Fälle. Dabei wird bei jedem Paarvergleich festgestellt, in welche Beziehung die Werte zueinander stehen. Sind beide Werte des ersten Falles größer als beide Werte des zweiten Falles oder sind umgekehrt beide niedriger, so spricht man von einem konkordanten

Paar. Von diskordanten Paaren spricht man, wenn der eine Wert des ersten Falles niedriger ist als der Wert des zweiten Falles auf dieser Variablen und bei der anderen Variablen der umgekehrte Fall vorliegt. Gebundene Paare liegen vor, wenn wenigstens einer der Werte gleich ist. Überwiegen die konkordanten Paare, dann liegt ein positiver Zusammenhang vor, überwiegen die diskordanten, ist der Zusammenhang negativ. In der Regel ist der Wert von Kendalls Tau etwas kleiner als der Wert von Spearman‘s Rho.

10.2.3 BEISPIELAUSWERTUNG

Kendalls Tau und der Rangkorrelationskoeffizient nach Spearman lassen sich analog zum Produkt-Moment-Korrelations-Koeffizient nach Pearson in der Dialogbox ***Bivariate Korrelationen*** unter der Menüfolge ***Analysieren → Korrelationen → Bivariat*** anwählen.

Aufgabe 10.3
Analysieren Sie den Zusammenhang zwischen dem „Anteil männlicher Ausländer (gruppiert)“ und „Gemeindegrößenklassen“ aus der SPSS-Datendatei „Bev_Rhein_Main“ mit Hilfe des Rangkorrelationskoeffizienten nach Spearman und Kendalls Tau. Gibt es Unterschiede bei den Ergebnissen der verschiedenen Zusammenhangmaße? Welches Maß würden Sie präferieren?

Aufgabe 10.4
Berechnen Sie, wie bereits in Aufgabe 10.1, ob es einen Zusammenhang zwischen *„Tage mit Niederschlag (Jahr) über 0,1mm“* und der *„Geograph. Breite“* gibt. Verwenden Sie dieses Mal sowohl den Korrelationskoeffizienten nach Pearson, als auch Kendalls Tau und Spearman. Inwieweit unterscheiden sich die Ergebnisse? Welchen Kennwert würden Sie bevorzugen?

10.3 NOMINALES SKALENNIVEAU

Bei der Berechnung der hier vorgestellten Zusammenhangsmaße auf nominalem Datenniveau vergleicht man die empirisch beobachteten Häufigkeiten mit den erwarteten Häufigkeiten, wenn man von einer vollständigen Unabhängigkeit der Merkmale ausgeht. Die Grundlage dieser Vergleiche stellt der so genannte χ2 (Chi^2)-Wert dar. Dieser ist für sich allerdings nicht für die Beschreibung eines Zusammenhanges geeignet, da er sowohl von der Stichprobengröße als auch von der Zahl der Freiheitsgrade abhängig ist. Jedes der im Folgenden vorgestellten Zusammenhangsmaße erweitert Chi^2 deshalb auf unterschiedliche Art.

Auf nominalem Datenniveau ist bei keinem der im Folgenden beschriebenen Zusammenhangsmaße eine Aussage über die Richtung des Zusammenhanges möglich. Lediglich die Stärke des Zusammenhangs lässt sich hier bemessen.

10.3.1 PHI-KOEFFIZIENT

Der Phi Koeffizient ergibt sich aus χ2 geteilt durch den Stichprobenumfang („n“) und der daraus gezogenen Wurzel und erfüllt hiermit die Aufgabe eines von der Größe der Stichprobe

unabhängigen Zusammenhangsmaßes. Für diese Maßzahl sind insbesondere Tabellen mit zwei Zeilen und zwei Spalten geeignet (so genannte 2x2-Tabellen).

Der Wertebereich des Phi-Koeffizienten liegt zwischen null und eins. Wobei man bei einem Wert von 0 von keinem Zusammenhang sprechen kann und bei einem Wert von 1 von einem perfekten Zusammenhang. Besteht die Kontingenztabelle jedoch aus mehr als zwei Zeilen und zwei Spalten, dann erreicht der Phi-Koeffizient Werte größer als eins und lässt sich nur schwer interpretieren. Hier greift man auf den Kontingenzkoeffizienten zurück.

10.3.2 KONTINGENZKOEFFIZIENT

Bei der Berechnung dieser Maßzahl wird $\chi 2$ nicht durch den Stichprobenumfang n geteilt, sondern durch $\chi 2 + n$. Der Kontingenzkoeffizient nimmt ebenfalls den Wert 0 an, wenn kein Zusammenhang besteht, geht jedoch (im Gegensatz zum Phi-Koeffizienten) niemals über den Wert 1 hinaus. Der Nachteil dieses Zusammenhangsmaßes ist hingegen, dass selbst bei einem perfekten Zusammenhang niemals der Wert von 1 erreicht werden kann. Der maximal erreichbare Wert hängt hierbei von der Zahl der Spalten der Tabelle ab. Hier hilft Cramers V.

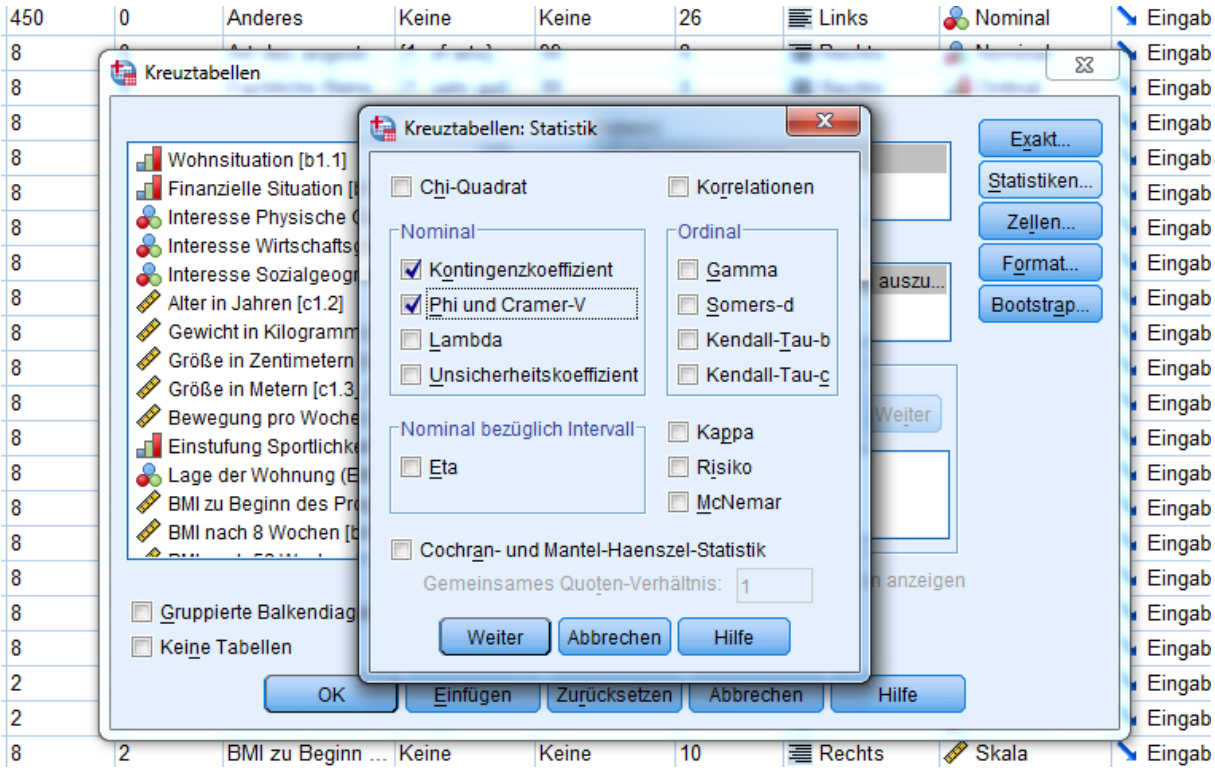
10.3.3 CRAMERS V

Cramers V ist eine weitere Variation, die auch bei größeren Tabellen und unterschiedlichen Zeilen- und Spaltenzahlen Verwendung finden kann und Werte zwischen 0 und 1 ausgibt. Zur Berechnung wird die Quadratwurzel aus dem Quotienten von $\chi 2$ und k gezogen, wobei k für den kleineren Wert der Anzahl der Zeilen oder Spalten steht.

10.3.4 BEISPIELAUSWERTUNG

Zur Berechnung eines Zusammenhangsmaßes auf nominalem Datenniveau wählen wir das bereits bekannte Kontextmenü für Kreuztabellen unter ***Analysieren → Deskriptive Statistiken → Kreuztabellen*** an. Nachdem Sie sich für eine Spalten- sowie eine Zeilenvariable entschieden haben, klicken Sie auf die Schaltfläche Statistik. Es öffnet sich die unter Abbildung 10.6 gezeigte Dialogbox.

Abbildung 10.6: Kontextmenü ***Statistik*** bei der Erstellung von Kreuztabellen



Wie zu erwarten war, finden wir unter der Spalte ***Nominal*** die hier vorgestellten Zusammenhangsmaße. Man wählt also die gewünschten Kennzahlen aus und bestätigt mit einem Klick auf ***Weiter***. Wir gelangen zurück ins Ursprungsmenü und schließen die Berechnung mit ***Ok*** ab.

