UNIVERSITÄT TRIER

ZIMK - Zentrum für
Informations-, Medien- und
Kommunikationstechnologie

Bernhard Baltes-Götz

Statistisches Praktikum mit IBM SPSS Statistics 28

2022 (Rev. 220325)

Herausgeber: Zentrum für Informations-, Medien- und Kommunikationstechnologie (ZIMK) an der Universität Trier
Universitätsring 15
D-54286 Trier
WWW: zimk.uni-trier.de
E-Mail: zimk@uni-trier.de
Autor: Bernhard Baltes-Götz
WWW: https://www.uni-trier.de/~baltes
E-Mail: baltes@uni-trier.de

Vorwort

Das seit Jahrzehnten bewährte und ständig aktualisierte Statistikprogramm **SPSS** (***S**tatistical **P**ackage for the **S**ocial **S**ciences*) trägt seit der Übernahme des Herstellers durch die Firma IBM den Namen **IBM SPSS Statistics**. Wir verwenden im Manuskript den kompakten Namen **SPSS** und reden dabei über ein relativ leicht zu bedienendes Statistikprogramm mit großem Funktionsumfang, das einen hohen Verbreitungsgrad besitzt (z. B. in den Sozial-, Wirtschafts- und Geowissenschaften) und alle wichtigen Betriebssysteme für Arbeitsplatzrechner unterstützt (Linux, macOS, Windows). AnwenderInnen[1] profitieren u. a. von der ...

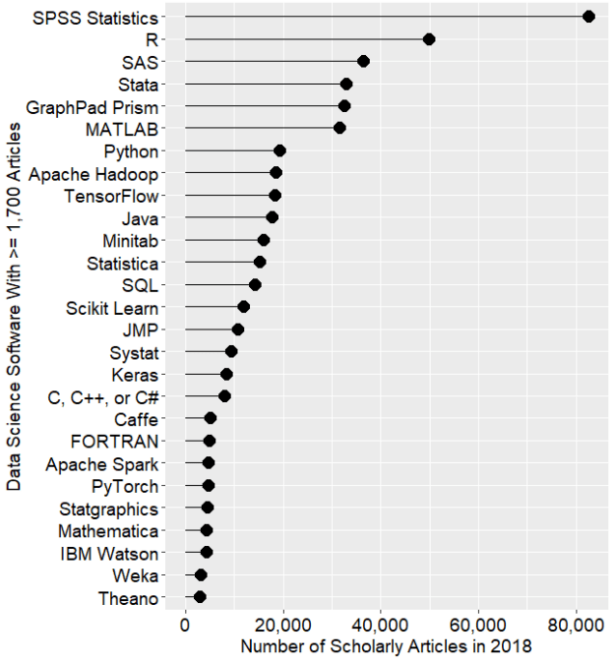
- guten Unterstützung des Programms durch die Statistikliteratur und frei verfügbare Anleitungen,
- guten Erweiterbarkeit des Programms, die zu zahlreichen Lösungen (Extensions, Makros etc.) geführt hat.

Im Vergleich zu alternativen Statistikprogrammen mit einem vergleichbaren breiten Einsatzspektrum (z. B. R, SAS, Stata) ist als Pluspunkt von SPSS vor allem die relativ einfache Bedienung zu nennen.

Robert A. Muenchen vergleicht auf seiner Webseite mit dem Thema

The Popularity of Data Science Software

die wichtigsten Statistikprogramme bzgl. diverser Kriterien (z. B. Stellenangebote, Nutzung in Forschungsartikeln).[2] In der Kategorie *Number of Scholarly Articles* war SPSS im Jahr 2018 wie in den 20 vorherigen Jahren einsame Spitze:



---

1 Zur Vermeidung von sprachlichen Umständlichkeiten beschränkt sich das Manuskript meist auf die männliche Form.

2 http://r4stats.com/articles/popularity/ (abgerufen am 20.03.2022)
Für die Jahre nach 2018 hat Robert A. Muenchen leider noch keine Daten berichtet.

Das erklärt Robert A. Muenchen auf überzeugende Weise so:

> SPSS is by far the most dominant package, as it has been for over 20 years. This may be due to its balance between power and ease-of-use. R is in second place with around half as many articles. It offers extreme power, though with less ease of use.

Die freie Statistik-Programmiersprache R muss nicht als *Konkurrenz* zu SPSS, sondern kann ebenso gut als *Ergänzung* zu SPSS angesehen werden. Über eine Erweiterungslösung kann man R-Funktionen ziemlich bequem in SPSS nutzen, und seit der SPSS-Version 28 wird eine kompatible R-Version gleich mitinstalliert.

Das vorliegende Manuskript bietet einen Einstieg in die statistische Datenanalyse mit der SPSS-Version 28 für Windows. Wesentliche Teile des Manuskripts sind wegen der weitgehend konsistenten Bedienungslogik auch für andere SPSS-Versionen ab 16 unter Windows oder alternativen Betriebssystemen verwendbar.

Das Manuskript dient primär als Begleitlektüre zum Kurs *Statistisches Praktikum mit SPSS*, den das *Zentrum für Informations-, Medien- und Kommunikationstechnologie* (*ZIMK*) an der Universität Trier anbietet, kann aber auch im Selbststudium verwendet werden.

Zielgruppe und Voraussetzungen

Der Kurs ist konzipiert für Personen, die *in wesentlichem Umfang* bei Forschungsarbeiten mit SPSS mitwirken wollen (z. B. im Rahmen einer eigenen Studie oder als Mitglied in einem Forschungsteam). Im Kurs wird eine *methodologische Grundausbildung* (in empirischer Forschung und Statistik) vorausgesetzt, wie sie üblicherweise in den Studiengängen von empirisch-statistisch forschenden Disziplinen vermittelt wird. Begriffe aus der univariaten Verteilungsanalyse (z. B. Mittelwert, Median, Varianz, Konfidenzintervall) sollten ebenso geläufig sein wie einfache Verfahren der Zusammenhangsanalyse (z. B. t-Test für (un)abhängige Stichproben, lineare Regression).

Kursinhalte und Lernziele

Wir konzentrieren uns darauf, in anderen Veranstaltungen (z. B. zur empirischen Forschung oder Statistik) erlernte Begriffe und Methoden *mit SPSS in der Praxis anzuwenden*. Zwar werden im Kursverlauf viele methodologische Themen (z. B. Verteilungsanalyse, Konfidenzintervall, Logik von Signifikanztests) in knapper Form erläutert, doch kann dabei eher vorhandenes Wissen aufgefrischt und vertieft, als neues Wissen erworben werden. Insbesondere kann die Diskussion und Anwendung der vielfältigen statistischen Modelle bzw. Auswertungsmethoden nur exemplarisch stattfinden. Eine gründliche, mehr oder weniger vollständige Behandlung ist nur bei wenigen Verfahren möglich (z. B. t-Test, lineare Regression, Kreuztabellenanalyse).[1]

---

[1] Zu vielen Auswertungsmethoden bietet das ZIMK spezielle Kurse an, in denen die wesentlichen methodologischen Grundlagen und natürlich auch die praktische Durchführung mit SPSS erläutert werden. Informationen über das ZIMK-Kursprogramm finden Sie u. a. auf dem Webserver der Universität Trier, z. B. zum Sommersemester 2022:

https://www.uni-trier.de/index.php?id=43557

Zu den meisten Kursen sind ausführliche Manuskripte entstanden, die Sie auf dem Webserver der Universität Trier folgendermaßen finden:

https://www.uni-trier.de/index.php?id=20047

Im Sinne einer *praxisnahen und projektorientierten Ausbildung* wird im Kurs bzw. Manuskript eine vollständige empirische Studie von der ersten Idee über die Untersuchungsplanung, sowie die Codierung, Erhebung, Erfassung, Kontrolle, Korrektur und Modifikation der Daten bis zur statistischen Auswertung und zur Aufbereitung der Ergebnisse für Forschungsberichte durchgeführt. Alle methodologischen Entscheidungen im Demonstrationsprojekt werden sorgfältig begründet (z. B. die Auswahl der statistischen Verfahren).

Insgesamt sollten mit dem Kurs bzw. Manuskript u. a. die folgenden Lernziele erreicht werden:

- **Korrekte Anwendung von elementaren statistischen Forschungsmethoden**
- **Rationelles Arbeiten mit SPSS**

Zwar werden die Datenverarbeitungs- und Analyse-Optionen von SPSS nicht annähernd vollständig behandelt, doch sollten Sie nach dem Kurs mit den erworbenen Kenntnissen und unter Verwendung der aufgezeigten Informationsmöglichkeiten selbständig und erfolgreich mit SPSS können.

Dateien zum Kurs bzw. Manuskript

Die aktuelle Version dieses Manuskripts ist als PDF-Dokument zusammen mit den im Kurs benutzen Dateien auf dem Webserver der Universität Trier zu finden:

https://www.uni-trier.de/?id=22552

Leser im Selbststudium können mit den zur Verfügung gestellten Dateien fast alle Arbeitsschritte im Beispielprojekt des Kurses konkret durchführen.

Kritik und Verbesserungsvorschläge zum Manuskript werden dankbar entgegen genommen (z. B. unter der Mail-Adresse baltes@uni-trier.de).

Trier, im März 2022 Bernhard Baltes-Götz

Inhaltsverzeichnis

1 Einstieg in IBM SPSS Statistics

Vor dem Einstieg in die Projektarbeit werden einige organisatorische und technische Informationen zu SPSS präsentiert.

1.1 SPSS-Produkte an der Universität Trier

An der Universität Trier steht SPSS für Linux, macOS und Windows mit einer kompletten Modul-Ausstattung zur Verfügung:

> Statistics Base, Regression, Advanced Statistics, Categories, Conjoint, Custom Tables, Data Preparation, Decision Trees, Exact Tests, Forecasting, Missing Values, Bootstrapping, Neural Networks, Direct Marketing, Complex Samples

Aus der SPSS-Produktfamilie ist außerdem noch das Programm **Amos** vorhanden, das Strukturgleichungsanalysen (z. B. konfirmatorische Faktorenanalysen) unterstützt und leider nur für Windows verfügbar ist.

SPSS und Amos können von Angehörigen der Universität Trier im Rahmen ihrer dienstlichen Tätigkeit bzw. Ausbildung auf folgende Weise genutzt werden:

a) ZIMK - Pool-PCs

Auf den vom ZIMK betreuten Pool-PCs unter dem Betriebssystem Windows finden Sie im Startmenü diese Programmgruppen:[1]

IBM SPSS Statistics 28

IBM SPSS Amos 26

b) Kostenlose Nutzung über die ZIMK-Lizenzserver (netzabhängig)

Über die Webseite

http://www.uni-trier.de/index.php?id=25191

stehen SPSS und Amos samt Installationsanleitung für Angehörige der Universität Trier zum Herunterladen bereit. Mit den bezogenen Dateien lassen sich SPSS und/oder Amos auf einem Rechner mit permanentem Internetzugang (an der Uni oder im Privatbereich) zur kostenlosen Nutzung der ZIMK-Lizenzserver installieren.

c) Kostenpflichtige individuelle Mietlizenz (netzunabhängig)

Für Rechner ohne permanenten Internetzugang (z.B. für einen an verschiedenen Orten verwendeten Laptop) kann eine befristete Einzelplatzlizenz erworben werden, die SPSS und Amos umfasst.

1.2 Programmstart und Benutzeroberfläche

1.2.1 SPSS starten

Nach erfolgreicher Anmeldung bei einem ZIMK - Pool-PC unter Windows 10 erreichen Sie SPSS 28 über das Desktop-Symbol

[1] Die Aktualisierung von SPSS Amos auf die aktuelle Version 28 erfolgt in Kürze.

oder über die folgende Programmgruppe im Startmenü:

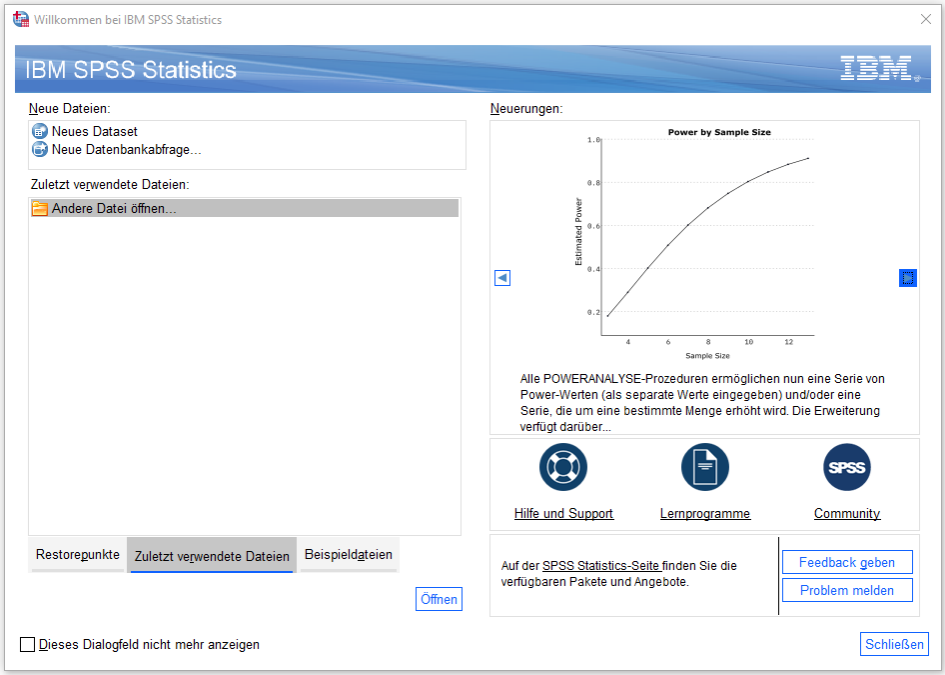
IBM SPSS Statistics 28

Nach einer Standardinstallation von SPSS 28 unter Windows 10 (z. B. auf einem privaten PC) finden Sie im Startmenü die Programmgruppe

IBM SPSS Statistics

mit einem gleichnamigen Item zum Starten von SPSS 28.

Der folgende Begrüßungsdialog erscheint bei jedem Start von SPSS 28, sofern Sie ihn nicht über das Kontrollkästchen in seiner unteren linken Ecke unterdrücken:[1]

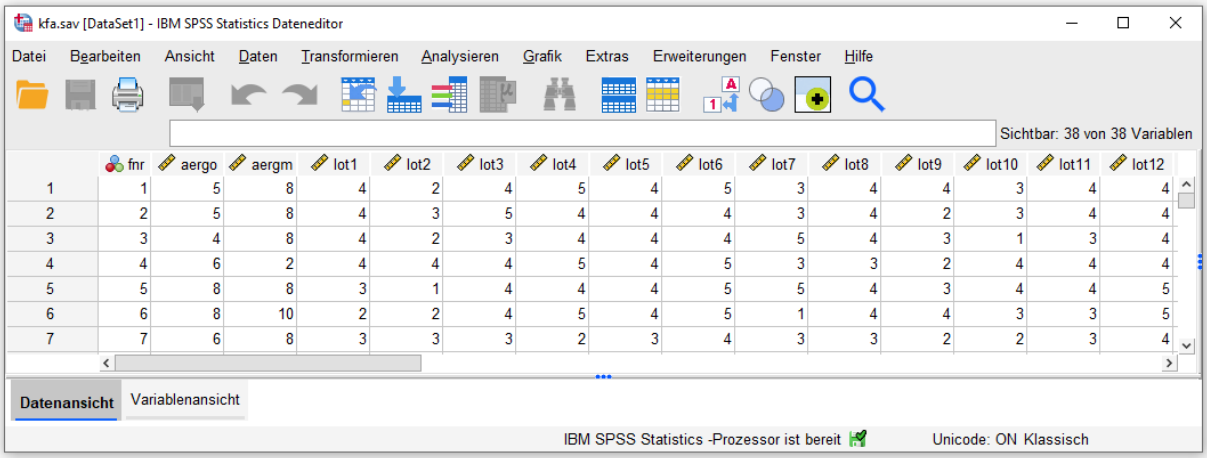


Er ermöglicht z. B. das bequeme Öffnen von in früheren Sitzungen benutzten Dateien.

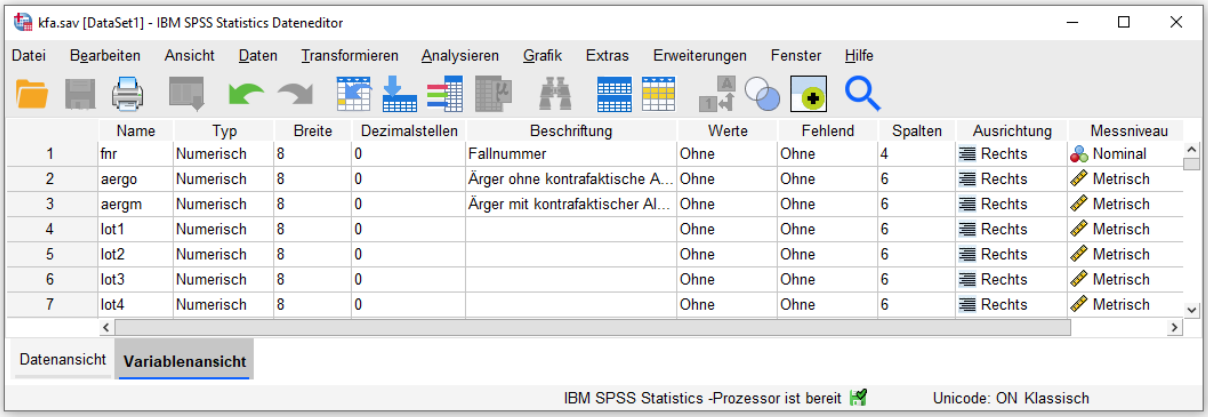
1.2.2 Die wichtigsten SPSS-Fenster

Das **Dateneditorfenster** dient zur Deklaration, Erfassung und Verwaltung von Daten. Dazu besitzt es die Registerkarten **Datenansicht**

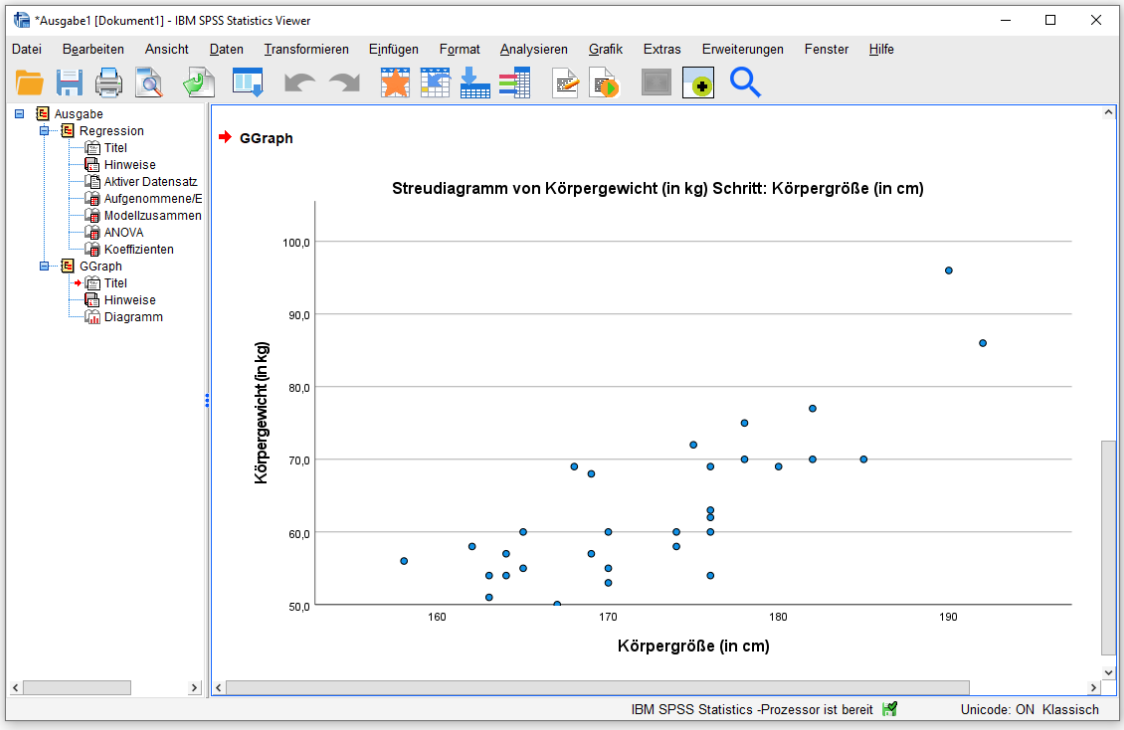
[1] Wenn Sie den Begrüßungsdialog im späteren Verlauf einer SPSS-Sitzung wiedersehen wollen, können Sie im Fenster des Dateneditors den Menübefehl **Datei > Begrüßungsdialogfeld** verwenden.



und **Variablenansicht**:



Angeforderte Ergebnistabellen und Diagramme erscheinen im **Ausgabefenster**, auch **IBM SPSS Statistics Viewer** genannt, z. B.:



Die SPSS-Fenster enthalten ...

- in der Kopfzone eine Menüzeile und eine Symbolleiste
- im Fußbereich eine Statuszeile mit Informationen über wichtige Programmzustände.

1.2.3 Was man mit SPSS so alles machen kann

Wir sind im Moment dabei, einen ersten Eindruck vom Arbeitsplatz *SPSS Statistics* zu gewinnen. Einen guten Überblick vermitteln die Optionen in der Menüzeile des Dateneditorfensters:

- **Datei**
  Hier finden Sie u. a. Befehle zum Öffnen bzw. Sichern von Daten-, Ausgabe oder Syntaxdateien sowie zum Beenden von SPSS.
- **Bearbeiten**
  Über das **Bearbeiten**-Menü erreichen Sie Editorbefehle zum Ausschneiden, Kopieren, Einfügen, Löschen und Suchen von Daten sowie die **Optionen**-Dialogbox zur Anpassung von diversen SPSS-Einstellungen. Außerdem können Sie hier Modifikationen des Datenfensters rückgängig machen.
- **Ansicht**
  Hier können Sie u. a. die Statuszeile sowie die Symbolleisten aus- bzw. einschalten sowie die Schriftart der angezeigten Daten festlegen.
- **Daten**
  Über das **Daten**-Menü sind u. a. Dialoge zur Auswahl einer Teilstichprobe, zum Zusammenführen von SPSS-Dateien (z. B. mit Daten aus verschiedenen Stichproben) sowie zum Sortieren und Gewichten der Fälle erreichbar.
- **Transformieren**
  Hier finden Sie z. B. die Befehle zum Recodieren von Variablen oder zum Berechnen von neuen Variablen aus bereits vorhandenen.
- **Analysieren**
  Dieser Menüpunkt erschließt die statistischen Auswertungsmethoden, mit denen wir letztlich unsere Forschungsfragen klären wollen.
- **Grafik**
  An dieser Stelle bietet SPSS vielfältige Möglichkeiten zur grafischen Präsentation von Datenstrukturen an.
- **Extras**
  Hier finden sich diverse Funktionen (z. B. Kommentieren einer Datendatei, Definition von Variablensets zur vereinfachten Verwendung von sehr umfangreichen Dateien).
- **Erweiterungen**
  Dieses mit der SPSS-Version 25 hinzu gekommene Menü unterstützt die Verwaltung von Erweiterungen und benutzerdefinierten Dialogfeldern.
- **Fenster**
  Über dieses Menü sind die offenen SPSS-Fenster erreichbar.
- **Hilfe**
  Die Hilfefunktion bietet neben systematischen Informationen über das SPSS-System auch ein Lernprogramm, Fallstudien (komplette Anwendungsbeispiele) und einen Statistik-Assistenten.

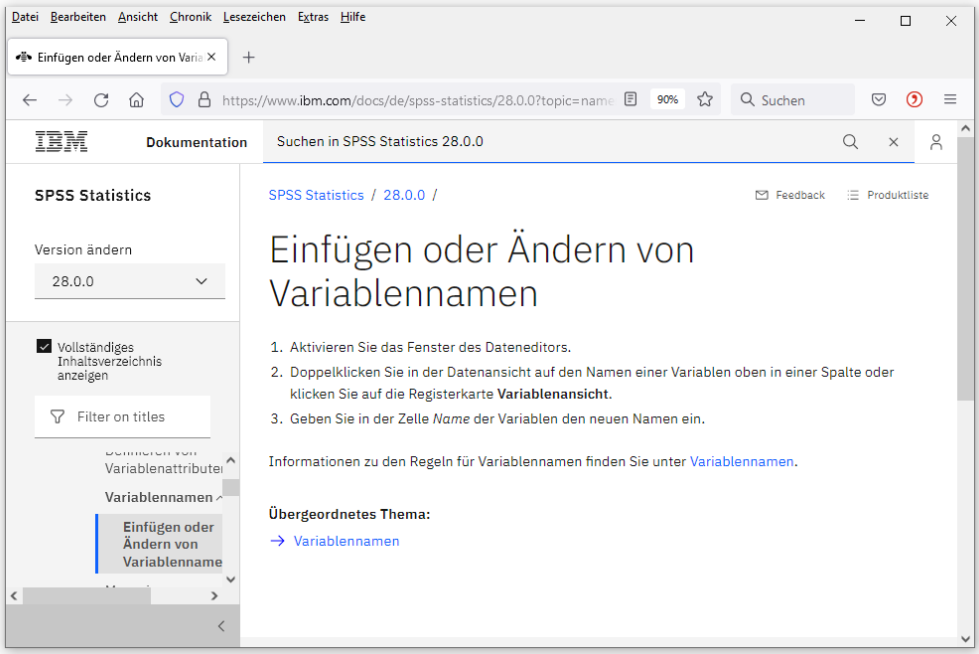
Die anderen SPSS-Fenster bieten angepasste Menüzeilen.

1.3 Hilfesystem

Bei der Arbeit mit SPSS können Sie stets auf ein Hilfesystem zurückgreifen, dessen wichtigste Möglichkeiten in diesem Abschnitt vorgestellt werden. Zu technischen Fragen (z. B. verfügbare numerische Funktionen, Syntax von Kommandos) informiert die Hilfe zuverlässig und umfassend. Aufschlüsse über die Hintergründe statistischer Verfahren oder über die Bedeutung von Optionen von Statistik-Prozeduren sind allerdings aus der Hilfe kaum zu beziehen.

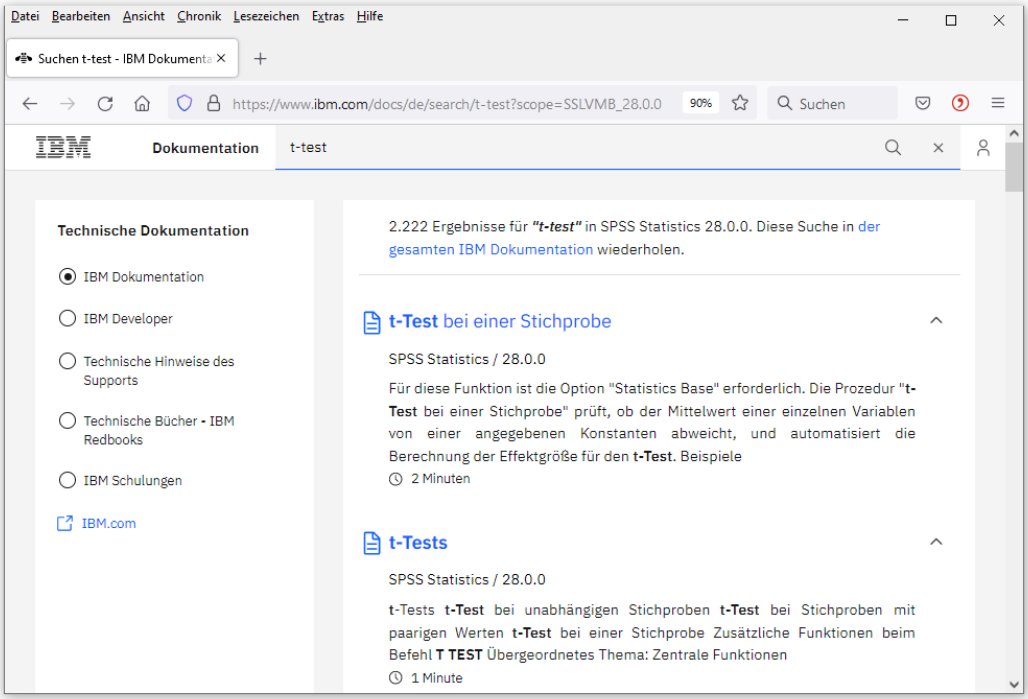
1.3.1 Systematische Informationen

Nach dem Menübefehl **Hilfe > Themen** finden Sie in einem Browser-Fenster Informationen in systematischer Form, z. B.:



1.3.2 Suche nach Begriffen

Das Browser-Fenster der SPSS-Hilfe ermöglicht die Suche nach Informationen zu bestimmten Begriffen, z. B.:



1.3.3 Kontextsensitive Hilfe zu den Dialogboxen

In fast jeder SPSS-Dialogbox können Sie mit der Standardschaltfläche **Hilfe** Informationen zu den verfügbaren Optionen anfordern.

1.4 Weitere Informationsquellen

Neben dem umfangreichen Hilfesystem sind zu SPSS noch weitere Informationsquellen vorhanden.

1.4.1 Handbücher und Manuskripte

Es stehen u. a. zur Auswahl:

- **Programmhandbücher des Herstellers**
  Mit SPSS wird eine umfangreiche Sammlung von PDF-Dokumenten zu den einzelnen Modulen und zu den statistischen Algorithmen ausgeliefert. Allein die Dokumentation der Kommandosprache, über die man die meisten Leistungen des SPSS-Systems abrufen kann (siehe z. B. Kapitel 6), umfasst ca. 2450 Seiten. Dieses Handbuch ist im Hilfe-Menü ebenso verlinkt (**Hilfe > Befehlssyntaxreferenz**) wie eine Webseite mit weiteren PDF-Dokumenten (**Hilfe > Dokumentation im PDF-Format)**. Allerdings sind die von SPSS zur Verfügung gestellten Handbücher keine geeignete Lektüre, wenn man eine statistische Methode neu erlernen möchte.
- **Sekundärliteratur**
  Im Buchhandel und in wissenschaftlich orientierten Bibliotheken finden sich zahlreiche Bücher zu SPSS. Besonders nützlich sind Bücher, die statistische Methoden anwendungsorientiert behandeln und die konkrete Realisation mit SPSS beschreiben (z. B. durch eine Erläuterung der Ergebnistabellen). Leider habe ich mir aus Zeitgründen von

den zahlreichen Statistik-Lehrbüchern mit SPSS-Unterstützung nur wenige Titel näher ansehen können, sodass die folgende Liste unvollständig ist:

Backhaus et al. (2008). *Multivariate Analysemethoden*

Backhaus et al. (2015). *Fortgeschrittene multivariate Analysemethoden*

Brosius (2018). *SPSS: Umfassendes Handbuch zu Statistik und Datenanalyse*

Bühl (2016). *SPSS 23. Einführung in die moderne Datenanalyse*

Bühner & Ziegler (2017). *Statistik für Psychologen und Sozialwissenschaftler*

Cohen et al. (2003). *Applied Multiple Regression/Correlation Analysis …*

Diehl & Staufenbiehl (2007). *Statistik mit SPSS für Windows. Version 15*

Faik (2018). *Statistik mit SPSS: Alles in einem Band für Dummies*

Field (2017). *Discovering Statistics Using IBM SPSS Statistics*

Norušis (2012a). *IBM SPSS Statistics 19 Guide to Data Analysis*

Norušis (2012b). *IBM SPSS Statistics 19 Statistical Procedures Companion*

Norušis (2012c). *IBM SPSS Statistics 19 Advanced Statistical Procedures*

Rasch, Friese, Hofmann & Naumann (2014). *Quantitative Methoden* (Band 1 und 2)

Tabachnik & Fidell (2013). *Using multivariate statistics*

Warner (2013). *Applied statistics*

Die vollständigen bibliografischen Angaben sind im Literaturverzeichnis zu finden.

- **Frei verfügbare Manuskripte**
  Im Internet sind viele frei verfügbare Manuskripte zur Anwendung statistischer Methoden mit SPSS zu finden, die in ihrer didaktischen Qualität teilweise mit den kommerziellen Angeboten des Buchhandels mithalten können. Auf die ZIMK-Manuskripte zur Verwendung spezieller Analysemethoden in SPSS wurde schon im Vorwort hingewiesen.

1.4.2 ZIMK - Service-Punkt

Angehörige der Universität Trier können sich bei Problemen mit der Anwendung von SPSS-Produkten an den ZIMK - Service-Punkt wenden:

Web:	http://helpdesk.uni-trier.de
Mail:	helpdesk@uni-trier.de
Tel.:	0651 - 201 - 4400
Ort:	Campus I, Foyer Gebäude E (Raum 43a)
Zeiten:	Siehe: http://www.uni-trier.de/index.php?id=19249

1.5 SPSS für Windows beenden

Die Beendigung einer SPSS-Sitzung wird mit

Datei > Beenden

eingeleitet. Falls Sie während der Sitzung Dokumente erstellt bzw. verändert und noch nicht gesichert haben (z. B. im Daten- oder im Ausgabefenster), werden Sie von SPSS an das Speichern erinnert.

2 Von der Theorie bis zu den SPSS-Variablen

In diesem Kurs bzw. Manuskript starten wir *nicht* mit fertigen Datendateien, um diese zu öffnen, eventuell noch leicht zu modifizieren und dann diverse statistische Analysemethoden auf die enthaltenen Variablen anzuwenden. Stattdessen nehmen wir den gesamten Prozess der statistisch-empirischen Forschung in den Blick, starten in einem Demonstrationsprojekt mit Überlegungen zum theoretischen Hintergrund, leiten daraus Hypothesen ab, nehmen eine Untersuchungsplanung vor, organisieren die Datenerhebung und führen alle weiteren Arbeitsschritte bis zum Ergebnisbericht durch (siehe Überblick in Abschnitt 2.2). Am Ende von Kapitel 2 wird als Zwischenergebnis feststehen, welche Informationen im Demonstrationsprojekt erhoben und welche SPSS-Variablen (im Sinn von Abschnitt 2.4.1) daraus erstellt werden sollen.

2.1 Wesen und Aufgaben der Statistik

Erfahrungswissenschaften bemühen sich um allgemeingültige Aussagen beschreibender, erklärender oder prognostizierender Art. Viele Wissenschaftsdisziplinen (z. B. Geo-, Sozial- oder Wirtschaftswissenschaften, Biologie, Medizin) können kaum mit *deterministischen* Gesetzen (z. B. Hebelgesetz der Mechanik) rechnen und benötigen daher eine Methodologie zur Erforschung **probabilistischer Gesetze**.

Eine beispielhafte Forschungsfrage aus dem Bereich der Medizin könnte lauten:

> Erhöht Rauchen das Risiko, an Lungenkrebs zu erkranken?

2.1.1 Kluge Entscheidungen trotz Unsicherheit

Wie wir wissen, hat das Rauchen auch bei identischer Dosierung keinesfalls für alle Personen dieselben Folgen. Die Frage nach dem Einfluss des Rauchens auf die Entstehung von Lungenkrebs ist anhand von wenigen Einzelbeobachtungen (z. B. bei einem steinalten Kettenraucher) *nicht* zu klären. In einer solchen, durch **Unsicherheit** geprägten Situation können *statistische* Methoden dabei helfen, rationale Entscheidungen zu treffen, denn Wallis & Roberts (1956, S. 1) stellen treffend fest:

> **Statistics is a body of methods for making wise decisions in the face of uncertainty.**

Bevor die Statistik bei der Klärung einer Forschungsfrage ihren Nutzen entfalten kann, sind relevante Daten in ausreichendem Umfang nach einer sorgfältigen **Untersuchungsplanung** zu beschaffen. Insbesondere ist die wichtige Entscheidung zwischen einem Experiment (mit seinen gravierenden Vorteilen für die kausale Interpretation der Ergebnisse) und einer Beobachtungsstudie zu treffen. Im Raucherbeispiel kann aus ethischen Gründen mit Menschen nur eine Beobachtungsstudie durchgeführt werden.

Eine bewährte Strategie der statistisch arbeitenden Forschung, um in einer empirischen Studie trotz Unsicherheit und beschränkter Forschungsmittel zu guten Entscheidungen zu kommen, besteht darin, zu einer Fragestellung **hinreichend viele, repräsentative** und **unabhängige** Untersuchungseinheiten einzubeziehen, um aus dieser **Stichprobe** Informationen über die zugrunde

liegende **Population** der potentiellen Untersuchungseinheiten zu gewinnen.[1] Im Kurs werden Sie an Beispielen erfahren, wie man eine erforderliche Stichprobengröße ermitteln und das verbleibende Maß an Unsicherheit quantifizieren kann.

Zur Untersuchung der Raucherproblematik wird man vielleicht 400 zufällig ausgewählte Erwachsene mit einem Alter ab 50 Jahren (**Untersuchungseinheiten**, **Merkmalsträger**, **Fälle**) auf Nikotinkonsum und das Vorliegen einer Lungenkrebserkrankung untersuchen, sodass bei vereinfachender Dichotomisierung die beiden nominalskalierten **Merkmale** *Raucher* und *Lungenkrebs* (jeweils mit den Ausprägungen *Ja* und *Nein*) resultieren, z. B. mit der folgenden gemeinsamen Stichprobenverteilung:

		Raucher		
		Ja	Nein	Gesamt
Lungenkrebs	Ja	30	15	45
	Nein	70	285	355
	Gesamt	100	300	400

2.1.2 Aufgabenstellungen und statistische Forschungsmethoden

Anschließend wird skizziert, welche Aufgabenstellungen sich mit statistischen Forschungsmethoden bearbeiten lassen.

2.1.2.1 Punkt- und Intervallschätzung

Im medizinischen Anwendungsbeispiel interessieren u. a. zwei Wahrscheinlichkeiten:

> Wie groß ist die Wahrscheinlichkeit, an Lungenkrebs zu erkranken, bei Rauchern bzw. Nichtrauchern?

Um z. B. die Lungenkrebswahrscheinlichkeit für Raucher (also einen **Populationsparameter**) anhand der obigen Stichprobendaten zu schätzen, verwendet man die relative Häufigkeit von Lungenkrebs in der Raucherteilstichprobe (= 0,3).

In der Regel wird man sich *nicht* auf eine **Punktschätzung** beschränken, sondern zusätzlich auch eine **Intervallschätzung** vornehmen. Dabei wird aus den Stichprobendaten für den fraglichen Populationsparameter ein **Vertrauensintervall** (synonym: **Konfidenzintervall**) ermittelt, das den wahren Wert mit einer gewünschten Wahrscheinlichkeit (von z. B.: 0,95) enthält. Obige Daten liefern für die 100 Raucher das 95% - Vertrauensintervall [0,217; 0,395].[2]

Für die 300 Nichtraucher resultiert zum Lungenkrebsrisiko die Punktschätzung 0,05 mit dem Vertrauensintervall [0,030; 0,079], das aufgrund der größeren Teilstichprobe präziser (kleiner) ausfällt.[3]

---

[1] Um aufkeimenden Missverständnissen entgegenzuwirken, sei schon an dieser Stelle erwähnt, dass *unabhängige* Untersuchungseinheiten (z. B. 200 aus der Population aller Schüler zufällig gezogene Probanden) zwar günstige (zumindest einfach handhabbare) Voraussetzungen für die Anwendung statistischer Analysemethoden schaffen, aber nicht unbedingt erforderlich sind. Mit (allerdings in der Regel aufwändigeren Methoden) können auch *abhängige* Untersuchungseinheiten (z. B. 500 Schüler aus insgesamt 30 Klassen) statistisch analysiert werden.

[2] Wie man dieses Vertrauensintervall mit SPSS ermittelt, ist in Abschnitt 5.8 zu erfahren.

[3] Die Breite des Vertrauensintervalls für eine geschätzte Wahrscheinlichkeit hängt auch vom wahren Wert ab. Je weiter die wahre Wahrscheinlichkeit vom neutralen Wert 0,5 entfernt ist, desto schmaler (präziser) wird bei fest vorgegebener Stichprobengröße das Vertrauensintervall.

Den Konfidenzintervallen wird in der methodologischen Literatur immer mehr Bedeutung zugemessen. Viele Autoren halten sie für erheblich informativer und wichtiger als die in Abschnitt 2.1.2.2 zu beschreibenden Hypothesentests (siehe z. B. Brandstätter 1999, Cohen et al. 2003). In Publikationsrichtlinien wird nachdrücklich verlangt, für jede wichtige Parameterschätzung auch das Vertrauensintervall zu berichten (z. B. APA 2010, S. 34).

Von Wahlprognosen sind Vertrauensintervalle sehr vertraut. Die *Forschungsgruppe Wahlen* beschreibt die Genauigkeit ihres Politbarometers im März 2022 folgendermaßen:[1]

> Die Interviews wurden in der Zeit vom 8. bis 10. März 2022 bei 1.345 zufällig ausgewählten Wahlberechtigten telefonisch erhoben. Dabei wurden sowohl Festnetz- als auch Mobilfunknummern berücksichtigt. Die Befragung ist repräsentativ für die wahlberechtigte Bevölkerung in Deutschland. Der Fehlerbereich beträgt bei einem Anteilswert von 40 Prozent rund +/- drei Prozentpunkte und bei einem Anteilswert von 10 Prozent rund +/-zwei Prozentpunkte.

Bei einem dichotomen Merkmal ist mit *einer* Wahrscheinlichkeit die gesamte Verteilung beschrieben. Hat ein nominalskaliertes Merkmal k (> 2) Ausprägungen, ist seine Verteilung durch die Wahrscheinlichkeiten von (k - 1) Kategorien bestimmt, sodass (k - 1) Parameter zu schätzen sind. Bei einem metrischen Merkmal sind in der Regel *viele* Ausprägungen vorhanden, und entsprechend viele Einzelwahrscheinlichkeiten liefern keine praktikable Beschreibung der Verteilung. In dieser Situation ist es sinnvoller, wichtige *Momente* der Verteilung zu schätzen, z. B. den Erwartungswert als Maß der zentralen Tendenz (erstes Moment), die Standardabweichung als Dispersionsmaß (zweites Moment) sowie die Schiefe und die Wölbung zur Beschreibung der Verteilungsgestalt (drittes bzw. viertes Moment).

Neben den gerade angesprochenen **univariaten Verteilungsparametern** sind in der statistisch-empirischen Forschung in der Regel auch **bivariate Verteilungsparameter** (z. B. Korrelationen) sowie **Modellparameter** (z. B. Regressionskoeffizienten) zu schätzen, was auch in unserem Kurs bzw. Demonstrationsprojekt geschehen wird.