Aufgabe 10.5
Berechnen sie alle vorgestellten ordinalen Zusammenhangsmaße für die Variable „Wohnsituation“ und „Finanzielle Situation“ aus der „Studentenstudie_1.sav“. Interpretieren sie das Ergebnis. Welches Maß würden Sie bevorzugen? Gibt es einen signifikanten Zusammenhang?

10.4 BISERIALE KORRELATION

Für den speziellen Fall der Berechnung eines Zusammenhangsmaßes für eine Variable auf nominalem und einer auf metrischem Datenniveau verwendet man den Eta-Koeffizienten. Dieser zeigt an, wie sehr sich die Mittelwerte der metrischen Variablen zwischen den verschiedenen Kategorien des nominalen Merkmals unterscheiden. Unterscheiden sich die Werte gar nicht, liegt Eta bei 0. Unterscheiden sie sich stark, geht Eta gegen 1. Dabei ist zu beachten, dass die nominale Variable als unabhängige und die metrische als abhängige Variable verstanden werden sollten. In der SPSS-Ausgabeansicht werden jedoch beide Varianten darstellt. Man sollte hier nur auf das Zusammenhangsmaß für die Annahme des metrischen Merkmals als unabhängige Variable achten.

Die Berechnung dieses Koeffizienten läuft über analog zum nominalen Skalenniveau über ***Analysieren → Deskriptive Statistiken → Kreuztabellen***.

Aufgabe 10.6
Berechnen sie, ob es einen Zusammenhang zwischen ***„Geschlecht“*** und ***„Gewicht in kg“*** in „Studentenstudie_1.sav“ gibt.

10.5 KORRELATIONEN UND KAUSALITÄT

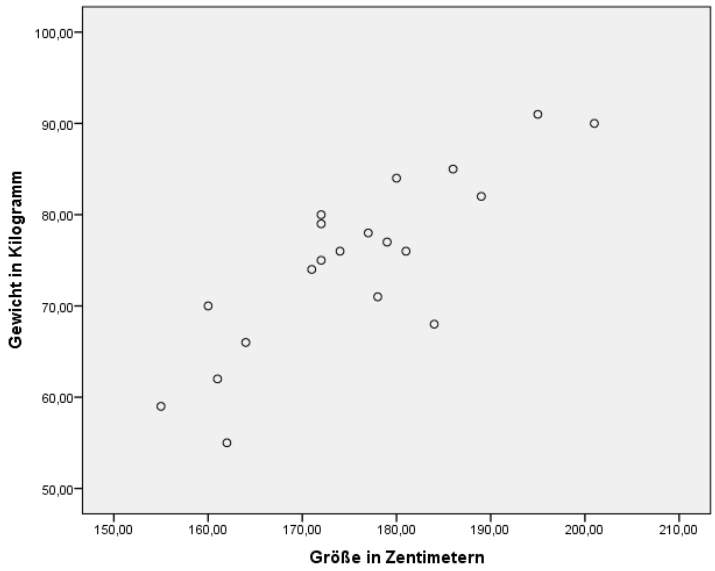
Auch wenn ein berechneter Korrelationskoeffizient besonders hoch ausfällt und die Signifikanz für eine geringe Fehlerwahrscheinlichkeit spricht, kann man nicht zwangsläufig auf einen „inneren“ bzw. „kausalen“ Zusammenhang zwischen zwei Variablen schließen. Mithilfe der Statistik wird zunächst nur eine Korrelation und keine Ursache-Wirkung-Beziehung festgestellt. Liegt eine Korrelation vor, aber es fehlt an einer Ursache-Wirkung-Beziehung, spricht man auch von einer Scheinkorrelation. Neben zufälligen Korrelationen werden auch so genannte *common cause*- oder *mediator variable*-Korrelationen als Scheinkorrelationen bezeichnet.

Aufgabe 10.7
Öffnen Sie die SPSS-Datendatei „Bev_Rhein_Main.sav“ und bestimmen Sie ein Zusammenhangsmaß zwischen dem *„Anteil männlicher Ausländer (gruppiert)“* und der *„Bevölkerung insgesamt“*. Interpretieren Sie das Ergebnis.

Aufgabe 10.8
Öffnen Sie die Datei „Studentenstudie_1.sav“ und bestimmen Sie für die Variablen *„Lage der Wohnung (Eltern oder eigene Wohnung)“* und *„Geschlecht“* ein passendes Zusammenhangsmaß und interpretieren Sie das Ergebnis.

Aufgabe 10.9
Öffnen Sie die SPSS-Datendatei „Klimastationen Europa.sav“ und bestimmen Sie für die Variablen *„Land“* und *„Höhe über NN in m“* sowie *„Mittl. Temperatur (Jahr) in °C“* und *„Tage mit Niederschlag (Jahr) über 0,1mm“* ein passendes Zusammenhangsmaß und interpretieren Sie das Ergebnis.

Aufgabe 10.10
Die folgende Abbildung beruht auf Daten, die in einem Tutorium mit 20 studentischen Teilnehmer_Innen in Osnabrück erhoben worden sind.
1. Benennen Sie die untersuchten Merkmale, Merkmalsträger und die Grundgesamtheit.
2. Auf welchen Skalen wurden die beiden Merkmale gemessen?
3. Worin besteht die Kernaussage der Grafik?



4. Welche Form eines Zusammenhanges können Sie aus den verschiedenen Grafiken ablesen? Erläutern sie den grafischen Analysebefund.
5. Welches Zusammenhangsmaß würden Sie zur Beschreibung der beiden Diagramme verwenden?

Aufgabe 10.11

Untersuchen Sie für die „„Studentenstudie_1.sav“ die folgende Hypothese: Für männliche Studenten sind die „Bewegung pro Woche“ und die „Einstufung der eigenen Sportlichkeit“ zwei voneinander unabhängige Merkmale.

Aufgabe 10.12

Ein Lehrer ermittelt einen Korrelationskoeffizienten nach Pearson von 0,83 für den Zusammenhang von der Qualität der Schreibschrift und der Schuhgröße von Schülern im Alter zwischen 7 und 14 Jahren. Was schließen Sie daraus?

C PRÜF– UND TESTSTATISTIK

11 EINFÜHRUNG IN DIE PRÜF- UND TESTSTATISTIK

Was bedeutet Prüfstatistik?

Was ist ein Signifikanzniveau?

Welche Fehlerformen habe ich zu beachten?

Wann kann ich welchen statistischen Test anwenden?

Die bereits beschriebenen Methoden der deskriptiven Statistik befassen sich mit der Beschreibung von Daten, meist einer Stichprobe. Oftmals sind deren Charakteristika jedoch nicht das, was wirklich interessiert. Man will eigentlich mehr über die Verteilung innerhalb der Grundgesamtheit erfahren, aus der die Stichprobe gezogen wurde. Man will die Daten verallgemeinern. Wir wollen bspw. von der erhobenen Stichprobe unter den Wahlberechtigten auf die Wahlentscheidungen aller Wahlberechtigten in Deutschland schließen oder von den erhobenen Daten verschiedener Klimastationen auf einen Zusammenhang zwischen Temperatur und Verdunstung weltweit. Hier helfen die Methoden der Prüfstatistik (auch Teststatistik oder Inferenzstatistik). Diese umfassen verschiedene Verfahren, die anhand einer Stichprobe (Teilmenge der Grundgesamtheit) Rückschlüsse auf oder Folgerungen für die Grundgesamtheit ermöglichen.

11.1 STATISTISCHE HYPOTHESEN

Die Prüfstatistik arbeitet mit Hypothesen, die mit einer gewissen Wahrscheinlichkeit angenommen oder verworfen werden können. Die Einräumung einer Fehlerwahrscheinlichkeit ist deswegen von großer Bedeutung, da man sich die Formulierung von Hypothesen auch sparen könnte, wenn man sich sicher wäre. Da wir aber im Normalfall keine vollständigen Informationen über die Parameter einer Grundgesamtheit haben, können wir auch niemals von vollständiger Sicherheit sprechen.

Statistische Hypothesen werden stets als Hypothesenpaar formuliert. Die sogenannte Nullhypothese (H_0) steht hierbei der Alternativhypothese (H_A) gegenüber und es ist die Aufgabe des Signifikanztests, diese Hypothesen gegeneinander abzuwiegen. Die Nullhypothese behauptet (i. d. R.), dass es zwischen den Gruppen oder Variablen keine Unterschiede oder Zusammenhänge gibt. Die Alternativhypothese spricht dagegen für einen Zusammenhang, Unterschied oder eine Veränderung. Damit beinhaltet sie zumeist das, was den Forscher interessiert und was er eigentlich nachweisen möchte.

11.2 SIGNIFIKANZ UND FEHLERFORMEN

Da wir aufgrund der unvollständigen Informationen über die Grundgesamtheit nicht sicher wissen können, was „wahr" ist, sind die getroffenen Aussagen jedoch unsicher – es handelt sich um Wahrscheinlichkeitsaussagen. Aus diesem Grund kann die Annahme der Alternativhypothese entweder richtig oder falsch sein und umgekehrt auch die der Nullhypothese. Hier können zwei Fehler auftreten:

Ein **α-Fehler** (Fehler erster Art) liegt vor, wenn H_A angenommen wird (wenn wir an einen Unterschied in der Population glauben), obwohl es ihn in der Grundgesamtheit nicht gibt.

Ein **β-Fehler** (Fehler zweiter Art) liegt vor, wenn H_0 angenommen wird (wenn wir an keinen Effekt in der Population glauben), obwohl sie in der Grundgesamtheit nicht gilt.

Die verschiedenen Möglichkeiten einer korrekten bzw. falschen Entscheidung lassen sich gut in einer Vierfeldertafel verdeutlichen (Abbildung 11.1):

Abbildung 11.1: Vierfeldertafel zur Verdeutlichung von α-Fehler und β-Fehler

		In „Wirklichkeit" gilt:	
		Nullhypothese	**Alternativhypothese**
Testentscheidung	**Nullhypothese**	Richtige Entscheidung	β-Fehler
	Alternativhypothese	α-Fehler	Richtige Entscheidung

Den α-Fehler kann man über die Festlegung des Signifikanzniveaus noch relativ einfach beeinflussen. Diese ist eine willkürlich gesetzte Grenze, die angibt, wie hoch die Wahrscheinlichkeit dafür höchstens sein darf, sich zu irren, wenn man die Alternativhypothese annimmt. Je mehr man sich davor schützen will, die Alternativhypothese fälschlicherweise anzunehmen, desto strenger muss das Signifikanzniveau sein. Man könnte die Grenze beispielsweise von 5 % auf 1 % oder sogar 0,01 % anheben. Der β-Fehler ist zum einen von der Höhe des Signifikanzniveaus abhängig (je höher das Signifikanzniveau, desto höher die Wahrscheinlichkeit die Alternativhypothese fälschlicherweise abzulehnen), zum anderen jedoch auch von der „Macht" des gewählten statistischen Tests. Unter dieser „Macht" versteht man das Vermögen eines statistischen Tests, eine in der Grundgesamtheit gültige Alternativhypothese zu erkennen.

Das Ergebnis eines statistischen Tests ist der so genannte „p-Wert". Dieser wird mit dem vorher bestimmten Signifikanzniveau verglichen. Ist der aus den vorliegenden Daten ermittelte p-Wert kleiner oder gleich dem Signifikanzniveau, entscheidet man sich zugunsten der

Alternativhypothese. In dem Fall, dass der p-Wert größer ist als das Signifikanzniveau, entscheidet man sich dagegen für die Nullhypothese. Mit anderen Worten: Wenn ein statistischer Test (bei gegebenem Signifikanzniveau) ein signifikantes Ergebnis liefert, heißt das, der Unterschied zwischen den beiden Gruppen (oder Variablen) wird wahrscheinlich dadurch verursacht, dass es tatsächlich Unterschiede in der Grundgesamtheit gibt.

11.3 DIE WAHL EINES PASSENDEN STATISTISCHEN VERFAHRENS

Für die Wahl des passenden statistischen Tests sind neben der eigentlichen Forschungsfrage folgende Faktoren von Bedeutung:

Handelt es sich um unabhängige oder um abhängige Stichproben?
Möchte man zwei oder mehrere Stichproben miteinander vergleichen?
Welches Skalenniveau haben die zu untersuchenden Merkmale?
Sind die Daten normalverteilt?
Verfügen die Daten über annähernd ähnliche Varianzen?

Stichproben gelten dann als unabhängig, wenn sie unterschiedliche Fälle (Untersuchungselemente, Personen) enthalten. Man vergleicht also zwei verschiedene Stichproben miteinander (Männer und Frauen). Abhängige Stichproben liegen vor, wenn man die gleichen Fälle mehrmals befragt. Beispielsweise könnte man aufeinanderfolgende Messungen zu Beginn, nach zwei Wochen und/oder nach vier Wochen durchführen. Die verschiedenen Skalenniveaus werden in Kapitel 2.3 behandelt. Eine Prüfung auf Normalverteilung bzw. Varianzgleichheit liefert der Kolmogorov-Smirnov- bzw. Levene-Test, die beide in Kapitel 12 weiter erläutert werden.

Aufgabe 11.1
In vielen Rechtsstaaten gilt die Unschuldsvermutung „Im Zweifel für den Angeklagten“. Ein Angeklagter ist so lange unschuldig, bis seine Schuld durch genügend Beweise nachgewiesen werden kann. Mit welchem Prinzip der Prüfstatistik kann man die Unschuldsvermutung vergleichen?

Aufgabe 11.2
Sie lesen in einer statistischen Publikation den Ergebnissatz: „Also trifft die Nullhypothese zu“. Wie hätten Sie den Satz formuliert?

12 VORAUSSETZUNGSTESTS UND F-TEST

Wie prüft man die beschriebenen Voraussetzungen einer vorliegenden Normalverteilung und einer relativen Varianzgleichheit?

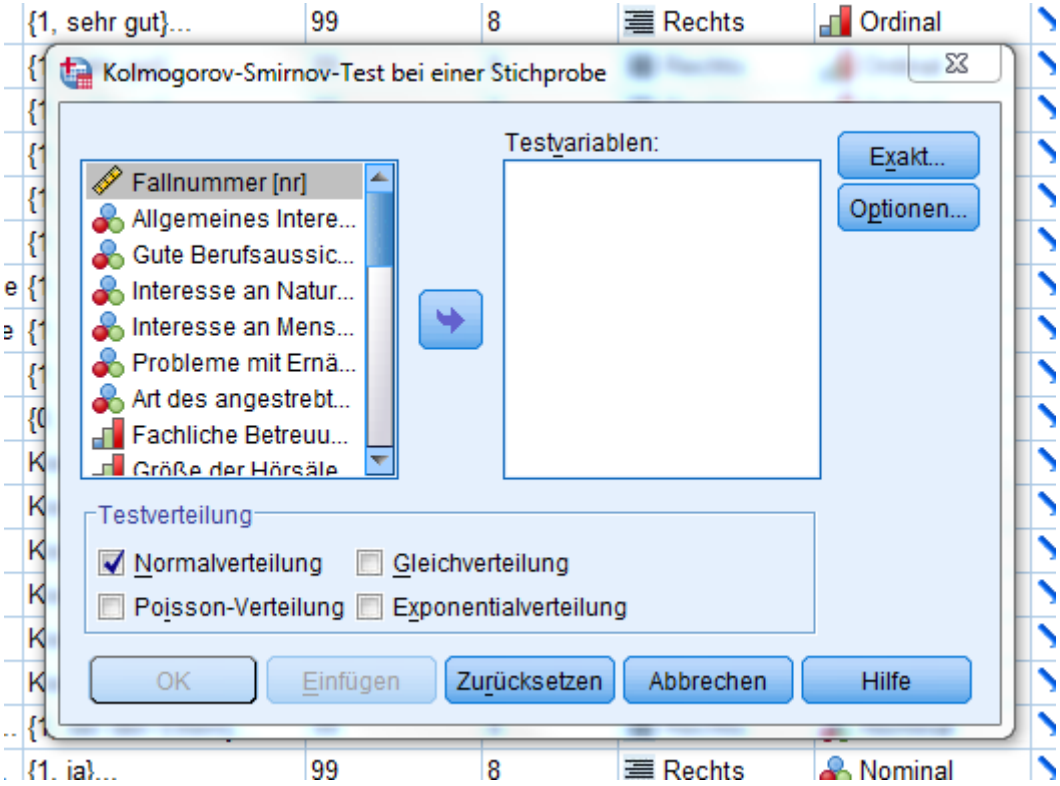
Wie setzt man einen F-Test in SPSS um?

Da das Vorliegen einer Normalverteilung bzw. relativ homogener Varianzen die Voraussetzungen für die Anwendung verschiedener statistischer Tests darstellen, sollen die Verfahren des Kolmogorov-Smirnov- bzw. des Levene-Tests an dieser Stelle erläutert werden. Außerdem deckt der Levene-Test den Anwendungsbereich des klassischen F-Tests ab.

12.1 KOLMOGOROV-SMIRNOV-TEST AUF NORMALVERTEILUNG

Der Kolmogorov-Smirnov-Test ist ein sogenannter nichtparametrischer Test, der Abweichungen in der Verteilung der Daten von einer Normalverteilung entdecken soll. Er prüft die Nullhypothese, dass die Stichprobe einer normalverteilten Grundgesamtheit entstammt. Die Alternativhypothese besagt, dass die Stichprobe nicht einer normalverteilten Grundgesamtheit entstammt.

Abbildung 12.1: Dialogbox **Kolmogorov-Smirnov-Test bei einer Stichprobe**



Hierzu wählen wir ***Analysieren → Nichtparametrische Tests → Alte Dialogfelder → K-S bei einer Stichprobe*** und es öffnet sich folgendes Kontextmenü (Abbildung 12.1). Man erkennt,

dass dieser Anpassungstext auch andere Verteilungen zugrunde legen kann als die Normalverteilung.

Da der Kolmogorov-Smirnov-Test hier bereits voreingestellt ist, bringen wir einfach die gewünschte Variable in das Variablenfeld und klicken auf ***Ok***.

Zur Veranschaulichung einer Interpretation für den Kolmogorov-Smirnov-Test betrachten wir die Ergebnisausgabe für die Betrachtung der Variable „Größe in cm“ aus der Datei Studentenstudie_1.sav (siehe Abbildung 12.2).