Bei allen wichtigen Paramterschätzungen (z.B. zu einem Erwartungswert oder Regressionskoeffizienten) sind die Vertrauensintervalle gefragt.

2.1.2.2 Hypothesenprüfung

Bei vielen Forschungsprojekten steht die Prüfung von vorab formulierten Hypothesen in Vordergrund, die sich aus theoretischen Analysen und/oder vorherigen, eher explorativ ausgerichteten Studien ergeben haben. Einige Beispiele:

- Mit der Entscheidungsbefugnis eines Mitarbeiters steigt seine Arbeitszufriedenheit.
- Das Ausmaß an sozialer Ungleichheit in einer Gesellschaft hat einen negativen Effekt auf das Bruttoinlandsprodukt.
- Durch Beweidung von Wiesenflächen erhöht sich die pflanzliche Artenvielfalt.

Im aktuell betrachteten medizinischen Beispiel ist u. a. die folgende Hypothese zu prüfen:

> Bei Rauchern ist das Lungenkrebsrisiko höher als bei Nichtrauchern.

Genau genommen sind bei einem statistischen Entscheidungsproblem stets *zwei komplementäre* Hypothesen im Spiel:

---

[1] https://www.forschungsgruppe.de/Aktuelles/Politbarometer/ (abgerufen am 20.03.2022)

- **Alternativhypothese (H_1)**
  Durch die Alternativhypothese werden in der Regel Populationsverhältnisse beschrieben, die man belegen möchte.
  Im medizinischen Beispiel wurde eben die Alternativhypothesenformulierung präsentiert.
- **Nullhypothese (H_0)**
  Durch die Nullhypothese werden in der Regel Populationsverhältnisse beschrieben, die man als falsch zurückweisen möchte.
  Im Beispiel: Das Lungenkrebsrisiko ist bei Rauchern *nicht* höher als bei Nichtrauchern.

Kann in einem Forschungsprojekt die Nullhypothese aufgrund ihrer geringen Plausibilität gegeben die beobachteten Daten verworfen werden, dann gewinnt die Alternativhypothese. Ansonsten wird die Nullhypothese beibehalten.

In den meisten Anwendungsfällen, sollte (wie im medizinischen Beispiel) ein **gerichtetes Hypothesenpaar** formuliert werden. Das ermöglicht einen gerichteten (einseitigen) Signifikanztest, und ein vorhandener Effekt (z.B. ein Mittelwertsunterschied) kann mit einer höheren Wahrscheinlichkeit entdeckt werden als durch einen ungerichteten bzw. zweiseitigen Test, der bei einem ungerichteten Hypothesenpaar angewendet werden muss.

Die Testentscheidung läuft so ab:

1) Für eine aus den Stichprobendaten berechenbare Teststatistik (Prüfgröße) *T*, die indikativ ist für Abweichungen der wahren Populationsverteilung von der Nullhypothesenbehauptung, wird die Stichprobenrealisation T_{emp} ermittelt.
2) Man bestimmt die sogenannte *Überschreitungswahrscheinlichkeit p*, bei gültiger Nullhypothese den Wert T_{emp} oder eine der Nullhypothese noch stärker widersprechende Ausprägung der Teststatistik zu erhalten.
3) Man vergleicht die Überschreitungswahrscheinlichkeit *p* mit der akzeptierten Irrtumswahrscheinlichkeit erster Art (mit dem sogenannten α-Fehlerrisiko, z. B. $\alpha = 0{,}05$):
   - Bei $p < \alpha$ wird die Nullhypothese abgelehnt, und man entscheidet sich für die Alternativhypothese. In dieser Situation liegt ein empirisches Ergebnis vor, das beim Ziehen einer Zufallsstichprobe aus einer Nullhypothesenpopulation nur sehr selten auftreten würde.
   - Bei $p \geq \alpha$ wird die Nullhypothese beibehalten, womit sie aber nicht bewiesen ist.

Bei einem korrekt durchgeführten Hypothesentest zeigt sich die oben versprochene Weisheit der statistischen Methodologie in der Realisation der folgenden Ziele:

- Eine *wahre* Nullhypothese wird nur mit der zuvor akzeptierten Irrtumswahrscheinlichkeit erster Art (α-Fehlerrisiko) fälschlicherweise verworfen.
- Eine *falsche* Nullhypothese wird mit einer möglichst großen Wahrscheinlichkeit verworfen (hohe Teststärke, kleine Irrtumswahrscheinlichkeit zweiter Art, geringes β-Fehlerrisiko). Zu einer konkreten Effektstärke in der Population kann die Wahrscheinlichkeit für ein signifikantes Testergebnis bei gegebener Stichprobengröße ermittelt werden.

Das eben nur kurz skizzierte binäre Entscheidungskonzept von Neyman und Pearson wird in Abschnitt 8.1 ausführlicher behandelt.

Auch zur Klärung der Frage, welcher Stichprobenumfang erforderlich ist, um einen Populationseffekt bestimmter Größe (z. B. Risiko(Raucher) = 0,3 und Risiko(Nichtraucher) = 0,05) durch

einen statistischen Test zum α-Fehlerrisiko 0,05 mit einer Wahrscheinlichkeit von 0,95 (d. h. bei einem β-Fehlerrisiko von 0,05) entdecken zu können, stehen statistische Methoden bereit (siehe Abschnitt 2.3).

Im Raucherbeispiel wird übrigens die Nullhypothese aufgrund der oben präsentierten Stichprobendaten deutlich verworfen ($p < 0{,}001$). Eine Stichprobenumfangskalkulation ergibt, dass bei der eben beschriebenen Effektlage (mit den Risiken 0,3 und 0,05) sowie einem Stichprobenanteil der Raucher von 0,25 bereits 155 Fälle genügen, um mit der gerichteten Variante des exakten Tests von Fisher zum Niveau $\alpha = 0{,}05$ den vorhandenen Effekt mit der Wahrscheinlichkeit 0,95 entdecken zu können (vgl. Abschnitt 14.4.5.1). Allerdings ist die Stichprobengröße von 400 keine sinnlose Fehlinvestition, weil z. B. das geschätzte 95% - Vertrauensintervall zum Lungenkrebsrisiko bei Rauchern ([0,217; 0,395]) für die meisten praktischen Zwecke zu ungenau ist. Man sollte also bei der Stichprobenumfangplanung nicht ausschließlich die Power eines zentralen Hypothesentests berücksichtigen, sondern z. B. auch die benötigte Präzision von wichtigen Vertrauensintervallen (vgl. Abschnitt 5.8.3).

2.1.2.3 Modellierung

Die zu schätzenden bzw. auf Signifikanz zu prüfenden Parameter stammen oft aus einem *mathematischen Modell*, das auf ein *empirisches System* angewendet wird. Im Raucherbeispiel kommt ein sehr einfaches Modell mit zwei binomialverteilten (dichotomen) Merkmalen zu Einsatz, das kaum als Modell in Erscheinung tritt. Daher betrachten wir an dieser Stelle noch das häufig verwendete Modell der *linearen Regression*, das in seiner einfachsten Form ein zu erklärendes metrisches Kriterium *Y*, einen metrischen oder dichotom-kategorialen Regressor *X* und ein normalverteiltes Residuum ε enthält:

$$Y = \beta_0 + \beta_1 X + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$$

Als Modellparameter treten der Ordinatenabschnitt β_0 und der Steigungskoeffizient β_1, die gemeinsam eine Regressionsgerade festlegen, sowie die Varianz σ^2 des Residuums auf. Mit der Formel

$$\varepsilon \sim N(0, \sigma^2)$$

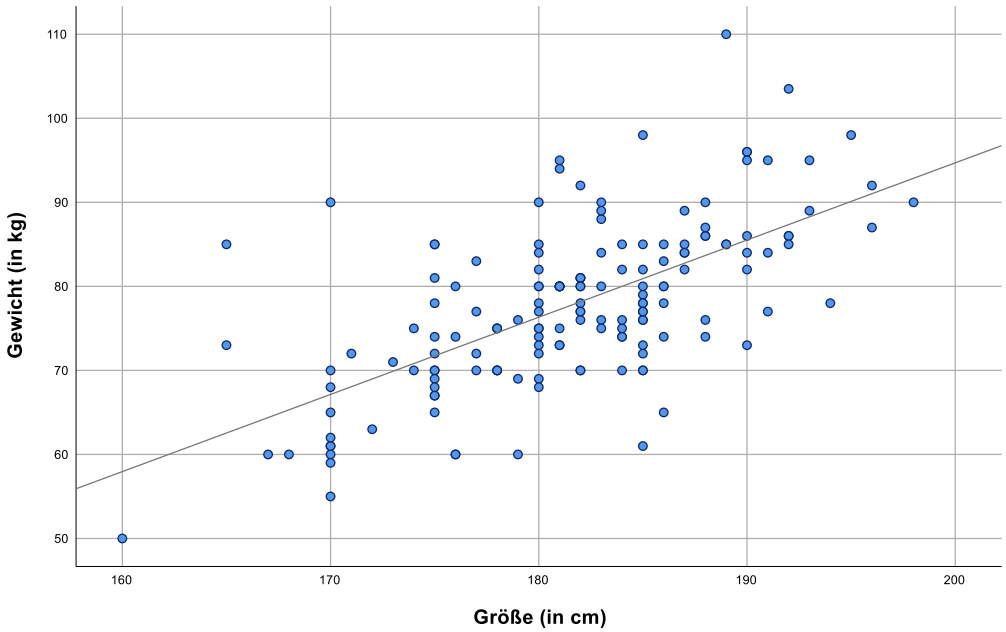
wird die Annahme formuliert, das Residuum ε folge für jede Beobachtung (also insbesondere für beliebige Werte des Regressors) einer Normalverteilung mit dem Mittelwert 0 und der Varianz σ^2.

Modellgemäß wird für Fälle mit dem konkreten Wert *x* des Regressors die mittlere *Y*-Ausprägung $\beta_0 + \beta_1 x$ erwartet, wobei die tatsächlichen Beobachtungswerte um diesen Erwartungswert herum normalverteilt sind mit der Varianz σ^2.

Das Modell der bivariaten linearen Regression eignet sich z. B. dazu, den Einfluss der Körpergröße von Personen auf das Körpergewicht zu beschreiben. Im folgenden Streudiagramm ist die gemeinsame empirische Verteilung der beiden Merkmale in einer Stichprobe mit 159 Männern zu sehen:[1]

---

[1] In diesem Beispiel hat die Beschränkung auf eine „Männergesellschaft“ statistische bzw. didaktische Gründe: In der Gesamtpopulation (mit Frauen und Männern) wird die Regression von Gewicht auf Größe nach einer im Kursverlauf noch zu prüfenden ernährungsphysiologischen Hypothese vom Geschlecht moderiert. Weil Frauen und Männer im Mittel unterschiedlich robust gebaut sind, wächst das Gewicht eventuell unterschiedlich schnell mit der Größe. Um diese Verhältnisse korrekt zu modellieren, ist ein relativ anspruchsvolles Modell mit Interaktionsterm erforderlich (siehe z. B. Baltes-Götz 2019a), das zum aktuellen Abschnitt nicht gut passen würde.



Offenbar ist im Beispiel die von den beiden Modellparametern β_0 (Schnittpunkt mit der Y-Achse) und β_1 (Steigung) definierte Regressionsgerade gut geeignet, den erwarteten *Y*-Wert für eine gegebene *X*-Ausprägung vorherzusagen. Die empirischen Residuen (Abstände der beobachteten Kriteriumswerte von der Regressionsgeraden in Y-Richtung) pendeln relativ varianzhomogen um die Modellprognosen (auf der Regressionsgerade), sodass man von einem gültigen Modell ausgehen darf.[1] Unter dieser Voraussetzung sind die Punktschätzung, das Konfidenzintervall und der Signifikanztest zum Steigungsparameter β_1 von zentralem Interesse. Bei einem gerichteten (einseitigen) Signifikanztest ist das folgende Entscheidungsproblem zu klären:

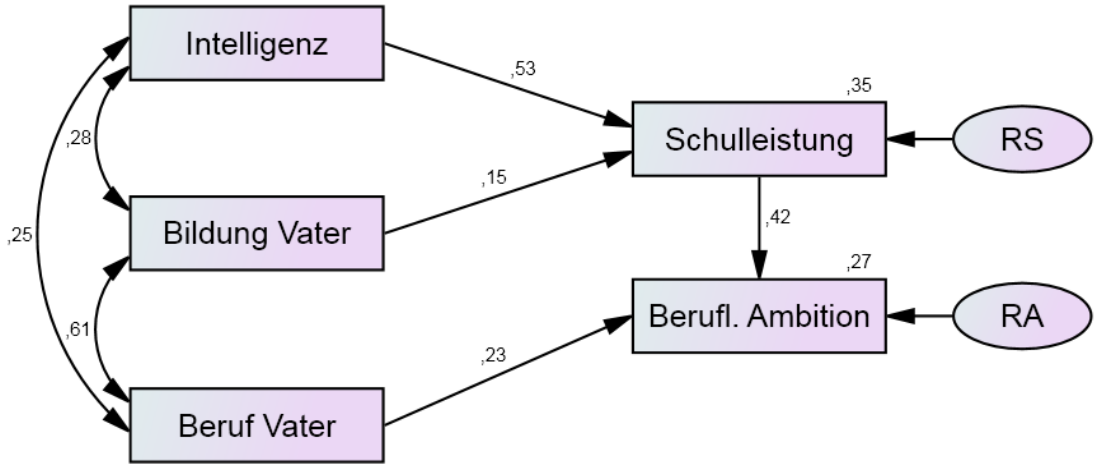
$$H_0: \beta_1 \leq 0 \text{ versus } H_1: \beta_1 > 0$$

Im Unterschied zum gerade präsentierten Beispiel sind bei den meisten Fragestellungen *mehrere* Einflussgrößen zu untersuchen und z. B. in ein *multiples* lineares Regressionsmodell aufzunehmen:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$$

Ein typisches Forschungsprojekt wird zudem nicht bei der Untersuchung *eines* Kriteriums stehenbleiben, sondern ein Modell zur Beschreibung und Erklärung eines empirischen Systems anstreben. Man kann z. B. versuchen, das Zusammenwirken aller relevanten Komponenten durch ein aus mehreren Gleichungen bestehendes Pfad- oder Strukturgleichungsmodell approximativ zu beschreiben. Dabei sind mehrere Parameterschätzungen und Hypothesentests beteiligt. Das folgende Pfadmodell zu den Bedingungen der beruflichen Ambitionen von Kindern (Kerchoff 1974)

[1] Eine Beurteilung der Residuen auf Normalverteilung ersparen wir uns an dieser Stelle (vgl. Abschnitt 10.1.1).



enthält ...

- ein Kriterium (berufliche Ambitionen eines Kindes),
- einen *Mediator* (Schulleistung des Kindes)
- sowie drei exogene Regressoren:
    - Intelligenz des Kindes
    - Bildung des Vaters
    - Beruf des Vaters

In diesem Beispiel ist die völlige Fixierung auf väterliche Einflüsse wohl durch die im letzten Jahrtausend noch mangelhafte Emanzipation zu erklären.

Das Pfadmodell besteht aus zwei Regressionsgleichungen:

- Die Schulleistung des Kindes wird zurückgeführt auf seine Intelligenz und die Bildung des Vaters.
- Die berufliche Ambition des Kindes wird erklärt durch seine Schulleistung und den Beruf des Vaters.

Modelle mit Mediatoren dienen zur Klärung der Frage, *wie* unabhängige Variablen ihre Wirkung auf eine abhängige Variable ausüben, welche vermittelnden Prozesse beteiligt sind (siehe z. B. Baltes-Götz 2020a; Hayes 2018).

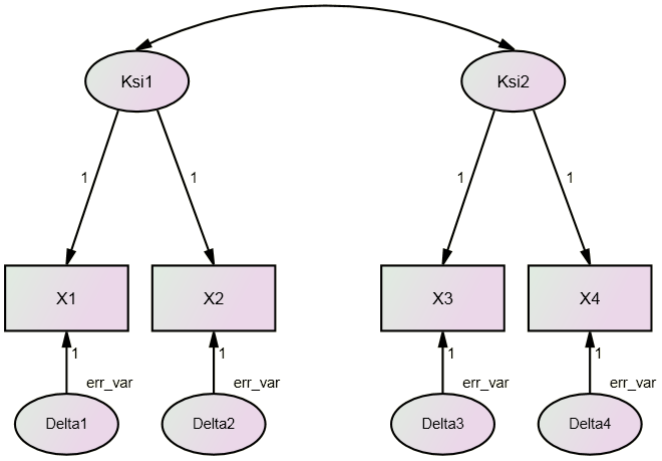
Ein Forschungsprojekt mit dem Ziel eines allgemeingültigen Modells wird meist eingestehen müssen, eine *vereinfachte* Sicht zu liefern, was beim Statistiker **George E. P. Box** zur folgenden desillusionierten, aber keinesfalls resignierten Einsicht geführt hat (Box 1979, S. 202):

All models are wrong but some are useful.

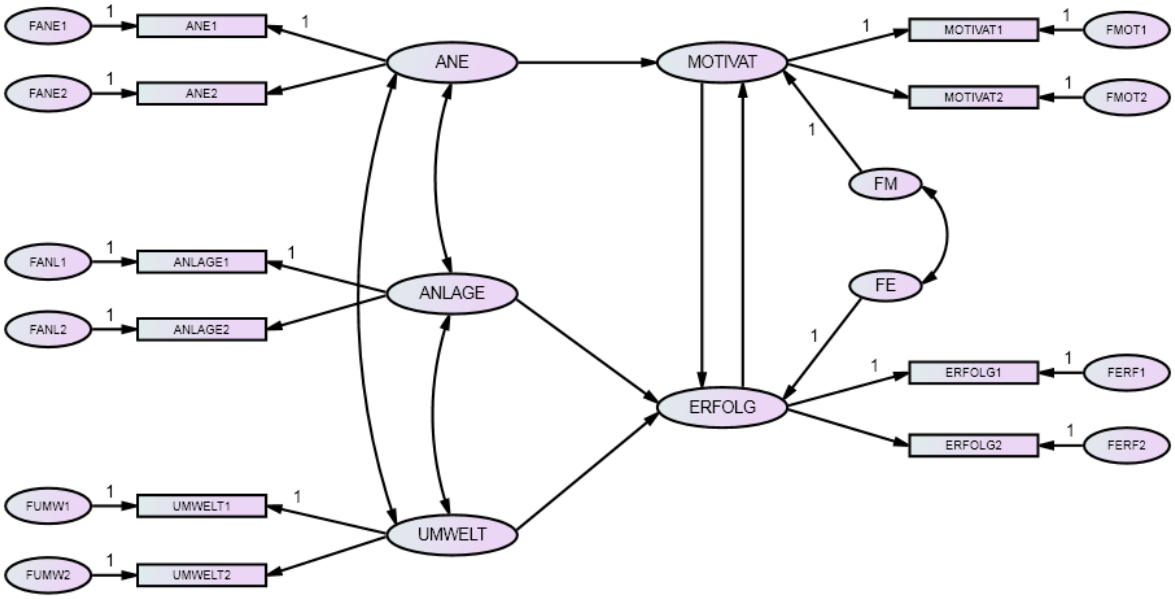
2.1.2.4 Weitere statistische Forschungsmethoden

Das statistische Methodenarsenal deckt neben der Parameterschätzung, Hypothesenprüfung und Modellierung noch weitere Aufgabenfelder ab, z. B.:

- Mit den Verfahren der **explorativen** oder **konfirmatorischen Faktorenanalyse** verfolgt man das Ziel, die hinter einer Anzahl von manifesten Variablen wirkenden und statistische Relationen stiftenden latenten Variablen zu identifizieren (siehe z. B. Bühner 2011; Eid et al. 2017). Während die explorative Faktorenanalyse von SPSS Statistics unterstützt wird, ist für die konfirmatorische Faktorenanalyse in der IBM/SPSS - Produktfamilie das Programm Amos zuständig. Hier ist ein simples, mit Amos erstelltes Modell aus dem Bereich der konfirmatorischen Faktorenanalyse zu sehen (Baltes-Götz 2015, S. 57ff):



- Werden für latente Variablen nicht nur Kovarianzen zugelassen, sondern auch Effekte modelliert, dann geht die konfirmatorische Faktorenanalyse in eine **Strukturgleichungsanalyse** über. Auch für diese flexible Kombination von dimensions- und regressionsanalytischen Verfahren ist in der IBM/SPSS - Produktfamilie das Programm Amos zuständig. Das folgende Beispiel enthält zwei Mediatoren, die sich wechselseitig beeinflussen (Baltes-Götz 2015, S. 82ff):[1]



[1] Modelle mit der (direkten oder indirekten) wechselseitigen Beeinflussung von zwei Merkmalen sind in der aktuellen Sozialforschung sehr selten anzutreffen und werden auch im Kurs nicht mehr auftauchen.

- Mit der **Clusteranalyse** oder der **latenten Klassenanalyse** (engl.: *latent class analysis*) sucht man nach einer taxonomischen Ordnung für einen Gegenstandsbereich. Während SPSS einige Varianten der Clusteranalyse unterstützt, wird für die (bislang selten eingesetzte) latente Klassenanalyse eine alternative Software benötigt (z.B. Mplus oder das R-Paket poLCA).
- Liegen zu einer Hypothese bereits zahlreiche Forschungsergebnisse vor, dann kann eine quantitative **Metaanalyse** indiziert sein, um aus der gesamten Befundlage kluge Schlussfolgerungen abzuleiten (siehe z. B. Urban & Fiebig 2015). Seit der Version 28 unterstützt SPSS quantitative Metaanalysen.

2.1.3 Lügen

Viel bekannter als die oben zitierte Aussage von Wallis & Roberts (1956) zum Wesen der Statistik ist leider die folgende, Mark Twain oder Benjamin Disraeli zugeschriebene Charakterisierung (Griffith 2010, S. 9):

Es gibt Lügen, verdammte Lügen und Statistik.

In dieser Formulierung kommt die reale Gefahr zum Ausdruck, dass Menschen bei der Anwendung statistischer Methoden auf komplexe, durch Unsicherheit geprägte Sachverhalte aufgrund unredlicher Motive oder durch Unvermögen scheitern können. Dabei resultieren fehlerhafte Forschungsergebnisse, die wiederum zu falschen Entscheidungen führen können. Dieser Gefahr ist durch statistische Sachkompetenz und ein gesundes Misstrauen zu begegnen.

2.1.4 Keine statistische Praxis ohne Informationstechnologie (IT)

Kehren wir kurz zum medizinischen Beispiel zurück. In einer realen Studie wird man sich nicht auf die beiden oben zur Illustration verwendeten dichotomen Merkmale (*Rauchen* und *Lungenkrebs*) beschränken, sondern Dauer und Ausmaß des (aktiven und passiven) Rauchens sowie den gesundheitlichen Status der Probanden genauer untersuchen und außerdem viele zusätzliche Merkmale erheben, z. B. Alter, Geschlecht, Bildung, Beruf, Schadstoffbelastung am Arbeitsplatz und in der Wohnung (z. B. Radon-Gas in der Raumluft). Eine praktikable Auswertung solcher Datenmengen ist nur mit IT-Hilfe möglich. Mit SPSS steht ein bequemes, leistungsfähiges und sehr bewährtes Analysesystem für die statistische Forschung zur Verfügung. Es bietet u. a. …

- fast alle wichtigen statistischen Verfahren
- gute grafische Darstellungsmöglichkeiten
- eine umfangreiche Unterstützung bei der Datenverwaltung und -aufbereitung
- diverse Verfahren zur Kooperation mit anderen Programmen (z. B. Programmierschnittstellen, Datenbankzugriff)

Weil SPSS auf allen wichtigen Plattformen (Linux, macOS, Windows) vertreten ist, und sein Datendateiformat weithin unterstützt wird, bestehen günstige Bedingungen für die kollegiale Kommunikation.

2.2 Planung und Durchführung einer empirischen Studie im Überblick

In diesem Abschnitt wollen wir uns einen Überblick über die Phasen einer empirischen Studie und damit auch über unser Kursprogramm verschaffen. Dabei werden zahlreiche Aufgaben und Methoden angesprochen, über die Sie sich im Bedarfsfall in den Lehrveranstaltungen oder in der Literatur zur empirischen Forschung informieren können (siehe z. B. Bortz & Döring 2016; Eid et al. 2017; Jacob et al. 2013, Pedhazur & Pedhazur Schmelkin 1991; Schnell et al. 2018; Wooldridge 2013). Die anschließende Darstellung ist relativ knapp gehalten. Ihr folgt unmittelbar die konkrete Anwendung auf unsere Beispielstudie.

Weil die dargestellten Aufgaben teilweise interdependent sind, existiert keine strenge, bei allen empirischen Studien gleichförmig und unidirektional ablaufende Sequenz.

2.2.1 Forschungsziele, Hypothesen und Modelle

Am Anfang einer wissenschaftlichen Studie steht in der Regel eine längere Phase der intensiven theoretischen Auseinandersetzung mit dem Thema. Daraus ergeben sich Forschungsinteressen, die – u. a. in Abhängigkeit vom Forschungsstand – eher von *explorativer* oder eher von *konfirmatorischer* Natur sind:

- Die explorative Forschung startet mit der Beschreibung empirischer Sachverhalte und führt oft zur Formulierung von Hypothesen bzw. Modellen.
- Die konfirmatorische Forschung überprüft Hypothesen bzw. Modelle.

Oft werden beide Forschungsstrategien vertreten sein. Die zu prüfenden Hypothesen sollten wegen ihrer Steuerungsfunktion für spätere Schritte möglichst exakt formuliert werden. Häufig werden sich die Hypothesen auf Parameter in einem **mathematischen Modell** (z. B. in einer linearen Regressionsgleichung) beziehen.

Um einen Gegenstandsbereich empirisch zu explorieren oder eine Theorie (ein Modell) bzw. eine Hypothesenfamilie empirisch zu prüfen, werden geeignete Daten benötigt, die aus vorhandenen Beständen und/oder aus eigener Forschungstätigkeit stammen können.

2.2.2 Nutzung vorhandener Daten

Besteht zur Klärung einer empirischen Forschungsfrage die Möglichkeit zur (Sekundär-)Analyse von vorhandenen Daten, sollte diese Chance aus wirtschaftlichen Gründen genutzt werden. Wichtige Bezugsquellen für öffentlich zugängliche sozial- und wirtschaftswissenschaftliche Datenbestände sind:

- **Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS)**
  Das GESIS-Institut für die Sozialwissenschaften befragt seit 1980 alle zwei Jahre eine repräsentative (jeweils neu gezogene) Stichprobe aus der erwachsenen deutschen Bevölkerung.[1] Seit 1986 geschieht dies im Rahmen des internationalen Kooperationsprogramms *International Social Survey Programme* (ISSP).

[1] http://www.gesis.org/allbus,
http://de.wikipedia.org/wiki/Allgemeine_Bevölkerungsumfrage_der_Sozialwissenschaften

- **Sozio-oekonomisches Panel (SOEP)**
  Das Deutsche Institut für Wirtschaftsforschung (DIW) führt seit 1984 eine jährliche sozio-oökonomische Befragung bei einer repräsentativen und (im Rahmen der organisatorischen Möglichkeiten) konstanten Stichprobe aus der deutschen Bevölkerung durch.[1]
- **Statistisches Bundesamt**
  Das statistische Bundesamt bietet z. B. in der GENESIS - Online-Datenbank (*GEmeinsames NEues Statistisches Informations-System*) aggregierte Daten aus diversen Bereichen der amtlichen deutschen Statistik an (z. B. volkswirtschaftliche Indizes, Bildung).[2]
- **Nationales Bildungspanel (NEPS)**
  Das Leibnitz-Institut für Bildungsverläufe führt im NEPS-Projekt (*Nationales Bildungspanel*) sechs Panelstudien mit Startkohorten unterschiedlichen Alters durch (von Neugeborenen bis zu Erwachsenen) und stellt der wissenschaftlichen Gemeinschaft die Daten zur Verfügung.[3]
- **General Social Survey (GSS)**
  Seit 1972 wird jährlich eine (jeweils neu gezogene) Stichprobe der US-Amerikanischen Bevölkerung zu einer breiten Themenpalette befragt.[4] Das GSS-Trägerinstitut gehörte zu den ISSP-Gründungsmitgliedern, und die ISSP-Module sind dementsprechend in den GSS-Fragenkatalog integriert.
- **European Social Survey (ESS)**
  Seit 2002 werden alle 2 Jahre in über 30 Europäischen Ländern bei jeweils neu gezogenen Stichproben sozialwissenschaftliche Daten erhoben. Es werden erwachsene Personen ab 15 Jahren befragt, bevorzugt in persönlichen Interviews.[5]
- **World Values Survey (WVS)**
  Seit 1981 werden in fast 100 Ländern, in denen insgesamt ca. 90% der Weltbevölkerung leben, mit einem gemeinsamen Fragebogen repräsentative Studien zum Wertewandel durchgeführt.[6]

Weitere Details zu den genannten und zu anderen Datenquellen finden sich z. B. in Baur & Fromm (2011, S. 61ff), Jacob et al. (2013, S. 237ff) sowie in Vlaeminck et al. (2015).

Etliche Webseiten pflegen Listen mit Datenquellen-Verzeichnissen, z. B.:

- https://www.nature.com/sdata/policies/repositories#social
- https://psychologie.de/forschung/studien-durchfuehren/forschungsdaten-repositorien/

Neben den für Forschungszwecke systematisch aufgebauten Datenbeständen und den amtlichen Statistiken existieren (z. B. in Behörden oder Firmen) weitere Aufzeichnungen von Interesse für spezielle Forschungsfragen.

---

1 http://www.diw.de/de/soep,
http://de.wikipedia.org/wiki/Sozio-oekonomisches_Panel

2 https://www-genesis.destatis.de/genesis/online

3 https://www.neps-data.de/

4 http://gss.norc.org/About-The-GSS
http://en.wikipedia.org/wiki/General_Social_Survey

5 http://www.europeansocialsurvey.org/about/
https://de.wikipedia.org/wiki/European_Social_Survey

6 http://www.worldvaluessurvey.org/WVSContents.jsp

Bei der Suche nach statistischen Daten aus diversen Quellen hilft das Statistikportal **Statista**. Um den von der Universitätsbibliothek Trier abgeschlossenen Campusvertrag beim Recherchieren nutzen zu können, muss die Verbindung zur Statista-Webseite (http://www.statista.com/) im Campusnetz der Uni Trier bzw. bei aktiver VPN-Verbindung mit dem Campusnetz von der Homepage der Uni Trier (http://www.uni-trier.de/) ausgehend folgendermaßen hergestellt werden:

Bibliothek > Suchen und finden > Datenbanken (DBIS) >
Fachliste > Wirtschaftswissenschaften > Statista

2.2.3 Untersuchungsplanung

Ist eine komplette eigene Studie (inkl. Datenerhebung) erforderlich, dann sind bei der Untersuchungsplanung zahlreiche Aufgaben zu lösen.

2.2.3.1 Untersuchungseinheiten, Population und Merkmale

In der Regel ergibt sich aus der Fragestellung unmittelbar, welche Untersuchungseinheiten (Merkmalsträger) in eine Studie einbezogen werden sollen (z. B. Personen, Volkswirtschaften, Planeten, Unternehmen, Pflanzschalen), und welche Merkmale bei jeder Untersuchungseinheit festgestellt werden sollen. Gelegentlich kommen aber z. B. zur Prüfung einer medizinischen Theorie sowohl Personen, als auch Tiere oder Zellkulturen als Untersuchungseinheiten in Frage.

Eng verknüpft mit der Wahl von Untersuchungseinheiten ist die Festlegung einer **Grundgesamtheit** bzw. **Population** von Untersuchungseinheiten, über die allgemeingültige Aussagen gewonnen werden sollen. Hier geht es um die angestrebte Generalisierbarkeit (externe Validität) der Forschungsergebnisse.

Gelegentlich sind **hierarchisch geschachtelte Untersuchungseinheiten** zu betrachten (siehe z. B. Baltes-Götz 2020b; Raudenbush & Bryk 2002; Snijders & Bosker 2012), wobei man auch von *Cluster-Stichproben* spricht. So hat man es etwa bei einer Studie zur Arbeitszufriedenheit und Produktivität von Arbeitnehmern aus verschiedenen Firmen in Abhängigkeit von Person- und Organisationsmerkmalen mit Untersuchungseinheiten und jeweiligen Merkmalen auf *zwei* Ebenen zu tun:

- Firmen
- Arbeitnehmer

Es können auch mehr als zwei Ebenen beteiligt sein (z. B. bei Schülern in Klassen, die zu Schulen gehören, welche sich in verschiedenen Ländern befinden). Wir beschränken unsere Betrachtungen an dieser Stelle auf Studien mit *zwei* Ebenen und können daher bequem von der *Makro-* und der *Mikroebene* sprechen. Bei der statistischen Auswertung ist zu beachten, dass traditionelle Methoden (z. B. die lineare Regressionsanalyse) *unabhängige Residuen* voraussetzen. Die bei einer hierarchischen Datenstruktur auf der Mikroebene (im Beispiel: auf der Ebene der Arbeitnehmer) naturgemäß anzutreffende Abhängigkeit der Beobachtungen aus derselben Makroeinheit muss in speziellen Modellen berücksichtigt werden, um gültige Vertrauensintervalle und Hypothesentests zu erhalten.

Das Demonstrationsprojekt im Kurs kommt mit einer flachen Datenstruktur aus, und die Behandlung der speziellen Optionen und Probleme von Cluster-Stichproben bleibt speziellen ZIMK-Kursen bzw. Manuskripten vorbehalten.[1]

2.2.3.2 Untersuchungsdesign

Man kann z. B. einen (quasi-)experimentellen Untersuchungsplan entwerfen oder eine Beobachtungsstudie konzipieren, die quer- und/oder längsschnittlich angelegt sein kann. Bei einer Beobachtungsstudie werden Merkmalsausprägungen festgestellt, doch finden keine (starken) Eingriffe in das zu analysierende empirische System statt. Sind die Beobachtungseinheiten Personen, dann werden oft schriftliche Befragungen, seltener Interviews durchgeführt.

Auf dem Weg zu der im Regelfall anzustrebenden kausalen Interpretation von Forschungsergebnissen gilt das Experiment nach wie vor als die ideale Methode (siehe z. B. Antonakis, et al. 2010; Eid et al. 2017), doch muss man z. B. in den Wirtschafts- und Sozialwissenschaften oft aus ehtischen und/oder praktischen Erwägungen mit Beobachtungsdaten arbeiten.

Bei (quasi-)experimentell manipulierten Bedingungen hat man gelegentlich die Wahl zwischen einem **Messwiederholungs- und einem Gruppierungsfaktor** (engl.: *within-* versus *between-subjects factor*). Bei einem Messwiederholungsfaktor wird jeder Fall nacheinander unter *mehreren* Bedingungen beobachtet (z. B. Gedächtnisleistung von Personen mit bzw. ohne Medikament zur zerebralen Durchblutungsförderung). Es resultieren entsprechend viele Messwerte, und interindividuelle Unterschiede können aus der Fehlervarianz des Analysemodells herausgehalten werden, sodass eine günstige Teststärke bei der Beurteilung des Faktors zu erwarten ist. Allerdings sind in der Regel Übertragungen zwischen den Bedingungen zu befürchten (z. B. durch Lern- oder Ermüdungseffekte), die oft durch eine zufällige Variation der Bedingungsreihenfolge neutralisiert werden können. Bei einem Gruppierungsfaktor wird jede Untersuchungseinheit nur unter *einer* Bedingung beobachtet, wobei eine Zufallszuordnung anzustreben ist, damit ein randomisiertes Experiment mit seinem kausalitätstheoretischem Interpretationsvorteil entsteht. Interindividuelle Unterschiede erschweren als Fehlervarianzquelle den Nachweis von Effekten. Erfolgt keine zufällige Zuweisung der Untersuchungseinheiten zu den Bedingungen, dann liegt ein Quasi-Experiment vor.

---

[1] Derzeit sind zwei Technologien zur Analyse von Cluster-Stichproben verbreitet:

- (Generalisierte) lineare gemischte Modelle
  Ein *lineares gemischtes Modell* (LMM, *Linear Mixed Model*) erklärt für ein metrisches Kriterium mit normalverteilten Residuen die Abhängigkeit der aus einem Cluster stammenden Beobachtungen durch clusterspezifische Zufallseffekte (siehe z. B. Baltes-Götz 2020b; Raudenbush & Bryk 2002; Snijders & Bosker 2012). Aus dem LMM entsteht das *generalisierte lineare gemischte Modell* (GLMM), wenn man …
  a) für die Residuen statt der Normalverteilung auch andere Verteilungen (z. B. die Binomial- oder die Poisson-Verteilung) zulässt,
  b) als Verbindung zwischen dem erwarteten Kriteriumswert und einer Prädiktorwertekombination neben der Identität auch andere Link-Funktionen erlaubt (z. B. die Logit- oder die Logarithmusfunktion).
- GEE-Modelle (*Generalized Estimating Equations*)
  Während bei einem (generalisierten) linearen gemischten Modell die Kovarianzmatrix der Beobachtungen zum Explanandum gehört und durch die Zufallseffekte im statistischen Modell erklärt werden soll, betrachtet die von Liang & Zeger (1986) eingeführte GEE-Methodologie (*Generalized Estimating Equations*) die Abhängigkeit der Beobachtungen als lästige, durch geeignete Maßnahmen zu kompensierende Störung (siehe z. B. Baltes-Götz 2016b). Ein GEE-Modell erlaubt bzgl. der Residualverteilung und der Link-Funktion dieselben Generalisierungen wie ein GLMM.

Unter den Beobachtungsstudien bietet die **Panelsstudie** relativ günstige Bedingungen für eine kausale Interpreration der Ergebnisse (siehe z. B. Baltes-Götz 2016a, Brüderl 2010). Dabei werden *mehrere Untersuchungseinheiten* zu *mehreren Zeitpunkten* beobachtet. In der Regel ist die Anzahl *N* der Untersuchungseinheiten relativ groß (z. B. 300) und die Anzahl *T* der Zeitpunkte relativ klein (z. B. 5).

2.2.3.3 Operationalisierung der zu untersuchenden Merkmale

Zur Operationalisierung von theoretischen Begriffen (z. B. sozioökonomischer Status, Optimismus) sind möglichst reliable und valide Messmethoden zu wählen bzw. zu entwerfen, die außerdem nicht zu aufwändig sind. Das Skalenniveau der Messmethoden muss die Voraussetzungen der geplanten statistischen Auswertungsverfahren (siehe unten) erfüllen.

Bei *quantitativen* Merkmalen (z. B. Alter) sollten die verfügbaren Informationen bei der Operationalisierung *nicht* durch eine *willkürliche* Klassenbildung reduziert werden (z. B. durch Bildung der Altersklassen < 20, 21 - 40, 41 - 60, > 60). Man bewegt sich je nach Wahl der Klassengrenzen vom metrischen Messniveau mehr oder weniger weit weg in Richtung auf ein geordnet-kategoriales Niveau. Häufig sind Modelle für metrische Daten einfacher und erfolgreicher als solche für geordnet-kategoriale Daten. Außerdem kann man mit SPSS zu einer metrischen Variablen nach Belieben klassifizierte Varianten erzeugen, wenn dies für spezielle Analysen wünschenswert erscheint. Eine Ausnahme von der Empfehlung zur Erhebung metrischer Informationen ist z. B. bei der Befragung von Personen nach ihrem Einkommen zu machen. Um bei dieser sensiblen Frage Widerstände zu vermeiden, muss man sich in der Regel auf die Erhebung von groben Einkommensklassen und damit auf das geordnet-kategoriale Messniveau beschränken.

Bei den Überlegungen zur Operationalisierung können auch die zur Datenerhebung verfügbaren technischen Hilfsmittel eine Rolle spielen. Mit Hilfe der Computer-Technik ist eine interaktive, individualisierte und dabei auch noch ökonomische Datenerhebung möglich. Zur Steuerung experimenteller Abläufe oder zur hochgenauen Messung von Reaktionszeiten werden traditionell spezielle Rechner im Forschungslabor verwendet. Neuerdings kommen aber auch Internet-basierte Lösungen für solche Zwecke zum Einsatz (siehe z. B. Kim, Gabriel & Gygax 2019). Für eine kontinuierliche, alltagsbegleitende Datenerfassung können oft Smartphones oder andere Rechner im Taschenformat genutzt werden. Schriftliche Befragungen werden mittlerweile routinemäßig via Internet realisiert, wenn die zu untersuchende Population auf diesem Weg erreichbar ist (siehe Abschnitt 3.2).

2.2.3.4 Empirisch prüfbare Hypothesen (über Modellparameter) formulieren

Aus einer in theoretischen Begriffen formulierten Hypothese ergibt sich im Verlauf der Untersuchungsplanung durch zahlreiche Konkretisierungen und Operationalisierungen eine in empirischen Begriffen formulierte und damit statistisch prüfbare Hypothese.

Von sehr einfachen Fällen abgesehen, werden sich die zu prüfenden Hypothesen einer Studie auf Parameter in einem **mathematischen Modell** beziehen, das also spätestens jetzt explizit zu formulieren ist.

Oft hat man die Wahl zwischen einer **gerichteten** und einer **ungerichteten Hypothesenformulierung**. Über den Steigungsparameter β_1 der bivariaten linearen Regression mit dem Kriterium *Y* und dem Regressor *X* (vgl. Abschnitt 2.1)

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

behauptet eine *gerichtete* Alternativhypothese z. B.

$$H_1: \beta_1 > 0$$

Dem steht die folgende Aussage der *ungerichteten* Alternativhypothese gegenüber:

$$H_1: \beta_1 \neq 0$$

Für die mangelnde Bereitschaft, sich auf eine Richtung festzulegen, wird man durch eine geringere Power beim Signifikanztest bestraft. Die Chance, einen vorhandenen Effekt zu entdecken, ist also geringer. Immerhin ist man berechtigt, nach einem signifikanten Ergebnis beim zweiseitigen Test doch noch eine Richtungsaussage zu machen.[1]

Selbstverständlich kann gelegentlich auch eine ungerichtete Hypothesenformulierung angemessen sein, weil z. B. theoretisches Vorwissen fehlt.

Manchmal ist eine gerichtete Formulierung prinzipiell unmöglich, z. B. beim Vergleich der Frauenanteile in den sechs Fachbereichen der Universität Trier (siehe Kapitel 14). Hier kommt nur die „zweiseitige“ Alternativhypothesenbehauptung ungleicher Frauenanteile in Frage.