Abbildung 12.2: Beispiel-Ausgabe **Kolmogorov-Smirnov-Anpassungstest**

Kolmogorov-Smirnov-Anpassungstest

		Größe in Zentimetern
N		20
Parameter der Normalverteilung[a,b]	Mittelwert	175,6500
	Standardabweichung	11,94406
Extremste Differenzen	Absolut	,099
	Positiv	,085
	Negativ	-,099
Kolmogorov-Smirnov-Z		,441
Asymptotische Signifikanz (2-seitig)		,990

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

Wie bereits erläutert, besagt die Nullhypothese des Kolmogorov-Smirnov-Tests, dass die untersuchte Variable in der Grundgesamtheit normalverteilt ist, weshalb ein nicht signifikantes Ergebnis in diesem Fall für eine Normalverteilung sprechen würde. Der ***p-Wert*** wird hier unter ***Asymptotischer Signifikanz*** aufgeführt. Der ***p-Wert*** für das oben angeführte Beispiel beträgt 0,99, sodass die Nullhypothese angenommen und bis zu einem Signifikanzniveau von 99 % beibehalten werden muss. Es kann also von einer Normalverteilung in der Grundgesamtheit ausgegangen werden.

Aufgabe 12.1
Können die Merkmale „Mittl. Niederschlag (Jahr) in mm“ und „Mittl. Temperatur (Jahr) in °C“ aus der SPSS-Datendatei „Klimastationen Europa.sav“ als normalverteilt angenommen werden?

12.2 LEVENE-TEST AUF VARIANZGLEICHHEIT (F-TEST)

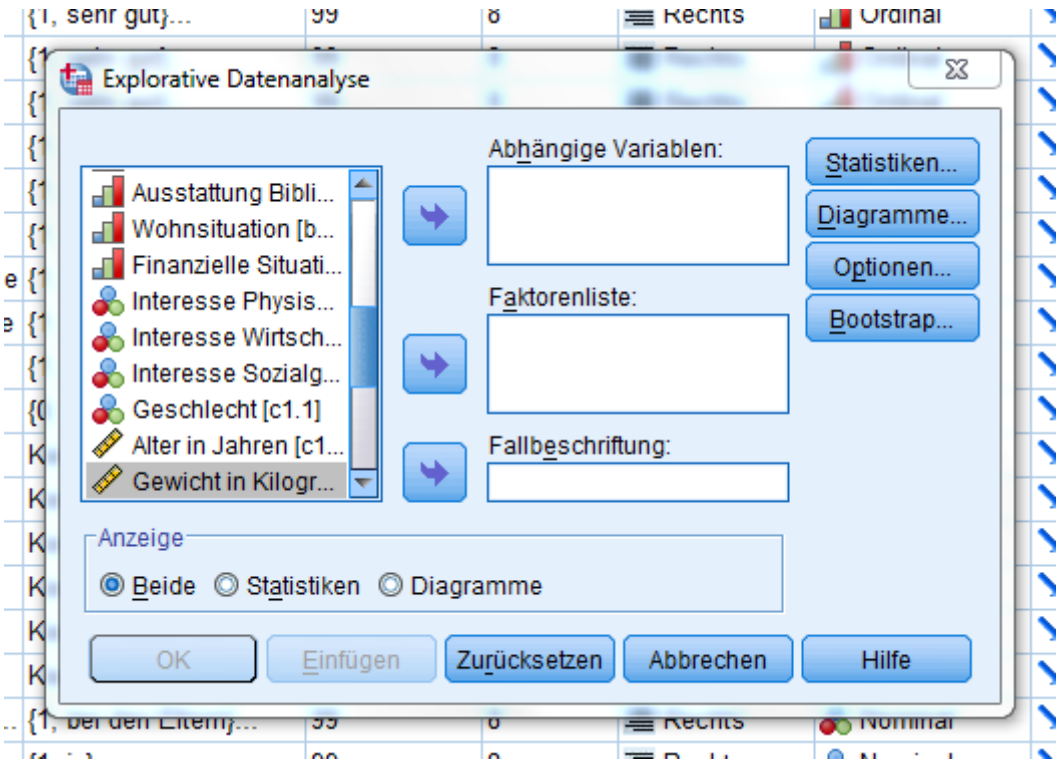
Der Levene-Test untersucht die Homogenität (Gleichheit) der Varianzen zwischen zwei Gruppen. Damit bietet er dem Nutzer zwei wichtige Funktionen: Zum einen ist die Varianzgleichheit eine wichtige Voraussetzung für verschiedene weiterführende statistische Tests (wie etwa den t-Test), zum anderen lässt sich hiermit überprüfen, ob zwei Gruppen von Pro-

banden aus derselben Grundgesamtheit stammen (da der Levene-Test eine Variante des F-Tests darstellt). Damit erfüllt der Levene-Test auch die Rolle des klassischen F-Tests.

Bei der Umsetzung eines Levene-Tests wird für jeden einzelnen Fall die absolute Abweichung vom Gruppenmittelwert gebildet. Danach wird eine Varianzanalyse der Varianz dieser Differenzen durchgeführt. Sollte die Nullhypothese zutreffen, dürfte sich die Streuung innerhalb der Gruppen nicht signifikant von der zwischen den Gruppen unterscheiden. Die Alternativhypothese zielt dann entsprechend auf einen signifikanten Unterschied der Varianzen gibt. Wir brauchen also zwei Variablen: zum einen die auf ihre Varianz hin zu untersuchende metrische Variable, zum anderen eine nominale oder ordinale Gruppierungsvariable.

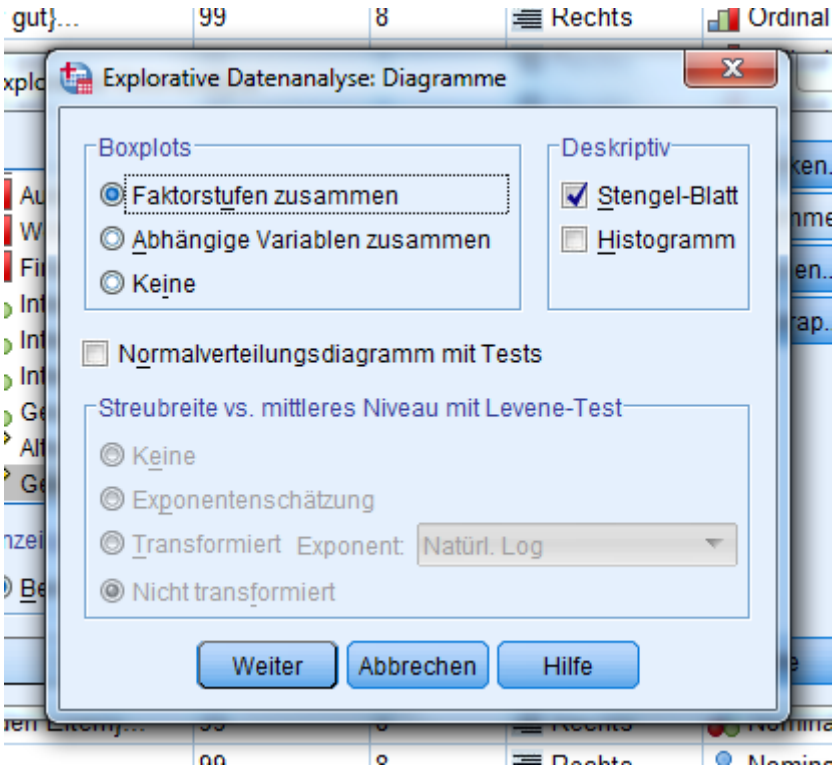
Um einen Levene-Test in SPSS durchzuführen, klicken wir auf ***Analysieren → Deskriptive Statistiken → Explorative Datenanalyse***. Es öffnet sich das bereits bekannte Kontextmenü (Abbildung 12.3).

Abbildung 12.3: Dialogbox ***Explorative Datenanalyse***



Wir wählen die abhängige Variable für das Feld ***Abhängige Variable*** (metrisches Datenniveau) und die Gruppierungsvariable für das Feld unter ***Faktorenliste*** (maximal ordinales Datenniveau). Nun klicken wir auf die Schaltfläche Diagramme und es öffnet sich folgendes Kontextmenü (siehe Abbildung 12.4).

Abbildung 12.4: Dialogbox **Explorative Datenanalyse: Diagramme**



In diesem neuen Untermenü setzt man unter ***Streubreite vs. mittleres Niveau mit Levene-Test*** einen Haken bei ***Nicht transformiert***. Dann wählen wir ***Weiter*** und ***Ok***. Abbildung 12.5 zeigt eine Beispielausgabe für die Variablen „Geschlecht“ und „Gewicht in kg“ aus der „Studentenstudie 1“.