2.2.3.5 Statistisches Entscheidungsverfahren

Für jede Hypothese ist ein statistisches Entscheidungsverfahren zu wählen, dessen Voraussetzungen an Skalenniveau und Verteilungsverhalten der beteiligten Merkmalsoperationalisierungen (voraussichtlich) erfüllt sind.

In der wirtschafts- und sozialwissenschaftlichen Forschung ist es eine verbreitete Praxis, Antworten von Personen auf einer Skala mit 5 oder mehr Stufen (oft als *Likert-Item* bezeichnet) als intervallskaliert zu behandeln, obwohl streng genommen nur ein geordnet-kategoriales Messniveau vorliegt. Für diese Praxis finden sich viele Befürworter (z.B. Sullivan & Artino Jr. 2013), aber auch Kritiker (z.B. Liddell & Kruschke 2018). Wer eine Merkmalsoperationalisierung mit Likert-Item - Format nicht als intervallskaliert behandeln möchte, kann zur Analyse z. B. die für geordnet-kategoriale Kriterien geeignete ordinale logistische Regression verwenden (siehe z. B. Baltes-Götz 2012).

Häufig wird aus einer Anzahl von Likert-Items ein Mittelwert berechnet und in der statistischen Analyse zur Messung des intendierten latenten Merkmals verwendet. Man spricht dann von einer *Likert-Skala*. In der methodologischen Literatur besteht eine weitgehende Übereinstimmung darin, dass für Item-Mittelwerte das metrische Messniveau beansprucht werden kann (siehe z. B. Warner 2013, S, 6ff).

Im Rahmen einer konfirmatorischen Studie ist zu jedem Hypothesentest das akzeptierte Fehlerrisiko erster Art (α-Fehler) festzulegen, also das Risiko, eine gültige Nullhypothese fälschlich zu verwerfen, wobei meist die 5% - Konvention übernommen wird.

[1] Dies garantiert ein mathematischer Satz über den sogenannten *Abschlusstest*.

2.2.3.6 Stichprobenrekrutierung

Es ist zu überlegen, wie eine *repräsentative* und zur aussichtsreichen Durchführung der geplanten Auswertungsverfahren *hinreichend große* Stichprobe aus der (möglichst präzise definierten) Population rekrutiert werden kann. *Alle* Fälle in der Population zu untersuchen, ist in der Regel weder realisierbar, noch erforderlich. Die Methoden der Inferenzstatistik machen es möglich, mit kontrollierter Irrtumswahrscheinlichkeit von Stichprobenergebnissen auf Populationsverhältnisse zu schließen.

Bei ausgeprägt konfirmatorisch angelegten Studien ist bei der Stichprobenumfangsplanung insbesondere das **Fehlerrisiko zweiter Art** (der β-Fehler) zu berücksichtigen, also das Risiko, einen vorhandenen Effekt zu übersehen.

Das Ideal einer streng zufallsgesteuerten Stichprobengewinnung, wobei aus einer klar definierten Population jedes Mitglied dieselbe Chance hat, in die Stichprobe zu gelangen, kann in den Wirtschafts- und Sozialwissenschaften nur selten realisiert werden. Oft wird eine **Bequemlichkeitsstichprobe** (engl. *convenience sample*) verwendet, in die bevorzugt leicht (bei minimalen Kosten) erreichbare Untersuchungseinheiten geraten (z. B. Personen in der Reichweite von (digitalen) sozialen Netzwerken oder Patienten in der Praxis eines forschenden Mediziners). In diesem Fall ist die externe Validität (Generalisierbarkeit) der Forschungsergebnisse in Frage gestellt. Z. B. dürfte ein erheblicher Teil der psychologischen Forschung mit Studierenden aus den Anfangssemestern durchgeführt werden, wobei in Bezug auf Alter, Bildung und andere soziodemografische Merkmal große Segmente der Population ausgeschlossen sind (Warner 2013, S. 4). Die Leser von Forschungsberichten müssen durch eine gründliche Beschreibung der Stichprobe und der Rekrutierungsprozedur über Einschränkungen der Generalisierbarkeit informiert werden.

Bei der Erforschung von *Merkmalszusammenhängen* (z. B. Effekt von Selbstvertrauen auf die Bereitschaft zur Unterstützung von Gewaltopfern) kann man hoffen, aus einer Stichprobe mit mangelnder Repräsentativität Ergebnisse zu gewinnen, die auch für die Gesamtbevölkerung von Relevanz sind. Wenn hingegen *univariate Verteilungseigenschaften* in einer Population untersucht werden sollen (z. B. Parteienpräferenzen, Arbeitslosenrate), dann *muss* eine strikte Zufallsauswahl oder eine stratifizierte Zufallsauswahl mit Gewichtungsverfahren zum Einsatz kommen, damit verwertbare Forschungsergebnisse resultieren.

2.2.3.7 Datendeklaration und Codierplan

Wer ganz sicher gehen will, dass die bei seiner Studie erhobenen Informationen sicher und bequem in die elektronische Datenverarbeitung übernommen werden können, sollte die Daten schon in der Planungsphase gegenüber der zuständigen Software deklarieren.

Beim Entwurf eines Formulars für eine Online-Erhebung geschieht die Datendeklaration gegenüber der verwendeten Software (also *vor* der Datenerhebung). Diese Software kann in der Regel die erfassten Merkmalsoperationalisierungen später in eine SPSS-Datendatei exportieren, sodass keine erneute Datendeklaration gegenüber SPSS erforderlich ist.

Auch in der heutigen Zeit werden noch viele Daten mit schriftlichen Untersuchungsdokumenten oder per Interview erhoben und anschließend manuell erfasst. Es kann nicht schaden, auch bei diesem Vorgehen die Daten schon *vor* der Erhebung gegenüber dem geplanten Erfassungsprogramm (z. B. SPSS-Dateneditor) zu deklarieren. Forschungsneulinge werden bei der Arbeit mit einem Computer-Programm, das die vorwiegend forschungslogisch und kaum durch IT-

Restriktionen diktierte Datenstruktur explizit einfordert, konzeptionelle Probleme eher entdecken als bei der schriftlichen Beschreibung des Forschungsvorhabens.

Bei den meisten Forschungsprojekten können die Daten in *einer* Tabelle (Matrix) mit den Fällen als Zeilen und den Merkmalen als Spalten untergebracht werden. Gelegentlich werden *mehrere* Tabellen benötigt, z. B. bei einer Untersuchung von Mitarbeitern und Kunden einer Einzelhandelskette.

Findet eine manuelle Datenerfassung auf der Basis von schriftlichen Untersuchungsdokumenten statt, dann ist (zumindest bei größeren Projekten) ein **Codierplan** als genaue Arbeitsvorschrift unverzichtbar. Hier wird z. B. festgelegt, dass beim Merkmal Geschlecht die Ausprägung *weiblich* durch eine Eins und die Ausprägung *männlich* durch eine Zwei erfasst werden muss. Mit dem Abschluss der Erfassung wird der Codierplan nicht überflüssig, sondern dient nun zur Dokumentation der entstandenen Datendatei gegenüber einem größeren Nutzerkreis.

Nach Abschluss der Planungs- und Vorbereitungsphase kann die empirische Phase mit der Datenerhebung stattfinden.

2.2.4 Daten erfassen, prüfen und korrigieren

Nach einer Datenerhebung mit einem in Papierform vorgelegten Fragebogen steht als nächster Schritt die Datenerfassung an. Das Eintragen der Rohdaten in eine Computer-Datei kann mit dem Dateneditor von SPSS geschehen oder mit einem speziellen Datenerfassungsprogramm. In jedem Fall ist bei der Erfassung der in der Planungsphase oder spätestens nach der Datenerhebung erstellte Codierplan genau einzuhalten. Hier ist für jedes Merkmal festgelegt, wie seine Ausprägungen codiert werden müssen.

Durch Fehler bei der Datenerhebung oder -erfassung können irreguläre Werte entstehen. Je fehleranfälliger die eingesetzten Techniken waren, desto mehr Aufwand ist bei der Datenprüfung und -korrektur erforderlich.

Bei einer Online-Datenerhebung entfällt die Datenerfassung. Durch Fehler bei der Datenerhebung oder -übertragung können aber trotzdem irreguläre Werte in eine Datendatei geraten, sodass in der Regel eine Kontrolle erforderlich ist. Im Abschnitt 3.2 folgen weitere Informationen zur Online-Datenerhebung.

2.2.5 Datentransformation

Nach der Erfassung und Prüfung liegen bei vielen Studien die Daten immer noch nicht in auswertbarer Form vor. Vielfach müssen Variablen modifiziert (z. B. umcodiert) oder aus Vorläufern neu berechnet werden (z. B. durch Mittelwertsbildung). Solche Transformationen nehmen bei vielen Projekten einen erheblichen Umfang an, wobei sowohl akribische Fleißarbeit als auch kreative Begriffsbildung gefragt sind.

2.2.6 Statistische Datenanalyse

Nach langer Vorbereitung können mit Hilfe von SPSS z. B. die gesuchten Schätzwerte (samt Konfidenzintervallen) ermittelt und die geplanten Hypothesentests durchgeführt werden. Bei einer eher explorativen Untersuchungsanlage ist eine längere, kreative Auseinandersetzung mit den Daten erforderlich, wobei zahlreiche Datentransformationen und statistische Analysen ausgeführt werden.

2.3 Theorie und Untersuchungsplanung im Demonstrationsprojekt

Um die im Rahmen einer empirischen Studie mit SPSS zu erledigenden Arbeiten in einem realistischen Kontext üben zu können, wird im Kursverlauf eine kleine psychologische Fragebogenstudie durchgeführt.[1] Dabei werden Sie alle Phasen der empirischen Forschung kennenlernen und die erforderlichen Arbeiten zum großen Teil selbständig durchführen. Als Beispiel wurde u. a. deshalb eine querschnittliche Fragebogenstudie gewählt, weil die Kursteilnehmer dabei in wenigen Minuten interessante empirische Daten selbst erzeugen können. Damit ist auch die Phase der Daten*erhebung* in den Übungsablauf einbezogen, die ansonsten aus Zeitgründen ausgespart bleiben müsste.

Bezogen auf die in Abschnitt 2.2 vorgestellte Übersicht mit den Phasen bzw. Teilaufgaben einer empirischen Untersuchung beschäftigen wir uns nun mit dem theoretischen Hintergrund unserer Beispielstudie und mit Fragen der Untersuchungsplanung.

2.3.1 Die allgemeinpsychologische KFA-Hypothese

Nach einer Theorie von Kahneman[2] & Miller (1986) hängt die Stärke unserer emotionalen Reaktion auf ein positives oder negatives Ereignis u. a. davon ab, welche *alternativen* (aber *nicht* eingetretenen) Ereignisse wir uns vorstellen können, mit anderen Worten: welche **kontrafaktischen Alternativen** mental verfügbar sind. Wir beschränken uns auf den Fall ungünstiger Ereignisse. Hierfür stellen Kahneman & Miller die folgende Hypothese auf:

> **Bei einem negativen Ereignis erhöht die mentale Verfügbarkeit (Vorstellbarkeit) einer kontrafaktischen (also positiven) Ereignisalternative den erlebten Ärger.**

Weil diese Hypothese für beliebige Personen Gültigkeit beansprucht, kann sie als *allgemeinpsychologisch* bezeichnet und von *differentialpsychologischen* Hypothesen unterschieden werden, die sich mit Unterschieden zwischen Personen beschäftigen (siehe Abschnitt 2.3.3).

Im weiteren Verlauf wollen wir unser Projekt kurz als *KFA-Studie* bezeichnen.

---

[1] Hierbei werden in stark vereinfachter Form Ideen aus einem ehemaligen Forschungsprojekt von Herrn Prof. Dr. Jochen Brandtstädter (Emeritus der Universität Trier) aufgegriffen, dem ich an dieser Stelle herzlich für die Erlaubnis und für die Überlassung von Untersuchungsmaterial danken möchte.

[2] Daniel Kahneman erhielt 2002 den Nobelpreis für Wirtschaft, womit vor allem seine erfolgreiche Anwendung psychologischer Erkenntnisse (u. a. zu Urteilen und Entscheidungen unter Unsicherheit) in wirtschaftswissenschaftlichen Theorien gewürdigt wurde.

2.3.2 Untersuchungsplanung

2.3.2.1 Untersuchungseinheiten, Population, Merkmale, Design und Operationalisierung

Hinsichtlich des Untersuchungsdesigns haben wir uns aufgrund praktischer Erwägungen bereits auf eine **querschnittlich angelegte Fragebogenstudie** mit den Kursteilnehmern als **Untersuchungseinheiten** festgelegt. Zentrales **Merkmal** ist der Ärger über ein negatives Ereignis. Er soll bei An- und Abwesenheit einer kontrafaktischen (also positiven) Ereignisalternative gemessen werden.

Eine Studie zum Ärger aufgrund bestimmter Situationsmerkmale bzw. Erlebnisinhalte würde vermutlich durch die physische Realisation von entsprechenden experimentellen Bedingungen eine hohe Überzeugungskraft erzielen. Um den Aufwand eines realen Experiments zu vermeiden, verlagern wir die experimentellen Bedingungen in die Köpfe der Untersuchungsteilnehmer und bitten diese, sich in eine Geschichte einzufühlen, bei der zwei Personen objektiv denselben Schaden erleiden, jedoch in unterschiedlichem Grad eine kontrafaktische (also günstige) Alternative vor Augen haben. Dann sollen die Probanden für jeden Geschädigten angeben, wie stark sie sich in dessen Lage ärgern würden. Die genaue Instruktion ist dem unten wiedergegebenen Fragebogen (Teil 1, siehe Abschnitt 2.3.6) zu entnehmen.

Es ist keinesfalls ungewöhnlich, durch Fragebogen-Items imaginierte Experimente zu realisieren. Folglich ist die Grenze zwischen dem Experiment und der Beobachtungsstudie nicht so streng und klar, wie es bisher dargestellt wurde.

Indem wir in der KFA-Studie jede Person den *beiden* imaginierten Behandlungen aussetzen, also einen **Messwiederholungsfaktor** KFA verwenden, gewinnen wir jeweils *zwei* Beobachtungswerte, die eine statistische Analyse der allgemeinpsychologischen Hypothese mit relativ hoher Teststärke (kleiner Irrtumswahrscheinlichkeit zweiter Art) ermöglich. Für jede Person wird die Differenz aus ihren beiden Ärgerwerten gebildet, wobei individuelle Besonderheiten (z. B. generell hohe oder niedrige Ärgerneigung, aktuelle Stimmung) aus der statistischen Analyse herausgehalten werden. Weil interindividuelle Unterschiede als Quelle von unaufgeklärter Varianz ausscheiden, kann der KFA-Effekt leichter nachgewiesen werden als bei Verwendung eines Gruppierungsfaktors (siehe Cohen 1988, S. 51).

Gegen diese Vorgehensweise lässt sich einwenden, dass durch die Präsentation der *beiden* Situationsvarianten ein Kontrast künstlich induziert, zumindest jedoch verstärkt wird. Um diese **Artefaktgefahr** zu vermeiden, könnte man statt des Messwiederholungsfaktors KFA einen **Gruppierungsfaktor** verwenden und jede Person zufallsabhängig nur zu *einer* Schädigungsvariante befragen. In einer Online-Studie ist die zufällige Aufteilung der Probanden auf zwei oder mehr Gruppen sehr einfach zu realisieren. Weil das Artefaktargument nicht zwingend und die Kursstichprobe aus organisatorischen Gründen sehr klein ist, hat das Teststärkeargument ein höheres Gewicht und wir verwenden einen Messwiederholungsfaktor mit zwei KFA-Ausprägungen.

Bei Verwendung eines Gruppierungsfaktors mit randomisierter Gruppenzuordnung wäre ein Experiment im engeren, klassischen Sinn realisiert. Der Versuchsplan mit Messwiederholung ist zweifellos eine anerkannte Forschungsmethode und wird oft auch unter den experimentellen Techniken aufgelistet.[1] Eine differenzierende Klärung des Begriffs *Experiment* ist in diesem Manuskript nicht erforderlich.

---

[1] https://dorsch.hogrefe.com/stichwort/experiment

Für die Ärgermessungen zu den beiden Situationsvarianten setzen wir Ratingskalen ein, wobei das Antwortformat der Anschaulichkeit halber an ein Thermometer mit den Ankerpunkten 0° und 100° erinnert. Wir gehen davon aus, dass die Ärgermessungen annähernd metrisches Messniveau (Intervallniveau) besitzen.

Unsere Studie soll aus praktischen Gründen mit der **studentischen Stichprobe** der Kursteilnehmer durchgeführt werden. Damit können wir unter induktivistischer Perspektive die Ergebnisse günstigstenfalls auf die Population der Studierenden generalisieren. Ernsthafte Bedenken bzgl. der externen Validität (Generalisierbarkeit) unserer Ergebnisse bestehen allerdings nur dann, wenn die Besonderheiten unserer Stichprobe (z. B. bzgl. Alter und Bildungsgrad) bei einer Fragestellung von Relevanz sind, wenn also Moderatoreffekte bestehen (siehe Eid et al. 2017). Dies ist bei der KFA-Hypothese kaum zu befürchten. Wir gehen also davon aus, trotz der Verwendung einer Bequemlichkeitsstichprobe (vgl. Abschnitt 2.2.3.6) einen allgemeinpsychologischen Forschungsbeitrag leisten zu können.

2.3.2.2 Formulierung und Illustration der empirisch prüfbaren Hypothese

In Abschnitt 2.3.1 wurde die KFA-Hypothese noch ohne Bezug auf unsere Untersuchungsplanung formuliert. Jetzt nehmen wir eine Konkretisierung vor durch …

- Verwendung von direkt beobachtbaren Begriffen
- Bezug auf die statistisch zu analysierenden Verteilungsparameter
  Zu Beginn von Kapitel 2 wurde betont, dass unsere Hypothesen in der Regel *probabilistischer* Natur sind. Auch bei einer allgemeinpsychologischen Hypothese wird man kaum auf einer Gültigkeit für *alle* Personen einer Population bestehen (womöglich sogar mit derselben Effektstärke). Die konkretisierte Hypothese sollte über die im statistischen Entscheidungsverfahren (siehe unten) tatsächlich analysierten Modell- bzw. Verteilungsparameter reden.

Außerdem soll hier der Klarheit halber (in einer für Forschungsberichte kaum zu empfehlenden Ausführlichkeit) dargelegt werden, dass bei einer inferenzstatistischen Hypothesenprüfung nach dem binären Entscheidungskonzept von Neyman und Pearson *zwei* konkurrierende und komplementäre Hypothesen beteiligt sind:

Nullhypothese (H_0)	Die Untersuchungsteilnehmer erleben in der Rolle des Geschädigten mit hochgradig mental verfügbarer kontrafaktischer, also positiver Ereignisalternative im Mittel *nicht* mehr Ärger als in der Rolle des Geschädigten mit „weit entfernter“ kontrafaktischer Alternative.
Alternativhypothese (H_1)[1]	Die Untersuchungsteilnehmer erleben in der Rolle des Geschädigten mit hochgradig mental verfügbarer kontrafaktischer Alternative im Mittel mehr Ärger als in der Rolle des Geschädigten mit „weit entfernter“ kontrafaktischer Alternative.

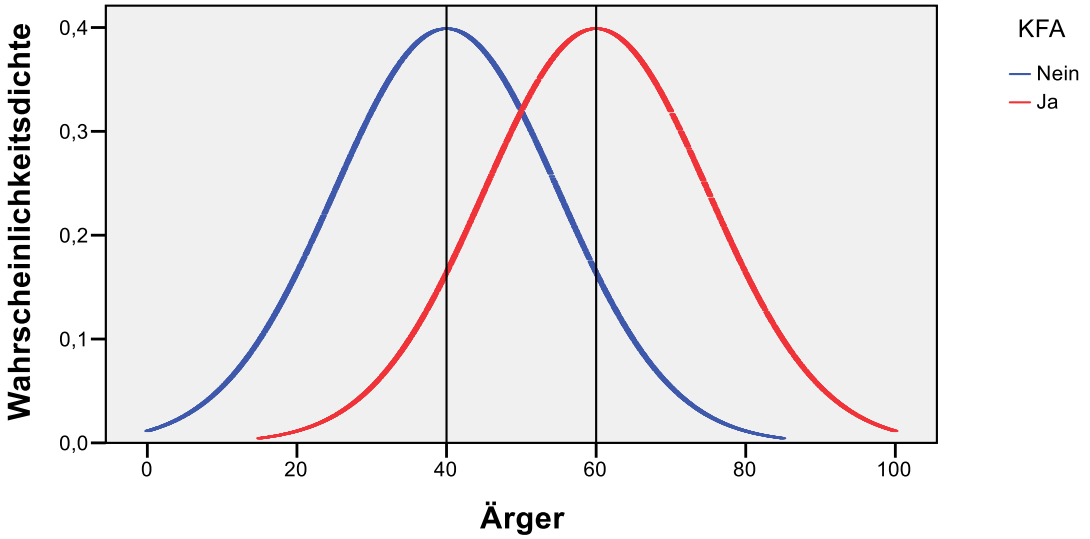
[1] Hier handelt es sich um einen statistischen Terminus, der nur zufällig mit unserer allgemeinpsychologischen Hypothese den Wortbestandteil *alternativ* gemeinsam hat.

In den Hypothesen werden die Erwartungs- bzw. Mittelwerte von zwei metrischen Merkmalen verglichen:

- Ärgerausprägung in der Situation mit weit entfernter kontrafaktischer Alternative
- Ärgerausprägung in der Situation mit hochgradig mental verfügbarer kontrafaktischer Alternative

Zur Vereinfachung der Ausdrucksweise sprechen wir ab jetzt vereinfachend über die Situationen *ohne* bzw. *mit* KFA.

Für die beiden Merkmale wird jeweils eine *Verteilung* der erlebten bzw. berichteten Ausprägungen angenommen. Das folgende Diagramm zeigt die Verteilungen des Ärgers aus den Situationen *ohne* (blau) bzw. *mit* KFA (rot) im Sinn der Alternativhypothese:



Zur Veranschaulichung der Alternativhypothese wird hier für beide Ärgervarianten eine spezielle eingipflige Verteilungsform, nämlich die Normalverteilung, verwendet. Allerdings ist in den Hypothesen *keine* Aussage über die Verteilungsform enthalten.

Mit dem Symbol μ_O für den Erwartungswert (Populationsmittelwert) des Ärgers *ohne* KFA und dem Symbol μ_M für den Erwartungswert des Ärgers *mit* KFA kann unser allgemeinpsychologisches KFA-Testproblem kompakt so formuliert werden:

$$H_0 : \mu_M \leq \mu_O \quad \text{versus} \quad H_1 : \mu_M > \mu_O$$

2.3.2.3 Entscheidungsverfahren

Wir wollen das eben beschriebene Entscheidungsproblem mit einem **t-Test für verbundene** (alias: **abhängige**) **Stichproben** lösen, sofern dessen Verteilungsvoraussetzung erfüllt ist. Dann ist der t-Test für verbundene Stichproben nicht nur zulässig, sondern auch sehr gut dazu geeignet, einen vorhandenen Effekt nachzuweisen. Er besitzt dann eine hohe Teststärke (engl.: *Power*).

Die Verteilungsvoraussetzung des t-Tests für abhängige Stichproben betrifft die *Differenz* der beiden Ärgerausprägungen, die in der Population normalverteilt sein muss (siehe Abschnitt 8.1). Ist diese Bedingung *nicht* erfüllt, ist der t-Test für verbundene Stichproben unzulässig und besitzt außerdem in der Regel *keine* gute Teststärke (Power, Entdeckungswahrscheinlichkeit). Wir wer-

den die Normalverteilungsannahme für die Ärgerdifferenz mit den Daten unserer Stichprobe überprüfen und bei ungünstigem Testergebnis ein alternatives Entscheidungsverfahren wählen.

Da gerichtete Hypothesen vorliegen, ist **einseitig** zu testen. Dabei wird eine Irrtumswahrscheinlichkeit erster Art in Höhe von **α = 0,05** akzeptiert.

Es wäre unzulässig und im Hinblick auf die Teststärke vor allem auch nachteilig, anstatt des t-Tests für abhängige Stichproben einen t-Test für *un*abhängige Stichproben durchzuführen, und dabei die beiden von einer Person stammenden Ärgermessungen als unabhängig zu betrachten. Messen z. B. die Variablen $X_i^{(O)}$ bzw. $X_i^{(M)}$ den von Person *i* gelieferten Ärger in der Situation ohne bzw. mit KFA, dann enthalten beide Variablen denselben Personeffekt (z. B. die generelle Ärgerneigung und die aktuelle Stimmung von Person *i*). Liegt z. B. $X_i^{(O)}$ bei einer Person *i* mit generell hoher Ärgerneigung über μ_O, dann liegt erwartungsgemäß auch $X_i^{(M)}$ über μ_M. Die beiden Residuen (Abweichungen von den Erwartungswerten) sind also korreliert, und die Unabhängigkeitsannahme des t-Tests für unabhängige Stichproben ist verletzt.

Beim t-Test für verbundene Stichproben wird für jede Person die Differenz aus den Variablen $X_i^{(O)}$ und $X_i^{(M)}$ gebildet. Es ist also pro Person nur noch *ein* Wert im Spiel, sodass die Abhängigkeitsproblematik verschwindet. Außerdem verschwinden durch die Differenzbildung auch die für unaufgeklärte Varianz sorgenden interindividuellen Unterschiede, sodass die Chance steigt, einen Mittelwertsunterschied im Sinne der Alternativhypothese ($\mu_M > \mu_O$) in den Stichprobendaten zu entdecken.

2.3.2.4 Stichprobenumfangsplanung

Da aus statistischer Sicht eine Stichprobe nie zu groß sein kann, sollen nach Möglichkeit *alle* Kursteilnehmer als Probanden gewonnen werden. Es ist aus praktischen Gründen nicht möglich, weitere Untersuchungsteilnehmer zu rekrutieren. Der Übung halber soll aber trotzdem an dieser Stelle eine β-Fehler - basierte Kalkulation des benötigten Stichprobenumfangs vorgenommen werden. Vor allem bei hohen Kosten pro Untersuchungseinheit möchte man wissen, wie groß die Stichprobe sein muss, damit ein angenommener Populationseffekt in einem Test zum Niveau α = 0,05 mit einer gewünschten Wahrscheinlichkeit zu einem signifikanten Testergebnis führt. Diese Entdeckungswahrscheinlichkeit bzw. Power ist das Komplement zum β-Fehlerrisiko, d. h.

Power = 1 - β-Fehlerrisiko

2.3.2.4.1 G*Power

Obwohl SPSS Statistics seit der Version 27 eine Prozedur zur Power-Analyse enthält (siehe Abschnitt 2.3.2.4.2), verwenden wir zunächst das exzellente Power-Analyse - Programm **G*Power 3.1** (Faul et al. 2007, 2009). Es verfügt über ein breiteres Einsatzspektrum als die SPSS-Prozedur, ist leicht zu bedienen und wird in der Literatur gut unterstützt, sodass es sich lohnt, G*Power als Option zur Power-Analyse zu kennen. Das Programm ist für macOS und Windows kostenlos über die folgende Webseite zu beziehen:

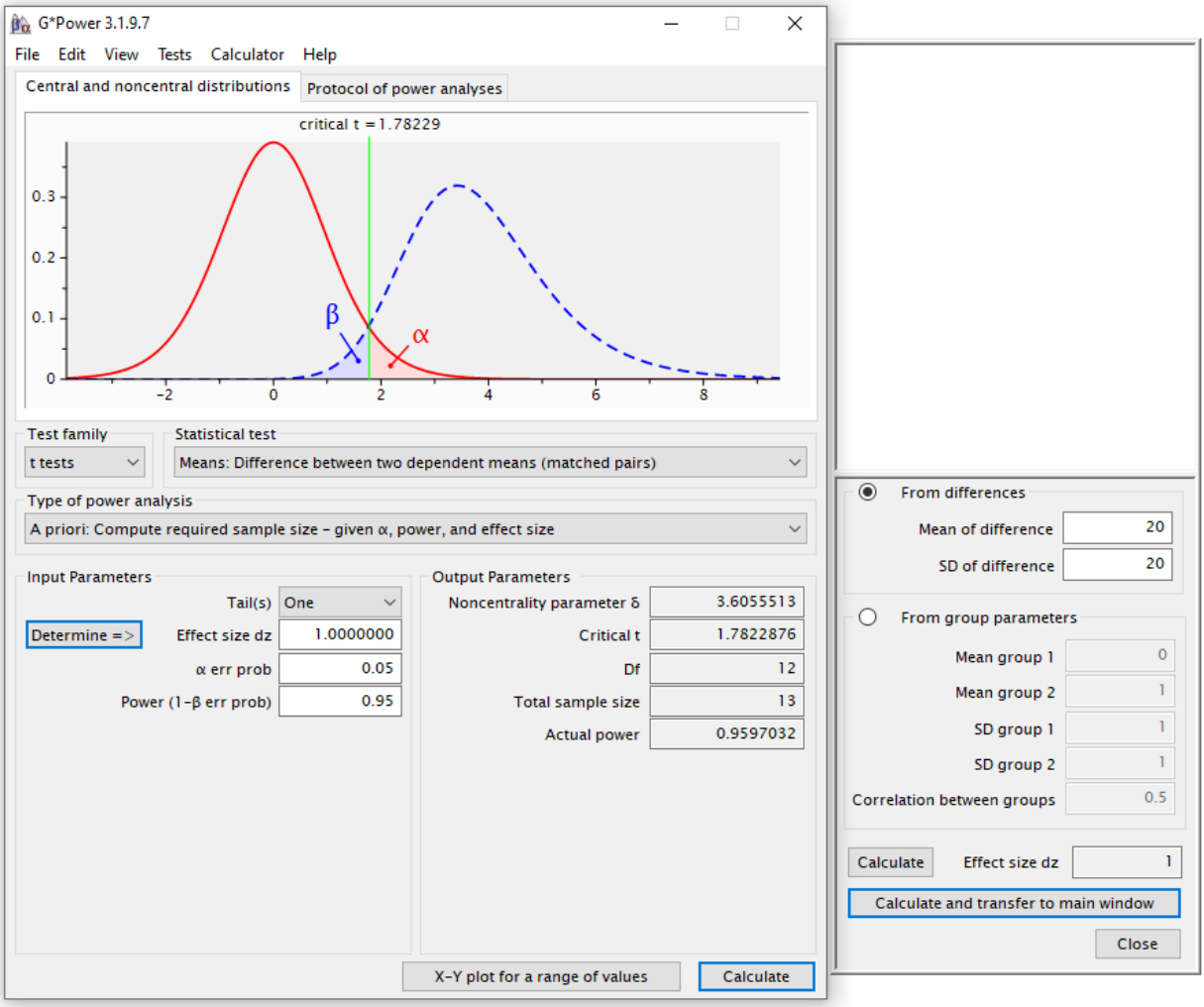
https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html

Auf den ZIMK-Pool-PCs an der Universität Trier unter dem Betriebssystem Windows befindet sich eine Verknüpfung zum Starten von G*Power 3.1 in der Startmenügruppe **Statistik**.

Zur Planung des Stichprobenumfangs für den t-Test zur KFA-Hypothese wählen wir in G*Power folgende Problembeschreibung:

- **Test family**: **t-Tests**
- **Statistical test**: **Means: Difference between two dependent means (matched pairs)**
- **Type of power analysis**: **A priori**

und öffnen über den Schalter **Determine** ein Zusatzfenster, um die Effektstärke in der Population aufgrund theoretischer Annahmen und/oder bisheriger empirischer Befunde festlegen zu können:



Unsere KFA-Hypothese handelt vom *Ärgerzuwachs* aufgrund einer mental gut verfügbaren positiven Alternative zum erlebten negativen Ereignis und kann über die *Differenz* der beiden Ärgermessungen beurteilt werden. Wir verwenden in G*Power 3.1 diese Sichtweise (**From differences**), um die unterstellte **Effektstärke** in der Population bequem festlegen zu können.

Beim geplanten t-Test für abhängige Stichproben und auch bei der nun anstehenden Stichprobenumfangsplanung wird für den *Ärgerzuwachs* eine Normalverteilung in der Population angenommen (siehe Abschnitt 8.1).

Die beim geplanten t-Test für abhängige Stichproben relevante Effektstärke d_z ist (wie beim letztlich zugrunde liegenden Einstichproben - t-Test) folgendermaßen definiert (vgl. Cohen 1988, S. 48; Faul et al. 2007, S. 182):

$$d_z := \frac{\mu_z}{\sigma_z}$$

Darin sind:

- μ_z Mittelwert für den Ärgerzuwachs in der Population
- σ_z Standardabweichung für den Ärgerzuwachs in der Population

Indem man den Abstand des Parameters μ_z von der Nullhypothesenbehauptung in *Standardabweichungseinheiten* angibt, erhält man einen *normierten* Effektstärkeindex mit guter Vergleichbarkeit zwischen Studien mit unterschiedlichen Ärger-Operationalisierungen. Z. B. ändert sich in der KFA-Studie die Effektstärke nicht, wenn man die Ärgertemperaturmessungen in Fahrenheit statt in Celsius vornimmt.[1]

Zur Beschreibung des Betrags von d_z - Ausprägungen hat Cohen (1988, S. 40) folgende Orientierungswerte vorgeschlagen:[2]

- kleiner Effekt: $d_z = 0{,}2$
- mittlerer Effekt: $d_z = 0{,}5$
- großer Effekt: $d_z = 0{,}8$

Für die Beispielstudie lässt sich eine konkrete Effektstärke begründen, sodass wir *nicht* auf die oft zu beobachtende Annahme einer *mittleren* Effektstärke zurückgreifen müssen. Beim Ärgerzuwachs wird ein Erwartungswert von 20 angenommen bzw. als theoretisch relevant und „entdeckungswürdig“ eingestuft. Als Ärgerzuwachs-Standardabweichung (Nebenparameter der KFA-Hypothese) vermuten wir aufgrund bisheriger Studien mit derselben Messmethode ebenfalls einen Wert von ca. 20. Mit dem Schalter

Calculate and transfer to main window

befördern wir die resultierende, sehr große Effektstärke von 1,0 in das G*Power-Hauptfenster.

Dort wählen wir …

- einen gerichteten (einseitigen) Test (**Tail(s)**: **One**)
- als akzeptiertes α-Fehlerrisiko (α **err prob**) den Wert 0,05
  Bei der Eingabe von Dezimalzahlen ist zu beachten, dass G*Power nur den *Punkt* als Dezimaltrennzeichen akzeptiert.
- eine gewünschte Teststärke (**Power**) von 0,95, also ein β-Fehlerrisiko von 0,05

Nach einem Mausklick auf den Hauptfensterschalter

Calculate

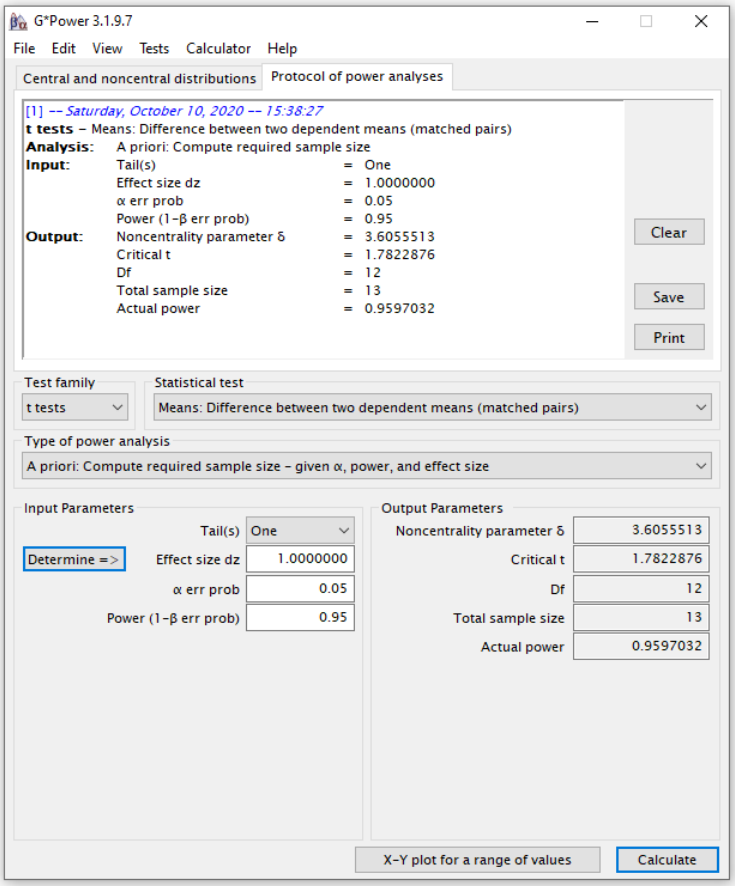
---

[1] Das Standardisieren verbessert die Vergleichbarkeit der Effektstärken aus zwei Studien mit unterschiedlichen Ärger-Operationalisierungen allerdings nur dann, wenn beide Studien in derselben Population durchgeführt werden. Beim Vergleich von Effektstärken aus verschiedenen Populationen mit unterschiedlichen Varianzen ist das Standardisieren eher ungünstig (siehe z. B. Eid et al. 2017).

[2] Nach diesem Vorschlag werden d_z - Werte genauso beurteilt wie *d*-Werte aus dem Vergleich von zwei unabhängigen Stichproben. Allerdings müssen d_z - Werte vor der Verwendung von Cohens Power-Tabellen für *d*-Werte nach folgender Formel in *d*-Werte transformiert werden: $d = d_z\sqrt{2}$ (siehe Cohen 1988, S. 46). Hier artikuliert sich die (bei gleicher Effektstärke) höhere Power des t-Tests für abhängige Stichproben gegenüber seinem Pendant für unabhängige Stichproben. Eid et al. (2017) interpretieren Cohen (1988) allerdings anders und werten z. B. schon den d_z - Wert $\frac{0{,}2}{\sqrt{2}} \approx 0{,}14$ als kleinen Effekt.

erhalten wir das beruhigende Ergebnis, dass lediglich 13 Fälle erforderlich sind. Sofern ein Effekt von der angenommenen (oder einer größeren) Stärke existiert, werden wir ihn mit großer Wahrscheinlichkeit entdecken, weil an der Kursstudie in der Regel ca. 25 Fälle teilnehmen.

Um die Ergebnisse der Stichprobenumfangskalkulation mit G*Power in einen Projektbericht zu übernehmen, wechselt man zur Registerkarte mit dem **Protocol of power analysis**:



Man kann das Protokoll in eine RTF-Datei (*Rich-Text-Format*) sichern (Schalter **Save**), drucken (Schalter **Print**) oder die markierten Bestandteile mit der Tastenkombination **Strg+C** in die Windows-Zwischenablage befördern, um den formatierten Text anschließend in ein Zieldokumentenfenster einzufügen (z. B. mit der Tastenkombination **Strg+V**).

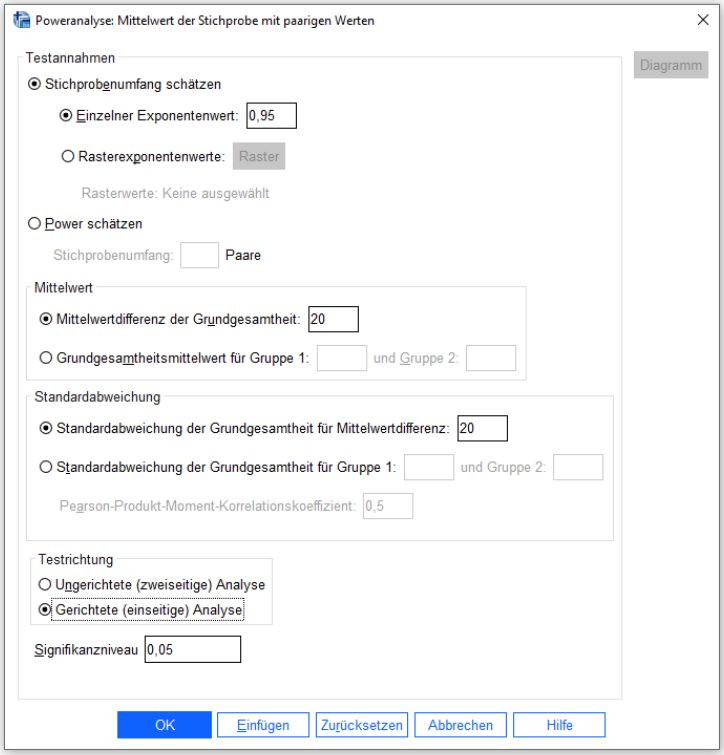
2.3.2.4.2 SPSS Statistics

Um die Stichprobenumfangsplanung für unsere KFA-Hypothese mit SPSS Statistics durchzuführen, ...

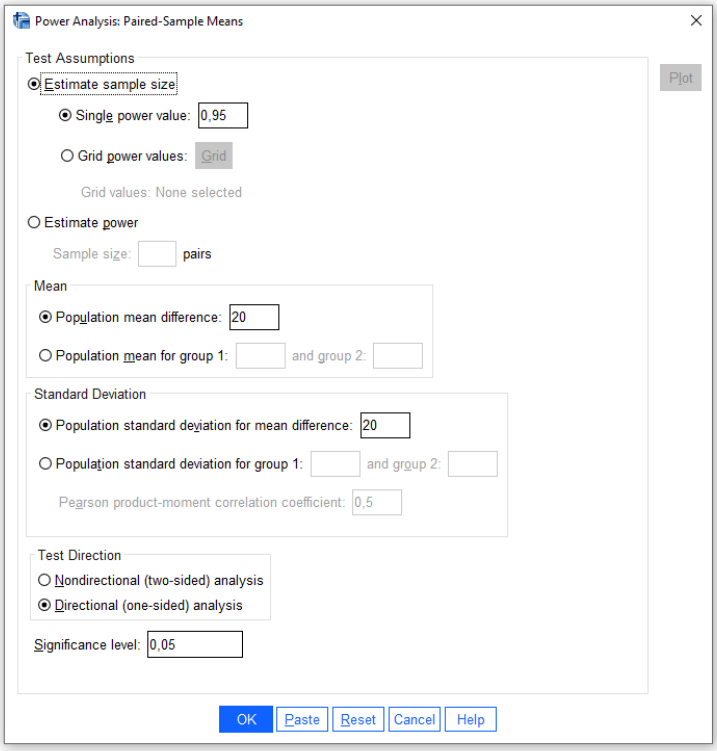
- starten wir SPSS (z. B. per Doppelklick auf das Desktop-Symbol),
- ignorieren den aktuell überflüssigen Begrüßungsdialog
- und wählen den Menübefehl:

  Analysieren > Poweranalyse > Mittelwerte > t-Test bei Stichproben mit paarigen Werten

Es erscheint ein Dialog,



der im Vergleich zum englischen Original



leider kuriose Übersetzungsfehler enthält. So wird z. B. *Power* durch *Exponent* übersetzt.