Abbildung 12.5: Beispielsausgabe eines Levene-Tests

Deskriptive Statistik

	Geschlecht			Statistik	Standardfehler
Gewicht in Kilogramm	männlich	Mittelwert		81,6000	1,83303
		95% Konfidenzintervall des Mittelwerts	Untergrenze	77,4534	
			Obergrenze	85,7466	
		5% getrimmtes Mittel		81,4444	
		Median		80,5000	
		Varianz		33,600	
		Standardabweichung		5,79655	
		Minimum		75,00	
		Maximum		91,00	
		Spannweite		16,00	
		Interquartilbereich		10,25	
		Schiefe		,552	,687
		Kurtosis		-1,047	1,334

	weiblich	Mittelwert		68,2000	2,50244
		95% Konfidenzintervall des Mittelwerts	Untergrenze	62,5391	
			Obergrenze	73,8609	
		5% getrimmtes Mittel		68,2778	
		Median		69,0000	
		Varianz		62,622	
		Standardabweichung		7,91342	
		Minimum		55,00	
		Maximum		80,00	
		Spannweite		25,00	
		Interquartilbereich		13,50	
		Schiefe		-,219	,687
		Kurtosis		-,688	1,334

Test auf Homogenität der Varianz

		Levene-Statistik	df1	df2	Signifikanz
Gewicht in Kilogramm	Basiert auf dem Mittelwert	,702	1	18	,413
	Basiert auf dem Median	,654	1	18	,429
	Basierend auf dem Median und mit angepaßten df	,654	1	15,767	,431
	Basiert auf dem getrimmten Mittel	,701	1	18	,413

Der klassische Levene-Test geht von der Abweichung der einzelnen Fälle vom arithmetischen Mittel aus. SPSS bietet jedoch auch weitere Optionen wie die Differenz vom Median oder die Differenz vom getrimmten arithmetischen Mittel. An den hohen Werten in der Spalte ***„Signifikanz“*** lässt sich ablesen, dass die Wahrscheinlichkeit, dass beide Gruppen aus derselben Grundgesamtheit stammen, sehr hoch ist. Wir müssen also die Nullhypothese beibehalten und gehen davon aus, dass die beiden Untersuchungsgruppen aus derselben Grundgesamtheit stammen. Je nach vorher festgelegtem Signifikanzniveau würde man erst bei einer Signifikanz von 0,05 und niedriger die Nullhypothese, der zufolge beide Gruppen eine ähnliche Varianz aufweisen, ablehnen.

Aufgabe 12.2
Führen Sie den Levene-Test für die Variablen „Ausländeranteil gruppiert 2“ und „Bevölkerung insgesamt“ aus der SPSS-Datendatei „Bev_Rhein_Main.sav“ durch und interpretieren Sie das Ergebnis.

13 t-TESTS

Wie kann man berechnen, ob die Mittelwertunterschiede zweier Stichproben signifikant sind?
Wie kann man ermitteln, ob sich die Mittelwerte einer Stichprobe im Zeitverlauf statistisch signifikant verändern?

13.1 t-TEST FÜR UNABHÄNGIGE STICHPROBEN

Mit dem t-Test für Mittelwertdifferenzen werden die Unterschiede der Mittelwerte zweier Gruppen auf ihre Signifikanz hin untersucht. Die Voraussetzungen für ein solches Verfahren sind:

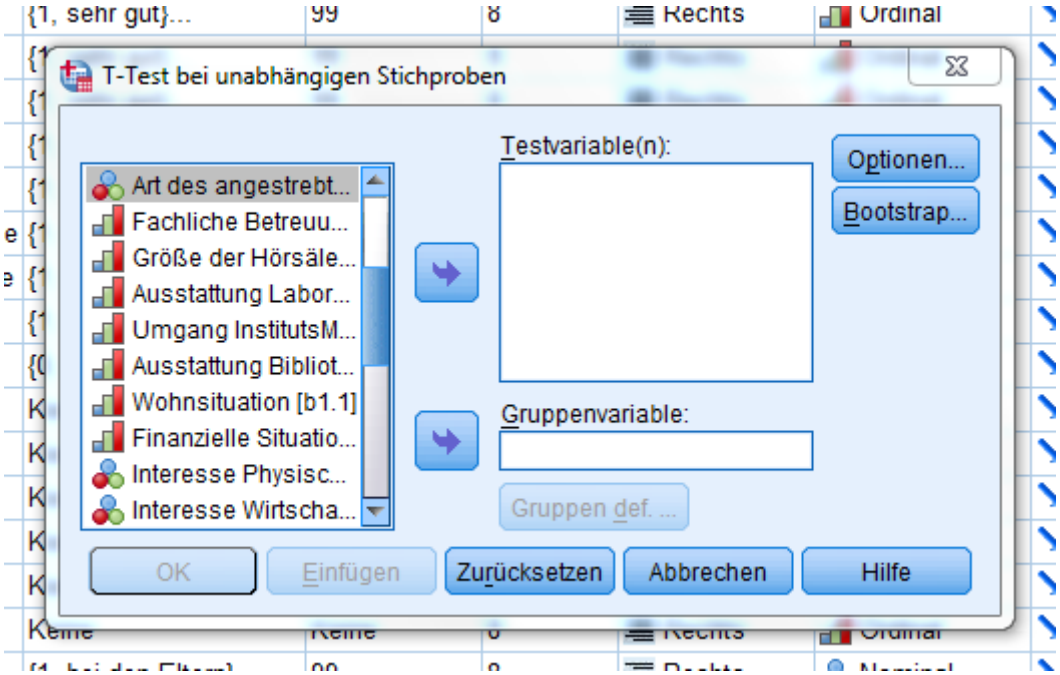
Metrisches Datenniveau
Normalverteilung
Varianzhomogenität
unabhängige Stichproben

Wollen wir Berechnungen ausführen, die das arithmetische Mittel beinhalten, benötigen wir zwangsweise Daten auf metrischem Datenniveau. Eine Stichprobe gilt dann als unabhängig, wenn die Vergleichsgruppen aus unterschiedlichen Fällen bestehen. Die Prüfung auf Normalverteilung und Varianzhomogenität muss in beiden Gruppen separat durchgeführt werden, verläuft aber ansonsten wie bereits in Kapitel 12 beschrieben.

Betrachten wir dazu ein Beispiel aus der Datei „Studentenstudie_1.sav“: Es sollen zwei unabhängige Stichproben hinsichtlich der abhängigen Variable „Körpergewicht in kg“ verglichen werden. Als unabhängige Variable (Gruppierungsvariable) dient die „Art des angestrebten Abschlusses“. Die Nullhypothese lautet hierbei: Die Mittelwerte der zwei Gruppen unterscheiden sich nicht. Die Alternativhypothese wäre: Die Mittelwerte unterscheiden sich zwischen den zwei Gruppen.

Dieser t-Test lässt sich über ***Analysieren > Mittelwerte vergleichen > t-Test bei unabhängigen Stichproben*** berechnen (siehe Abbildung 13.1).

Abbildung 13.1: Dialogbox **t-Test bei unabhängigen Stichproben**



Zunächst wählen wir die Testvariable ***„Körpergewicht in kg“***. Außerdem müssen wir eine Gruppierungsvariable benennen, die beide Stichproben voneinander trennt. Wir wählen die ***„Art des angestrebten Abschlusses“***. Da wir die Gruppen nicht im Vorfeld voneinander getrennt haben, müssen wir nun noch die beiden Stichproben definieren und klicken auf ***Gruppen definieren*** und geben im folgenden Menü (siehe Abbildung 13.2) den Wert 1 für die Gruppe 1 und den Wert 2 für Gruppe 2 an.

Abbildung 13.2: Dialogbox ***Gruppen definieren***

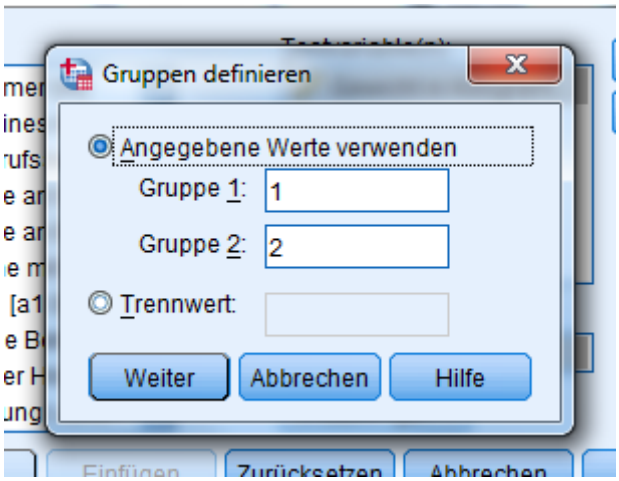


Abbildung 13.3 veranschaulicht ein Beispieloutput zur Interpretation der Ergebnisse des t-Tests.

Abbildung 13.3: Beispiel-Output eines t-Tests für unabhängige Stichproben

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	95% Konfidenzintervall der Differenz	
									Untere	Obere
Gewicht in Kilogramm	Varianzen sind gleich	,043	,838	1,157	18	,263	5,04167	4,35915	-4,11657	14,19991
	Varianzen sind nicht gleich			1,142	14,486	,272	5,04167	4,41586	-4,39972	14,48305

Die Tabelle ***Gruppenstatistiken*** liefert zunächst einige deskriptive Statistiken. Das eigentliche Ergebnis des t-Tests beinhaltet jedoch die Tabelle ***Test bei unabhängigen Stichproben***. Wie unter Kapitel 12.2 bereits beschrieben, benötigen wir für den t-Test eine Varianzhomogenität in der Grundgesamtheit. Nur wenn hier die Nullhypothese angenommen werden kann, sind auch die Ergebnisse des t-Tests verwendbar. Zur Annahme der Nullhypothese muss der ***p-Wert*** in der Spalte ***Signifikanz*** größer sein als 0,05 (bzw. 5 %). Das ist hier mit einem Wert von 0,838 der Fall ist. Deshalb werfen wir einen Blick auf die Zeile ***Varianzen sind gleich*** und ermitteln das Ergebnis des t-Tests. Wieder gibt die Zeile ***Signifikanz („Sig (2-seitig)“)*** den zugehörigen p-Wert aus, der anzeigt, ob die Stichprobendifferenz zufällig ist oder nicht. Beträgt der p-Wert hier höchstens 5 %, ist die Stichprobendifferenz nicht per Zufall entstanden (sondern resultiert z. B. daraus, dass die beiden Gruppen unterschiedlichen Grundgesamtheiten angehören, die sich hinsichtlich der Untersuchungsvariable unterscheiden); daraus resultiert die Annahme von H_A. Im vorliegenden Beispiel liegt die Signifikanz jedoch bei 0,263 und wir entscheiden uns für die Beibehaltung der Nullhypothese: Die Mittelwerte zwischen den zwei Gruppen unterscheiden sich nicht signifikant.

Aufgabe 13.1
Sie haben vor der eigentlichen Erhebung eine Hypothese über die „Studentenstudie 1“ aufgestellt: Sie vermuten, dass es zwischen den Probanden, die an gute Berufsaussichten nach dem Studium glauben und denen, die das nicht tun, ein Unterschied hinsichtlich des Geschlechts besteht. Formulieren Sie die entsprechenden Hypothesen H_0 und H_A, prüfen Sie auf Normalverteilung und Varianzhomogenität und testen Sie auf dem Signifikanzniveau von 5 %.

13.2 t-TEST FÜR ABHÄNGIGE STICHPROBEN

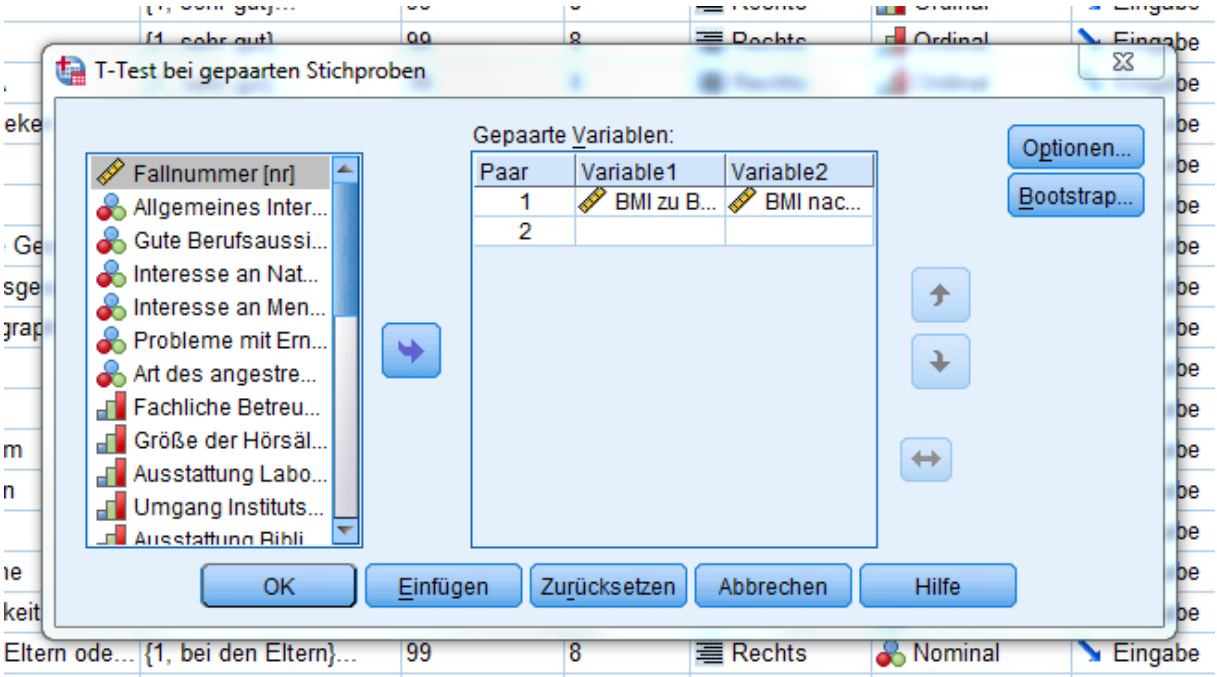
Bestehen die Vergleichsgruppen aus denselben Fällen, für die eine Variable mehrfach gemessen wurde, wendet man den t-Test für abhängige Stichproben an. Dieses Verfahren findet vor allem bei einer Betrachtung im Zeitverlauf Anwendung. Dabei werden die gleichen Probanden zu verschiedenen aufeinanderfolgenden Zeitpunkten befragt (abhängige Stichprobe), um beispielsweise einen „Vorher/Nachher-Vergleich“ zu ermöglichen. Die Voraussetzungen für ein solches Verfahren sind:

Metrisches Datenniveau
Normalverteilung
Varianzhomogenität
abhängige Stichproben

Das Prinzip hinter dieser Berechnung ist die Bildung von Differenzen zwischen den verschiedenen Messwerten der einzelnen Fälle. Nehmen wir zur Verdeutlichung abermals ein Beispiel aus der „Studentenstudie 1“ zur Hand. Die in der „Studentenstudie 1“ beteiligen Probanden haben nach dem Ausfüllen des Fragebogens ein Sport- und Ernährungsprogramm absolviert. Dabei wurde deren BMI-Wert (Body-Maß-Index) zu Beginn des Programmes und nach acht Wochen erhoben. Somit haben wir zwei abhängige Stichproben, die man im Rahmen eines „Vorher/Nachher-Vergleiches“ analysieren könnte. Wenn das Sportprogramm tatsächlich Auswirkungen hatte, muss die durchschnittliche Differenz zwischen den beiden Gewichtsmessungen über alle Personen hinweg deutlich erkennbar sein. Wir würden bei einem signifikanten Unterschied die Alternativhypothese annehmen. Die Nullhypothese spricht für geringe Gewichtsdifferenzen der Probanden und damit für ein nicht wirkendes Sportprogramm.

Wir berechnen den t-Test über ***Analysieren > Mittelwerte vergleichen > t-Test bei verbundenen Stichproben***. Es öffnet sich das folgende Kontextmenü (Abbildung 13.4).

Abbildung 13.4: Dialogbox **t-Test bei gepaarten Stichproben**



Übertragen Sie nun durch Anklicken aus der Quellvariablenliste die erste der zu untersuchenden Variablen „BMI zu Beginn des Programms“ und die zweite Variable „BMI nach 8 Wochen“ dann entsprechend unter ***Variable 2***. Durch einen Klick auf ***Ok*** ergibt sich folgende Ausgabe (Abbildung 13.5).

Abbildung 13.5: Output-Beispiel eines t-Tests für abhängige Stichproben

Statistik bei gepaarten Stichproben

		Mittelwert	N	Standardabweichung	Standardfehler des Mittelwertes
Paaren 1	BMI zu Beginn des Programms	24,2551	20	1,89823	,42446
	BMI nach 8 Wochen	24,0490	20	2,08508	,46624

Korrelationen bei gepaarten Stichproben

		N	Korrelation	Signifikanz
Paaren 1	BMI zu Beginn des Programms & BMI nach 8 Wochen	20	,976	,000

Test bei gepaarten Stichproben

		Gepaarte Differenzen					T	df	Sig. (2-seitig)
		Mittelwert	Standardabweichung	Standardfehler des Mittelwertes	95% Konfidenzintervall der Differenz Untere	Obere			
Paaren 1	BMI zu Beginn des Programms - BMI nach 8 Wochen	,20614	,47733	,10673	-,01726	,42953	1,931	19	,068

Wir erkennen in der ersten Tabelle in Abbildung 13.5, dass der Mittelwert des Gewichtes vorher größer ist als nach dem Sportprogramm (ein erster Hinweis auf einen möglichen Mittelwertunterschied). Die zweite Tabelle gibt die Korrelation zwischen den beiden Variablen an. Je höher diese Korrelation ist, desto eher müssen wir von einem fehlenden Effekt des Sportprogramms ausgehen. Die untere Tabelle liefert schließlich die aussagekräftigsten Ergebnisse und den ***p-Wert*** für den t-Test. Hier wird das arithmetische Mittel der Differenzen vor und nach dem Sportprogramm ausgegeben. Der p-Wert (***Sig (2-seitig)***) liegt mit 0,68 weit über der gebräuchlichen 5 % Grenze, weswegen wir H_0 beibehalten müssen: Das Sportprogramm hat keinen signifikanten Effekt auf den BMI der Probanden.

Aufgabe 13.2
Untersuchen Sie, ob das Sportprogramm nach 52 Wochen einen Effekt hat.

14 WEITERE STATISTISCHE TESTS

Wie kann man ermitteln, ob eine Beziehung zwischen zwei nominalen (ordinalen) Variablen vorliegt?

Wie kann man den „wahren“ Mittelwert der Grundgesamtheit bestimmen?