Über den Menübefehl

Bearbeiten > Optionen > Sprache

kann man in SPSS die Sprache der Bedienoberfläche und die Sprache der Ausgabeobjekte (Tabellen und Diagramme) ändern, wobei Englisch wohl in der Regel die beste Wahl zur Klärung statistischer Begriffe ist.

Wir wählen im **Poweranalyse** - Dialog:

- die Schätzung des **Stichprobenumfangs**
- 0,95 als gewünschte Power
- die **Mittelwertsdifferenz** 20 und die **Standardabweichung** 20
- einen **gerichteten** Test
- als **Signifikanzniveau** (akzeptiertes α-Fehlerrisiko) den Wert 0,05

Im Ausgabefenster erscheint eine Tabelle mit dem erwarteten Stichprobenumfang 13:

Poweranalysetabelle.

	N[b]	Tatsächlicher Exponent[c]	Testannahmen			
			Exponent	Std.-Abw.[d]	Effektgröße	Sig.
Test für Mittelwertdifferenz[a]	13	,960	,95	20	1,000	,05

a. Einseitiger Test.
b. Anzahl der Paare.
c. Basierend auf nicht zentraler t-Verteilung.
d. Standardabweichung der Mittelwertdifferenz.

Mit dem *tatsächlichen Exponenten* ist die tatsächliche Power gemeint, wie die englische Ausgabevariante zeigt:

Power Analysis Table

	N[b]	Actual Power[c]	Test Assumptions			
			Power	Std. Dev.[d]	Effect Size	Sig.
Test for Mean Difference[a]	13	,960	,95	20	1,000	,05

a. One-sided test.
b. Number of group pairs.
c. Based on noncentral t-distribution.
d. Standard deviation of the mean difference.

2.3.3 Eine differentialpsychologische Hypothese

Neben der im Fokus stehenden KFA-Hypothese soll in unserer Studie noch die folgende, auf Überlegungen von Scheier & Carver (1985) zurückgehende Hypothese überprüft werden:

> **Der durch ein negatives Ereignis ausgelöste Ärger wird durch dispositionellen Optimismus gedämpft.**

Begründung: Dispositioneller Optimismus führt zur Verwendung günstiger Bewältigungsstrategien (z. B. positive Reinterpretation von negativen Erfahrungen), die den Ärger bremsen und zu einer schnellen Reduktion beitragen.

Während unsere allgemeinpsychologische KFA-Hypothese für eine beliebig aus der Allgemeinbevölkerung herausgegriffene Person einen Effekt postuliert, geht es nun um Differentialpsychologie, also um Verhaltens- bzw. Erlebens*unterschiede* in Folge von relativ stabilen Personmerkmalen.

Nun ist die Untersuchungsplanung um die Behandlung der differentialpsychologischen Hypothese zu erweitern. Als Quasiereignis soll der schon zur Prüfung der allgemeinpsychologischen Hypothese verwendete imaginierte Schadensfall dienen (Fragebogenteil 1, siehe Abschnitt 2.3.6). Das arithmetische Mittel der für beide Situationsvarianten angegebenen Ärgerausprägungen soll als Ärgermaß verwendet werden. Zur Erfassung von dispositionellem Optimismus wird der von

Scheier & Carver (1985) entwickelte *Life Orientation Test* (LOT) eingesetzt (siehe Fragebogenteil 2). Wie aus den Antworten auf die zwölf Fragen dieses Tests ein Optimismus-Messwert zu ermitteln ist, wird später erläutert. Wir gehen jedenfalls davon aus, dass diese Messmethode annähernd metrisches Messniveau (Intervallniveau) besitzt.

Nach dieser **Operationalisierung** der theoretischen Begriffe kann die folgende **empirisch prüfbare Alternativhypothese** formuliert werden:

> **Je höher der LOT-Wert einer Person, desto weniger Ärger (Mittel aus beiden Situationsvarianten) berichtet sie über den imaginierten Schadensfall.**

Die Nullhypothese ergibt sich durch Negation der Alternativhypothese und muss daher nicht notiert werden.

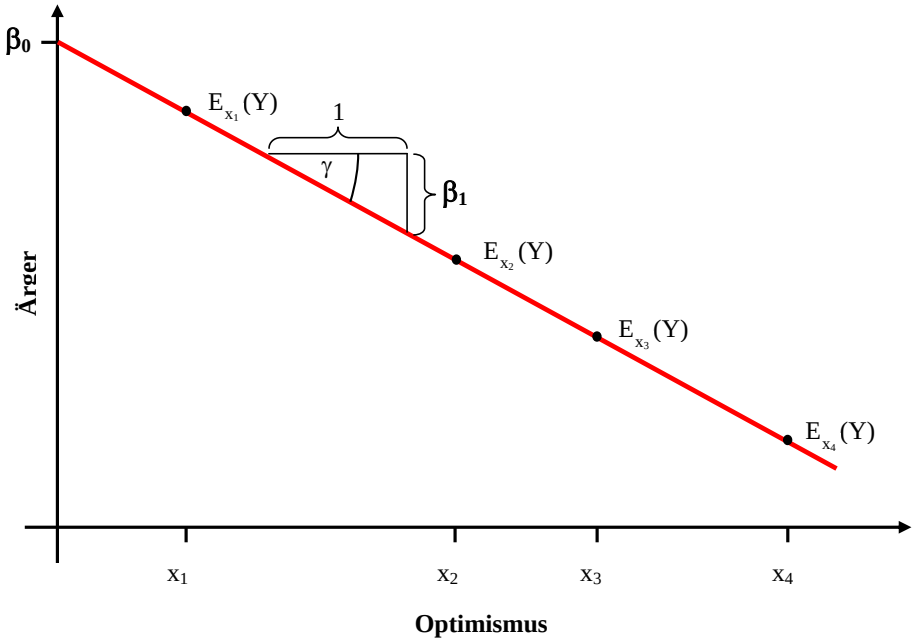
Weil die Messungen zum Ärger und zum Optimismus vermutlich metrisches Skalenniveau besitzen, kann die differentialpsychologische Hypothese mit einer **einfachen (bivariaten) linearen Regressionsanalyse** geprüft werden, sofern deren Modell- und Verteilungsvoraussetzungen annähernd erfüllt sind (siehe Abschnitt 8.2). Die Modellgleichung für ein zu erklärendes Kriterium Y, einen Regressor X und ein Residuum ε war schon in Abschnitt 2.1 zu sehen:

$$Y = \beta_0 + \beta_1 X + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$$

Als Modellparameter treten der Ordinatenabschnitt β_0 und der Steigungskoeffizient β_1, die gemeinsam eine Regressionsgerade festlegen, sowie die Varianz des Residuums auf. Unserer differentialpsychologischen Alternativhypothese entspricht eine Regressionsgerade mit negativer Steigung, weil wir eine Ärgerreduktion bei zunehmendem Optimismus erwarten. Für jede Ausprägung x des Regressors befindet sich die modellgemäße Erwartung

$$E_x(Y) = \beta_0 + \beta_1 x$$

auf der Regressionsgeraden:



Zur Interpretation des Koeffizienten β_1: Erhöht man X um eine Einheit, so sinkt modellgemäß der Erwartungswert $E_x(Y)$ um β_1 Einheiten.

Die statistisch zu prüfende Alternativhypothese lässt sich im Rahmen dieses Modells sehr präzise formulieren:

> **In der linearen Regression von Ärger (Mittel aus beiden Situationsvarianten) auf den LOT-Testwert ergibt sich ein negativer Steigungskoeffizient (β_1).**

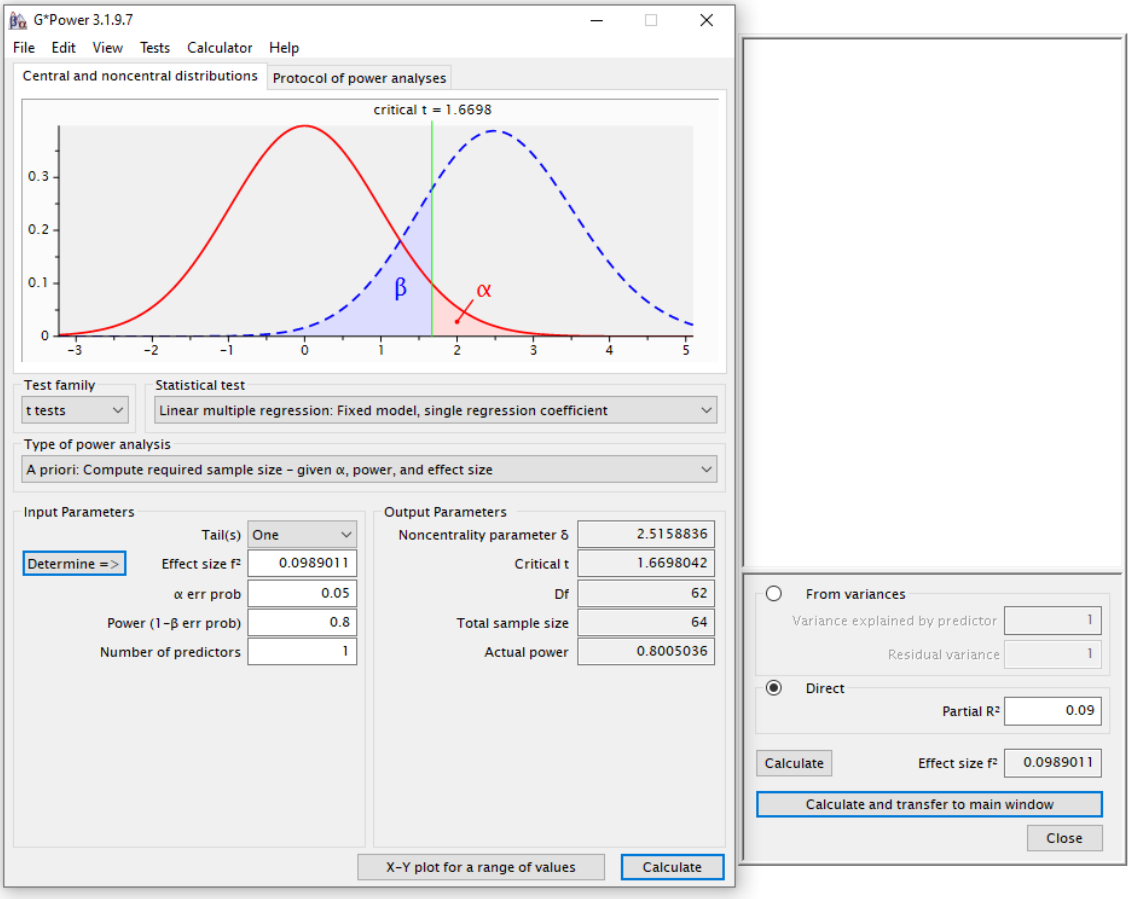
Die Hypothese ist *gerichtet* (*einseitig*) formuliert und soll bei einem α-Fehler - Risiko von 0,05 mit dem t-Test zum Regressionskoeffizienten β_1 geprüft werden:

$$H_0: \beta_1 \geq 0 \quad \text{versus} \quad H_1: \beta_1 < 0$$

Zur Berechnung des bei diesem Test für eine angestrebte Power erforderlichen Stichprobenumfangs wählen wir im Teststärkenanalyseprogramm G*Power 3.1 (vgl. Abschnitt 2.3.2):

- **Test family**: **t-Tests**
- **Statistical test**: **Linear multiple regression: Fixed model, single regression coefficient**
- **Type of power analysis**: **A priori**

Für den geplanten einseitigen t-Test mit einem α-Fehler - Risiko von 0,05 in einem Modell mit *einem* Prädiktor wählen wir die von Cohen (1988, S. 56) als Standardwert empfohlene Power (Entdeckungswahrscheinlichkeit) von 0,8 (entspricht einem β-Fehlerrisiko von 0,2):

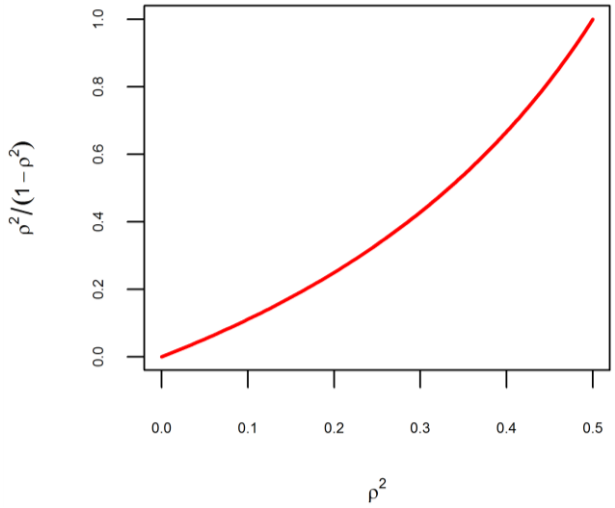


Das von G*Power verwendete Effektstärkemaß f^2 steht bei einer bivariaten Regression in folgender Beziehung zur quadrierten Korrelation ρ^2 zwischen Kriterium und Regressor in der Population:

$$f^2 = \frac{\rho^2}{1-\rho^2}$$

Die quadrierte Korrelation ρ^2 bezeichnet man auch als *Determinationskoeffizienten*, weil sie den vom Regressor erklärten Anteil der Kriteriumsvarianz quantifiziert.

Wie die folgende Abbildung zeigt, steht f^2 in einer monoton wachsenden, positiv beschleunigten Beziehung zu ρ^2, die im meist relevanten Wertebereich von 0 bis 0,5 fast linear ausfällt:



Wir nehmen eine LOT-Ärger - Korrelation von -0,3 und damit einen Determinationskoeffizienten von $\rho^2 = 0{,}09$ an, was zu einer Effektstärke f^2 von ca. 0,1 führt:

$$\frac{0{,}09}{1-0{,}09} = \frac{0{,}09}{0{,}91} \approx 0{,}1$$

In G*Power öffnen wir mit dem Schalter **Determine** das Seitenfenster zur Effektstärkenspezifikation, wählen dort die Option **Direct**, tragen im Textfeld **Partial R^2** den angenommenen Determinationskoeffizienten ein und ermitteln den resultierenden f^2 - Wert mit dem Schalter **Calculate and transfer to main window**. Beachten Sie bitte, dass G*Power als Dezimaltrennzeichen einen Punkt verlangt.

Wie nach einem Klick auf den Hauptfensterschalter **Calculate** zu erfahren ist, resultiert aus unserer Problembeschreibung ein erforderlicher Stichprobenumfang von 64 Fällen. Weil die Kursstichprobe in der Regel kleiner ist, stehen unsere Chancen, einen Effekt von der vermuteten Stärke zu entdecken, also eher schlecht. Bei einer gewünschten Power von 0,95 (β-Fehler 0,05) werden sogar 111 Fälle benötigt. In einem realen Forschungsprojekt zur Klärung der differentialpsychologischen Hypothese müsste der Stichprobenumfang folglich erhöht werden.

Bei einem *zweiseitigen* Test werden bei der oben angenommenen Effektstärke und $\alpha = \beta = 0{,}05$ sogar 134 Fälle benötigt. Wer den Unterschied zwischen gerichteten und ungerichteten Hypothesen ignoriert und den bei Statistikprogrammen üblicherweise voreingestellten zweiseitigen Test verwendet, muss also mehr Fälle einbeziehen bzw. verliert bei identischem Stichprobenumfang an Teststärke.

Zur Beschreibung der Effektstärke im Model der bivariaten linearen Regression nennt Cohen (1988, S. 77ff) folgende Orientierungswerte:

Effektstärke in der Population	ρ^2	f^2
klein	0,01	0,01
mittel	0,09	0,10
groß	0,25	0,33

Damit ein *großer* Effekt bei einseitiger Testung zum α-Niveau 0,05 mit einer Wahrscheinlichkeit von 0,8 zu einem signifikanten Ergebnis führt, sind nur 21 Fälle erforderlich, sodass für die differentialpsychologische Hypothese auch in der sehr kleinen Kursstichprobe noch Anlass zur Hoffnung besteht.

Bei der Stichprobenumfangsplanung ist es oft schwierig, eine begründete Effektstärke anzugeben. Ein häufig gewählter, wenig überzeugender Ausweg besteht darin, von einer *mittleren* Effektstärke auszugehen.

Wird ein Regressor aus einem *multiplen* linearen Regressionsmodell (mit zwei oder mehr Regressoren) betrachtet, dann geht statt der oben betrachteten quadrierten Kriteriumskorrelation ρ^2 die quadrierte *partielle* Korrelation zwischen dem betrachten Regressor und dem Kriterium bei Kontrolle der restlichen Regressoren in die Effektstärkenberechnung ein (siehe **Determine** - Seitenfenster in G*Power).

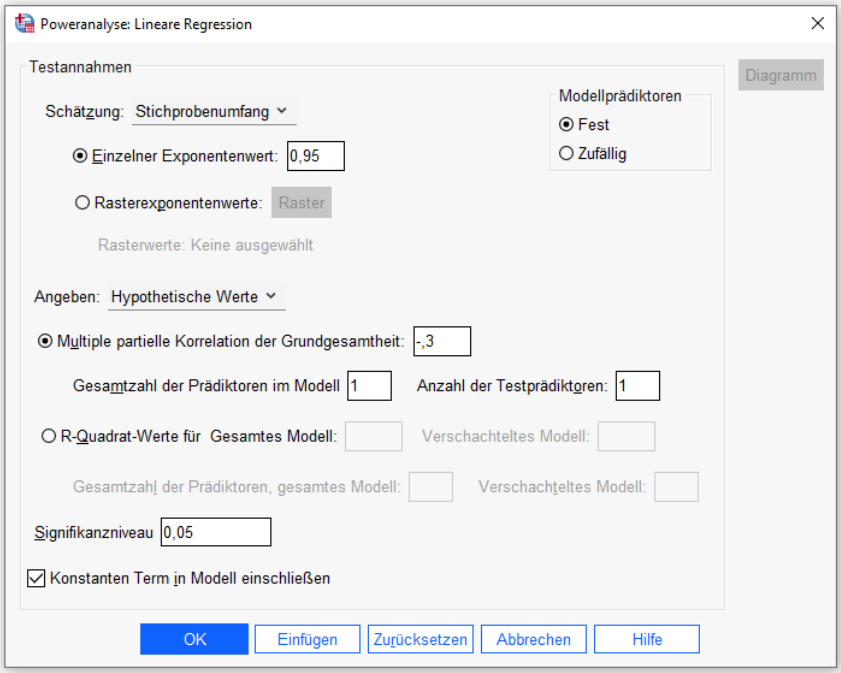
Wir wiederholen mit SPSS nach

Analysieren > Poweranalyse > Regression > Univariat linear

die Power-Analyse für die differentialpsychologische Hypothese. Wir wählen als exemplarische Aufgabenstellung im Power-Analyse - Dialog:

- die Schätzung des **Stichprobenumfangs**
- ein Modell mit **festen** Effekten (= Voreinstellung)
- 0,95 als gewünschte Power (falsch beschriftet mit **Exponentenwert**)
- 0,3 als **Korrelation**
- ein Modell mit *einem* Regressor, der folglich auch getestet werden soll
- als **Signifikanzniveau** (akzeptiertes α-Fehlerrisiko) den Wert 0,05
- ein Modell mit **konstantem Term** (= Voreinstellung)

Dass die Power-Analyse in SPSS beim t-Test für abhängige Stichproben die gerichtete Fragestellung unterstützt (siehe Abschnitt 2.3.2.4.2), beim t-Test zum Steigungskoeffizienten in der bivariaten linearen Regression jedoch *nicht*, ist als Mangel zu kritisieren:



Es ist ein leider immer noch oft anzutreffender Irrglaube, dass die wichtige Unterscheidung zwischen ein- und zweiseitigen Tests in der linearen Regression nicht anwendbar sei. G*Power ist von diesem Problem nicht betroffen.

SPSS liefert exakt dasselbe Ergebnis (N = 134), das wir auch von G*Power für die zweiseitige Testung erhalten haben:

Poweranalysetabelle.

	N	Tatsächlicher Exponent[b]	Prädiktoren		Testannahmen		
			Gesamt	Test	Exponent	Partiell[c]	Sig.
F-Test vom Typ III[a]	134	,951	1	1	,95	-,3	,05

a. Ein konstanter Term wird einbezogen.

b. Es wird davon ausgegangen, dass Prädiktoren festgelegt werden.

c. Multipler partieller Korrelationskoeffizient.

2.3.4 Demografische Merkmale

Auf die Erfassung demografischer Merkmale (siehe Fragebogenteil 4 in Abschnitt 2.3.6) kann man in keiner Studie mit Personen als Untersuchungseinheiten verzichten, auch wenn sich keine expliziten Hypothesen darauf beziehen. Man benötigt sie auf jeden Fall zur Beschreibung der Stichprobe, damit sich später die Leser von Berichten ein Urteil über die Interpretier- bzw. Generalisierbarkeit der Ergebnisse bilden können. Wir werden darüber hinaus einige demografische Merkmale auf Zusammenhänge mit unseren zentralen Projektvariablen untersuchen. Insofern finden sich auch in unserem überwiegend konfirmatorisch (hypothesenprüfend) angelegten Projekt einige explorative Elemente.

Wie Jacob et al. (2013, S. 138) plausibel begründen, sollten demografische Fragen *nicht* am Anfang einer Befragung stehen:

- Die Auskunftspersonen werden zunächst mit relativ langweiligen Aufgaben behelligt und erfahren den Zweck der Befragung verspätet.
- Manche demografische Fragen (etwa nach Bildung oder Einkommen) können als heikel erlebt werden. Dies kann die Teilnahmemotivation reduzieren und/oder das Antwortverhalten beeinflussen.

Damit ist das Ende eines Fragebogens in der Regel der angemessene Ort für demografische Fragen.

Anmerkungen zu einigen Erhebungstechniken im Beispielfragebogen:

- Statt nach dem Alter wird nach dem Geburtsjahr gefragt, weil manche Auskunftspersonen diese Information leichter und genauer liefern können.
- Nach der generell relevanten Bildung zu fragen, wäre in der diesbezüglich sehr homogenen Kursstichprobe eine Vergeudung von Befragungszeit.
- Indem die Probanden (allesamt Studierende an der Universität Trier) auch nach ihrem primären Fachbereich befragt werden, lassen sich Geschlechtsunterschiede bei der Studienfachpräferenz untersuchen. Diese Fragestellung werden wir in Kapitel 14 als Anwendungsbeispiel für die Kreuztabellenanalyse verwenden.

Über Standards zur Erfassung von soziodemografischen Daten informieren z. B. Jacob et al. (2013, S. 151ff) und das statistische Bundesamt (2016).

2.3.5 Zu Übungszwecken erhobene Merkmale

Zu Übungszwecken und ohne inhaltlichen Bezug zu den oben beschriebenen Hypothesen werden im Beispielfragebogen zusätzlich die folgenden Informationen erhoben:

- Motive zur Kursteilnahme (siehe Fragebogenteil 3 in Abschnitt 2.3.6)
  Mit diesen Merkmalen lässt sich üben, wie die aus Mehrfachwahlfragen und offenen Fragen resultierenden Informationen deklariert, erfasst und analysiert werden können.
- Größe und Gewicht (siehe Fragebogenteil 4 in Abschnitt 2.3.6)
  Mit diesen metrischen Merkmalen lassen sich manche statistische und grafische Verfahren gut demonstrieren. Außerdem werden wir einfache ernährungswissenschaftliche Studien durchführen:
  - Vergleich des Körpergewichts mit dem Idealgewicht nach der veralteten Formel „Größe - 100“
  - Geschlechtsunterschied beim Body Mass Index

2.3.6 Der Fragebogen

1) Fragen zur Reaktion in ärgerlichen Situationen

Versetzen Sie sich bitte möglichst gut in folgende Situation:

Herr Meier und Herr Schulze waren mit demselben Taxi auf dem Weg zum Flughafen. Sie sollten zur selben Zeit, aber mit verschiedenen Maschinen abfliegen. Durch einen Stau kommen sie erst eine halbe Stunde nach der planmäßigen Abflugzeit am Flughafen an.

***Herr Meier** erfährt, dass seine Maschine pünktlich vor einer halben Stunde gestartet ist.*

***Herr Schulze** erfährt, dass seine Maschine Verspätung hatte und erst vor zwei Minuten gestartet ist.*

Wie sehr würden Sie sich **ärgern**, wenn Sie in der Situation von ...

Herrn Meier wären?	0	10	20	30	40	50	60	70	80	90	100
Herrn Schulze wären?	0	10	20	30	40	50	60	70	80	90	100

Betrachten Sie bitte die Antwortskala als "Ärgerthermometer".

2) Aussagen zur Selbsteinschätzung

Teilen Sie bitte für die folgenden Selbstbeschreibungen durch Ankreuzen einer Antwortkategorie mit, inwiefern die Aussagen auf Sie persönlich zutreffen.

	völlig falsch	falsch	unent-schieden	stimmt	stimmt genau
1. Auch in unsicheren Zeiten rechne ich im Allgemeinen damit, dass sich alles zum Besten wendet.	- -	-	o	+	+ +
2. Ich kann mich leicht entspannen.	- -	-	o	+	+ +
3. Wenn etwas schief gehen kann, dann passiert es mir auch.	- -	-	o	+	+ +
4. Bei allem sehe ich stets die negative Seite.	- -	-	o	+	+ +
5. Ich blicke kaum einmal mit Zuversicht in die Zukunft.	- -	-	o	+	+ +
6. Ich bin gern mit Freunden zusammen.	- -	-	o	+	+ +
7. Ich muss mich immer mit etwas beschäftigen.	- -	-	o	+	+ +
8. Ich habe stets die Hoffnung, dass die Dinge in meinem Sinne gehen.	- -	-	o	+	+ +
9. Die Dinge laufen immer so, wie ich es mir wünsche.	- -	-	o	+	+ +
10. Ich bin nicht leicht aus der Ruhe zu bringen.	- -	-	o	+	+ +
11. Ich glaube an den sprichwörtlichen "Silberstreifen am Horizont".	- -	-	o	+	+ +
12. Dass mir einmal etwas Gutes widerfährt, damit rechne ich kaum.	- -	-	o	+	+ +

3) Ihre Motive für die Teilnahme am SPSS-Kurs

a) Kreuzen Sie bitte in der folgenden Liste möglicher Motive für die Teilnahme am SPSS-Kurs alle für Sie zutreffenden Aussagen an und/oder nennen Sie Ihre sonstigen Motive.

Ich möchte SPSS kennenlernen, ...

- ❒ um eine eigene empirische Studie damit auszuwerten.
- ❒ weil in vielen Stellenanzeigen SPSS-Kenntnisse verlangt werden.
- ❒ weil ich mich um eine Stelle als Hilfskraft in der Forschung bewerben will.
- ❒ weil ich mich für Computer-Technik interessiere.
- ❒ weil ich mich für Statistik interessiere.

❒ Andere Motive: _____

_____

b) Möchten Sie im Kurs bestimmte statistische Methoden besonders gerne üben? Ja ○ Nein ○

Wenn „Ja“, welche? _____

_____

_____

4) Angaben zur Person

Geschlecht	Frau ○ Mann ○ Divers ○
Geburtsjahr	__ __ __ __
Primärer Fachbereich	○ I ○ II ○ III ○ IV ○ V ○ VI ○ VII
Körpergröße	__,__ __ m
Körpergewicht	__ __ __ kg

2.4 Strukturierung und Codierung der Daten

Wir werden die mit unserem Fragebogen erhobenen Informationen später manuell mit dem SPSS-Dateneditor erfassen und erstellen daher einen **Codierplan** mit genauen Handlungsanweisungen für die Erfassung.

Bei einer *Online*-Erhebung wird *kein* Codierplan als Arbeitsvorschrift für Datenerfasser benötigt, jedoch kann auch hier eine Dokumentation der Daten nützlich sein (z. B. für die Kooperation in einer Arbeitsgruppe). Die in Abschnitt 2.4 zu beschreibenden Festlegungen (z. B. zur Codierung von Merkmalsausprägungen) werden bei einer Online-Erhebung teilweise bei der Projektdeklaration gegenüber dem zuständigen Programm geregelt und teilweise von dieser Software entschieden. Bei manchen Aufgaben sind aber Urteilsvermögen und Handarbeit von Menschen durch keine Software zu ersetzen, z. B. bei der Behandlung der Antworten auf offene Fragen (siehe Abschnitt 2.4.2.4). Einige Überlegungen (z. B. zu Mehrfachwahlfragen) sind relevant für das Design eines Online-Fragebogens (siehe Abschnitt 2.4.2.3). Insgesamt kann daher der Abschnitt 2.4 auch solchen Lesern zur Lektüre empfohlen werden, die zu einer Online-Erhebung tendieren.

2.4.1 Fälle und Merkmale in SPSS

Wir haben bereits daran erinnert, dass in einer empirischen Studie bei den einbezogenen Fällen bzw. Untersuchungseinheiten die Ausprägungen von Merkmalen bzw. Merkmalsoperationalisierungen festgestellt werden (siehe z.B. Abschnitt 2.3.2.1). Nun wollen wir uns ansehen, wie die Merkmalsausprägungen der Fälle im SPSS-System gespeichert werden. Die konkrete Demonstration von KFA-Beispieldaten im **SPSS-Dateneditorfenster** wird das Verständnis der anschließenden, wieder eher allgemein-methodologisch geprägten, Ausführungen unterstützen. U. a. werden dabei auch einige zentrale SPSS-Begriffe erläutert.

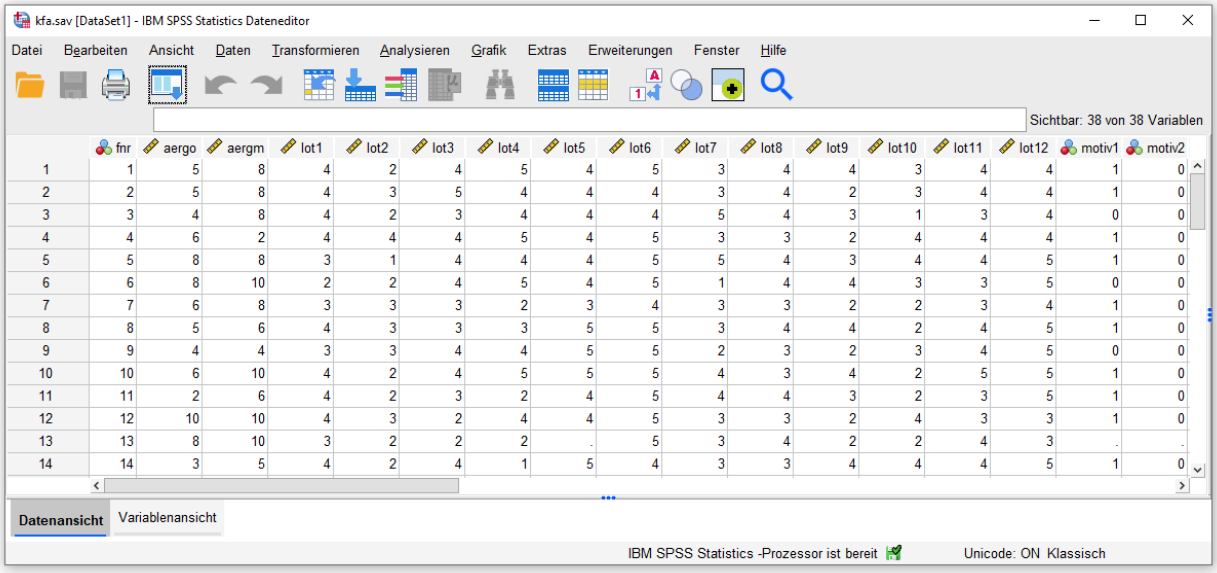
a) Variable

Der Begriff *Variable* wird in der Literatur zur statistischen Datenanalyse häufig synonym zum Begriff *Merkmal* gebraucht. Wir wollen ihn SPSS-konform in einer etwas technischeren Bedeutung verwenden: Schreibt man für ein Merkmal die Ausprägungen aller Fälle in der Stichprobe untereinander, so entsteht ein Spaltenvektor, und einen solchen Spaltenvektor wollen wir als *Variable* bezeichnen. Genau genommen gehören zu einer SPSS-Variablen auch noch etliche Attribute, z. B. ein Name und ein Messniveau.

Zwar resultieren Variablen meist (wie eben beschrieben) aus jeweils *einem* Merkmal, doch kann z. B. das Bemühen um eine rationelle Datenerfassung zu Ausnahmen führen. In Kürze wird eine Technik vorgeschlagen, die zur Erfassung von 100 Merkmalen mit Hilfe von sieben Variablen führt.

b) Datenmatrix und Dateneditor

Schreibt man alle Variablen eines Projekts nebeneinander, so entsteht eine (Fälle × Variablen) - **Datenmatrix** (Datentabelle). Sie kann in einem Fenster des **SPSS-Dateneditors** aufgebaut und auch dort während der Auswertungsarbeit ständig eingesehen und modifiziert werden. Die folgende Abbildung zeigt ein Dateneditorfenster mit den KFA-Beispieldaten aus einem früheren statistischen Praktikum:



Die zugrunde liegende Stichprobe dient im Manuskript häufig zur Demonstration und wird fortan als *Manuskriptstichprobe* bezeichnet.

Jede Variable, d. h. jede Spalte der Datenmatrix, besitzt einen eindeutigen **Variablennamen**, über den sie bei der Anforderung von Analysen und Diagrammen angesprochen werden kann.

Nachdem Sie einen exemplarischen Eindruck vom *Ziel* der Strukturierungs- und Codierungsbemühungen gewonnen haben, werden wir nun einige Details behandeln und einen Codierplan für unser Demonstrationsprojekt erstellen. Dabei soll u. a. angestrebt werden, den Aufwand und die Fehlergefahr bei der manuellen Datenerfassung möglichst gering zu halten.

2.4.2 Strukturierung

Welche SPSS-Variablen im oben definierten Sinn sollen zur Aufnahme der mit unserem Fragebogen erfassten Informationen definiert werden? Obwohl die Antwort auf diese Frage trivial zu sein scheint, sind doch zu einigen Themen kurze Erläuterungen angebracht.

2.4.2.1 Variablen zur Fallidentifikation

Über die empirischen Variablen hinaus sollten in die Datenmatrix stets organisatorische Variablen zur Fallidentifikation aufgenommen werden, was u. a. folgende Vorteile bietet:

- In der Regel besteht für die Identifikationsvariablen Eindeutigkeitszwang, sodass **Dubletten** (doppelt vorhandene Fälle) leicht identifiziert werden können.
- Sind zu den Fällen schriftliche oder sonstige Untersuchungsdokumente vorhanden, sollten diese eindeutig mit den Daten im Rechner **verknüpft** sein. Eine solche Korrespondenz ist für eventuelle spätere Kontrollen oder Korrekturen der Daten erforderlich.

Meist verwendet man zur Fallidentifikation eine *einzelne* Variable, die z. B. FNR (für *Fallnummer*) genannt werden kann. Natürlich muss die Fallidentifikation ggf. auch auf den schriftlichen oder sonstigen Untersuchungsdokumenten eingetragen werden. Bei personbezogenen Daten wählt man aus Datenschutzgründen zur Fallidentifikation z. B. eine zufällig vergebene Nummer.

Möglicherweise erscheint Ihnen bei der Erfassung schriftlicher Untersuchungsdokumente das Eintippen einer Identifikationsvariablen überflüssig, weil im SPSS-Dateneditor (siehe Abbildung in Abschnitt 2.4.1) die Zeilen bzw. Fälle ohnehin fortlaufend nummeriert sind. Die Nummern der Datenfensterzeilen stellen jedoch die gewünschte Korrespondenz zwischen den Datensätzen im Rechner und den nummerierten schriftlichen (oder sonstigen) Untersuchungsdokumenten *nicht zuverlässig* her. Die Nummerierung der Datenfensterzeilen kann sich nämlich leicht ändern, z. B. wenn ein Sortieren der Fälle nötig wird, oder wenn Fälle gelöscht oder eingefügt werden.

2.4.2.2 Abgeleitete Variablen gehören nicht in den Codierplan

In der Regel sind in einem Forschungsprojekt nicht nur die direkt erhobenen *Rohvariablen* von Interesse, sondern auch darauf aufbauende Variablen. Im KFA-Projekt soll etwa der Optimismus der Untersuchungsteilnehmer durch ihre mittlere Antwort auf die Fragen im Life Orientation Test (LOT) geschätzt werden. SPSS verfügt über leistungsfähige Dialogboxen bzw. Befehle zur Berechnung neuer Variablen aus bereits vorhandenen, sodass derartige Routinearbeiten keinesfalls während der Datenerfassung (z. B. per Taschenrechner) erledigt werden sollten. Freilich müssen nach diesem Vorschlag *alle* Ausgangsvariablen erfasst werden, was aber vielfach ohnehin erforderlich ist (z. B. zur Überprüfung messtechnischer Eigenschaften der Ausgangsvariablen per Faktoren- und/oder Reliabilitätsanalyse). Erfassen Sie also ausschließlich Rohvariablen, und führen Sie alle erforderlichen Transformationen später mit SPSS-Techniken durch. Wir

werden uns im weiteren Kursverlauf mit den SPSS-Transformationsmethoden ausführlich beschäftigen.

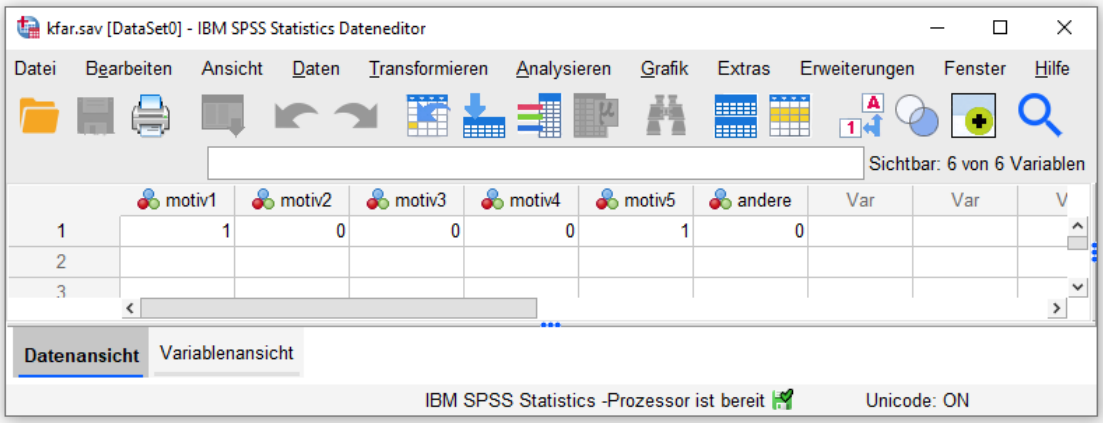
2.4.2.3 Mehrfachwahlfragen

Im Teil 3a unseres Fragebogens teilen die Untersuchungsteilnehmer für fünf konkrete Motive und eine Restkategorie mit, ob sie bei der Entscheidung für die Kursteilnahme relevant waren. Damit erfahren wir von jeder Person sechs eigenständige Merkmalsausprägungen und benötigen (ohne Komprimierungsverfahren, siehe unten) folglich in der SPSS-Datenmatrix sechs Variablen, um die Antworten aufzunehmen, die wir z. B. durch die Zahlen 1 (für ein markiertes Motiv) und 0 (für ein nicht markiertes Motiv) codieren können.

Beim Umgang mit einer Mehrfachwahlfrage müssen Sie vor allem von der fixen Idee fernhalten, die Informationen zu allen Merkmalen sollten oder könnten in *eine* Variable verpackt werden. Das käme dem unsinnigen Versuch gleich, für jede Person *mehrere* Werte (z. B. Zahlen) in *eine* Zelle der SPSS-Datenmatrix einzutragen. So absurd das eben beschriebene Ziel dem Leser (hoffentlich) erscheinen mag, geistert es doch durch manche Köpfe.

2.4.2.3.1 Vollständige Sets aus dichotomen Variablen

In unserem Beispiel aus dem Fragebogenteil 3a führt also eine Mehrfachwahlfrage zu sechs dichotomen SPSS-Variablen, die jeweils die Information darüber enthalten, ob ein bestimmtes Motiv (bzw. ein sonstiges Motiv) vorlag oder nicht. Das folgende Datenfenster zeigt die sechs Variablen, hier mit den Namen MOTIV1 bis MOTIV5 und ANDERE, bei einem Fall mit dem Antwortmuster (1, 0, 0, 0, 1, 0):



Wir werden in Kapitel 16 ein sogenanntes **Mehrfachantwortset** bestehend aus diesen sechs Variablen definieren und mit seiner Hilfe eine gemeinsame Auswertung der Variablen vornehmen. An dieser Stelle müssen Sie jedoch unbedingt akzeptieren, dass wir es mit *sechs* Merkmalen und demzufolge mit *sechs* Variablen zu tun haben, die eine inhaltliche Verwandtschaft und ein gemeinsames dichotomes Format besitzen.

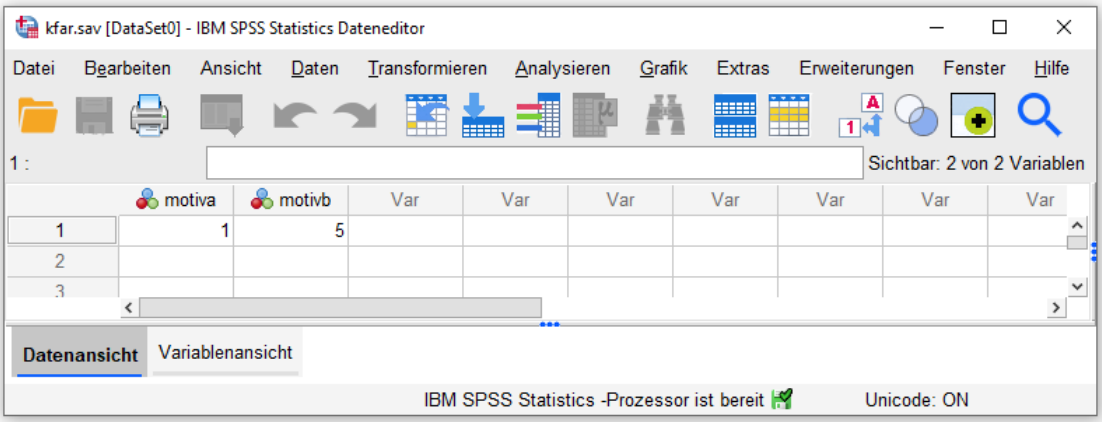
2.4.2.3.2 Sparsame Sets aus kategorialen Variablen

Das im letzten Abschnitt beschriebene Standardverfahren zur Übersetzung einer Mehrfachwahlfrage in SPSS-Variablen ist angemessen, sofern nicht zu viele Antwortmöglichkeiten im Spiel sind. Wenn Sie aber z. B. in einer touristischen Umfrage eine Liste mit 100 möglichen Freizeitaktivitäten präsentieren, dann führt das Schema zur Definition von 100 SPSS-Variablen. Unter der Annahme, dass jeder einzelne Untersuchungsteilnehmer maximal sieben Optionen wählen wird, ist das Schema bei der Datenerfassung mit dem SPSS-Dateneditor unpraktisch. Für solche Situationen bietet sich ein alternatives Vorgehen an, das im eben konstruierten Freizeitbeispiel lediglich sieben Variablen bzw. Spalten in der SPSS-Datenmatrix benötigt.