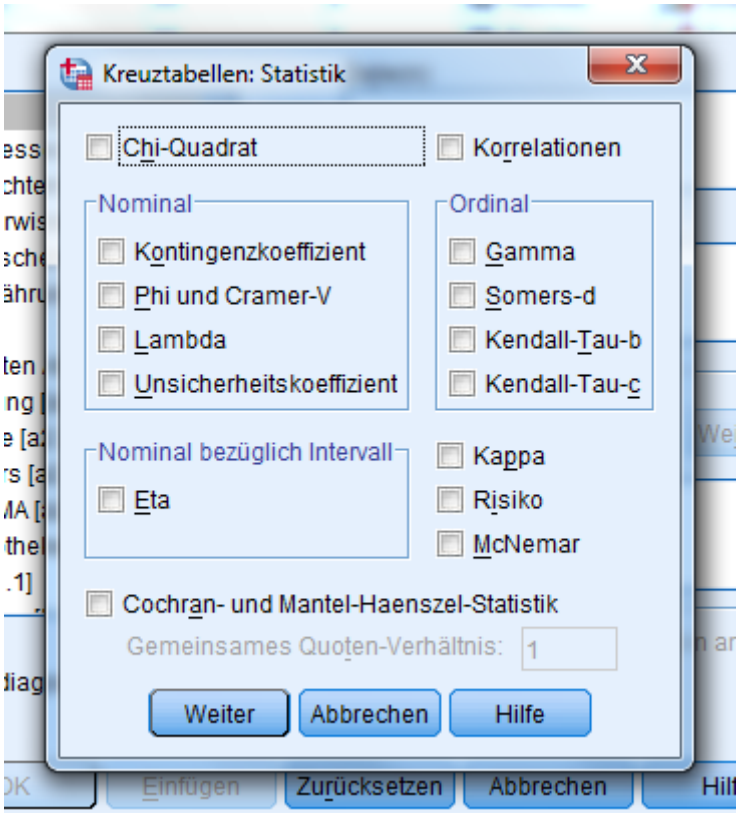
14.1 VIERFELDER- ODER CHI-QUADRAT-UNABHÄNGIGKEITSTEST

Im Rahmen des Chi-Quadrat-Tests wird die empirisch erhobene Verteilung mit einer erwarteten Verteilung verglichen. Die erwartete Verteilung entspricht dabei einer Verteilung, bei der man von einem fehlenden Zusammenhang zwischen den Variablen ausgeht. Die erwarteten Häufigkeiten werden hierbei aus den Randverteilungen ermittelt. Der Chi-Quadrat-Test hat den Vorteil, dass er keine Ansprüche an die Verteilung der vorliegenden Daten stellt und sogar auf nominalem Datenniveau ausgeführt werden kann.

Die Nullhypothese lautet in diesem Fall, dass keine Beziehung zwischen den untersuchten Merkmalen besteht. Die Alternativhypothese geht vom Vorliegen eines Zusammenhang aus. Die Entscheidung für oder gegen eine Hypothese beruht auf der Prüfgröße des Chi^2-Wertes, für den eine Wahrscheinlichkeitsverteilung (Chi^2-Verteilung) bekannt ist. Das Signifikanzniveau (üblicherweise 0,05) und die Freiheitsgrade [(Zahl der Spalten -1) x (Zahl der Zeilen -1)] bestimmen dann, in welchem kritischen Bereich der Werte der Prüfgröße die Alternativhypothese noch angenommen werden kann. Liegt die Prüfgröße außerhalb des kritischen Bereiches (also innerhalb des Vertrauensbereiches, d. h. sie ist kleiner als der kritische Chi^2-Wert), muss dagegen die Nullhypothese beibehalten werden.

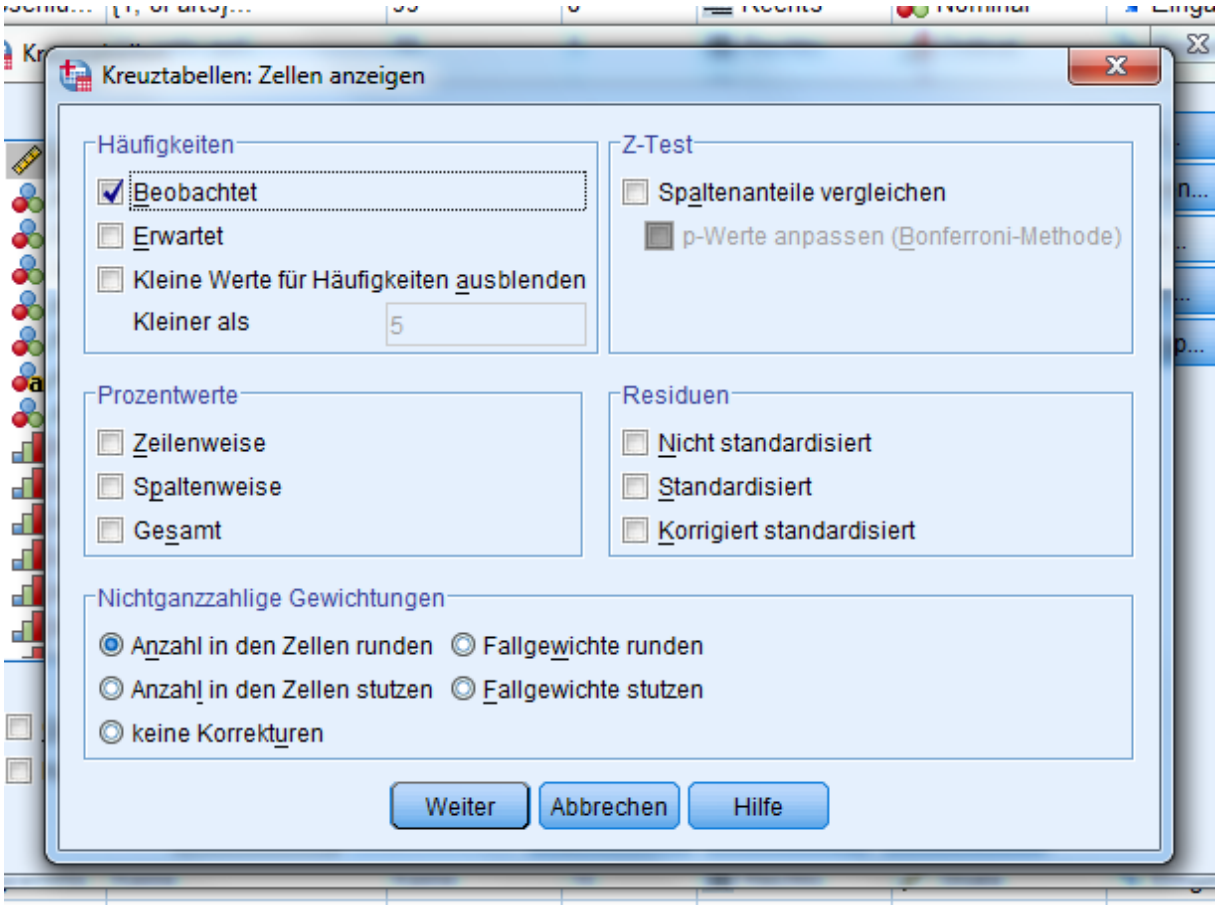
Die Berechnung eines Chi^2-Tests erfolgt über das bereits bekannte Kontextmenü ***Kreuztabellen***, dass über ***Analysieren → Deskriptive Statistik → Kreuztabellen*** angefordert werden kann. Klicken wir hier auf die Schaltfläche ***Statistik***, öffnet sich die dargestellte Dialogbox (Abbildung 14.1).

Abbildung 14.1: Dialogbox ***Kreuztabellen: Statistik***



Hier wählt man das Kontrollkästchen unter ***Chi-Quadrat*** und bestätigt mit ***Weiter***. Es lohnt sich außerdem, unter der Schaltfläche ***Zellen*** die Anzeige der erwarteten Werte auszuwählen (siehe Abbildung 14.2).

Abbildung 14.2: Dialogbox **Kreuztabellen: Zellen anzeigen**



Wieder bestätigen wir mit ***Weiter*** und schließen die Berechnung mit ***Ok*** ab. Betrachten wir nun ein Ausgabebeispiel für die „Studentenstudie 1“ zur Verdeutlichung der Interpretation der Ergebnisse. Wir wollen untersuchen, ob es einen Zusammenhang zwischen dem geäußerten ***„Interesse am Studienfach“*** und der ***„Erwartungen von guten Berufsaussichten“*** gibt (siehe Abbildung 14.3).

Abbildung 14.3: Output-Beispiel eines Chi^2-Tests

Allgemeines Interesse * Gute Berufsaussichten Kreuztabelle

			Gute Berufsaussichten		Gesamt
			nein	ja	
Allgemeines Interesse	nein	Anzahl	4	2	6
		Erwartete Anzahl	3,9	2,1	6,0
	ja	Anzahl	9	5	14
		Erwartete Anzahl	9,1	4,9	14,0
Gesamt		Anzahl	13	7	20
		Erwartete Anzahl	13,0	7,0	20,0

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	,010[a]	1	,919		
Kontinuitätskorrektur[b]	,000	1	1,000		
Likelihood-Quotient	,011	1	,918		
Exakter Test nach Fisher				1,000	,664
Zusammenhang linear-mit-linear	,010	1	,921		
Anzahl der gültigen Fälle	20				

a. 3 Zellen (75,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 2,10.

b. Wird nur für eine 2x2-Tabelle berechnet

Die zuerst dargestellte Tabelle beinhaltet die eigentliche Kreuztabelle, ergänzt um die Erwartungswerte. Bereits bei der Betrachtung dieser Verteilung können wir erkennen, dass sich die erwarteten Werte und die beobachteten Werte nahezu gleichen. Auch die Werte des Chi^2-Tests sprechen für keinen Zusammenhang zwischen den beiden Variablen. Die uns interessierende ***Asymptotische Signifikanz (2-seitig)*** (die anderen Angaben sind weitere Varianten des Chi^2-Tests) liegt bei 0,919 und die Chi^2-Prüfgröße bei 0,1. Wir verwerfen also die Alternativhypothese und nehmen die Nullhypothese an, die besagt, dass es keinen Zusammenhang zwischen dem ***„Interesse am Studienfach“*** und der ***„Erwartungen von guten Berufsaussichten“*** gibt. *ACHTUNG: Auch der Chi-²-Test hat gewisse Voraussetzungen, die hier in den Fußnoten a und b der unteren Tabelle von Abbildung 14.3 teilweise angedeutet werden. Die minimale erwartete Häufigkeit muss in allen Klassen >1 sein (erfüllt), die erwartete Häufigkeit darf bei maximal 20 % aller Klassen <5 sein (nicht erfüllt, der Test ist hier also nicht zulässig bzw. aussagekräftig!).*

Aufgabe 14.1
Untersuchen Sie den Zusammenhang zwischen dem „Geschlecht“ und der Antwort auf die Variable „Persönliche Probleme mit der Ernährung“ aus der „Studentenstudie 1“.

14.2 BERECHNUNG EINES KONFIDENZINTERVALLS ÜBER DIE EXPLORATIVE DATENANALYSE

Will man im Rahmen der deskriptiven Statistik nicht nur statistische Kennwerte für die Stichprobe berechnen, sondern auch mehr über den Mittelwert oder die Streuung innerhalb der Grundgesamtheit erfahren, muss man wahrscheinlichkeitstheoretische Überlegungen mit einbeziehen. Da im Normalfall keine Vollerhebungen vorliegen, können wir nicht mit vollständiger Sicherheit von den statistischen Kennwerten der Grundgesamtheit sprechen. Man kann aber Intervalle bestimmen, innerhalb derer der „wahre“ Wert mit einer anzugebenden Wahrscheinlichkeit liegt. Wir sprechen hierbei von einem Konfidenzintervall. SPSS bietet

Operationen für die Ermittlung des Konfidenzintervalls des arithmetischen Mittels, der Schiefe und der Wölbung sowie des Regressionskoeffizienten an. Unter dem Menü ***Diagrammerstellung*** lassen sich auch Konfidenzintervalle für die Spannweite und die Standardabweichung angeben. Wir konzentrieren uns im Folgenden jedoch auf die Ermittlung des arithmetischen Mittels der Grundgesamtgesamtheit.

Zur Berechnung eines Konfidenzintervalls für das arithmetische Mittel gehen wir auf ***Analysieren → Deskriptive Statistik → Explorative Datenanalyse***. Unter Statistiken setzten wir einen Haken bei ***Deskriptive Statistik*** und bestätigen mit ***Weiter*** und ***Ok***.

Abermals sehen wir uns eine Beispielausgabe für die Berechnung eines Konfidenzintervalls in SPSS an. Wieder verwenden wir die „Studentenstudie 1“ und betrachten die Variable ***„Gewicht in Kilogramm“***. Es ergibt sich folgende Ausgabe (Abbildung 14.4).

Abbildung 14.4: Beispielhafte Ausgabe einer Konfidenzintervall-Berechnung

Deskriptive Statistik

			Statistik	Standardfehler
Gewicht in Kilogramm	Mittelwert		74,9000	2,15443
	95% Konfidenzintervall des Mittelwerts	Untergrenze	70,3907	
		Obergrenze	79,4093	
	5% getrimmtes Mittel		75,1111	
	Median		76,0000	
	Varianz		92,832	
	Standardabweichung		9,63491	
	Minimum		55,00	
	Maximum		91,00	
	Spannweite		36,00	
	Interquartilbereich		13,00	
	Schiefe		-,344	,512
	Kurtosis		-,204	,992

Uns interessieren vor allem die ersten drei Zeilen. Es ist zu erkennen, dass der Mittelwert der Stichprobe bei 74,9 kg liegt. Der Mittelwert der Grundgesamtheit liegt mit einer Wahrscheinlichkeit von 95 % zwischen 70,3907 als Untergrenze und 79,4093 als Obergrenze.

Aufgabe 14.2
Berechnen Sie das Konfidenzintervall für den „wahren“ Mittelwert bezogen auf die Variable „Höhe über NN in m“ in der SPSS-Datendatei „Klimastationen Europa“.

15 LITERATURVERZEICHNIS

Auswahlliteratur zum Umgang mit SPSS:

Angele, G. (2012): SPSS Statistics 20 – Eine Einführung. http://www.uni-bamberg.de/ filead-min/uni/service/rechenzentrum/serversysteme/dateien/spss/skript.pdf (29.09.2012).

Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (2006): Multivariate Analysemethoden – Eine anwendungsorientierte Einführung. D.C.: Berlin, Heidelberg.

Benesch, M., Raab-Steiner, E. (2008): Der Fragebogen – Von der Forschungsidee zur SPSS-Auswertung. UTB: Wien.

Bühl, A. (2012): SPSS 20 - Einführung in die moderne Datenanalyse. Pearson Studium: Mün-chen.

Bühl, A., Zöfel, P. (2005): SPSS 12 – Einführung in die moderne Datenanalyse unter Windows. Pearson Studium: München.

Cleff, T. (2008): Deskriptive Statistik und moderne Datenanalyse – Eine computergestützte Einführung mit Excel, SPSS und STATA. Gabler: Wiesbaden.

Diehl, J.M., Staufenbiel, T. (2002): Statistik mit SPSS – Version 10+11. Verlag Dietmar Klotz: Eschborn.

Duller, C. (2006): Einführung in die Statistik mit Excel und SPSS – Ein anwendungsorientiertes Lehr- und Arbeitsbuch. Physica-Verlag: Heidelberg.

Elsner, F. (2009): Statistische Datenanalyse mit SPSS für Windows – Grundlegende Konzepte und Techniken. Version 2.7. Rechenzentrum der Universität: Osnabrück http://www.home.uni-osnabrueck.de/elsner/Skripte/spss.pdf (29.09.2014).

Heumann, C., Toutenburg, H. (2008): Deskriptive Statistik – Eine Einführung in Methoden und Anwendungen mit R und SPSS. Springer: Berlin, Heidelberg.

Heumann, C., Toutenburg, H. (2008): Induktive Statistik – Eine Einführung mit R und SPSS. Springer: Berlin, Heidelberg.

Janssen, J. Laatz, W. (2007): Statistische Datenanalyse mit SPSS für Windows – Eine anwen-dungsorientierte Einführung in das Basissystem und das Modul Exakte Tests. Springer: Berlin, Heidelberg, New York.

Niketta, R. (2009): SPSS-Skripte. http://www.home.uni-osnabrueck.de/rniketta/method/ html/spss-skripte.html (30.09.2014).

Rodeghier, M. (1997): Marktforschung mit SPSS – Analyse, Datenerhebung und Auswertung. Thomson Publishing: Bonn.

Untersteiner, H. (2007): Statistik – Datenauswertung mit Excel und SPSS. UTB: Wien.

Verbundprojekt des Verbunds Norddeutscher Universitäten unter Beteiligung der Universitäten Bremen, Greifswald, Hamburg und Rostock: Methodenlehre-Baukasten - Ein interaktives Lehr-Lernprogramm zur Statistik. http://methodenlehre-baukasten.de/web/php/index.php (29.09.2014 – leider kostenpflichtig).

Eine aktuelle Liste mit Originaldokumentation zu SPSS findet man unter: http://www-947.ibm.com/support/entry/portal/documentation/software/spss/spss_statistics (30.09.2014).

Statistische Grundlagenwerke:

Bahrenberg, G., Giese, E., Nipper, J., Mevenkamp, N. (2010): Statistische Methoden in der Geographie. Band 1: Univariate und bivariate Statistik. 5. Aufl. Borntraeger: Stuttgart.

Bahrenberg, G., Giese, E., Mevenkamp, N., Nipper, J. (2008): Statistische Methoden in der Geographie. Band 2: Multivariate Statistik. Gebrüder Borntraeger Verlagsbuchhandlung: Berlin, Stuttgart.

Bortz, J., Döring, N. (2009): Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Springer-Medizin-Verlag: Heidelberg.

Ernste, H. (2010): Angewandte Statistik in Geografie und Umweltwissenschaften. Vdf: Zürich.

Krämer, W. (2011): So lügt man mit Statistik. Piper: München.

Raithel, J. (2008): Quantitative Forschung – Ein Praxiskurs. VS Verlag: Wiesbaden.

Rogerson, P.A. (2011): Statistical Methods for Geography – A Student´s Guide. SAGE: London.

Voß, W. (2000): Taschenbuch der Statistik. Carl Hanser Verlag: München, Wien.

Aufgabensammlungen:

Baur, N., Fromm, S. (2008): Datenanalyse mit SPSS für Fortgeschrittene – Ein Arbeitsbuch. Verlag für Sozialwissenschaften: Wiesbaden.

Caputo, A., Fahrmeir, L., Künstler, R., Lang, S., Pigeot-Kübler, I., Tutz, G. (2009): Arbeitsbuch Statistik. Springer: Berlin, Heidelberg.

Eckstein, P. P. (2014): Datenanalyse mit SPSS: Realdatenbasierte Übungs- und Klausuraufgaben mit vollständigen Lösungen. 4. Aufl. Gabler: Wiesbaden.