Auch dieses Komprimierungsverfahren soll an unserem Motivbeispiel demonstriert werden, obwohl es hier (bei nur sechs Antwortmöglichkeiten) definitiv ungeeignet ist. Unter der Annahme, dass pro Person maximal *zwei* verschiedene Motive zur Kursteilnahme zutreffen werden, definiert man die beiden SPSS-Variablen MOTIVA und MOTIVB, die jeweils folgende Werte annehmen können:

1 für das Motiv *Eigene empirische Studie*
2 für das Motiv *Orientierung am Arbeitsmarkt*
3 für das Motiv *Bewerbung als Hilfskraft*
4 für das Motiv *Interesse an Computer-Technik*
5 für das Motiv *Interesse an Statistik*
6 für andere Motive

Mit den Variablen MOTIVA und MOTIVB stehen für jede Person *zwei* Möglichkeiten zur Verfügung, um die Nummern von markierten Motiven zu erfassen. Das Antwortmuster (1, 0, 0, 0, 1, 0) wird folgendermaßen übertragen:



Im Prinzip kann man im Beispiel die beiden Werte 1 und 5 auch in umgekehrter Reihenfolge eintragen (MOTIVA = 5, MOTIVB = 1). Wesentlich ist nur, dass die Nummer jedes markierten Motivs bei MOTIVA oder MOTIVB als Wert auftritt. Von einer Person, die zwei Motive markiert hat, wissen wir *nicht*, welchem Motiv sie die größte Bedeutung beimisst. Daher können auch die resultierenden Variablen eine solche subjektive Ranginformation nicht enthalten. Allerdings wird man beim Erfassen der Systematik halber so vorgehen, dass in MOTIVA die Nummer des ersten markierten Motivs landet usw. (bei Anordnung von oben nach unten).

Wir sparen vier Variablen ein, wobei kein Informationsverlust eintritt, wenn tatsächlich pro Person maximal zwei Motive markiert werden. Erweist sich ein sparsames Set während der Erfas-

sung als unterdimensioniert, kann es bei Verwendung des SPSS-Dateneditors problemlos erweitert werden (z. B. um die Variable MOTIVC).

Auch bei der komprimierten Informationsanordnung kann man mit SPSS nach erfolgter Datenerfassung z. B. für jedes Motiv ermitteln, wie viel Prozent der Kursteilnehmer es markiert haben. Vor einer solchen Auswertung ist wiederum ein Mehrfachantwortset zu definieren, diesmal bestehend aus den beiden Variablen MOTIVA und MOTIVB. Bei manchen Auswertungen ist es aber erforderlich, über Transformationsanweisungen aus dem sparsamen kategorialen Set das vollständige dichotome Set (mit *einer* Variablen pro Merkmal) herzustellen (siehe Abschnitt 16.4).

2.4.2.4 Offene Fragen

Eine offene Frage bewirkt vielfältige Antworten. Der übliche Weg zur Systematisierung und Erfassung besteht darin, eine **Kategorienliste** zu entwickeln, die möglichst kurz ist und trotzdem alle Antworten der Teilnehmer gut repräsentiert. Anschließend kann man bei jedem Fall die vorhandenen bzw. fehlenden Nennungen von Kategorien (Listenelementen) analog zu den Antworten für die Items einer Mehrfachwahlfrage behandeln. Jedem Element der Kategorienliste entspricht ein dichotomes Merkmal.

Im Beispiel unseres Fragebogenteils 3b ist also durch Inspektion der ausgefüllten Fragebögen aller Teilnehmer eine Liste mit speziell gewünschten statistischen Auswertungsverfahren zu erstellen, z. B. mit dem Ergebnis:

Lineare Regressionsanalyse
Kreuztabellenanalyse
Faktorenanalyse
Logistische Regressionsanalyse

Zur Übernahme der Daten in SPSS-Variablen wird man bei einer relativ kurzen Kategorienliste ein vollständiges Set mit dichotomen Variablen verwenden, ansonsten ein sparsames Set mit kategorialen Variablen (siehe Abschnitt 2.4.2.3). Aus der obigen vierelementigen Liste mit speziellen methodischen Interessen entsteht also ein vollständiges Set mit dichotomen Variablen, z. B. mit den Namen:

LINREG	für die lineare Regressionsanalyse
KT	für die Kreuztabellenanalyse
FAKT	für die Faktorenanalyse
LOGREG	für die logistische Regressionsanalyse

Als Wert für die Variable LINREG ist bei der sich anbietenden (0/1) - Codierung eine 1 einzutragen, wenn ein Fall auf die offene Frage hin die lineare Regressionsanalyse angegeben und damit sein Interesse an dieser Methode signalisiert hat. Anderenfalls wird eine 0 notiert, die aber *nicht* als explizit bekundetes Desinteresse an der linearen Regressionsanalyse zu interpretieren ist.

Beim Erstellen einer Kategorienliste sind zu enge Kategorien (mit geringer Wahlfrequenz, z. B.: *Spearmans Rangkorrelation*) ebenso ungeeignet wie zu breite Kategorien (mit geringem Informationsgehalt, z. B.: *Zusammenhangsanalysen*). Vielfach wird man aber mit einer Restkategorie arbeiten (z. B.: *Sonstige Methoden*), um bei vertretbarem Aufwand möglichst alle Äußerungen der Probanden berücksichtigen zu können.

Das beschriebene Vorgehen erfordert zum Erstellen der Kategorienliste eine (speziell bei großen Stichproben) lästige Vorauswertung der Fragebögen, die sich mit folgendem Trick vermeiden lässt: Man verwendet eine **dynamisch wachsende Kategorienliste** in Verbindung mit einem sparsamen Set kategorialer Variablen. In unserem Beispiel kann man sich über ein sparsames Set mit den drei Variablen METH1 bis METH3 darauf vorbereiten, für jeden Fall maximal drei spezielle Auswertungsinteressen festzuhalten. Die Kategorienliste wird erst während der Datenerfassung entwickelt, indem man bei jedem Fall entscheidet, welche bereits definierten oder neu in die Liste aufzunehmenden Kategorien er im Fragebogenteil 3b angesprochen hat. Die Liste kann dynamisch um beliebig viele Kategorien erweitert werden, weil die drei Variablen beliebig viele verschiedene Kategoriennummern als Werte annehmen können. Selbstverständlich müssen die neu aufgenommenen Kategorien mit den vergebenen Nummern sorgfältig dokumentiert werden. Für die simultane Datenerfassung durch *mehrere* Personen sind dynamisch wachsende Kategorienlisten offenbar weniger geeignet.

Mit dem bislang vorgeschlagenen Verfahren lassen sich aus den Antworten auf eine offene Frage nur *dichotome* Merkmale extrahieren. Um ein *geordnet-kategoriales* Merkmal zu gewinnen, kann man eine Serie von Kategorien als Stufen dieses Merkmals festlegen und bei jedem Fall schlussendlich die höchste aufgetretene Stufe als Merkmals- bzw. Variablenausprägung festhalten.

Offene Fragen sind sicher sinnvoll, weil sie Informationen zutage fördern können, an die bei der Untersuchungsplanung niemand gedacht hat. Gelegentlich sind die Antworten jedoch so spärlich oder so schlecht strukturierbar, dass sich eine *statistische* Analyse nicht lohnt. So werden erfahrungsgemäß im Teil 3a des Beispielfragebogens kaum individuelle Motive zur Kursteilnahme angegeben, und wir ignorieren diese offene Frage im weiteren Projektverlauf.

Bilden in einer Studie längere Textpassagen (z. B. transkribierte Interviews) einen wesentlichen Bestandteil des empirischen Datenmaterials, dann sollte die Erstellung und Anwendung eines Kategoriensystems durch eine Spezialsoftware aus dem Bereich der sogenannten **qualitativen Datenanalyse** unterstützt werden. Auf einigen Pool-PCs an der Universität Trier steht für solche Fälle das Programm **MAXQDA** zur Verfügung.[1]

2.4.3 Codierung

Für jedes erhobene Merkmal muss festgelegt werden, wie die einzelnen Ausprägungen codiert werden sollen. Dabei ist eine Codierung durch einfach aufgebaute Werte anzustreben (z. B. durch positive, ganze Zahlen). Welche Werte (z.B. Zahlen oder Zeichenfolgen) bei einer Variablen erlaubt sind, hängt von ihrem Typ ab. Daher müssen wir bei den Überlegungen zur Codierung berücksichtigen, welche Variablentypen von SPSS unterstützt werden.

[1] http://www.maxqda.de/

2.4.3.1 Die wichtigsten Variablentypen in SPSS

An dieser Stelle beschränken wir uns auf die wichtigsten Variablentypen, mit denen die meisten Projekte auskommen:

- **Numerische Variablen**
  Werte: Zahlen (mit oder ohne Nachkommastellen)
  Z. B. geeignet für die Merkmale:
  - Ärgerausprägung
  - Geburtsjahr
  - Gewicht
- **Eingeschränkt numerische Variablen**
  Die sogenannten *eingeschränkt numerischen Variablen* beschränken sich auf nichtnegative ganze Zahlen. Sie respektieren führende Nullen und erlauben trotzdem numerische Funktionen.
  Werte: Nichtnegative ganze Zahlen, die links bis zu einer definierten Länge mit Nullen aufgefüllt werden
  Z. B. geeignet für das Merkmal: Postleitzahl des Wohnorts, z. B.: 01067
- **Zeichenfolgenvariablen (synonym: alphanumerische Variablen, String-Variablen)**
  Werte: Folgen von Zeichen (Buchstaben, Ziffern, Sonderzeichen) mit einer maximalen Länge von 32767 Zeichen.
  Z. B. geeignet für die Merkmale:
  - Lieblingsautor
  - Man könnte das Merkmal Geschlecht alphanumerisch kodieren mit den Werten **f**, **m** und **d**.
- **Datumsvariablen**
  Werte: Datumsangaben
  Z. B. geeignet für das Merkmal: Untersuchungsdatum

Anwendungsfälle für Datumsvariablen dürften in der Regel klar erkennbar sein.

Bei Merkmalen mit **mindestens ordinalem Messniveau** kommen nur numerische Variablen in Frage.

Bei Merkmalen mit **nominalem Messniveau** hat man die Wahl zwischen numerischer und alphanumerischer Codierung der Merkmalsausprägungen.

Beispiel Geschlecht:
- numerische Codierung: **1** für Frauen, **2** für Männer, **3** für divers
- alphanum. Codierung: **f** für Frauen, **m** für Männer, **d** für divers

Beim Arbeiten mit SPSS empfiehlt es sich, auch nominalskalierte Merkmale mit einer überschaubaren Anzahl von Ausprägungen numerisch zu kodieren, weil manche Auswertungsverfahren auch dort numerische Variablen verlangen, wo aus statistischer Sicht lediglich nominales Messniveau erforderlich ist (z. B. die Prozeduren REGRESSION, DISCRIMINANT, NPAR TESTS).

Für nominalskalierte Merkmale mit sehr vielen unterschiedlichen Ausprägungen (z. B. Lieblingsautor, letztes Urlaubsland) können Zeichenfolgenvariablen eine gute Lösung sein. Außerdem werden sie benötigt, wenn mit SPSS längere Texte erfasst werden sollen (z. B. von Probanden zu einem präsentierten Bild erfundene Geschichten).

2.4.3.2 Fehlende Werte

Trotz aller Sorgfalt sind in vielen Forschungsprojekten bei manchen Fällen einige Merkmalsausprägungen unbekannt, z. B. wegen verweigerter Antworten oder technischer Fehler. Bei der Codierungsplanung muss für alle betroffenen Variablen festgelegt werden, welche Ersatzwerte an Stelle fehlender oder ungültiger Werte in die Zellen der Datenmatrix eingetragen werden sollen. Diese Ersatzwerte bezeichnet man häufig als *MD-Indikatoren*, wobei *MD* für *Missing Data* steht. Wir konzentrieren uns bei der anschließenden Behandlung des Themas auf nicht eingeschränkte numerische Variablen.

2.4.3.2.1 Benutzerdefinierte MD-Indikatoren

Grundsätzlich kann jede Zahl als MD-Indikator für eine numerische Variable vereinbart werden, z. B. der Wert 9 bei einer verweigerten Auskunft über das in fünf Stufen erfasste und durch die Werte 1 bis 5 codierte Einkommen. Gelegentlich sind bei einer Variablen sogar *mehrere* MD-Indikatoren nötig. Wenn z. B. die Besucher einer touristischen Einrichtung per Ratingskala nach der Zufriedenheit mit ihrer Ferienunterkunft befragt werden, dann kann man für die resultierende SPSS-Variable die folgenden Codierungsregeln vereinbaren:

- Vorhandene Beurteilungen werden durch die Zahlen 1 bis 5 codiert.
- Ersatzwerte für fehlende Antworten:
  - Tagesgäste haben keine Ferienunterkunft und erhalten den MD-Indikator 8 (*Frage trifft nicht zu*).
  - Liegt bei Übernachtungsgästen keine Beurteilung vor, wird der MD-Indikator 9 vergeben.

Beachten Sie bei der Verwendung von benutzerdefinierten MD-Indikatoren folgende Regeln:

- Es ist klar, dass alle MD-Indikatoren zu einer Variablen außerhalb des validen Wertebereichs liegen müssen. So wäre z. B. die Zahl 99 kein geeigneter MD-Indikator für die im Demonstrationsprojekt vorhandene Variable Körpergewicht (gemessen in kg).
- Wählen Sie möglichst prägnante oder extreme Werte (also z. B. bei einer Variablen mit den validen Werten 1 und 2 den MD-Indikator 9). Dies bewirkt warnend auffällige Ergebnisse, falls Fälle mit fehlenden Werten nicht ordnungsgemäß von einer Analyse ausgeschlossen wurden.
- Der Einfachheit halber sollte für alle Variablen mit ähnlichem Wertebereich derselbe MD-Indikator verwendet werden.
- **Besonders wichtig: Die benutzerdefinierten MD-Indikatoren müssen dem SPSS-System unbedingt bekanntgemacht werden (siehe Abschnitt 4.2.2).**

Die für sogenannte **Residualkategorien** (z. B. „Weiß nicht“, „Keine Angabe“) vergebenen Codierungswerte müssen zumindest für bestimmte Auswertungen (z. B. Mittelwertberechnungen) ebenfalls als MD-Indikatoren deklariert werden.

2.4.3.2.2 System-Missing (SYSMIS)

Neben den vom Benutzer variablenspezifisch vereinbarten MD-Indikatoren verwendet SPSS für alle numerischen Variablen automatisch einen weiteren MD-Indikator, der mit *System-Missing*, *systemdefiniert fehlend* oder *SYSMIS* bezeichnet wird. Er kommt immer dann zum Einsatz, wenn SPSS auf eines von den folgenden Problemen trifft:

- Im Dateneditor oder beim Lesen einer Datendatei (z. B. im Textformat) findet SPSS im Feld einer als numerisch definierten Variablen unzulässige Zeichen oder überhaupt keinen Eintrag.
- Beim Neuberechnen einer Variablen per Transformationsanweisung (siehe Kapitel 7) fehlt ein Argument, oder der Funktionswert ist nicht definiert (z. B. bei einer Division durch null).

Gerade war u. a. zu erfahren, dass man bei der Datenerfassung mit dem SPSS-Dateneditor für eine numerische Variable den Ersatzwert SYSMIS ganz einfach dadurch vereinbaren kann, dass man in die betroffene Zelle *nichts* einträgt.

Tipp: Bei der Datenerfassung mit dem SPSS-Dateneditor können Sie für numerische Variablen routinemäßig SYSMIS als MD-Indikator verwenden, bei Bedarf ergänzt durch zusätzliche benutzerdefinierte MD-Indikatoren. Weil SYSMIS automatisch als Initialisierungswert für numerische Datenzellen verwendet wird, resultiert dieser Wert, wenn man eine frisch initialisierte Zelle unverändert lässt. Eine Deklaration dieses bei numerischen Variablen generell definierten MD-Indikators ist weder nötig noch möglich und kann daher auch nicht vergessen werden.

Im Datenfenster und in der Ergebnisausgabe wird SYSMIS durch einen Punkt dargestellt (siehe Abbildung in Abschnitt 2.4.1, Variable LOT5 bei Fall 13).

2.4.3.2.3 Fehlende Werte bei Zeichenfolgenvariablen

Aus den in Abschnitt 2.4.3.1 genannten Gründen werden wir in unserer Beispielstudie keine Zeichenfolgevariablen verwenden. In anderen Studien können sie jedoch sinnvoll sein. Als Indikatoren für fehlende Werte kommen in Frage:

- Leere Zeichenfolge
  Eine leere Zeichenfolge ist per Voreinstellung ein *gültiger* Wert. Soll sie als fehlender Wert interpretiert werden, ist eine entsprechende Definition erforderlich (siehe Abschnitt 4.2.2.2). Bei Zeichenfolgenvariablen gibt es *keine* Entsprechung zum SYSMIS-Wert der numerischen Variablen.
- Sonstige benutzerdefinierte Indikatoren
  SPSS erlaubt bei Zeichenfolgenvariablen benutzerdefinierte MD-Indikatoren mit maximal 8 Zeichen. Man kann z. B. den MD-Indikator „-88“ verwenden, wenn diese Zeichenfolge nicht als valider Wert auftritt.

2.4.3.2.4 Fehlende Werte bei Mehrfachwahl-Fragen und offenen Fragen

Nachdem der Sinn und die Verwendung von MD-Indikatoren geklärt sind, ist unser nächstes Thema eine spezielle Interpretationsunsicherheit im Zusammenhang mit fehlenden Werten, die bei Mehrfachwahlfragen auftreten kann. Das aus der Verwendung eines teilnehmer-freundlichen Antwortformats resultierende potentielle Problem lässt sich aber durch ein korrektes Design ver-

hindern. Im Fragebogenteil 3a zu den Motiven für die Kursteilnahme sorgt die sechste Markieralternative (*Andere Motive*) durch das **Komplettieren der Antwortmöglichkeiten** dafür, dass eine redliche Auskunftsperson mindestens eines der sechs Kästchen markieren muss. Wir dürfen annehmen, dass hinter einem aufwändigen Verhalten (Statistisches Praktikum mit SPSS besuchen) mindestens *ein* Motiv steht.

Ohne diese Restkategorie könnten wir bei einem zu erfassenden Fragebogenexemplar mit *fünf* leeren Motivkästchen die folgenden Möglichkeiten *nicht* unterscheiden:

- Bei der Auskunftsperson trifft tatsächlich keines der fünf vorgegebenen Motive zu.
- Die Person hat den Fragebogenteil 3a nicht bearbeitet (fehlende Daten).

Ursache für die Interpretationsunsicherheit ist das bei Mehrfachwahlfragen übliche Antwortformat, das pro Motiv nur *ein* Kästchen verwendet, statt jeweils ein Ja- *und* ein Nein-Kästchen vorzugeben. Damit ersparen wir den Untersuchungsteilnehmern zahlreiche Nein-Markierungen. Dies ist sinnvoll, damit die Kooperationsbereitschaft nicht überstrapaziert wird.

Wenn fehlende Antworten nicht von Nein-Antworten unterschieden werden können und infolgedessen als Nein-Antworten behandelt werden müssen, kommt es zu einer mehr oder weniger gravierenden Überschätzung des Nein-Anteils. Ein solches Artefakt ist als schwerwiegender Fehler unbedingt zu vermeiden.

Durch das Komplettieren der Antwortmöglichkeiten wird im Fragebogenteil 3a dafür gesorgt, dass Nein-Antworten sicher von fehlenden Antworten unterschieden werden können. Liegen bei einem Probanden sechs leere Kästchen vor, dann ist von fehlenden Antworten auszugehen.

Die Empfehlung, bei einer Mehrfachwahlfrage die Liste der Antwortmöglichkeiten zu komplettieren, ist offenbar auch bei Online-Studien relevant.

Bei der offenen Frage in Teil 3b wird durch die vorgeschaltete Frage, ob überhaupt spezielle Methoden gewünscht sind, dafür gesorgt, dass bei Fragebögen ohne eingetragene Methodeninteressen folgende Möglichkeiten unterschieden werden können:

- Die Person hat kein Interesse an speziellen Auswertungsmethoden.
- Die Person hat den Fragebogenteil 3b nicht bearbeitet (fehlende Daten).

2.4.3.2.5 Vereinfachung der Erfassung durch Datentransformationstechniken

Im Zusammenhang mit dem MD-Problem bei den Variablen zu unserem Fragebogenteil 3 wage ich nun einige Vorschläge, die zwar dem Datenerfasser das Leben erleichtern, aber zugegebenermaßen die Kursteilnehmer beim ersten Entwurf eines Codierplans durch zusätzliche Überlegungen belasten. Bei der Mehrfachwahlfrage nach den Kursmotiven haben wir geschickt durch die sechste Ankreuzalternative *Andere Motive* dafür gesorgt, dass Personen mit fehlenden Werten sicher zu identifizieren sind. Wir könnten nun die Erfasser per Codierplan beauftragen:

- Schreibe bei den Variablen MOTIV1 bis MOTIV5 und ANDERE den Wert 1, wenn das zugehörige Kästchen markiert ist, sonst eine 0.
- Ist aber *keines* der sechs Kästchen markiert, dann erhalten die Variablen MOTIV1 bis MOTIV5 und ANDERE den MD-Indikator SYSMIS.

Die im zweiten Satz enthaltene Regel lässt sich in SPSS mit später anzuwendenden Datentransformationen bequem automatisieren, sodass wir die Erfasser nicht damit belasten sollten. Damit wird die Lösung des MD-Problems zugunsten einer möglichst einfachen Erfassung in die spätere

Projektphase der Datentransformation verschoben. Schlussendlich soll für die Variablen MOTIV1 bis MOTIV5 und ANDERE folgende Codierung sichergestellt sein:

0	=	Nein
1	=	Ja
SYSMIS	=	Wert unbekannt

Zur Erfassung der Informationen im Fragebogenteil 3b wollen wir eine dynamische Kategorienliste mit einem zugehörigen sparsamen Set kategorialer Variablen METH1 bis METH3 (vgl. Abschnitt 2.4.2.4) entwickeln. Der damit schon reichlich belastete Erfasser soll folgendermaßen vorgehen (bei Verwendung des SPSS-Dateneditors):

- Die Antwort auf die Frage, ob spezielle Methodenwünsche bestehen, wird konventionell durch eine Variable mit dem Namen SMG (**S**pezielle **M**ethoden **G**ewünscht) und folgender Codierungsvorschrift erfasst:

0	=	Nein
1	=	Ja
System-Missing	=	keine Antwort

- In die Dateneditorzellen zu den Variablen METH1 bis METH3 sollen die Kategoriennummern der gewünschten Methoden eingetragen werden. Bei weniger als drei Nennungen soll in die überflüssigen Zellen nichts eingetragen werden, was zum MD-Indikator SYSMIS führt. Diese Regel erleichtert die Erfassung und hat noch einen weiteren Vorteil: Sollte sich herausstellen, dass zusätzliche Variablen METH4 etc. benötigt werden, können wir diese ergänzen, ohne bei bereits erfassten Fällen irgendwelche Ersatzwerte (z. B. Nullen) nachtragen zu müssen.

Bei den Variablen METH1 bis METH3 ist später mit Datentransformationen dafür zu sorgen, dass ihre Ausprägungen zuverlässig folgendermaßen interpretiert werden können:

0	=	Von der *i*-ten (i = 1,..,3) Option zur Nennung einer interessierenden Methode wurde kein Gebrauch gemacht.
natürliche Zahl ≥ 1	=	Die Methode mit dieser Kategoriennummer wurde angegeben.
SYSMIS	=	Wert unbekannt

Dazu müssen unter den verschiedenen Wertekonstellationen der Variablen SMG und METH1 bis METH3 die folgenden Anpassungen vorgenommen werden:

		Mindestens eine speziell interessierende Methode angegeben?	
		Ja	Nein
SMG	1	METH1 ... METH3: SYSMIS → 0 Bem.: Korrektes Antwortverhalten. Variablen zu nicht benutzten Optionen (gem. Codierplan auf SYSMIS) werden auf 0 gesetzt.	SMG: 1 → SYSMIS Bem.: Irreguläres Antwortverhalten. METH1 bis METH3 behalten SYMIS. SMG wird ebenfalls auf SYMIS gesetzt.
	0	SMG: 0 → 1 METH1 ... METH3: SYSMIS → 0 Bem.: Leicht irreguläres Antwortverhalten. Wir sind großzügig und setzen SMG auf 1 sowie die Variablen zu nicht benutzten Optionen auf 0.	METH1 ... METH3: SYSMIS → 0 Bem.: Korrektes Antwortverhalten. Die Variablen zu allen Optionen (gem. Codierplan auf SYSMIS) werden auf 0 gesetzt.
	SYSMIS	SMG: SYSMIS → 1 METH1 ... METH3: SYSMIS → 0 Bem.: Leicht irreguläres Antwortverhalten. Wir sind großzügig und setzen SMG auf 1 sowie die Variablen zu nicht benutzten Optionen auf 0.	Bem.: Irreguläres Antwortverhalten. Alle Variablen behalten den Wert SYSMIS.

Vermutlich kam beim Lesen der letzten Ausführungen wenig Freude auf. Das MD-Problem verursacht oft erheblichen Aufwand, wobei auch Ermessenentscheidungen gefragt sind. Jedenfalls sind die vorgeschlagenen Methoden zur Erfassung der Informationen aus dem Fragebogenteil 3 recht simpel und praktikabel.

2.4.3.3 Fehlerquellen bei der manuellen Datenerfassung minimieren

Wenn die Daten manuell erfasst werden, ist bei den Codierungsvereinbarungen darauf zu achten, dass dem Erfasser keine zeitaufwändigen und fehleranfälligen Arbeiten zugemutet werden, z. B.:

- Treten gebrochene Zahlen als Werte auf (z. B. bei unserer Frage nach der Körpergröße), so kann man durch Wechsel der Maßeinheit das lästige Dezimaltrennzeichen eliminieren.
  Beispiel: 1,65 m → 165 cm
- Bei bipolaren Skalen (z. B. bei unseren LOT-Fragen) empfiehlt sich eine Codierung durch ausschließlich positive Werte z. B.:

  \- - → 1
  \- → 2
  o → 3
  \+ → 4
  \+ \+ → 5

  Durch die Vermeidung negativer Werte spart man Tipparbeit und macht keine Fehler durch vergessene Vorzeichen.

- Wurden einige Fragen aus messtechnischen Gründen umgepolt (negativ formuliert), was im KFA-Projekt bei einigen LOT-Fragen geschehen ist, dann sollte diese Umpolung keinesfalls während der Erfassung rückgängig gemacht werden. Dies gelingt sehr viel bequemer und ohne Fehlerrisiko mit den Transformationsmöglichkeiten von SPSS (siehe Kapitel 7).
- Dass Erfasser nicht durch die Berechnung von abgeleiteten (z. B. als Mittelwert oder Differenz definierten) Variablen belastet werden sollen, wurde schon in Abschnitt 2.4.2.2 erläutert.

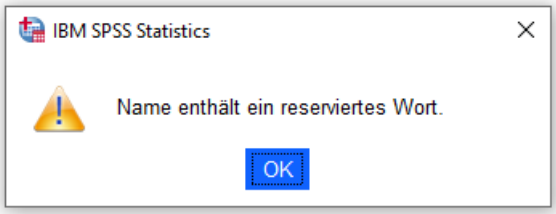
2.4.3.4 Variablennamen

In den bald zu erstellenden Codierplan (siehe Abschnitt 2.4.3.5) sollen auch die SPSS-Variablennamen aufgenommen werden. Daher beschäftigen wir uns im aktuellen Abschnitt mit Regeln und Empfehlungen zur Bildung von Variablennamen:

- Maximal 64 Zeichen
  Die frühere Beschränkung von SPSS-Variablennamen auf 8 Zeichen ist längst überwunden, doch sollte man sich weiterhin möglichst kurzfassen. Lange Namen belegen viel Platz (z. B. in der Kopfzeile des Dateneditorfensters) und sind beim Einsatz von SPSS-Syntax (siehe z. B. Kapitel 7) umständlich. Außerdem existieren immer noch SPSS-Erweiterungen von Drittanbietern, die bei Variablennamen mit mehr als 8 Zeichen Probleme haben. Lange Namen werden von solchen Erweiterungen entweder abgelehnt oder auf 8 Zeichen gekürzt, wobei die Eindeutigkeit verloren gehen kann. Im Manuskript werden daher im Sinne maximaler Kompatibilität Variablennamen mit mehr als 8 Zeichen vermieden.
- Das erste Zeichen muss ein Buchstabe sein.
- An den restlichen Positionen sind folgende Zeichen zugelassen: Buchstaben, Ziffern sowie die Symbole @, #, _ und $. Von der zweiten bis zur vorletzten Position ist außerdem der Punkt erlaubt.
- Aus den eben genannten Regeln ergibt sich insbesondere, dass Leerzeichen in Variablennamen verboten sind.
- Die von älteren SPSS-Versionen abgelehnten Umlaute in Variablennamen werden mittlerweile akzeptiert. Seit der SPSS-Version 16 sind auch beim Übergang zu einem alternativen Betriebssystem keine Zeichensatzprobleme bei Variablennamen mehr zu befürchten. Bei Beteiligung anderer Programme, die z.B. eine von SPSS erstellte Datendatei lesen sollen, sind aber weiterhin Pannen durch Umlaute in Variablennamen möglich. Im Manuskript werden daher im Sinne maximaler Kompatibilität Umlaute in Variablennamen vermieden.
- Die folgenden Schlüsselwörter der SPSS-Kommandosprache dürfen nicht als Variablennamen verwendet werden: ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO, WITH.

- Die Groß-/Kleinschreibung ist irrelevant hinsichtlich der *Identifikation* von Variablen. Erscheint ein Variablenname in der *Ausgabe* (d. h. in Tabellen oder Diagrammen), dann verwendet SPSS die Schreibweise aus der Variablendefinition. Erhält eine Variable neben dem obligatorischen Namen auch eine optionale *Beschriftung* (siehe unten), dann erscheint in der Ausgabe die Beschriftung statt des Namens. Die Groß-/Kleinschreibung im Variablennamen wirkt sich also nur bei fehlender Beschriftung sich auf die Ausgabe aus. In Manuskript werden SPSS-Variablennamen zur Hervorhebung in Großbuchstaben geschrieben.

Beim Versuch, einen irregulären Variablennamen zu vereinbaren, erhalten Sie im SPSS-Dateneditor eine meist informative Fehlermeldung, z. B.:



Tipps zur Benennung:

- Bilden Sie möglichst *informative* Namen wie z. B. GEBJ und FB für *Geburtsjahr* und *Fachbereich* an Stelle technischer Bezeichnungen wie T4_F2 (Teil 4, Frage 2) und T4_F3.
- Die eben genannte Regel sollte in einem speziellen Fall relativiert werden: Bei Serien verwandter Variablen, z. B. resultierend aus den 12 Fragen des Life Orientation Tests (LOT) im Teil 2 unseres Fragebogens, ist es in der Regel umständlich und wenig rentabel, entsprechend viele individuell-vollinformative Variablennamen zu bilden. Hier ist meist eine Indexschreibweise günstiger, bei der an einen informativen Namensstamm eine fortlaufende Nummer angehängt wird, z. B. LOT1, LOT2, ...

2.4.3.5 Codierplan

Die Festlegungen zur Strukturierung und Codierung der Projektdaten sollten (zumindest bei größeren Projekten) in einem **Codierplan** dokumentiert werden. Er hat zwei Funktionen:

- Während der Erfassung aufgrund von schriftlichen Untersuchungsdokumenten regelt er, wie die Daten eines Falles ins SPSS-Dateneditorfenster einzutragen bzw. mit einem anderen Programm zu erfassen sind.
- Später dient er als kompakte Beschreibung der entstandenen Datendatei.

In unserem KFA-Projekt kann für die geplante Erfassung mit dem SPSS-Dateneditor der folgende Codierplan verwendet werden:

Merkmal	SPSS-Var.-name	Codierung	Bemerkungen
Fallnummer	FNR	MD-Indikator: entfällt	
Ärger als Herr Meier (ohne KFA)	AERGO	0 = 0 1 = 10 . . . 10 = 100 MD-Indikator: SYSMIS	Wir sparen uns per Division durch Zehn viel Schreibarbeit und haben dabei eine bei Intervallskalen zulässige Transformation vorgenommen.
Ärger als Herr Schulze (mit KFA)	AERGM	0 = 0 1 = 10 . . . 10 = 100 MD-Indikator: SYSMIS	
LOT-Fragen	LOT1 bis LOT12	1 = -- 2 = - 3 = o 4 = + 5 = ++ MD-Indikator: SYSMIS	
Kursmotive	MOTIV1 bis MOTIV5, ANDERE	0 = Nicht markiert 1 = Markiert	SYSMIS wird **nicht** vergeben! Die MD-Behandlung erfolgt später per Datentransformation.
Spezielle Methoden gewünscht?	SMG	0 = Nein 1 = Ja MD-Indikator: SYSMIS	
Gewünschte statistische Methoden	METH1 bis METH3	1 = Meth.-Kat. 1 gewünscht 2 = Meth.-Kat. 2 gewünscht . . . Bei weniger als drei Nennungen: SYSMIS-Initialisierungen belassen	Die Kategorienliste wird während der Erfassung nach Bedarf entwickelt und dokumentiert. Die MD-Behandlung erfolgt später per Datentransformation.
Geschlecht	GESCHL	1 = Frau 2 = Mann 3 = Divers MD-Indikatoren: SYSMIS	
Geburtsjahr	GEBJ	**vierstellige** Eingabe (z. B. 1984)! MD-Indikator: SYSMIS	
Fachbereich	FB	1 = I (Pädag., Philos., Psychol.) 2 = II (Sprachen) 3 = III (Hist. und polit. Wiss.) 4 = IV (BWL, Ethnol., Inform., Mathe, Soziol., VWL, Wirtsch.-Inf.) 5 = V (Jura) 6 = VI (Geowissenschaften) 7 = VII (Theologie) MD-Indikator: SYSMIS	
Körpergröße	GROESSE	Eingabe in **cm**! MD-Indikator: SYSMIS	
Körpergewicht	GEWICHT	Eingabe in kg MD-Indikator: SYSMIS	

Dieser Codierplan ist bei der Datenerfassung einfach zu handhaben und leistet damit einen Beitrag zur Qualität der Daten.

Bei der Erfassung mit dem SPSS-Dateneditor startet man mit einer Variablendeklaration, wobei einige Regeln des Codierplans einfließen (vgl. Abschnitt 4.2.2). Damit stellt sich die Frage, ob man auf einen Codierplan verzichten und das Regelwerk direkt im Deklarationsteil einer SPSS-Datendatei unterbringen kann. Allerdings enthält unser Beispiel viele Vorschriften (z. B. vierstellige Erfassung des Geburtsjahrs, Erfassung der Körpergröße in cm), die per SPSS-Variablendeklaration *nicht* hinreichend klar dokumentiert werden können.

Sind mehrere Personen bei der Erfassung und/oder Auswertung der Daten beteiligt, ist ein schriftlicher Codierplan unbedingt erforderlich. Wenn die Daten jedoch nur von *einer* Person analysiert werden, die zudem auch die Erfassung übernimmt, dann kann auf einen Codierplan verzichtet werden.

3 Durchführung der Studie (inklusive Datenerhebung)

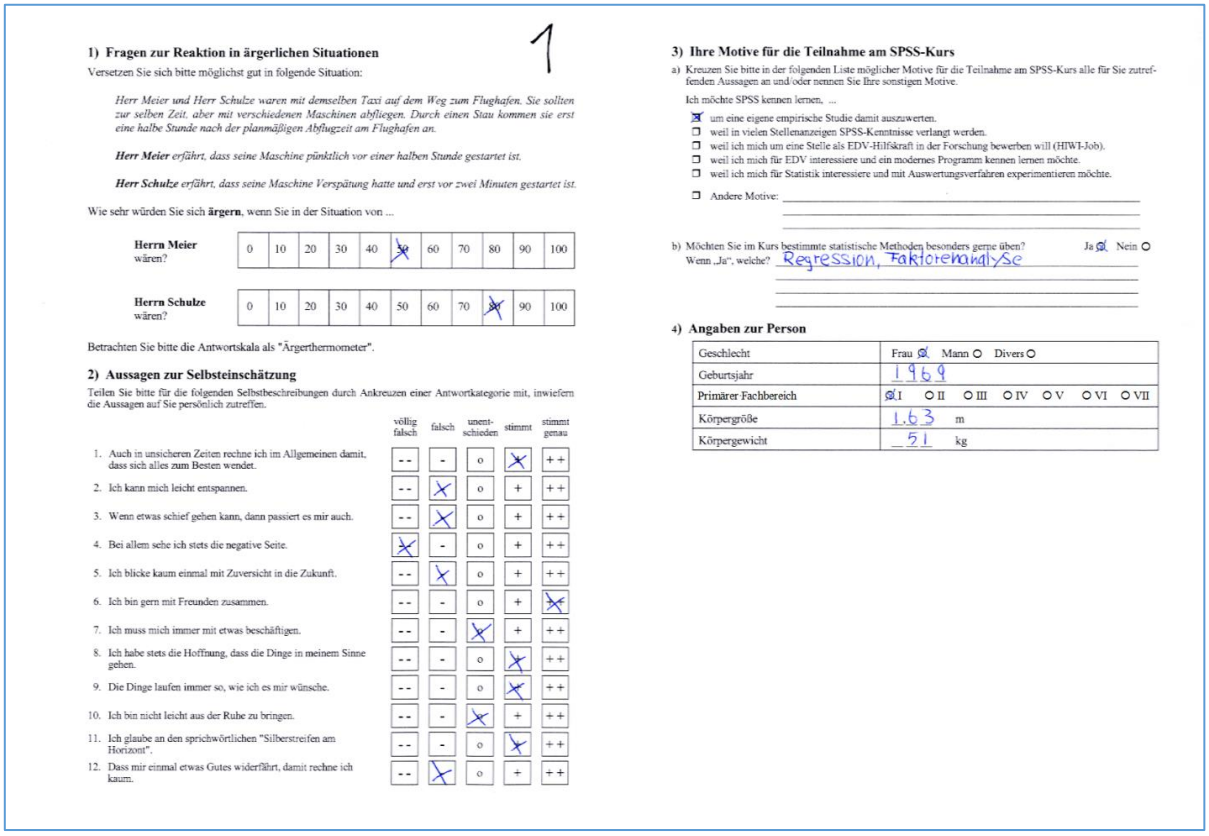
Bei den in Abschnitt 2.4 beschriebenen Überlegungen zur Strukturierung und Codierung der Daten hat sich ergeben, dass der in Abschnitt 2.3 präsentierte Fragebogen ohne Korrekturen eingesetzt werden kann. Damit steht der Durchführung der Befragung nichts mehr im Weg.

Oft ist eine Vorstudie (ein Pretest) angemessen, um den Untersuchungsplan sowie die Erfassungsinstrumente auf Schwächen zu untersuchen (z. B. schwer verständliche Frageformulierungen). Im Pretest zu einer Online-Studie erfährt man z. B., bei welcher Fragebogenseite die meisten Abbrüche auftreten.

Im realen Kursverlauf haben die Teilnehmer noch im Zustand der „naiven Unbefangenheit“ (ohne Kenntnis der KFA-Theorie) die Rolle der Probanden übernommen und so ihre eigenen Daten produziert. Die Leser(innen) im Selbststudium werden wohl aus praktischen Gründen in der Regel auf die Durchführung einer eigenen KFA-Erhebung verzichten. Im weiteren Verlauf des Manuskripts werden die in einem früheren Kurs erhobenen Daten analysiert. Die zugehörigen Dateien können via Internet bezogen werden (siehe Vorwort).

3.1 Das gute alte Papier

Im Manuskript wird auch noch der besonders arbeitsintensive Fall einer Studie mit papiergestützter Erhebung berücksichtigt, obwohl in den realen Kursen aus Zeitgründen eine Online-Erhebung stattfindet. Anschließend ist der ausgefüllte Fragebogen derjenigen Untersuchungsteilnehmerin zu sehen, die bei der zufälligen Vergabe einer Fallidentifikation (vgl. Abschnitt 2.4.2.1) die Nummer 1 erhalten hat:



1

1) Fragen zur Reaktion in ärgerlichen Situationen

Versetzen Sie sich bitte möglichst gut in folgende Situation:

Herr Meier und Herr Schulze waren mit demselben Taxi auf dem Weg zum Flughafen. Sie sollten zur selben Zeit, aber mit verschiedenen Maschinen abfliegen. Durch einen Stau kommen sie erst eine halbe Stunde nach der planmäßigen Abflugzeit am Flughafen an.

***Herr Meier** erfährt, dass seine Maschine pünktlich vor einer halben Stunde gestartet ist.*

***Herr Schulze** erfährt, dass seine Maschine Verspätung hatte und erst vor zwei Minuten gestartet ist.*

Wie sehr würden Sie sich **ärgern**, wenn Sie in der Situation von ...

	0	10	20	30	40	50	60	70	80	90	100
Herrn Meier wären?						X					
Herrn Schulze wären?									X		

Betrachten Sie bitte die Antwortskala als "Ärgerthermometer".

2) Aussagen zur Selbsteinschätzung

Teilen Sie bitte für die folgenden Selbstbeschreibungen durch Ankreuzen einer Antwortkategorie mit, inwiefern die Aussagen auf Sie persönlich zutreffen.

	völlig falsch	falsch	unentschieden	stimmt	stimmt genau
1. Auch in unsicheren Zeiten rechne ich im Allgemeinen damit, dass sich alles zum Besten wendet.	--	-	o	X	++
2. Ich kann mich leicht entspannen.	--	X	o	+	++
3. Wenn etwas schief gehen kann, dann passiert es mir auch.	--	X	o	+	++
4. Bei allem sehe ich stets die negative Seite.	X	-	o	+	++
5. Ich blicke kaum einmal mit Zuversicht in die Zukunft.	--	X	o	+	++
6. Ich bin gern mit Freunden zusammen.	--	-	o	+	X
7. Ich muss mich immer mit etwas beschäftigen.	--	-	X	+	++
8. Ich habe stets die Hoffnung, dass die Dinge in meinem Sinne gehen.	--	-	o	X	++
9. Die Dinge laufen immer so, wie ich es mir wünsche.	--	-	o	X	++
10. Ich bin nicht leicht aus der Ruhe zu bringen.	--	-	X	+	++
11. Ich glaube an den sprichwörtlichen "Silberstreifen am Horizont".	--	-	o	X	++
12. Dass mir einmal etwas Gutes widerfährt, damit rechne ich kaum.	--	X	o	+	++

3) Ihre Motive für die Teilnahme am SPSS-Kurs

a) Kreuzen Sie bitte in der folgenden Liste möglicher Motive für die Teilnahme am SPSS-Kurs alle für Sie zutreffenden Aussagen an und/oder nennen Sie Ihre sonstigen Motive.

Ich möchte SPSS kennen lernen, ...

- ☒ um eine eigene empirische Studie damit auszuwerten.
- ☐ weil in vielen Stellenanzeigen SPSS-Kenntnisse verlangt werden.
- ☐ weil ich mich um eine Stelle als EDV-Hilfskraft in der Forschung bewerben will (HIWI-Job).
- ☐ weil ich mich für EDV interessiere und ein modernes Programm kennen lernen möchte.
- ☐ weil ich mich für Statistik interessiere und mit Auswertungsverfahren experimentieren möchte.
- ☐ Andere Motive: _____

b) Möchten Sie im Kurs bestimmte statistische Methoden besonders gerne üben? Ja ☒ Nein ○
Wenn „Ja“, welche? Regression, Faktorenanalyse

4) Angaben zur Person

Geschlecht	Frau ☒ Mann ○ Divers ○
Geburtsjahr	1969
Primärer Fachbereich	☒ I ○ II ○ III ○ IV ○ V ○ VI ○ VII
Körpergröße	1.63 m
Körpergewicht	51 kg

Diese Nummer wurde nachträglich von der Untersuchungsleitung auf den Fragebogen geschrieben.

Das Papier ist in manchen Erhebungssituationen immer noch im Vorteil, weil keine Hardware, keine sonstigen IT-Ressourcen (z.B. Netzwerkverbindung) und keine IT-Kompetenzen der Probanden benötigt werden.

In manchen Abschnitten des Manuskripts leidet die Lesbarkeit möglicherweise unter der Berücksichtigung von zwei verschiedenen Erhebungsmethoden (Papier und Online). Ab Kapitel 5 spielt die Datenerhebung keine Rolle mehr.

3.2 Online-Datenerhebung

Wenn die nötigen technischen Voraussetzungen gegeben sind, dann sollte eine Online-Datenerhebung eingesetzt werden. Hiermit sind Verfahren gemeint, bei denen die Untersuchungsteilnehmer ihre Daten (aktiv oder passiv) direkt an eine digitale Technik übergeben:

- Schriftliche Befragungen werden mittlerweile routinemäßig via Internet realisiert, wenn die zu untersuchende Population auf diesem Weg erreichbar ist.
- Zur Steuerung experimenteller Abläufe oder zur hochgenauen Messung von Reaktionszeiten werden traditionell spezielle Rechner im Forschungslabor verwendet. Neuerdings kommen aber auch internet-basierte Lösungen für solche Zwecke zum Einsatz (siehe z. B. Kim, Gabriel & Gygax 2019).
- Für eine kontinuierliche, alltagsbegleitende Datenerfassung kommen Smartphones oder andere Rechner im Taschenformat in Betracht.

Nach Abschluss der Datenerhebung kann die Auswertung in günstigen Fällen sofort beginnen, weil die Daten automatisch in einer Datei landen, die von SPSS gelesen werden kann. Auf eine gelegentliche Kontrolle (z. B. wegen möglicher Defekte im Aufzeichnungsverfahren) sollte man aber nicht verzichten. Die Daten*erfassung* als eigenständige Arbeitsphase entfällt jedenfalls bei den Online-Verfahren.

Von den Online-Datenerhebungsmethoden hat die schriftliche Befragung im Internet die mit Abstand größte Bedeutung. Sie bietet im Vergleich zu Papier/Bleistift - Methoden oder Interviews erhebliche methodische und wirtschaftliche Vorteile:

- **Geringe Kosten**
  Es fallen deutlich geringere Kosten an (z. B. keine Druck- und Portokosten, keine Kosten für Interviewer). Die für eine Online-Befragung erforderlichen Ressourcen (z. B. Software, Server) können bei diversen Dienstleistern preiswert gemietet werden.
- **Schnelle Abwicklung**
  Zur Beschleunigung tragen einige Faktoren bei:
  - Es werden schnelle Internet-Kommunikationstechniken genutzt (z. B. zum Einladen der Probanden per E-Mail).
  - Die zeitaufwändige Datenerfassung entfällt.
  - Eingeladene Probanden entscheiden sich sehr schnell für oder gegen die Teilnahme an einer Online-Befragung, während ein ausgehändigter oder per Post zugestellter Fragebogen oft zur späteren Bearbeitung deponiert wird (Jacob et al 2013, S. 110).

- **Benutzerfreundliche Filterführung**
  Ein Proband sieht nur die ihn betreffenden Fragen, während er bei einem Papierfragebogen ggf. die Sprunginstruktionen korrekt beachten muss.
- **Experimentelle Versuchspläne**
  Bei einem Experiment mit Gruppierungsfaktoren können die Teilnehmer mit geringem Aufwand zufallsgesteuert auf die verschiedenen Bedingungen (z. B. realisiert durch unterschiedliche Instruktionen) verteilt werden. Bei Messwiederholungsfaktoren ist es leicht möglich, bei jedem Probanden die Faktorstufen in zufälliger Reihenfolge zu applizieren.
- Einbindung von **Medien** (Bilder, Video, Audio)
- Optionen zur **randomisierten Präsentation** von Fragen und/oder Items
  So lassen sich Reihenfolgeeffekte neutralisieren. Allerdings ist bei Standardinstrumenten aus Gründen der Normierung bzw. Vergleichbarkeit oft eine fixe Reihenfolge der Items vorgeschrieben.
- **Plausibilitätskontrollen** zur Steigerung der Datenqualität
- **Zusatzinfos** über die Auskunftspersonen (z. B. Verweildauer auf einzelnen Seiten)

Als Nachteile von Online-Umfragen sind zu nennen:

- Schlechte Erreichbarkeit mancher Personengruppen
  Die aktuelle ARD/ZDF - Onlinestudie ergab, dass 2021 ca. 94% aller Personen ab 14 Jahren das Internet genutzt haben.[1] Allerdings kommt eine Online-Umfrage z. B. noch nicht in Betracht, wenn die Zufriedenheit von Altersheimbewohnern untersucht werden soll.
- Abhängigkeit von technischen Ressourcen in der Befragungssituation
  Jeder Teilnehmer benötigt einen Computer, und der Internetzugang muss klappen. Bei einer Untersuchung von Schülern kann es z. B. passieren, dass zu wenige Computer verfügbar sind, oder dass Probleme mit dem WLAN (drahtlosen lokalen Netzwerk) der Schule auftreten.

Angehörige der Universität Trier haben über das ZIMK Zugriff auf das professionelle Online-Umfragesystem **Enterprise Feedback Suite Survey** (EFS Survey) der Firma **Tivian XI**, die unter dem Namen **Unipark** günstige Lizenzbedingungen für Hochschulen bietet (siehe Baltes-Götz 2021).

Hier ist die erste Seite der Online-Realisation des Beispielfragebogens zum Kurs zu sehen:

[1] https://www.ard-zdf-onlinestudie.de/ardzdf-onlinestudie/infografik/

Wer an der Universität Trier eine Online-Datenerhebung mit EFS Survey durchführen möchte, findet Informationen zum Erwerb der Zugangsberechtigung auf der folgenden Webseite:

https://www.uni-trier.de/index.php?id=53873

4 Manuelle Datenerfassung und SPSS-Dateneditor

Wie in unserer Kursstudie liegen auch in vielen anderen Projekten nach Abschluss der Datenerhebung schriftliche Untersuchungsdokumente vor, die nun erfasst, d. h. in eine Computer-Datei übertragen werden müssen. Bevor in Abschnitt 4.2 die Erfassung der KFA-Daten mit dem SPSS-Dateneditor beschrieben wird, sollen in Abschnitt 4.1 einige alternative Erfassungsmethoden vorgestellt werden.

4.1 Methoden zur manuellen Datenerfassung

4.1.1 Datenerfassung nach einer papier-gestützten Datenerhebung

Bei einer papiergestützten Datenerhebung müssen die Daten in der Regel per Tastatur unter Beachtung eines Codierplans in eine Computer-Datei befördert werden. Beim Entwurf des Codierplans ist darauf zu achten, dass dem Erfasser keine unnötigen und fehleranfälligen Arbeiten zugemutet werden (siehe Abschnitt 2.4).

Von den möglichen manuellen Erfassungsmethoden werden in diesem Manuskript vorgestellt:

- **Erfassung mit dem SPSS-Dateneditor**
  Der SPSS-Dateneditor ist ein integraler Bestandteil des SPSS-Systems, sodass wir uns mit seiner Bedienung auf jeden Fall vertraut machen müssen. Er ist nicht perfekt geeignet für die Erfassung *großer* Datenmengen, kann aber in kleinen bis mittleren Projekten verwendet werden. Relativ ähnliche Arbeitsbedingungen für die Datenerfassung bieten Tabellenkalkulationsprogramme wie z. B. LibreOffice Calc oder Microsoft Excel.
- **Einsatz eines speziellen Datenerfassungsprogramms**
  Ein spezielles Datenerfassungsprogramm bietet Vorteile gegenüber dem SPSS-Dateneditor (z.B. Plausibilitätskontrollen für die erfassten Daten), erfordert aber auch zusätzlichen Einarbeitungsaufwand.

Aufgrund des relativ geringen Datenaufkommens in unserem KFA-Projekt ist der SPSS-Dateneditor das optimale Erfassungswerkzeug. Weil in Abschnitt 4.2 die Erfassung der KFA-Daten mit dem SPSS-Dateneditor ausführlich beschrieben wird, können wir uns im aktuellen Abschnitt auf Erläuterungen zu den spezialisierten Datenerfassungsprogrammen beschränken.

Wenn bei *größeren* Projekten (mit mehr als ca. 200 Fällen) eine manuelle Datenerfassung unumgänglich ist, dann sollte in der Regel ein spezielles Datenerfassungsprogramm verwendet werden:

- Sofern Arbeitsplätze mit permanenter Internet-Verbindung zur Verfügung stehen, kann ein Online-Umfragesystem auch für die manuelle Dateneingabe eingesetzt werden (vgl. Abschnitt 3.2). Diese Lösung hat den Vorteil, dass an den Erfassungsplätzen als Software nur ein Betriebssystem und ein Web-Browser benötigt werden, sodass keine Software installiert werden muss.
- Wenn eine lokal installierte (internet-unabhängige) Software bevorzugt wird, kommt ein Datenbankprogramm mit der Möglichkeit zur Formulardefinition in Frage (z. B. Microsoft Access).

Bei Verwendung eines spezialisierten Datenerfassungsprogramms arbeitet man meist mit einer *Erfassungsmaske*, die einen *einzelnen* Fall in übersichtlicher Form präsentiert. Außerdem sorgen die folgenden Techniken für eine bequeme Erfassung und eine geringe Fehlerrate:

- **Filterführung** (*Skip & Fill*)
  In Abhängigkeit vom erfassten Wert einer Filtervariablen verzweigen die Datenerfassungsspezialisten zu unterschiedlichen Folgevariablen und versorgen dabei übersprungene Variablen mit einem MD-Indikator.
- **Plausibilitätsprüfungen**
  Bei Variablen mit einem beschränkten Wertebereich lassen sich ungültige Eingaben verhindern, z. B. beim Geschlecht die nicht zur zulässigen Wertemenge {1, 2, 3} passenden Eingaben. Bei metrischen Variablen können unplausible Eingaben abgewiesen werden, z. B. beim Geburtsjahr die Werte < 1900.

Allerdings entstehen beim Einsatz eines speziellen Datenerfassungsprogramms auch Kosten:

- Es muss ein zusätzliches Programm erlernt werden.
- Für jedes Projekt sind einige Konfigurationsarbeiten erforderlich (z. B. Gestaltung der Erfassungsmaske, Definition der Regeln zur Plausibilitätskontrolle).

4.1.2 Interviews mit sofortiger Datenerfassung

Bei bestimmten Befragungsthemen oder -kontexten kommen Interviews mit den Probanden in Frage (vgl. Abschnitt 2.1.2.1 zur Wählerbefragung durch die *Forschungsgruppe Wahlen*). In der Regel ist dann eine sofortige digitale Datenerfassung anzustreben, um den Umweg über Papier zu vermeiden. Nach der Befragungssituation sind zu unterscheiden:

- Vor-Ort-Interview (*CAPI, Computer Assisted Personal Interviewing*)
- Befragung per Telefon (*CATI, Computer Assisted Telephone Interviewing*)

Neben spezialisierten Computerprogrammen, die z. B. bei Telefon-Interviews auch das Wählen übernehmen können, eignen sich zur Datenerfassung im Interview auch die in Abschnitt 4.1.1 beschriebenen Datenerfassungslösungen. Sie stellen durch eine automatische Filterführung den korrekten Ablauf der Interviews sicher und reduzieren irreguläre Daten durch Plausibilitätskontrollen.

4.2 Erfassung mit dem SPSS-Dateneditor

Für die nächsten Schritte im KFA-Projekt benötigen Sie eine SPSS-Sitzung mit einem leeren Datenfenster. Diese Situation liegt z. B. vor, nachdem Sie SPSS gestartet und den Begrüßungsdialog über den Schalter **Schließen** verlassen haben. Nötigenfalls können Sie ein leeres Datenfenster mit dem folgenden Menübefehl anfordern:

Datei > Neu > Daten

Im realen statistischen Praktikum mit SPSS steht nun die Variablendeklaration und die Datenerfassung mit dem SPSS-Dateneditor an. Wenn Sie dieses Manuskript im Selbststudium lesen, können und sollten Sie trotzdem die Arbeitsschritte zur Variablendeklaration konkret nachvollziehen und die Daten des im Manuskript enthaltenen ersten Falles eintragen (siehe Seite 59). Alle Projektarbeiten nach der Datenerfassung können durch Verwendung von SPSS-Datendateien aktiv nachvollzogen werden, die auf der Webseite zum Manuskript zur Verfügung stehen (siehe Vorwort).

4.2.1 Dateneditor, Datenblatt und Arbeitsdatei

SPSS speichert zu analysierende Daten während der Sitzung in einer temporären Datei, die als **Datenblatt** oder **DataSet** bezeichnet wird. Zur Bearbeitung dient ein Fenster des **SPSS-Dateneditors**, das im Manuskript meist als **Datenfenster** bezeichnet wird. Ein Datenblatt enthält:

- Eine **rechteckige (Fälle × Variablen) - Datenmatrix**
  Diese wird auf dem **Datenansicht**-Registerblatt des Datenfensters bearbeitet.
- Einen **Deklarationsteil**, auch **Datenlexikon** genannt
  Jede Variable besitzt mehrere verarbeitungsrelevante Attribute, z. B. einen eindeutigen Namen und ein Messniveau. Zur Verwaltung der Attribute dient das **Variablenansicht**-Registerblatt des Datenfensters.

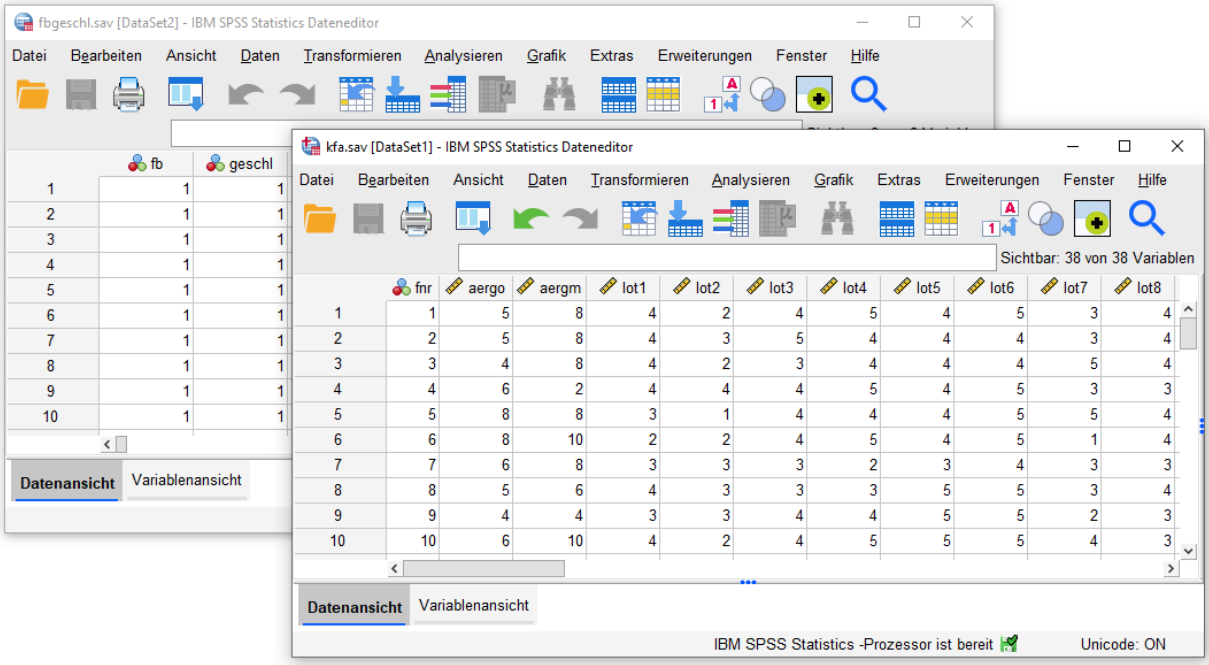
Mit Hilfe des Dateneditors oder durch Transformationskommandos (siehe unten) können u. a. die folgenden Modifikationen an einem Datenblatt vorgenommen werden:

- Definition von neuen Variablen
- Änderung der Attribute von vorhandenen Variablen
- Manuelle Erfassung von neuen Fällen
  Wir werden in unserem Demoprojekt die Daten manuell erfassen.
- Löschen von Variablen oder Fällen
- Berechnung neuer Variablen aus bereits vorhandenen Variablen oder Modifikation von Variablen über Datentransformationsbefehle
- Einlesen von Daten aus einer vorhandenen Datei mit einem unterstützten Format (z. B. SPSS, Text, Microsoft Excel, SAS, Stata) in ein neues Datenblatt.

Wenn ein Datenblatt über das Ende der Sitzung hinaus erhalten bleiben soll, muss es explizit gesichert werden (in der Regel in eine SPSS-Datendatei, siehe Abschnitt 4.2.3).

Wenn der Inhalt eines Datenfensters in eine Datendatei gesichert oder von dort gelesen worden ist, dann ist das Datenfenster mit dieser Datei verbunden, und der Dateiname erscheint in der Titelzeile des Fensters.

SPSS unterstützt die simultane Verwendung *mehrerer* Datenfenster, z. B.:

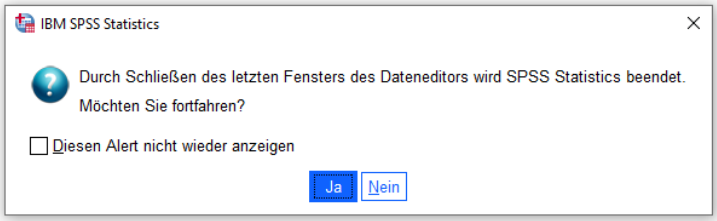


Das Datenblatt im *aktiven* Datenfenster wird als **Arbeitsdatei** bezeichnet und z. B. bei Analyseanforderungen per Menüsystem verwendet. Um ein Datenblatt zur Arbeitsdatei zu befördern, muss man das zugehörige Datenfenster per Mausklick oder **Fenster**-Menü in den Vordergrund holen. Das Datenfenster mit der Arbeitsdatei ist an einem Pluszeichen im Symbol zum Systemmenü zu erkennen (siehe linken Rand der Titelzeile).

Jedes Datenblatt hat einen Namen, welcher in der Titelzeile seines Dateneditorfensters durch eckige Klammern begrenzt hinter dem Namen der verbundenen Datendatei erscheint (siehe oben) und über den folgenden Menübefehl zu ändern ist:

Datei > Dataset umbenennen

Mit dem Schließen des *letzten* Dateneditorfensters beendet man SPSS:



4.2.2 Variablen definieren

Wie in Abschnitt 4.2.1 erwähnt, verwaltet SPSS für jede Variable eines Datenblatts zahlreiche verarbeitungsrelevante Attribute (z. B. das Messniveau und Indikatoren für fehlende Werte). Diese werden im Deklarationsteil des Datenblatts gespeichert und können via Datenfenster oder Syntax geändert werden (siehe unten). Da SPSS für alle Attribute geeignete Voreinstellungen benutzt, setzt die Datenerfassung nicht unbedingt eine Variablendefinition voraus,[1] doch wird

[1] Da in SPSS der Variablentyp *numerisch* voreingestellt ist, müssten wir vor dem Erfassen von Daten mit einem anderen Typ auf jeden Fall eine Variablendefinition vornehmen. Allerdings sind solche Variablen in unserem Codierplan nicht vorgesehen.

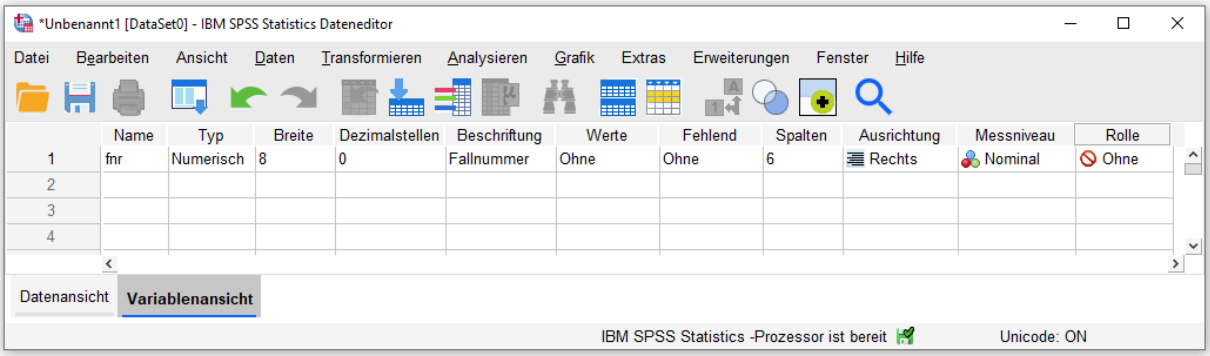
das Erfassen und die spätere Auswertungsarbeit z. B. durch benutzerdefinierte Variablennamen anstelle der automatisch generierten und wenig aussagekräftigen Namen VAR00001, VAR00002 usw. erleichtert. Daher liegt es nahe, dem SPSS-System die in unserem Codierplan beschriebenen Variablen vor dem Eintragen der Daten bekannt zu machen.

4.2.2.1 Das Datenfenster-Registerblatt Variablenansicht

Ein Datenfenster besitzt *zwei* Registerblätter zur Anzeige bzw. Bearbeitung eines Datenblatts:

- das Registerblatt **Datenansicht** zur Anzeige und Modifikation der (Fälle × Variablen)-Datenmatrix
- das Registerblatt **Variablenansicht** zur Anzeige und Modifikation der Variablenattribute

In einer Zeile der **Variablenansicht** wird eine Variable beschrieben, wozu in den Spalten insgesamt elf Attribute zur Verfügung stehen. Für unsere erste Variable (FNR) eignen sich z. B. die folgenden Angaben:



Um eine neue Variable anzulegen, trägt man einen zulässigen Namen in die nächste freie Zeile der Tabelle ein. Nach dem Verlassen der Namenszelle (z. B. per Mausklick auf eine andere Zelle oder per Tabulatortaste) werden die restlichen Attribute der neuen Variablen automatisch mit Voreinstellungen versorgt, die man nach Bedarf ersetzt. Darüber hinaus kann man Variablen einfügen, löschen oder verschieben (siehe unten).

4.2.2.2 Die SPSS-Variablenattribute

Bevor wir die Variablen unserer KFA-Studie deklarieren, sollen die SPSS-Variablenattribute erläutert werden:

- **Name**
  Die wesentlichen Regeln für SPSS-Variablennamen wurden schon im Zusammenhang mit dem Codierplan genannt (siehe Seite 55).
- **Typ**
  Die wichtigsten SPSS-Variablentypen sind schon benannt: Numerisch, Zeichenfolge und Datum (siehe Seite 49). In der Regel empfiehlt sich auch bei nominalskalierten Merkmalen eine numerische Codierung (siehe Abschnitte 2.4.3.1 und 4.2.2.6), sodass der voreingestellte numerische Variablentyp meist beibehalten werden kann.

- **Breite**
  Bei einer regulären (nicht eingeschränkten) numerischen Variablen beeinflusst dieses Attribut lediglich ihre voreingestellte Breite (inkl. Vorzeichen und Dezimaltrennzeichen) bei der Ausgabe in eine Textdatendatei (z. B. über das Kommando WRITE). Für die Arbeit mit dem Daten- oder mit dem Ausgabefenster ist die Breite der Variablen daher wenig relevant. Allerdings muss die Variablenbreite stets größer sein als die Anzahl der Dezimalstellen (siehe unten).

  Bei einer *eingeschränkten* numerischen Variablen, die nichtnegative ganze Zahlen als Werte annehmen kann, legt die **Breite** die maximale Anzahl der Ziffern fest. Schöpft ein Wert die maximale Breite nicht aus, wird er links mit Nullen aufgefüllt. Ist ein neu aufzunehmender Wert zu lang, werden am linken Rand Ziffern gestrichen.

  Bei einer Zeichenfolgenvariablen legt die **Breite** die maximale Anzahl der Zeichen fest. Ist ein Wert zu lang, werden am rechten Rand Zeichen gelöscht. Das passiert auch bei einer nachträglichen Reduktion der **Breite**.

- **Dezimalstellen**
  Bei einer normalen (nicht eingeschränkten) numerischen Variablen können Sie festlegen, welche Anzahl von Dezimalstellen bei der *Anzeige* ihrer Werte im Daten- und im Ausgabefenster verwendet werden soll. Diese Angabe betrifft *nicht* die Speichergenauigkeit, sondern nur die Anzeige. Bei einer eingeschränkt-numerischen Variablen oder einer Zeichenfolgenvariablen ist das Attribut irrelevant und auf den Wert 0 fixiert.

- **Beschriftung**
  Zu jeder Variablen kann optional eine Beschriftung (synonym: ein *Variablenlabel*) vereinbart werden. Ist eine Beschriftung vorhanden, dann erscheint diese in Ergebnistabellen und Diagrammen an Stelle des aus praktischen Erwägungen möglichst kurz gewählten und mit Syntaxrestriktionen (z. B. Leerzeichenverbot) belasteten Variablennamens, z. B.:

Variablenname	Variablenbeschriftung
FB	Fachbereich an der Universität Trier
GROESSE	Körpergröße (in cm)

  Während wir die Variablen*namen* in SPSS der Einfachheit halber meist klein schreiben, ist bei den Variablenbeschriftungen eine publikationsreife Groß/Kleinschreibung angemessen.

- **Werte**
  Hier kann man Wertbeschriftungen (synonym: *Wertelabels*) mit maximal 120 Zeichen definieren, die in Ergebnistabellen und Diagrammen per Voreinstellung an Stelle der Werte einer Variablen erscheinen. SPSS empfiehlt Wertbeschriftungen bei nominalen und ordinalen Variablen. Bei numerisch codierten nominalskalierten Merkmalen sind sie fast unverzichtbar, z. B.:

Variablenname	Werte	Wertbeschriftungen
GESCHL	1	Frau
	2	Mann
	3	Divers

  Man erhält nicht nur informative Beschriftungen (z. B. in einem Balkendiagramm), sondern beeinflusst unter bestimmten Umständen auch die Berücksichtigung von Kategorien: Soll eine unbesetzte Kategorie (mit der Häufigkeit null) in einem Diagramm er-

scheinen (z. B. als Balken mit der Höhe null), dann muss der zugehörige (in den Daten nicht vorhandene) Wert eine Beschriftung erhalten.

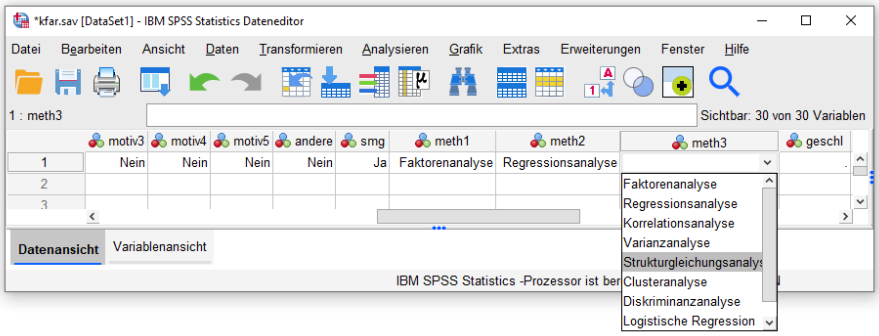
Bei metrischen Variablen sind Wertbeschriftungen in der Regel nur für benutzerdefinierte MD - Indikatoren relevant.

In der **Datenansicht** bietet der Dateneditor über den Menübefehl

Ansicht > Wertbeschriftungen

oder den Symbolschalter  einige Unterstützungen für die Wertbeschriftungen:

- Sie werden an Stelle der Werte angezeigt.
- Alternativ zur Werteingabe per Tastatur kann man per Drop-Down - Menü eine Wertbeschriftung wählen, z. B.:

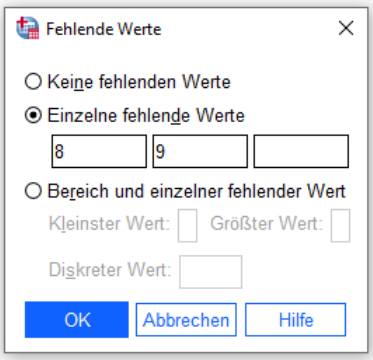


Es ist aber *nicht* möglich, durch die Vergabe von Wertbeschriftungen die Menge der gültigen Werte einer Variablen zu definieren und eine Plausibilitätskontrolle für die Erfassung per Dateneditor einzurichten. Trotz obiger Wertbeschriftungsvereinbarung wird der SPSS-Dateneditor z. B. bei der Variablen GESCHL beliebige Zahlen akzeptieren.

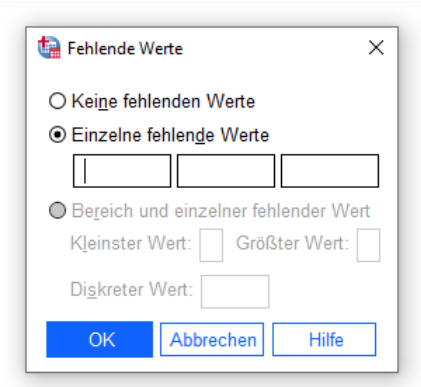
- **Fehlend**

Wenn Sie bei einer Variablen *benutzerdefinierte* MD-Indikatoren verwenden (siehe Abschnitt 2.4.3.2.1), müssen Sie diese unbedingt deklarieren, weil sie sonst wie gültige Werte behandelt werden (z. B. bei einer Mittelwertsbildung). Gehen Sie dabei folgendermaßen vor:

- Markieren Sie bei der betroffenen Variablen die Zelle zum Attribut **Fehlend**.
- Nach einem Mausklick auf den nun vorhandenen Erweiterungsschalter  erscheint eine Dialogbox, in der man entweder bis zu drei Einzelwerte oder aber ein Intervall samt zusätzlichem Einzelwert als MD-Indikatoren vereinbaren kann, z. B.:

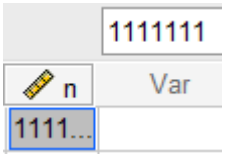


- Soll bei einer *Zeichenfolgenvariablen* die leere Zeichenfolge als MD-Indikator deklariert werden, ist ein einzelnes Leerzeichen einzutragen, z. B.:



- **Spalten** und **Ausrichtung**
  Wie breit soll die Spalte einer Variablen im Datenfenster sein? Wie sollen die Werte ausgerichtet werden (linksbündig, zentriert, rechtsbündig)? Die Attribute **Spalten** und **Ausrichtung** wirken sich nur auf die Darstellung einer Variablen im Datenfenster aus.

  Genügt die **Spalten**-Zahl nicht, signalisiert SPSS die Platznot durch Pünktchen, z. B.:



  Bei der **Ausrichtung** hält sich SPSS an die Konvention, Texte linksbündig und Zahlen rechtsbündig zu schreiben, sodass es kaum einen Grund gibt, dieses Attribut zu ändern.

- **Messniveau**
  Bei diesem methodischen Attribut sind die Werte **Metrisch** (Intervall- bzw. Rationalskalenniveau), **Ordinal** und **Nominal** erlaubt. Bislang spielt das deklarierte Messniveau der Variablen bei den meisten Statistikprozeduren in SPSS noch keine Rolle. Bei der Diagrammerstellung (siehe Abschnitt 11.1.1) hängt die Behandlung einer Variablen jedoch vom deklarierten Messniveau ab, sodass zumindest für alle an einem Diagramm beteiligten Variablen das Messniveau korrekt angegeben werden muss.

- **Rolle**
  Man kann den Variablen eine Rolle zuweisen, die in einigen Dialogfeldern zur Variablenvorauswahl genutzt wird:
  - **Eingabe**
    Die Variable kommt als unabhängige Variable (Regressor) in Frage.
  - **Ziel**
    Die Variable kommt als abhängige Variable (Kriterium) in Frage.
  - **Beides**
    Die Variable kommt als Regressor (unabhängige Variable) und als Kriterium (abhängige Variable) in Frage.
  - **Ohne**
    Die Variable erhält keine Rolle.
  - **Partitionieren**
    Die Variable dient zur Zerlegung der Stichprobe, z. B. in Modellierungs- und Kreuzvalidierungsfälle.
  - **Aufteilen**
    Diese Rolle ist im Hinblick auf das Produkt SPSS Modeler vorhanden und in SPSS Statistics irrelevant.

4.2.2.3 Variablendefinition durchführen

Aktivieren Sie nötigenfalls die **Variablenansicht** des Datenfensters, und tragen Sie für die erste Variable (zur Fallidentifikation) den im Codierplan vorgesehenen Namen FNR ein. Nach dem Markieren der zugehörigen Zelle können Sie sofort mit dem Eintippen des Namens beginnen. Die Groß/Kleinschreibung ist bzgl. der Identifikation von Variablen irrelevant. Die folgenden Namen bezeichnen alle dieselbe Variable:

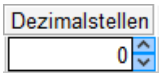
fnr Fnr FNR

Im Manuskript erscheinen die Variablennamen aus darstellungstechnischen Gründen in Großbuchstaben.

Sobald Sie die Zelle mit dem Variablennamen verlassen (z. B. per Mausklick auf eine andere Zelle oder per Tabulatortaste) wird eine neue Variable mit dem gewünschten Namen in die Arbeitsdatei aufgenommen, sofern gegen den Variablennamen keine Einwände bestehen. Die restlichen Attribute der neuen Variablen werden mit den jeweiligen Standardwerten initialisiert.

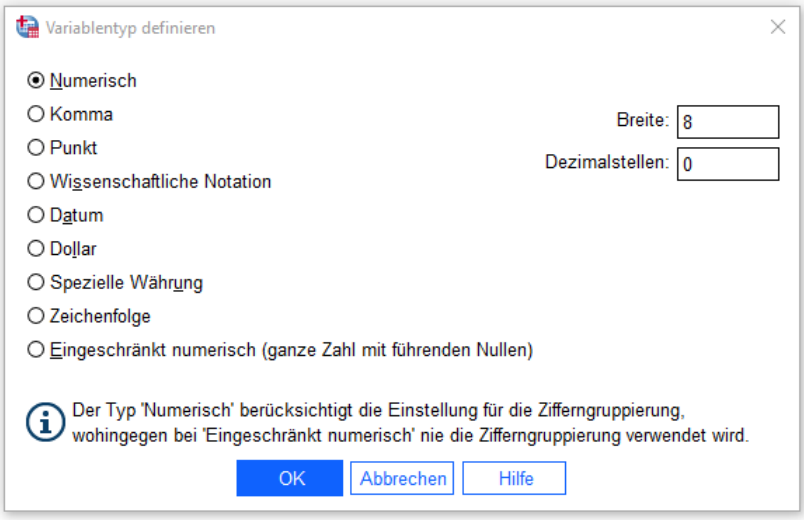
Als **Typ** übernehmen wir bei FNR die Voreinstellung **Numerisch**. Im Demonstrationsprojekt werden wir das auch bei allen anderen Variablen tun.

Nach dem Markieren der Zelle **Dezimalstellen** kann man die gewünschte Anzahl von Dezimalstellen durch Eingabe einer Zahl oder per Up-Down - Regler wählen. Bei FNR eignet sich der Wert 0:

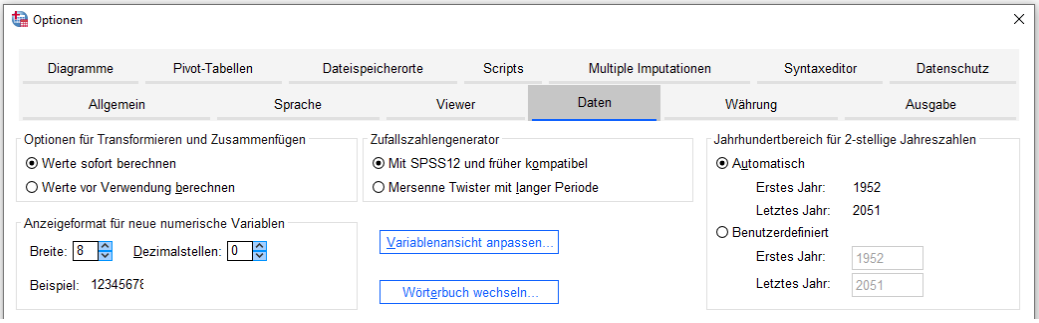


Analog wird auch das Attribut **Breite** festgelegt, das allerdings bei der von uns geplanten Arbeitsweise keine nennenswerte Rolle spielt (siehe Abschnitt 4.2.2.2). Wir belassen der Einfachheit halber im Demonstrationsprojekt generell den Voreinstellungswert 8.

Eine alternative Möglichkeit zur Veränderung der Attribute **Typ**, **Breite** und **Dezimalstellen** bietet die Dialogbox **Variablentyp definieren**, die nach einen Mausklick auf den Erweiterungsschalter ... in der markierten **Typ**-Zelle erscheint:



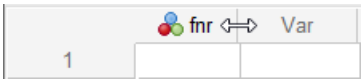
Tipp: Wenn in einem Projekt das voreingestellte Anzeigeformat für numerische Variablen (Breite = 8, Dezimalstellen = 2) häufig durch eine bestimmte Alternative ersetzt werden muss, kann zur Vereinfachung der Deklaration die Voreinstellung entsprechend geändert werden. Dazu öffnet man mit **Bearbeiten > Optionen** die Dialogbox **Optionen**, wechselt hier zum Registerblatt **Daten** und nimmt im Rahmen **Anzeigeformat für neue numerische Variablen** die gewünschten Einstellungen vor, z. B.:



Wenngleich die Variable FNR im Ausgabefenster nicht allzu oft auftauchen wird, tragen wir in die Zelle zum Attribut **Beschriftung** den Text *Fallnummer* ein.

Statt die Breite der FNR-Spalte im Datenfenster über eine gut geschätzte **Spalten**-Angabe festzulegen, können Sie bei aktiviertem Registerblatt **Datenansicht** auch folgendermaßen vorgehen:

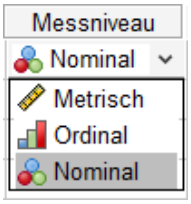
- Setzen Sie den Mauszeiger auf den rechten Rand der Zelle mit dem Variablennamen, woraufhin der Zeiger eine neue Form und dementsprechend eine neue Funktion erhält:



- Nun lässt sich der rechten Rand der aktuellen Spalte verschieben: Linke Maustaste drücken, ziehen und an der gewünschten Position wieder loslassen.

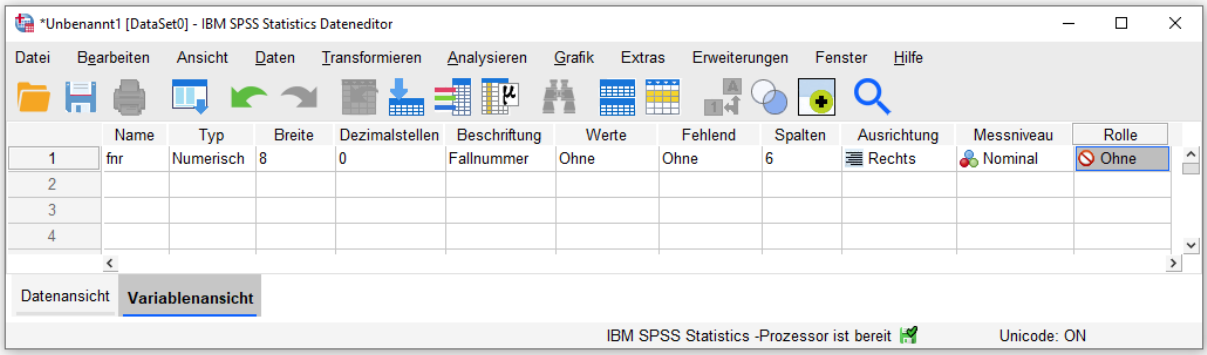
Eine so festgelegte Spaltenbreite wird von SPSS als **Spalten**-Wert übernommen.

Öffnen Sie in der markierten **Messniveau**-Zelle das Drop-Down-Menü, um für die Fallnummer ein nominales Skalenniveau zu deklarieren:



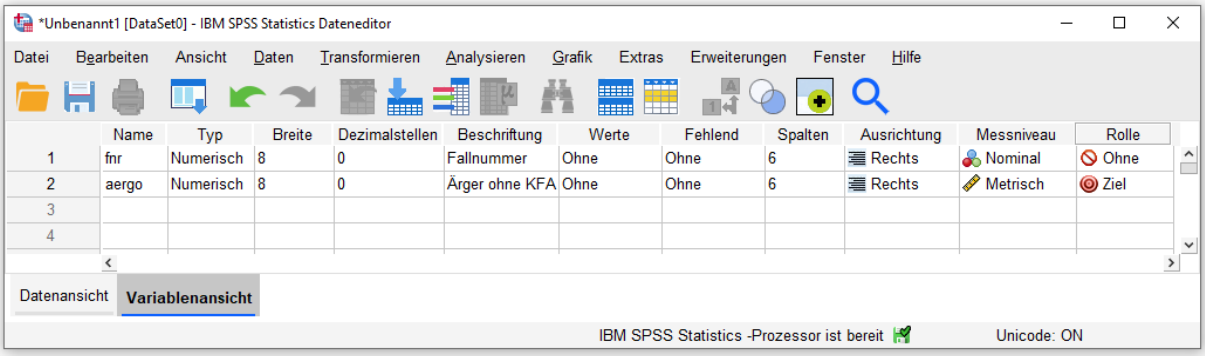
Öffnen Sie in der markierten Zelle zur Entscheidung über eine **Rolle** das Drop-Down-Menü, um der Variablen FNR eine Rolle zu verweigern:

MD-Indikatoren und Wertelabels sind bei der momentan bearbeiteten Fallnummernvariablen irrelevant, und das Attribut **Ausrichtung** übernehmen wir stets unverändert. Daher können wir die Deklaration der Variablen FNR beenden:



Bei Bedarf sind Anpassungen jederzeit möglich.

Vereinbaren Sie nun in der zweiten Zeile der **Variablenansicht** für den Ärger ohne KFA-Wirkung den Variablennamen AERGO, eine Anzeige mit 0 Dezimalstellen und die Variablenbeschriftung *Ärger ohne KFA*. Wir gehen von Intervallskalenniveau aus und vergeben die zugehörige **Messniveau**-Attributausprägung **Metrisch**. Wie ein Kurzbesuch auf dem Registerblatt **Datenansicht** zeigt, genügt die **Spalten**-Breite 6. Als mutmaßlich dominante **Rolle** der Variablen AERGO legen wir **Ziel** fest, sodass sich im Dateneditor das folgende Bild zeigt:



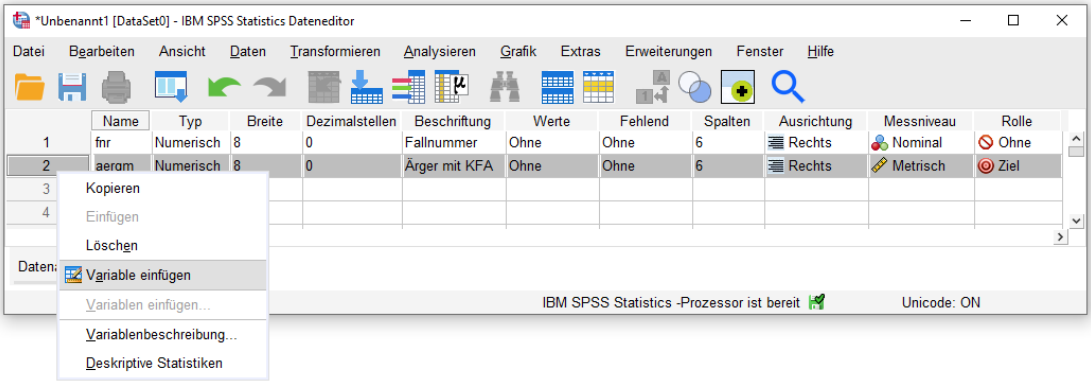
4.2.2.4 Variablen einfügen, löschen oder verschieben

Bei der Variablendefinition kann sich die Notwendigkeit ergeben, Variablen einzufügen oder zu löschen.

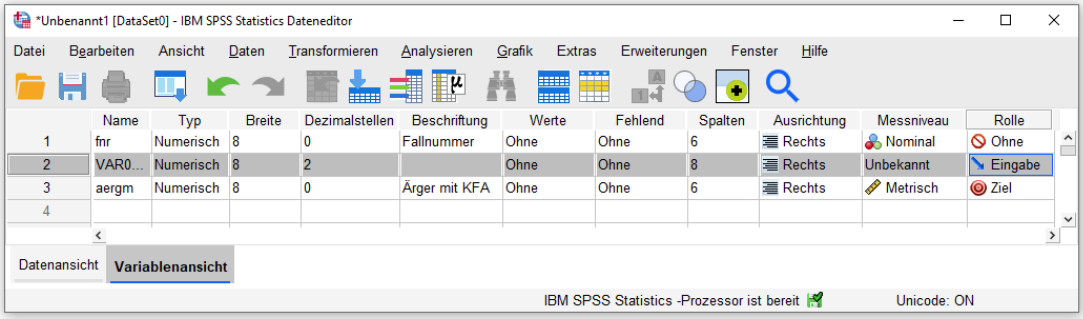
4.2.2.4.1 Variablen einfügen

Wenn wir z. B. nach FNR die Variable AERGM definiert, also die Variable AERGO vergessen hätten, könnten wir das Missgeschick in der Variablenansicht folgendermaßen korrigieren:

- Rechter Mausklick auf die Nummer der AERGM-Zeile (am linken Rand der Tabelle).
- Kontextmenü-Option **Variable einfügen** wählen



Daraufhin stellt SPSS vor AERGM eine neue Variable mit voreingestellten Attributausprägungen zur Verfügung, die nun beliebig angepasst werden können:



Auf analoge Weise lässt sich eine neue Variable auch in der Datenansicht einfügen:

- Setzen Sie einen rechten Mausklick auf die Beschriftung der AERGM-Spalte im Kopfbereich der Tabelle.
- Wählen Sie die Option **Variable einfügen** aus dem Kontextmenü.

4.2.2.4.2 Variablen löschen

Gehen Sie in der Variablenansicht folgendermaßen vor, um eine Variable zu löschen:

- Setzen Sie einen rechten Mausklick auf die Zeilennummer der betroffenen Variablen in der Nummerierungsspalte am linken Fensterrand.
- Wählen Sie aus dem Kontextmenü die Option **Löschen**.

Auf analoge Weise lässt sich eine Variable auch in der Datenansicht löschen.

4.2.2.4.3 Variablen verschieben

Das Verschieben einer Variablen kann z. B. zur Erleichterung der Orientierung in einer Datendatei bzw. in einem Datenblatt sinnvoll sein. Allerdings ist zu beachten, dass vorhandene SPSS-Syntax durch das Verschieben von Variablen ungültig werden kann, wenn dort Variablenbereiche über das Schlüsselwort TO angesprochen werden (siehe z. B. Abschnitt 19.1.5.1).

a) Verschieben per Drag & Drop

Gehen Sie in der Variablenansicht folgendermaßen vor, um Variablen per Drag & Drop (Ziehen und Ablegen) über eine kurze Distanz (z. B. innerhalb des sichtbaren Datenblattsegments) zu verschieben:

- Markieren Sie die zu verschiebenden Variablen auf windows-übliche Weise über Mausaktionen in der Nummerierungsspalte am linken Fensterrand, ggf. ergänzt durch die Strg - und/oder ⇧ - Taste. Lassen Sie anschließend die Maustaste wieder los.
- Klicken Sie in der Nummerierungsspalte auf die zu verschiebende Variablenauswahl, und halten Sie dabei die linke Maustaste gedrückt.
- Bewegen Sie bei gedrückter Maustaste den Mauszeiger zum Ziel der Verschiebungsaktion. Der aktuell anvisierte Zielort wird von SPSS durch eine rote Linie gekennzeichnet.
- Wenn Sie die Maustaste loslassen, erscheinen die Variablen am neuen Ort.

Auf analoge Weise lässt sich eine Variablenauswahl auch in der Datenansicht verschieben.

Bei einem ungeteilten Datenfenster wird ein Verschieben über eine größere Distanz (z. B. von der Position 800 auf die Position 3) zur Geduldsprobe. Muss zum Ansteuern der neuen Position der Datenfensterinhalt gerollt werden, dann verläuft die Bewegung sehr langsam. In dieser Lage kann die per Drag & Drop zu überwindende Distanz in der Datenansicht durch zwei nebeneinander stehende Ansichten auf das Datenblatt reduziert werden (siehe Abschnitt 4.2.7).

b) Verschieben per Cut & Paste

Per Drag & Drop lassen sich Variablen flott über größere Distanzen verschieben, doch können dabei Bedienungsfehler zu Datenverlusten führen:

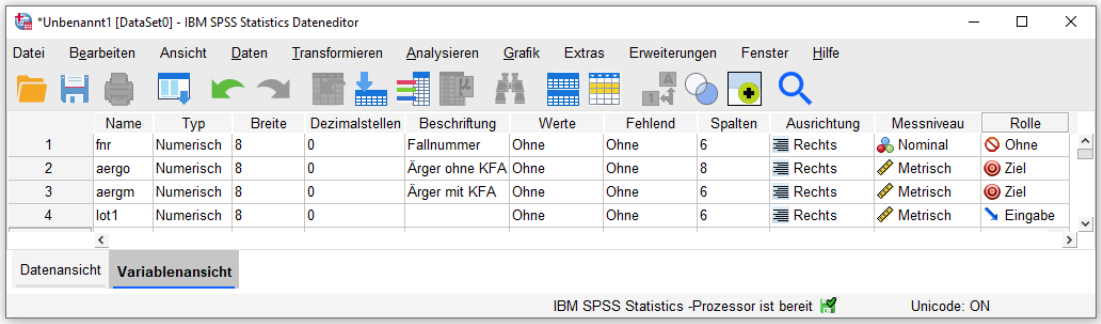
- Wechseln Sie zur **Datenansicht**.
- Markieren Sie die zu verschiebenden Variablen auf windows-übliche Weise per Mausklick auf die Titelzeile, ggf. ergänzt durch die Strg - und/oder ⇧ - Taste.
- Wählen Sie aus dem Kontextmenü der markierten Variablen das Item **Ausschneiden**, um die Variablen in die Windows-Zwischenablage zu befördern und am alten Ort zu löschen.
- Fügen Sie an der Zielposition so viele neue, leere Variablen ein, dass alle Umzügler Platz finden. Das geht am einfachsten über das Item **Variable einfügen** aus dem Kontextmenü zu derjenigen Variablen, die zum neuen rechten Nachbarn der Umzügler werden soll.
- Markieren Sie die neuen, noch leeren Spalten, und wählen Sie aus deren Kontextmenü das Item **Einfügen**.

4.2.2.5 Attributausprägungen auf andere Variablen übertragen

4.2.2.5.1 Variablendeklaration vervielfältigen

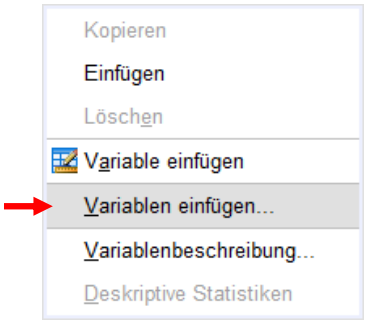
Für unsere zwölf LOT-Fragen sollen natürlich alle Variablenattribute mit Ausnahme des Namens identisch sein. Erfreulicherweise müssen wir nicht zwölf fast identische Variablendefinitionen erstellen, sondern wir können nach einer ersten Definition die Serienproduktion an SPSS delegieren. Beim anschließend beschriebenen Verfahren erhalten die zusätzlichen Variablen geeignete Namen und übernehmen die restlichen Attribute von der Mustervariablen:

- Deklarieren Sie die Variable LOT1 mit geeigneten Attributen, z. B.:



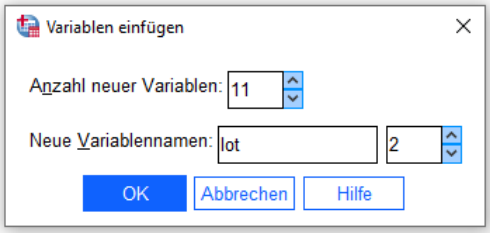
  Das voreingestellte metrische Messniveau kann beibehalten werden, obwohl die fünfstufigen Variablen LOT1 bis LOT12 (jeweils einzeln betrachtet) ziemlich grobe Indikatoren für das angenommene latente Merkmal Optimismus sind (sogenannte *Likert-Items*, siehe Abschnitt 2.2.3.5). In den geplanten Auswertungen werden wir nicht die zwölf Rohvariablen, sondern eine daraus abgeleitete Mittelwertsvariable verwenden, für die (als sogenannte *Likert-Skala*) ein approximativ metrisches Messniveau angenommen werden darf.

- Öffnen Sie das Kontextmenü zur Variablen LOT1 per Mausrechtsklick auf ihre Zeilennummer in der Nummerierungsspalte am linken Fensterrand, und wählen Sie das Item **Kopieren**, um alle Attribute der Variablen in die Windows-Zwischenablage zu befördern.
- Setzen Sie einen rechten Mausklick auf die Nummer der nächsten freien Zeile der Variablenansicht, und wählen Sie aus dem Kontextmenü die Option **Variablen einfügen** *mit den drei Punkten* am Ende der Beschriftung:



  Diese Option ist nur dann verfügbar, wenn sich eine komplette Variablenbeschreibung in der Windows-Zwischenablage befindet.

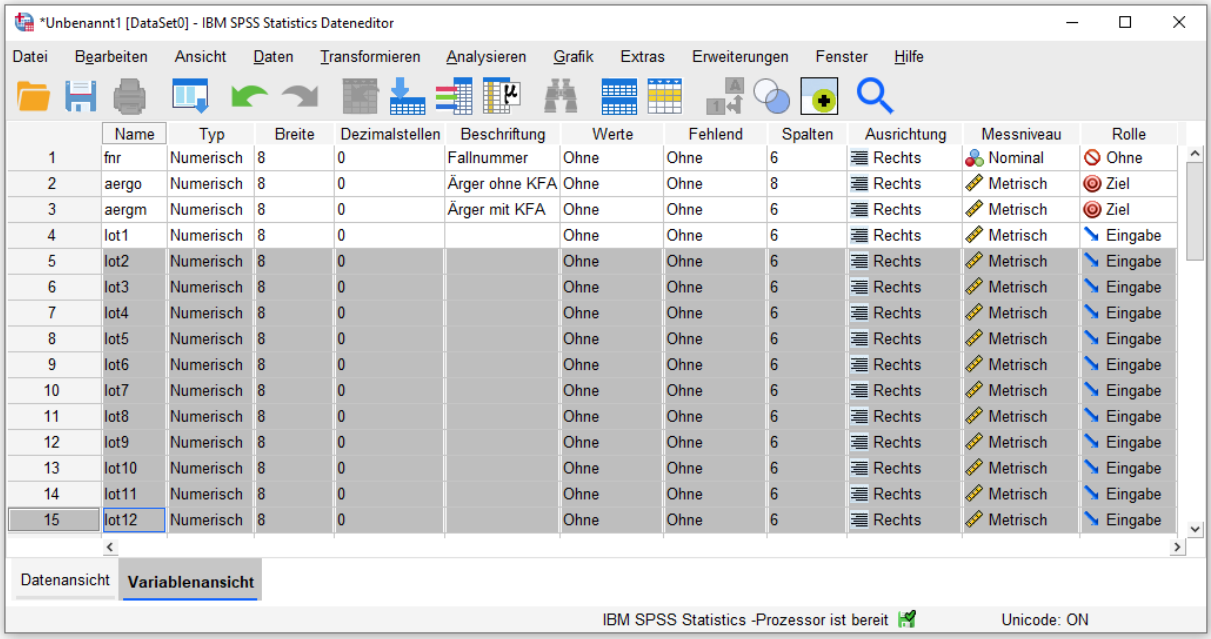
- In der folgenden Dialogbox



  können Sie nun festlegen, …

  - wie viele neue Variablen benötigt werden,
  - welche gemeinsame Wurzel die neuen Variablennamen haben sollen,
  - durch welchen Indexwert der erste Variablenname komplettiert werden soll.

Nach dem Quittieren der obigen Dialogbox entstehen elf neue Variablen mit den gewünschten Namen und Attributen:



	Name	Typ	Breite	Dezimalstellen	Beschriftung	Werte	Fehlend	Spalten	Ausrichtung	Messniveau	Rolle
1	fnr	Numerisch	8	0	Fallnummer	Ohne	Ohne	6	Rechts	Nominal	Ohne
2	aergo	Numerisch	8	0	Ärger ohne KFA	Ohne	Ohne	8	Rechts	Metrisch	Ziel
3	aergm	Numerisch	8	0	Ärger mit KFA	Ohne	Ohne	6	Rechts	Metrisch	Ziel
4	lot1	Numerisch	8	0		Ohne	Ohne	6	Rechts	Metrisch	Eingabe
5	lot2	Numerisch	8	0		Ohne	Ohne	6	Rechts	Metrisch	Eingabe
6	lot3	Numerisch	8	0		Ohne	Ohne	6	Rechts	Metrisch	Eingabe
7	lot4	Numerisch	8	0		Ohne	Ohne	6	Rechts	Metrisch	Eingabe
8	lot5	Numerisch	8	0		Ohne	Ohne	6	Rechts	Metrisch	Eingabe
9	lot6	Numerisch	8	0		Ohne	Ohne	6	Rechts	Metrisch	Eingabe
10	lot7	Numerisch	8	0		Ohne	Ohne	6	Rechts	Metrisch	Eingabe
11	lot8	Numerisch	8	0		Ohne	Ohne	6	Rechts	Metrisch	Eingabe
12	lot9	Numerisch	8	0		Ohne	Ohne	6	Rechts	Metrisch	Eingabe
13	lot10	Numerisch	8	0		Ohne	Ohne	6	Rechts	Metrisch	Eingabe
14	lot11	Numerisch	8	0		Ohne	Ohne	6	Rechts	Metrisch	Eingabe
15	lot12	Numerisch	8	0		Ohne	Ohne	6	Rechts	Metrisch	Eingabe

4.2.2.5.2 Alle Attribute einer Variablen auf andere Variablen übertragen

Gehen Sie z. B. folgendermaßen vor, um in der **Variablenansicht** *alle* Attribute einer Variablen (mit Ausnahme des Namens) auf andere, bereits vorhandene Variablen zu übertragen:

- Öffnen Sie das Kontextmenü zur Quellvariablen per Mausrechtsklick auf ihre Zeilennummer in der Nummerierungsspalte am linken Fensterrand, und wählen Sie das Item **Kopieren**, um alle Attribute der Variablen in die Zwischenablage zu befördern.
- Markieren Sie *eine* Zielvariable per Mausklick auf ihre Zeilennummer oder *mehrere* Zielvariablen durch Mausklicks in Kombination mit der ⇧ - und/oder Strg - Taste.
- Übertragen Sie die in der Zwischenablage gespeicherten Attribute über das Kontextmenü-Item **Einfügen**, mit der Tastenkombination Strg **+ V** oder mit dem Menübefehl

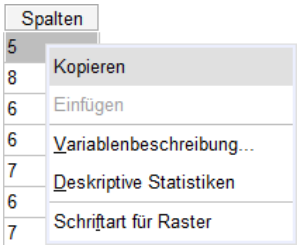
  Bearbeiten > Einfügen

  auf alle markierten Variablen.

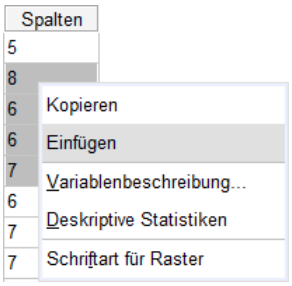
4.2.2.5.3 Einzelne Attribute einer Variablen auf andere Variablen übertragen

Es ist auch möglich, ein *einzelnes* Attribut von einer Variablen auf andere zu übertragen:

- Kontextmenü zur Quell-Attributzelle durch einen rechten Mausklick öffnen und das Item **Kopieren** wählen, z. B.:



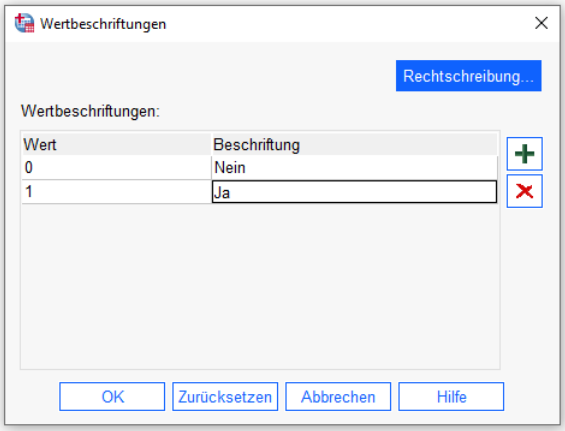
- Zu verändernde Attributzellen markieren, Kontextmenü zur Markierung öffnen und den Attributwert über das Kontextmenü-Item **Einfügen** aus der Zwischenablage übernehmen:



4.2.2.6 Numerische Codierung auch bei nominalskalierten Merkmalen

Vereinbaren Sie für die Variable zur ersten Motivoption (eigene Studie) den Variablennamen MOTIV1, eine numerische Codierung und eine Anzeige ohne Dezimalstellen. Dabei wird die Empfehlung aus Abschnitt 4.2.2.2 befolgt, auch bei den Variablen zur Erfassung von nominalskalierten Merkmalen eine *numerische* Codierung zu verwenden.

Nach der numerischen Codierung eines nominalskalierten Merkmals ist es sehr empfehlenswert, die letztlich willkürliche Zuweisung von Zahlen zu den Kategorien durch Wertbeschriftungen zu dokumentieren. Öffnen Sie daher für die Variable MOTIV1 mit einem Mausklick auf den Erweiterungsschalter ... in der markierten **Werte**-Zelle die folgende Dialogbox:



Legen Sie die Wertbeschriftungen folgendermaßen fest:

- Klicken Sie auf das Symbol +.
- Tragen Sie den **Wert** 0 ein, und quittieren Sie mit der **Enter**-Taste.
- Tragen Sie die **Beschriftung** *Nein* ein.
- Klicken Sie erneut auf das Symbol +.
- Tragen Sie den **Wert** 1 ein, und quittieren Sie mit der **Enter**-Taste.
- Markieren Sie die Beschriftungszelle zum Wert 1, und tragen Sie hier *Nein* ein.

4.2.2.7 Übung

Definieren Sie die laut Codierplan aus dem Teil 3 unseres Fragebogens resultierenden Variablen. Vermutlich werden Sie die 6 Motivvariablen ebenso rationell deklarieren, wie wir es in Abschnitt 4.2.2.5.1 mit den 12 LOT-Variablen getan haben. Die Mustervariable (im Beispiel: MO-

TIV1) sollte in der Regel vor der Vervielfachung noch keine Variablenbeschriftung erhalten. Diese Beschriftung würde auf die generierten Variablen übertragen und müsste dort wieder geändert werden. Nach der Vervielfachung lohnt sich aber speziell bei den Motivvariablen die Definition von Beschriftungen.

Definieren Sie für das nominalskalierte Merkmal Geschlecht:

- den Variablennamen GESCHL
- eine numerische Codierung
- eine Anzeige ohne Dezimalstellen
- die Variablenbeschriftung *Geschlecht*
- Wertbeschriftungen (1 = *Frau*, 2 = *Mann*, 3 = *Divers*)
- eine geeignete **Spalten**-Breite
- das **nominale** Messniveau
- die Rolle **Eingabe**

Definieren Sie die restlichen Variablen aus dem Fragebogenteil 4.

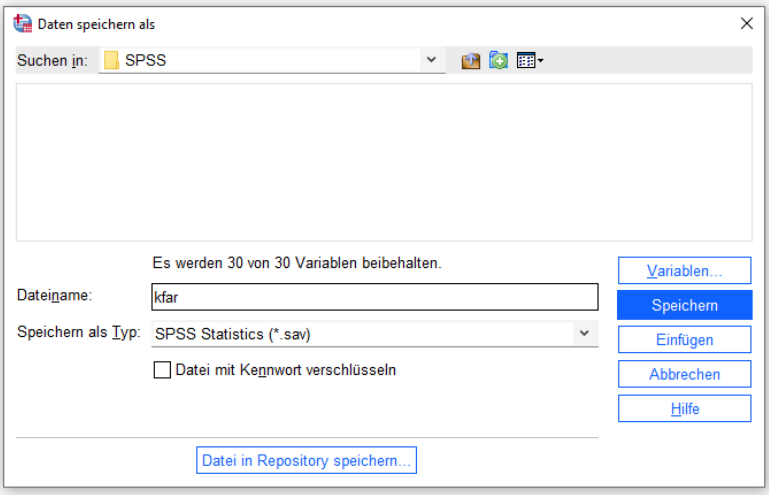
4.2.3 Sichern eines Datenblatts als SPSS-Datendatei

Wenn ein neu erstelltes Datenblatt über das Ende der Sitzung hinaus erhalten bleiben soll, muss es explizit auf einen permanenten Datenträger gesichert werden, wobei das **SPSS-Datendateiformat** (mit der Namenserweiterung **.sav**) voreingestellt und meist sinnvoll ist. In späteren Sitzungen kann durch Öffnen dieser Datendatei der gesicherte Zustand des Datenblatts wiederhergestellt werden.

Zwar enthält Ihre aktuelle Arbeitsdatei (das aktive und vermutlich einzige Datenblatt) noch keine Daten, aber im Datenlexikon (Deklarationsteil) befinden sich bereits wertvolle Informationen. Daher sollten Sie schon jetzt die temporäre Arbeitsdatei in eine permanente SPSS-Datendatei sichern, indem Sie den folgenden Menübefehl wählen:

Datei > Speichern unter...

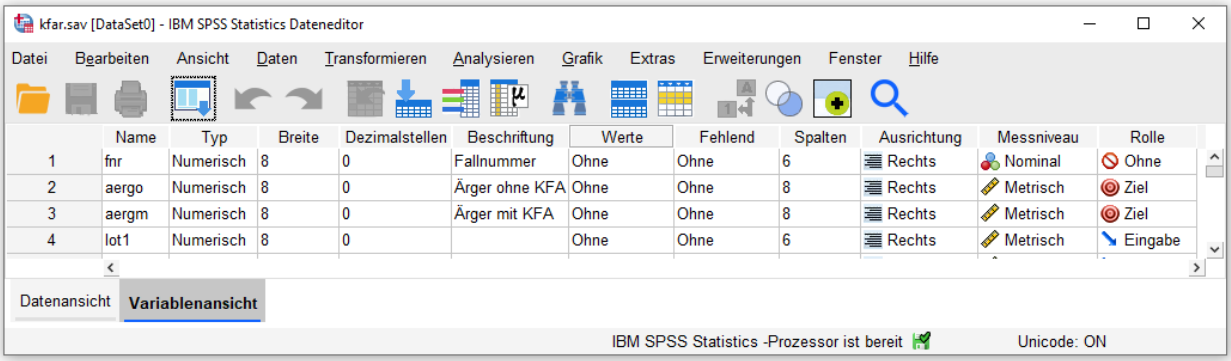
In der erscheinenden Dialogbox ist für die zu erzeugende SPSS-Datendatei ein Verzeichnis, ein **Dateiname** und ein **Typ** anzugeben:



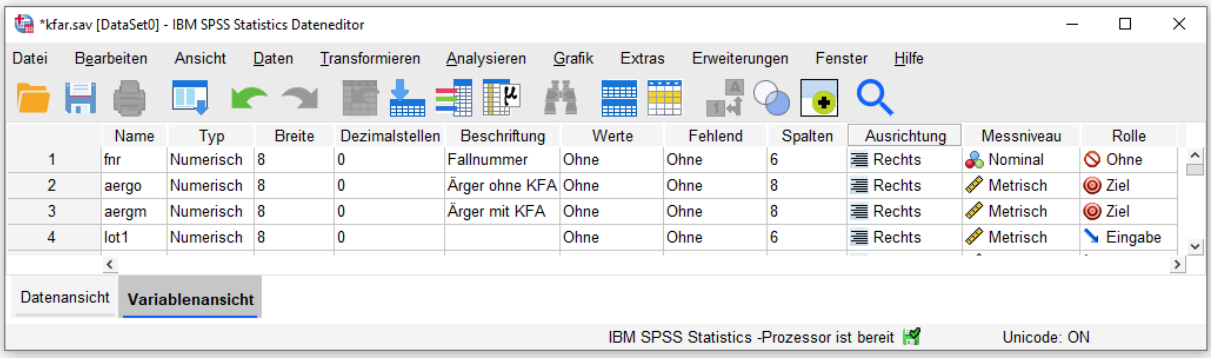
Wenn Sie keine Namenserweiterung angeben, wird die Voreinstellung **.sav** verwendet, und das spätere Öffnen der Datendatei geht besonders bequem.

Als Name für unsere Beispieldatei wird **kfar.sav** vorgeschlagen, verbunden mit der Versicherung, die Begründung für das **r** im nächsten Abschnitt nachzuliefern.

Nach dem **Speichern** zeigt die Titelzeile des Datenfensters neben dem Datenblattnamen den Namen der nunmehr zugeordneten Datendatei, in unserem Fall also **kfar.sav**:

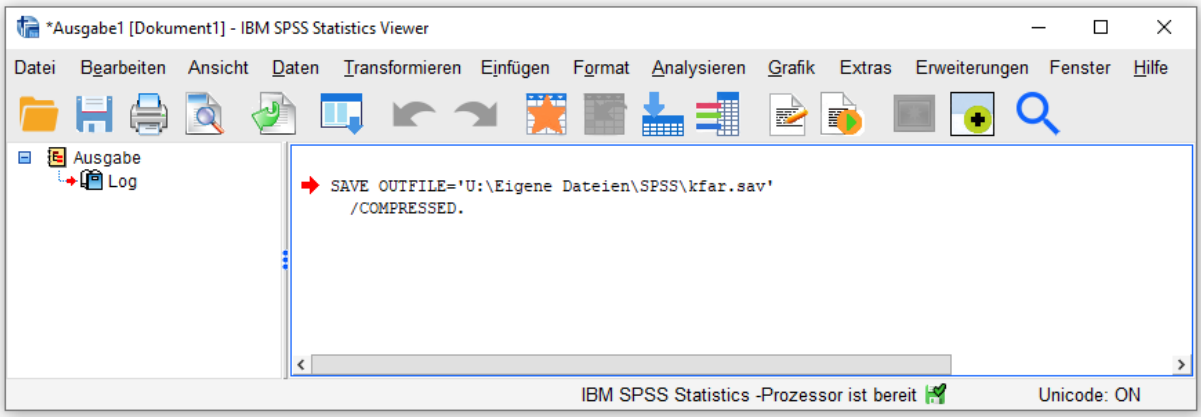


Sobald ein Datenblatt gegenüber dem zuletzt gespeicherten Zustand geändert worden ist, erscheint ein Sternchen vor dem Dateinamen, und die Symbolleisten-Schaltfläche zum Speichern (mit dem Diskettensymbol) wird aktiv, z. B.:



Beim Speichern führt der SPSS-Prozessor das Kommando SAVE aus, was später noch zu erläutern ist (siehe Abschnitt 7.7.1).

Von SPSS Versionen < 28 werden ausgeführte Kommandos per Voreinstellung im Ausgabefenster protokolliert, z.B. von SPSS 27:[1]



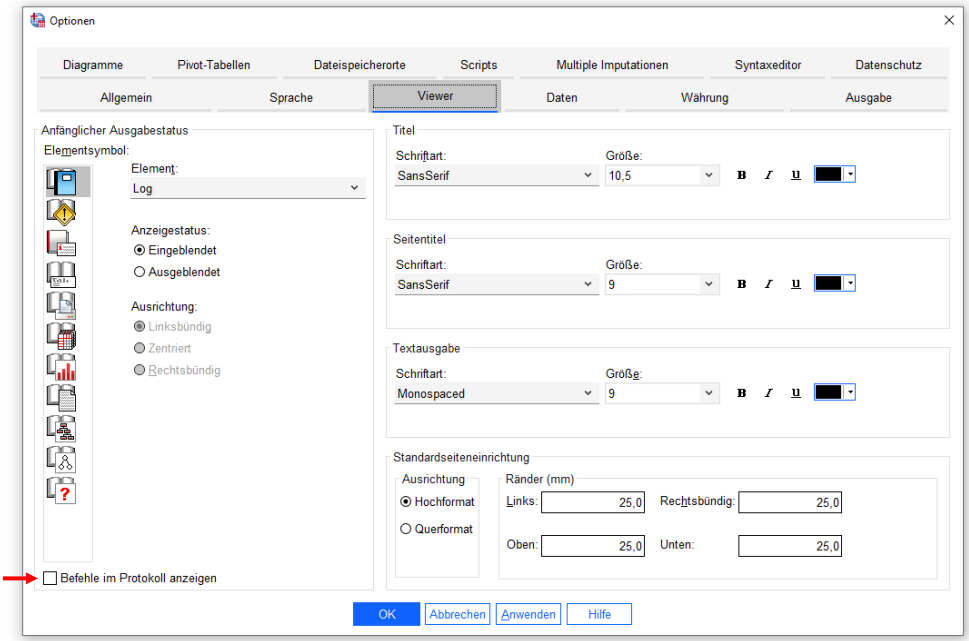
[1] Nähere Informationen zu den Ausgabefenstern folgen in den Abschnitten 5.4 und 10.4.

In SPSS 28 wurde diese Voreinstellung, die für überraschende Auftritte des Ausgabefensters gesorgt hat, revidiert.

Nach dem Menübefehl

Bearbeiten > Optionen

kann man im **Optionen**-Dialog auf der Registerkarte **Viewer** das Protokollieren der Kommandos im Ausgabefenster ab- bzw. einschalten:



Beim Speichern eines Datenblatts können auch alternative Dateiformate gewählt werden (z. B. EXCEL, SAS, Stata, Text).

Zum späteren Sichern in eine bereits zugeordnete Datei dient der Befehl:

Datei > Speichern

Alternativ können Sie auf das Symbol klicken oder die Tastenkombination **Strg+S** benutzen.

4.2.4 Rohdatendatei, Transformationsprogramm und Fertigdatendatei

Möglicherweise haben Sie sich beim Lesen des letzten Abschnitts gefragt, was das **r** im vorgeschlagenen Dateinamen **kfar.sav** bedeuten soll. Bei der Beantwortung dieser Frage sind leider einige Vorgriffe auf spätere Abschnitte nötig. Versuchen wir es trotzdem. Das **r** soll signalisieren, dass in dieser Datei die nach den Vorschriften des Codierplans erfassten **R**ohdaten stehen. In **kfar.sav** sollen ausschließlich folgende Arbeitsschritte einfließen:

- Variablendeklaration gemäß Codierplan
- Datenerfassung
- Nötigenfalls spätere Korrekturen von Erfassungsfehlern

Damit ist diese Datei für viele im Demoprojekt geplante Auswertungsschritte noch nicht geeignet. Es fehlt z. B. der Optimismus-Schätzwert, der aus den zwölf LOT-Fragen zu berechnen ist.

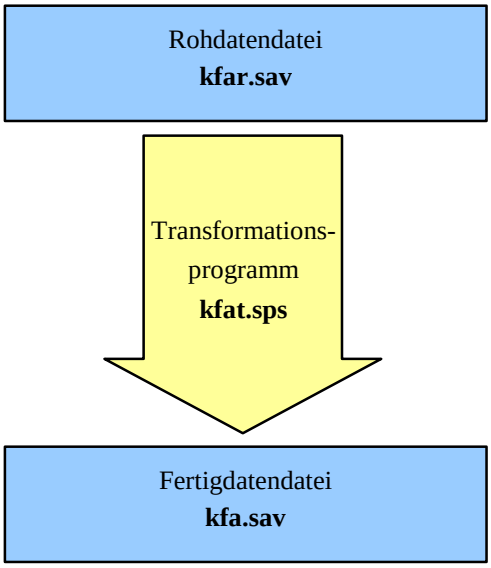
Aus der Rohdatendatei werden wir bald durch Variablenmodifikationen und -neuberechnungen die **Fertigdatendatei** erstellen, die alle potentiell mehrfach benötigten Variablen enthalten soll, sodass eine bequeme Datenbasis für statistische und grafische Analysen vorhanden sein wird.

In fast jedem Forschungsprojekt sind Variablenmodifikationen und -neuberechnungen in erheblichem Umfang erforderlich. Profis basteln dabei nicht so lange an der Rohdatendatei herum, bis die Fertigdatendatei entstanden ist, sondern sie erstellen (z. B. durch Konservieren von bearbeiteten Dialogboxen) ein **SPSS-Programm** (siehe unten), das alle Transformationen erledigt, und das bei Bedarf wiederholt ausgeführt werden kann.

Die zweistufige Projektdatenverwaltung mit Roh- und Fertigdatendatei verhindert in Kombination mit dem vermittelnden SPSS-Transformationsprogramm, dass bei jeder Änderung der Rohdaten die erwähnten Transformationen zur Fertigdatei wiederholt werden müssen. Solche Änderungen der Rohdaten (z. B. durch Fehlerkorrekturen oder Stichprobenerweiterungen) kommen nicht selten vor.

Weil die Kommandos des Transformationsprogramms auch mit Hilfe von Dialogboxen erstellt werden können, erfordert die professionelle Vorgehensweise kaum Programmierkenntnisse.

Es wird also folgende Struktur für die Verwaltung der Projektdaten vorgeschlagen:



Neben der bereits bekannten Rohdatendatei **kfar.sav** sind hier zu sehen:

- die Datei **kfat.sps** mit dem Transformationsprogramm, also mit den SPSS-Kommandos zur Modifikation bzw. zur Erstellung von Variablen
- die durch Anwendung des Transformationsprogramms auf die Rohdatendatei entstehende Fertigdatendatei **kfa.sav**

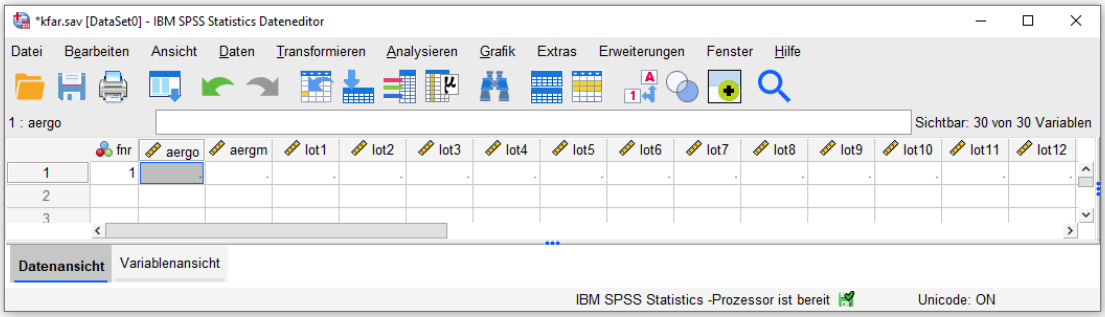
Die Erläuterungen in diesem Abschnitt werden vermutlich erst dann voll verständlich, wenn Sie sich mit Variablentransformationen und SPSS-Programmen auskennen.

Nach diesem Vorausblick wenden wir uns der nächsten Aufgabe im Demonstrationsprojekt zu: Wir tragen die erhobenen Daten in die Rohdatendatei **kfar.sav** ein.

4.2.5 Dateneingabe

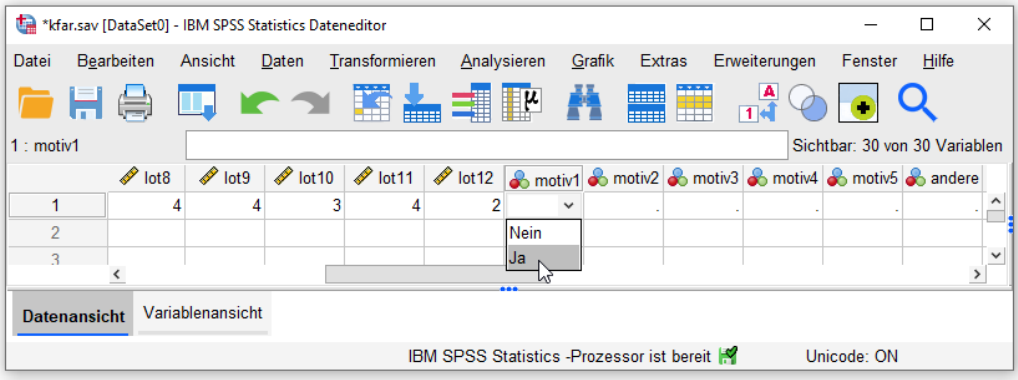
Wechseln Sie bitte bei Bedarf zur **Datenansicht** des (vermutlich noch einzigen) Datenfensters, und geben Sie die Daten des ersten Falles ein:

- Aktivieren Sie nötigenfalls die Zelle zur ersten Variablen, und tippen Sie den zugehörigen Wert ein.
- Drücken Sie die Tabulatortaste ⇤ oder die Taste mit dem Rechtspfeil →, um den eingetippten Wert zu quittieren und die Zellenmarkierung um eine Spalte nach *rechts* zu verschieben (zur nächsten Variablen):



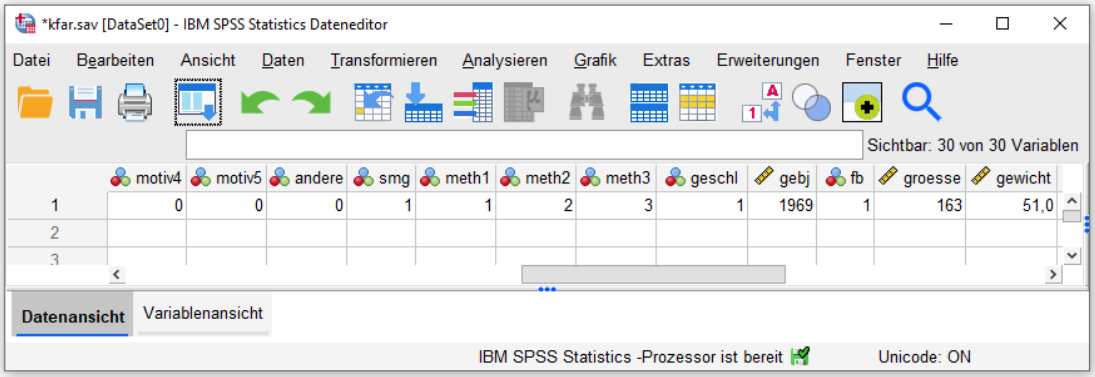
  Auch die **Enter**-Taste quittiert den eingetippten Wert, bewegt jedoch anschließend die Zellenmarkierung um eine Zeile nach *unten* (zum nächsten Fall), was in unserer jetzigen Lage weniger praktisch ist. Wenn Sie auf Abwege geraten sind, können Sie die Zellenmarkierung natürlich (z. B. per Mausklick) neu positionieren.

- Sobald für einen neuen Fall die erste Variablenausprägung eingetragen und quittiert worden ist, erhält er für die restlichen Variablen den Initialisierungswert SYSMIS (dargestellt durch einen Punkt).
- Wenn auf dem Registerblatt **Datenansicht** über den Menübefehl **Ansicht > Wertbeschriftungen** oder den Symbolschalter die Anzeige von Wertbeschriftungen aktiviert worden ist, dann lässt sich für jede Variable mit mindestens einer Wertbeschriftung per Doppelklick auf eine Datenzelle ein Drop-Down - Menü zur Unterstützung der Werteingabe öffnen, z. B.:



  So ist eine Datenerfassung ohne Kenntnis der Codierungsvorschriften möglich, wobei allerdings der Zeitaufwand im Vergleich zur direkten numerischen Werteingabe steigt. Es lohnt sich also, die Codierungsvorschriften im Kopf zu haben. Selbstverständlich ist auch bei aktivierter Beschriftungsanzeige die rationelle numerische Eingabe möglich.

- Tragen Sie die restlichen Werte des ersten Falles ein, jeweils quittiert mit der Tabulatortaste. So sieht der vollständig erfasste erste Fall der Manuskriptstichprobe (siehe Seite 59) im Datenfenster aus (bei abgeschalteter Wertbeschriftungsanzeige):



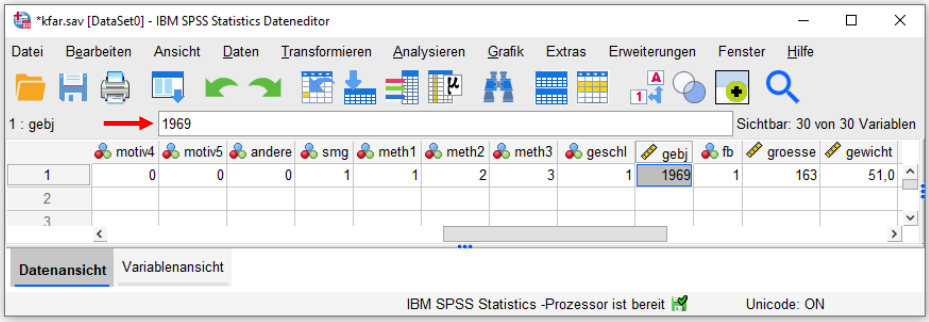
- Wenn Sie den Wert der letzten Variablen eines Falles mit der Tabulatortaste quittieren, setzt SPSS freundlicherweise die Zellenmarkierung in die erste Datenzelle des nächsten Falles, sodass Sie die Dateneingabe unmittelbar fortsetzen können.

4.2.6 Daten korrigieren

4.2.6.1 Wert einer Zelle ändern

Natürlich können die Eintragungen in einer Zelle jederzeit korrigiert werden:

- Wert ersetzen:
  - Zelle markieren
  - neuen Wert eintippen, wobei der alte überschrieben wird
- Wert in der Eingabezone editieren:
  - Zelle markieren
  - Wert in der Eingabezone oberhalb der Datenmatrix editieren



- Wert in der Zelle editieren:
  - Doppelklick auf die Zelle
  - Nun kann der Wert in der Zelle geändert werden.

4.2.6.2 Einen Fall einfügen

Gehen Sie folgendermaßen vor, um einen Fall einzufügen:

- Setzen Sie in der Nummerierungsspalte am linken Fensterrand einen rechten Mausklick auf die Zeilennummer desjenigen Falles, *vor* dem ein neuer Fall eingefügt werden soll. Daraufhin wird die gesamte angeklickte Zeile markiert, und es erscheint ein Kontextmenü.
- Wählen Sie aus dem Kontextmenü die Option **Fälle einfügen**

Der neue Fall erhält bei allen Variablen den Initialisierungswert SYSMIS.

4.2.6.3 Einen Fall löschen

Gehen Sie folgendermaßen vor, um einen Fall, d. h. eine Zeile der Datenmatrix, komplett zu löschen:

- Setzen Sie in der Nummerierungsspalte am linken Fensterrand einen rechten Mausklick auf die Zeilennummer des überflüssigen Falles. Daraufhin wird die gesamte angeklickte Zeile markiert, und es erscheint ein Kontextmenü.
- Wählen Sie aus dem Kontextmenü das Item **Löschen**

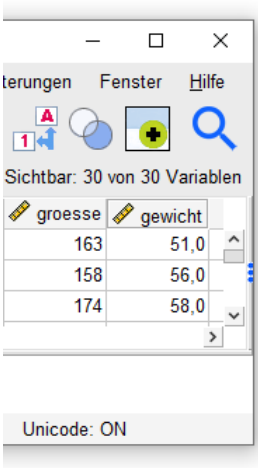
4.2.6.4 Fälle verschieben

Gehen Sie folgendermaßen vor, um Fälle per Drag & Drop (Ziehen und Ablegen) zu verschieben:

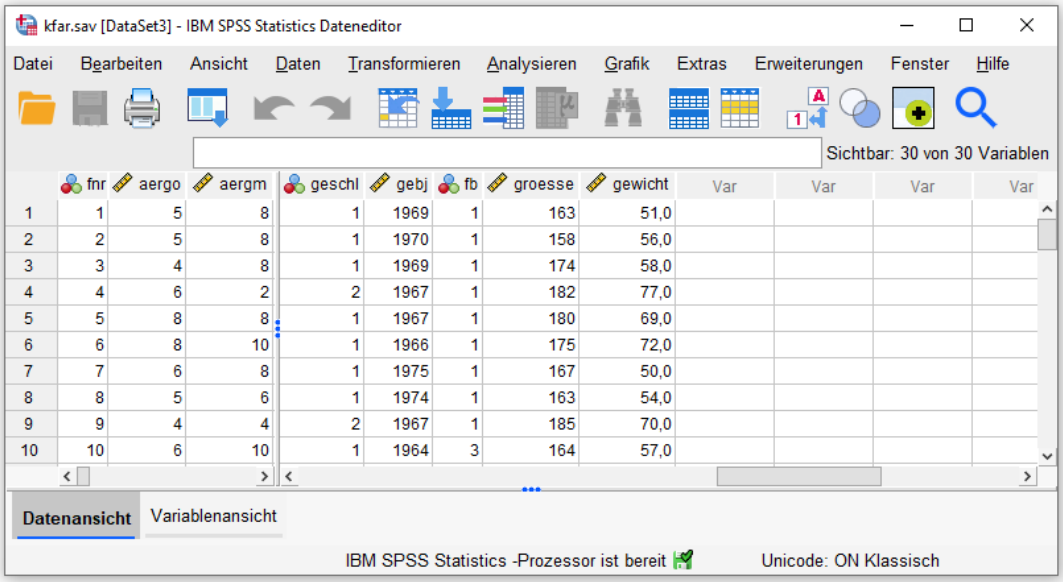
- Markieren Sie die zu verschiebenden Fälle auf windows-übliche Weise über Mausaktionen in der Nummerierungsspalte am linken Fensterrand, ggf. ergänzt durch die Strg - und/oder ⇧ - Taste. Lassen Sie anschließend die Maustaste wieder los.
- Klicken Sie in der Nummerierungsspalte erneut auf die zu verschiebende Fallauswahl, und halten Sie dabei die linke Maustaste gedrückt.
- Bewegen Sie bei gedrückter Maustaste den Mauszeiger zum Ziel der Verschiebungsaktion. Der augenblicklich eingestellte Zielort wird von SPSS durch eine rote Linie markiert.
- Wenn Sie die Maustaste loslassen, erscheinen die Fälle am neuen Ort.

4.2.7 Neben- oder übereinander stehende Ansichten auf ein Datenblatt

In der Datenansicht sind neben- oder übereinander stehende Ansichten auf ein Datenblatt realisierbar. Am rechten Rand befindet sich eine vertikale Begrenzung mit drei blauen Punkten:

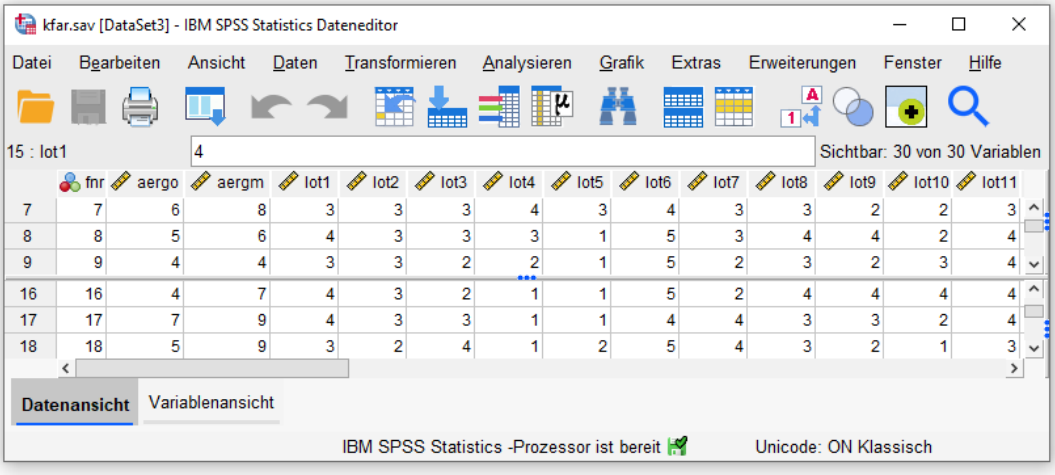


Diese Begrenzung kann per Maus nach links gezogen werden, um zwei parallele Ansichten auf das Datenblatt zu öffnen, z. B.:

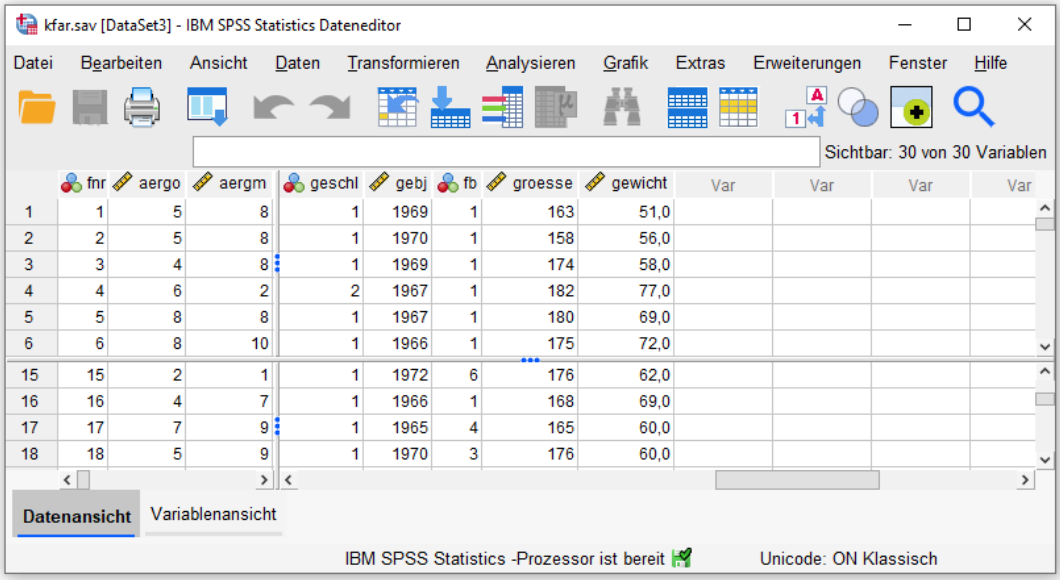


Die beiden Segmente können unabhängig fokussiert werden, sodass man beliebige Variablen nebeneinander positionieren kann, ohne das Datenblatt ändern zu müssen.

Analog lassen sich zwei übereinander stehende Ansichten auf das Datenblatt erzielen, sodass man beliebige Fälle übereinander positionieren kann, ohne das Datenblatt ändern zu müssen:



Die beiden Aufteilungen sind kombinierbar:



Als mögliche Anwendungen der aufgeteilten Ansichten sind zu nennen:

- Man kann Variablen oder Fälle zu Vergleichszwecken neben- bzw. übereinander positionieren, ohne das Datenblatt verändern zu müssen.
- Man kann Variablen oder Fälle per Drag & Drop zügig über größere Distanzen verschieben.

4.2.8 Weitere Möglichkeiten des Dateneditors

Über die beschriebenen Methoden hinaus bietet der Dateneditor u. a. die Möglichkeit, beliebige rechteckige Segmente einer Datenmatrix auszuschneiden, zu kopieren und einzufügen (auch zwischen verschiedenen Datenblättern).

Wer derartige, relativ fehleranfällige Modifikationen vornimmt, wird gelegentlich von der Möglichkeit profitieren, mit der Tastenkombination **Strg+Z**, über den Symbolschalter  oder mit dem Menübefehl:

Bearbeiten > Rückgängig

die letzte Änderung rückgängig machen zu können.

In Abschnitt 5.7 wird beschrieben, wie man im Datenfenster nach Variablenausprägungen suchen kann.

4.2.9 Übung

Für die Teilnehmer(innen) des realen statistischen Praktikums mit SPSS steht nun die Erfassung der erhobenen Daten an. Geben Sie alle Fälle ein, und sichern Sie (auch zwischendurch) in die zugeordnete Datendatei (z. B. **U:\Eigene Dateien\SPSS\kfar.sav**).

Wer dem Vorschlag aus Abschnitt 2.4.2.4 folgend zur Erfassung der Antworten auf die offene Frage im Fragebogenteil 3b eine dynamische Kategorienliste in Kombination mit einem sparsamen Set aus kategorialen Variablen vorgesehen hat (z. B. METH1 bis METH3), der muss nicht nur mechanisch Daten eintippen, sondern auch gelegentlich mit Kreativität und Ordnungssinn

neue Methodenkategorien definieren und dokumentieren. Beim Erfassen der Manuskriptstichprobe entstand die folgende Liste:

Kategorie	Code
Faktorenanalyse	1
Regressionsanalyse	2
Korrelationsanalyse	3
Varianzanalyse	4
Strukturgleichungsanalyse	5
Clusteranalyse	6
Diskriminanzanalyse	7
Logistische Regression	8
Conjoint-Analyse	9

Diese Tabelle vervollständigt den Codierplan (vgl. Abschnitt 2.4.3.5). Es ist empfehlenswert, die Definitionen der Variablen METH1 bis METH3 durch entsprechende Wertbeschriftungen zu komplettieren (vgl. Abschnitt 4.2.2.3).

5 Univariate Verteilungs- und Fehleranalysen

Nach etlichen methodischen und technischen Vorbereitungen werden wir in diesem Kapitel endlich statistische und grafische Analysen durchführen. Im Demonstrationsprojekt stehen aktuell zwei Aufgaben an:

- Kontrolle der erfassten Daten
  Bei der im Projekt praktizierten manuellen Datenerfassung sind Fehler nahezu unvermeidlich. Irreguläre Werte können aufgrund von Konfigurationsfehlern oder technischen Pannen aber auch bei einer Online-Datenerhebung auftreten.
- Inspektion der univariaten Verteilungen
  Die univariaten Verteilungen enthalten wichtige Informationen über die Zusammensetzung und das Verhalten der Stichprobe. Im Demonstrationsprojekt sind wir bei einigen Variablen besonders neugierig auf die Verteilungen (z. B. bei AERGO und AERGM).

Eine *Beschränkung* auf univariate Analysen ist der Regel *nicht* sinnvoll, weil dabei keine Ursachen aufgeklärt oder Prognosemodelle entwickelt werden können. Dazu sind multivariate Zusammenhangsanalysen erforderlich.

Auch bei univariaten Analysen greift die pure Deskription zu kurz, weil man in der Regel Informationen über die *Population* gewinnen möchte, aus der die Stichprobe stammt. So muss etwa die Wahlforschung zu den geschätzten Stimmanteilen der Parteien in der Regel auch Vertrauensintervalle liefern, also Inferenzstatistik betreiben.

5.1 Fehlerhafte Werte aufspüren

Speziell bei der manuellen Datenerfassung sind Fehler praktisch unvermeidbar. Manche Fehler sind als Verstöße gegen Gültigkeitsregeln leicht aufzuspüren:

Beispiel: Wenn bei der Variablen GESCHL nur die Werte 1 (für Frauen), 2 (für Männer) und 3 (Divers) erlaubt sind, dann ist z. B. der Wert 5 sofort als Fehler zu erkennen.

Von den *logisch* unmöglichen Werten (z. B. GESCHL = 5) sind die *empirisch* unmöglichen zu unterscheiden (z. B. Alter eines Menschen über 130 Jahre).

Wie die von einem irregulären Wert betroffenen Fälle in der Arbeitsdatei aufzuspüren sind, wird in Abschnitt 5.7 beschrieben.

Weit schwieriger sind Fehler zu entdecken, die keine allgemeine Gültigkeitsregel verletzen:

Beispiel: Wenn unter der oben angegebenen GESCHL-Codierungsvorschrift für den Untersuchungsteilnehmer Kurt Müller versehentlich der Wert 1 eingegeben wurde, dann kann dieser Fehler nur gefunden werden, wenn ...

- schriftliche oder sonstige Untersuchungsdokumente vorhanden sind
- und aufwändige Handarbeit investiert wird.

Welcher Aufwand bei der Datenprüfung erforderlich bzw. sinnvoll ist, hängt von der verwendeten Erhebungs- bzw. Erfassungsmethode (und damit von der Fehlerwahrscheinlichkeit) sowie von der Stichprobengröße ab.

Ein entdeckter Fehler ist nach Möglichkeit (z. B. aufgrund von schriftlichen Untersuchungsdokumenten) zu korrigieren. Ist dies nicht möglich, bestehen folgende Optionen:

- Man kann den fehlerhaften Wert durch einen MD-Indikator ersetzen, z. B. durch Löschen auf SYSMIS setzen.
- Man kann den fehlerhaften Wert als zusätzlichen MD-Indikator deklarieren (siehe Abschnitt 4.2.2.2). Der Wert wird damit neutralisiert, aber nicht eliminiert.

Bei einem Fall mit *zahlreichen* fehlerhaften und/oder fehlenden Werten kommt auch das komplette Löschen des Falles in Betracht.

Weitere Details zu Ursachen und Behandlungsmethoden für fehlerhafte Daten finden sich bei Lück (2011, S. 66ff).

Im Zusammenhang mit der Fehlerbereinigung soll noch das SPSS-Modul *Data Preparation* erwähnt werden, das die routinemäßige Untersuchung von strukturgleichen Datendateien durch passend definierbare Regeln unterstützt (siehe Handbuch IBM Corp. 2021a und Menübefehl **Daten > Validierung**). Im Vergleich zum anschließend vorgestellten Verfahren, die Suche nach unzulässigen Werten parallel zur univariaten Verteilungsanalyse zu betreiben, erlauben die Prozeduren im Modul *Data Preparation* eine schnellere Identifikation von Fällen mit unzulässigen Werten. Die Definition eines entsprechenden Regelpakets lohnt sich z. B. dann, wenn in einer Forschungseinrichtung oder Firma regelmäßig Datendateien mit identischen Variablen eintreffen, die einer Qualitätskontrolle unterzogen werden müssen.

5.1.1 Suche nach unzulässigen Werten

Über ein Online-Werkzeug erfasste Daten sind bei einer sorgfältig konfigurierten Eingangskontrolle frei von unzulässigen Werten.[1] Auch bei der manuellen Dateneingabe mit einem spezialisierten Erfassungsprogramm (vgl. Abschnitt 4.1.1) lassen sich unzulässige Werte von der Datendatei fernhalten.

Bei der manuellen Erfassung mit dem SPSS-Dateneditor findet eine derartige Eingangskontrolle *nicht* statt. Das Definieren von Wertbeschriftungen hat zwar nützliche Effekte auf Tabellen und auf das Verhalten von einigen Dialogboxen bzw. Assistenten in SPSS (z. B. bei der Diagrammerstellung). Man kann auf diese Weise aber *nicht* die Menge der zulässigen Werte definieren. Eine per Dateneditor erfasste Datendatei muss daher systematisch nach Werten außerhalb der zulässigen Bereiche durchsucht werden. Dies kann allerdings ohne großen Zusatzaufwand im Rahmen der aus wissenschaftlichen Gründen ohnehin erforderlichen univariaten Verteilungsanalyse geschehen.

Gelegentlich ist eine *bivariate* Verteilungsanalyse erforderlich, um unzulässige, d. h. aus logischen oder empirischen Gründen unmögliche, Werte zu identifizieren. So sind z. B. Studierende im 17. Semester oder mit einem Alter von 19 Jahren nicht ungewöhnlich, doch wäre eine Kombination der beiden Merkmalsausprägungen eindeutig als irregulär zu erkennen. Wir werden im Kurs nicht systematisch mit bivariaten oder noch komplexeren Techniken auf Fehlersuche gehen. Wenn jedoch z. B. ein bivariates Streudiagramm zur grafischen Analyse bzw. Präsentation von Daten zum Einsatz kommt (vgl. Kapitel 11), dann sollten irreguläre Wertekombinationen einem wachsamen Auge nicht entgehen.

---

[1] Allerdings fehlen in einer so entstandenen Datendatei eventuell auch einige Basketballspieler, die empört die Studienteilnahme angebrochen haben, nachdem ihre Körpergröße von z. B. 235 cm als unzulässig abgelehnt wurde. Bei Plausibilitätskontrollen in Online-Erfassungswerkzeugen darf man also nicht zu kleinlich sein.

5.1.2 Einzelprüfung aller Werte

Fehler, die gegen keine Gültigkeitsregel verstoßen, lassen sich nur mit Fleißarbeit entdecken, wobei z. B. die erfassten Daten Wert für Wert mit schriftlichen Unterlagen zu vergleichen sind.

Eine aufwändige Prüfmethode ist *bei kleinen Stichproben* (bis zu 50 Fälle) empfehlenswert, denn:

- Erfassungsfehler wirken sich hier besonders stark aus.
- Der Zeitaufwand ist erträglich.

Wir wollen exemplarisch den Effekt von Erfassungsfehlern auf die Varianz (Unsicherheit) eines Stichprobenmittelwerts als Schätzer für den zugehörigen Populationserwartungswert untersuchen. Für N erfasste Werte X_i (i = 1, ..., N) nehmen wir an, dass sie jeweils mit einem Erfassungsfehler F_i belastet sind, wobei die Erfassungsfehler den Erwartungswert 0 haben sowie untereinander und von den korrekten Beobachtungswerten T_i unabhängig sein sollen:

$$X_i = T_i + F_i, \quad \mathrm{E}(F_i) = 0, \quad \mathrm{E}(X_i) = \mathrm{E}(T_i) =: \mu$$
$$\mathrm{Var}(T_i) = \sigma^2, \quad \mathrm{Var}(F_i) = \sigma_F^2$$

Der Stichprobenmittelwert $\overline{X}$ aus den fehlerhaft erfassten Werten hat denselben Erwartungswert μ wie der Stichprobenmittelwert $\overline{T}$ aus den korrekt erfassten Werten.

Für die Varianz von $\overline{T}$ gilt:

$$\mathrm{Var}(\overline{T}) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} T_i\right) = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(T_i) = \frac{1}{N^2}\,N\,\sigma^2 = \frac{\sigma^2}{N}$$

Für die Varianz von $\overline{X}$ erhalten wir:

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}(T_i + F_i)\right) = \frac{1}{N^2}\sum_{i=1}^{N}\left(\mathrm{Var}(T_i) + \mathrm{Var}(F_i)\right) = \frac{1}{N^2}\,N\left(\sigma^2 + \sigma_F^2\right) = \frac{\sigma^2}{N} + \frac{\sigma_F^2}{N}$$

Im Stichprobenmittel, das den Erwartungswert der Population schätzt, hängt offenbar der Präzisionsverlust von der Erfassungsfehlervarianz σ_F^2 und von der Stichprobengröße N ab. Während sich in einer großen Stichprobe der niedrige Ausgangswert $\frac{\sigma^2}{N}$ der Unsicherheit nur unwesentlich erhöht, kommt es in einer kleinen Stichprobe mit ihrem bereits ungünstigen Ausgangsniveau zu einem erheblichen Präzisionsverlust.

Vor allem in kleineren Stichproben beeinträchtigen fehlerbehaftete Messwerte die Präzision von Parameterschätzungen und die Power von Signifikanztests.

Obwohl bei unserer kleinen Stichprobe eine Einzelprüfung aller Werte angemessen wäre, verzichten wir aus Zeitgründen darauf.

5.2 Öffnen von Datendateien

Vermutlich haben Sie nach der anstrengenden Datenerfassung eine Pause eingelegt und SPSS verlassen, sodass wir jetzt offiziell das Öffnen einer Datendatei üben.

5.2.1 SPSS-Datendateien

Starten Sie SPSS, und öffnen Sie Ihre vorhandene Rohdatendatei **kfar.sav** entweder mit Hilfe des Begrüßungsdialogs oder über den Menübefehl

Datei > Zuletzt verwendete Daten

Während die bisher beschriebenen Verfahren auf dem Erinnerungsvermögen von SPSS basieren, klappt der folgende Menübefehl generell:

Datei > Öffnen > Daten

Ist eine SPSS-Datendatei in einem geöffneten Fenster des Windows-Explorers für Mausaktionen zugänglich, dann stehen weitere Techniken zum Öffnen der Datei zur Verfügung:

- Doppelklick
  Wenn SPSS nicht läuft wird es durch den Doppelklick gestartet.
- Drag & Drop: Ziehen und auf einem beliebigen SPSS-Fenster fallen lassen

Beim Öffnen einer Datendatei legt SPSS ein neues Datenblatt an, kopiert die eingelesenen Daten samt Variablendeklarationen dorthin und benennt das Datenblatt (z. B. mit: **DataSet1**). Wenn die Datendatei bisher noch nicht geöffnet war, dann wird sie mit dem Datenblatt verbunden.

Damit die Veränderungen an einem Datenblatt die aktuelle SPSS-Sitzung überdauern, muss das Datenblatt in eine permanente Datendatei gesichert werden, z. B. ...

- über den Menübefehl **Datei > Speichern** in die mit dem Datenblatt verbundene SPSS-Datendatei
- oder über den Menübefehl **Datei > Speichern unter** in eine andere SPSS-Datendatei.

5.2.2 Fremde Dateiformate

Auch bei Datendateien in Fremdformaten, stellt das Einlesen durch SPSS selten ein Problem dar, denn SPSS unterstützt beim Datenimport zahlreiche Formate (z. B. Textdateien, Microsoft Excel, SAS, Stata). Das Einlesen von Textdateien wird in Kapitel 17 beschrieben.

SPSS übernimmt die eingelesenen Daten in ein neues, benanntes Datenblatt. Bei Bedarf können Sie anschließend Variablendeklarationen vornehmen (vgl. Abschnitt 4.2.2) und das Ergebnis mit dem Menübefehl

Datei > Speichern unter

in eine neue SPSS-Datendatei sichern.

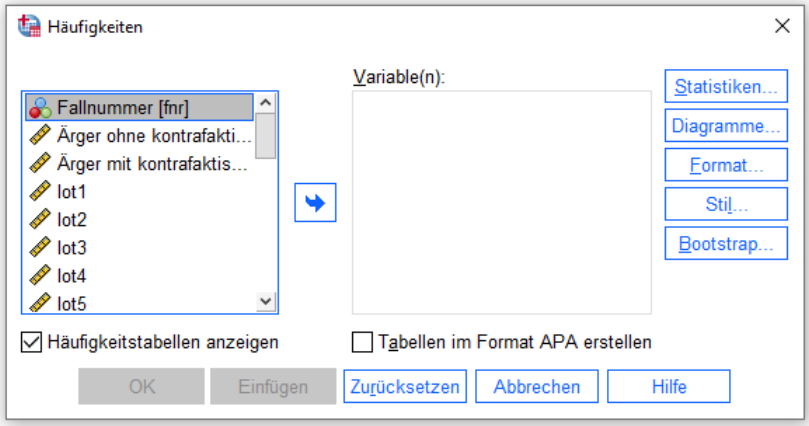
5.3 Verteilungsanalysen für kategoriale Variablen

Da wir unsere Daten mit dem SPSS-Dateneditor erfasst haben, der keine Plausibilitätskontrolle bei der Eingabe vornimmt, müssen wir aufgrund der Überlegungen aus Abschnitt 5.1 systematisch nach unzulässigen Werten suchen. Dies kann ohne nennenswerten Zusatzaufwand im Rahmen der univariaten Verteilungsanalysen geschehen, die wir aus Gründen wissenschaftlicher Sorgfalt und Neugier routinemäßig durchführen.

Wir untersuchen zunächst die Verteilungen der nominalskalierten Variablen GESCHL und FB mit Hilfe von Häufigkeitstabellen und Balkendiagrammen. Mit dem Menübefehl

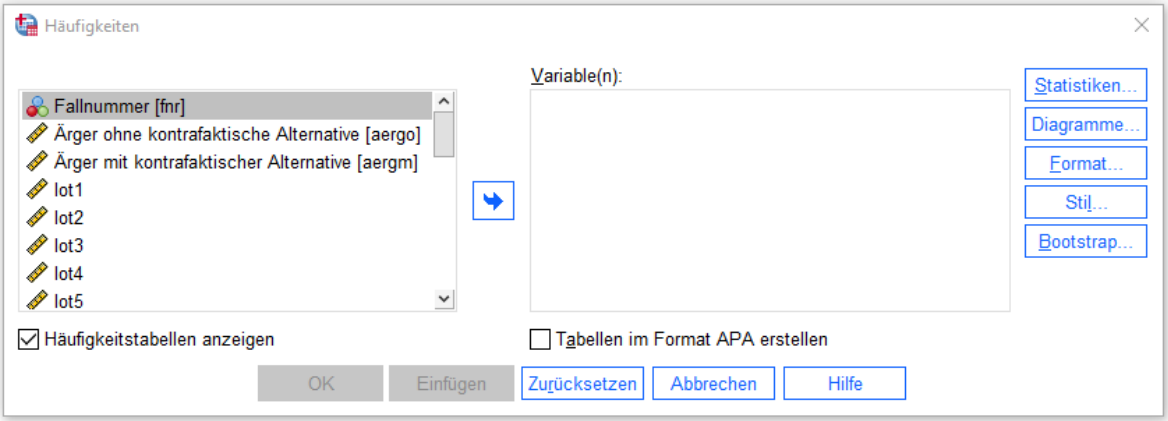
Analysieren > Deskriptive Statistiken > Häufigkeiten

erhalten wir die folgende Dialogbox zur Anforderung von Häufigkeitsanalysen:

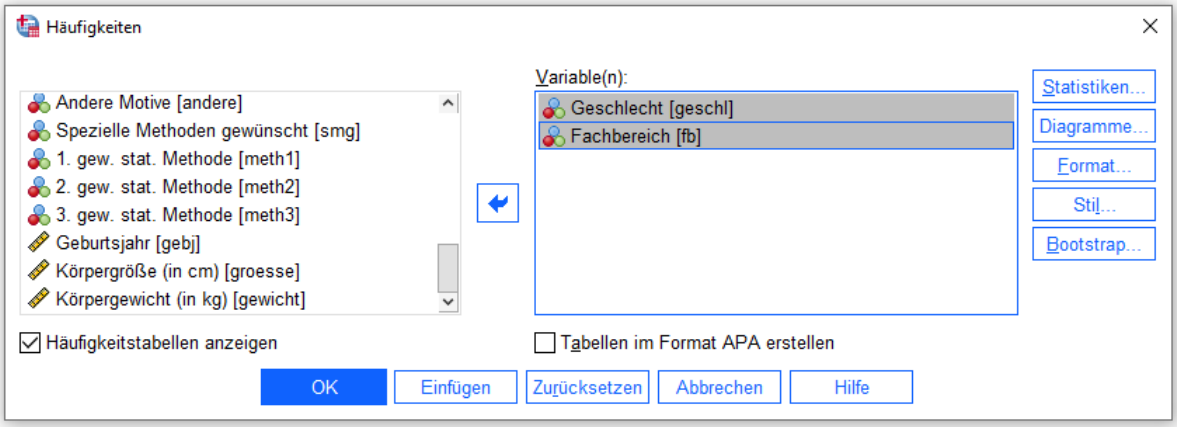


Zur bequemen Spezifikation der im aktuellen Prozeduraufruf zu analysierenden Variablen dienen die beiden Variablen-Auswahlbereiche. Links stehen alle Variablen der Arbeitsdatei, die derzeit *nicht* für die Analyse ausgewählt sind (*Anwärterliste*). Rechts daneben, unter dem Etikett **Variable(n)**, stehen die Ausgewählten (*Teilnehmerliste*). Dazwischen befindet sich ein Transportschalter (bzw.), mit dem sich markierte Variablen aus der linken Liste nach rechts bzw. aus der rechten Liste nach links verschieben lassen. Selbstverständlich ist es möglich, in eine Häufigkeitsanalyse *zwei oder mehr* Variablen einzubeziehen, wobei alle numerische Variablen (unabhängig vom deklarierten Messniveau) identisch behandelt werden.

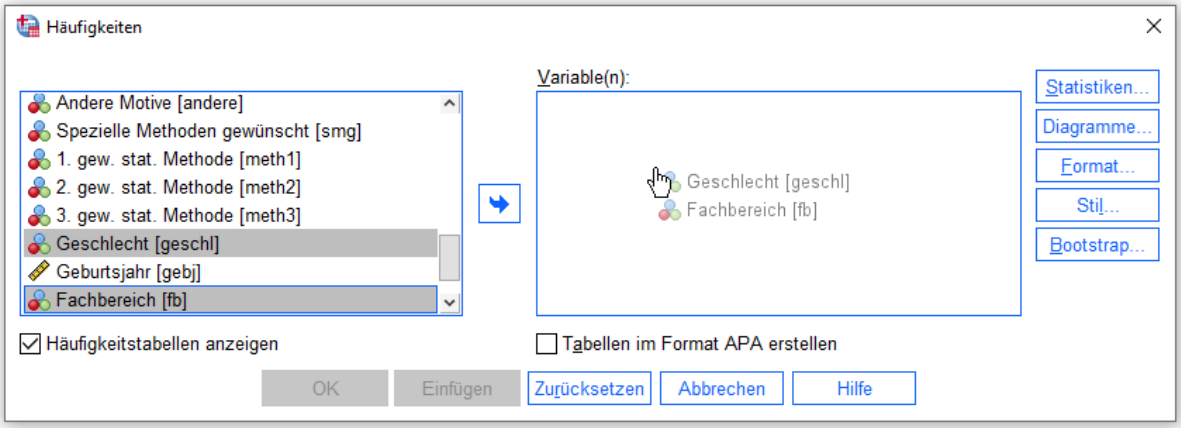
Alle SPSS-Dialogboxen lassen sich in der Größe verändern, wobei der verfügbare Platz auf die Bedienelemente dynamisch aufgeteilt wird. Daher kann man im konkreten Beispiel für die komplette Anzeige der Variablenbeschriftungen und der dahinter (zwischen eckigen Klammern) folgenden Variablennamen sorgen:



Markieren Sie in der Anwärterliste (links) die Variablen GESCHL und FB, und befördern Sie diese per Mausklick auf den Transportschalter in die Teilnehmerliste (rechts):

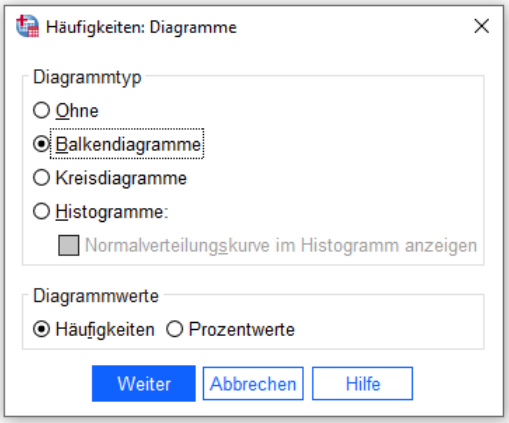


Statt den Schalter  zu benutzen, können Sie in SPSS solche Transportaufgaben auch per Drag & Drop (Ziehen und Ablegen) erledigen:



Bei einer längeren Variablenliste ist es sehr hilfreich, dass SPSS beim Eintippen einer Variablenbezeichnung das nächste kompatible Listenelement markiert, wobei der Name oder die Variablenbeschriftung (falls vorhanden) einzugeben ist. Die Eingabe kann stoppen, sobald das gewünschte Listenelement markiert ist.

Begeben Sie sich anschließend in den Subdialog **Diagramme**, indem Sie einen Mausklick auf den gleichnamigen Schalter setzen. Wählen Sie im Rahmen **Diagrammtyp** die für nominalskalierte Merkmale angemessene Option **Balkendiagramme**:



Wer Erläuterungen zu den verfügbaren Diagrammtypen wünscht, kann diese per Mausklick auf den Schalter **Hilfe** anfordern.

Quittieren Sie die Subdialogbox **Diagramme** mit **Weiter** und die Hauptdialogbox mit **OK**. Daraufhin präsentiert SPSS die Ergebnisse im Ausgabefenster (betitelt mit: **IBM SPSS Statistics Viewer**), das sich in den Vordergrund drängt.

Wir erfahren in der ersten Tabelle, dass bei den untersuchten Variablen alle Werte vorhanden waren:

Statistiken

		Geschlecht	Fachbereich
N	Gültig	31	31
	Fehlend	0	0

Bei Anforderung einer Häufigkeitsanalyse produziert SPSS per Voreinstellung für jede Variable eine Tabelle mit den aufsteigend geordneten Beobachtungswerten, die für jeden aufgetretenen Wert *x* eine Zeile mit den folgenden Angaben enthält:

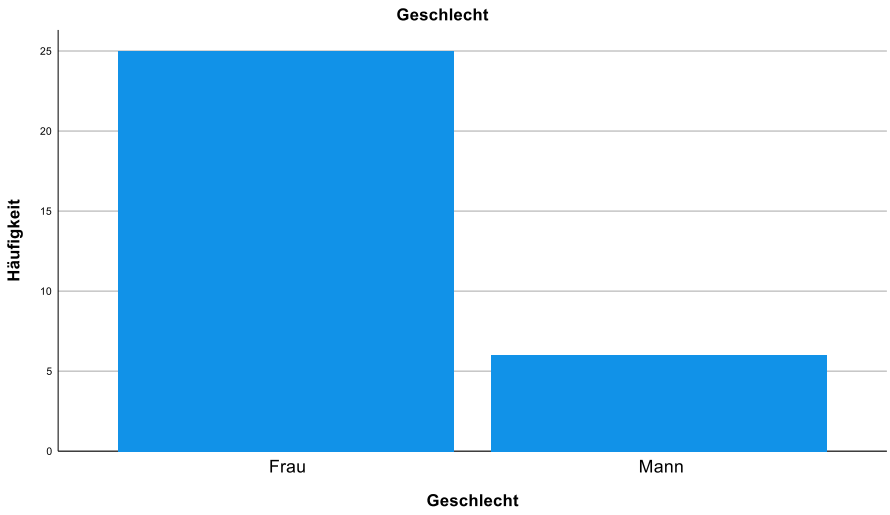
- Absolute Häufigkeit von *x*
- Prozentanteil von *x* bezogen auf den Stichprobenumfang
- Prozentanteil von *x* bezogen auf die Anzahl aller Beobachtungen mit validen Werten
- Prozentanteil von validen Werten kleiner oder gleich *x* bezogen auf die Anzahl aller Beobachtungen mit validen Werten

Für GESCHL erhalten wir diese Häufigkeitstabelle:

Geschlecht

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Frau	25	80,6	80,6	80,6
	Mann	6	19,4	19,4	100,0
	Gesamt	31	100,0	100,0	

und das folgende Balkendiagramm:



Zunächst beobachten wir, dass bei der Variablen GESCHL kein unzulässiger Wert vorliegt.

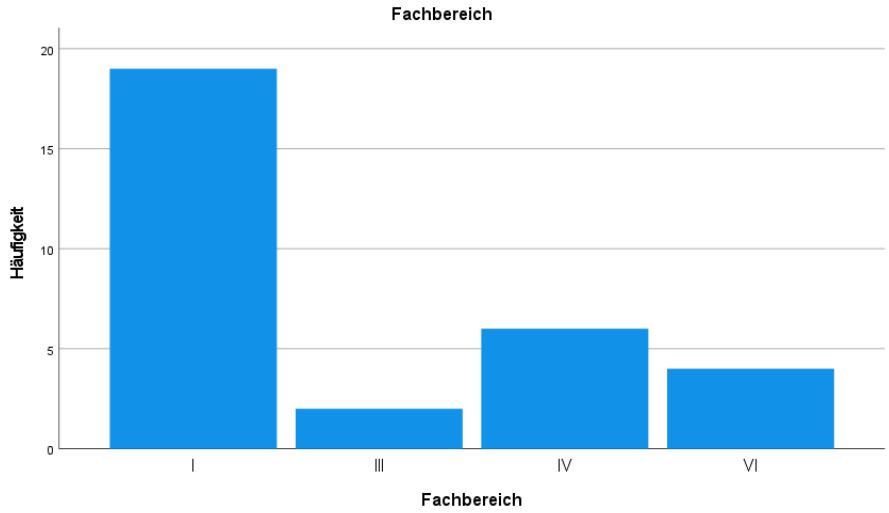
Bei der Geschlechtsverteilung stellen wir einen sehr hohen Frauenanteil fest ($\hat{p}$ = 0,81), der als wesentliches Merkmal unserer Stichprobe berichtet werden muss. Bei potentiell geschlechtsabhängigen Ergebnissen müssen wir besonders vorsichtig interpretieren bzw. generalisieren.

Das Vertrauensintervall zu einer Proportion kann die SPSS-Prozedur zur Häufigkeitsanalyse leider *nicht* liefern. Wie es mit der SPSS-Prozedur NPTESTS zu ermitteln ist, erfahren Sie in Abschnitt 5.8.

Erste Hinweise zur möglichen Ursache der hohen Frauenquote in der Manuskriptstichprobe liefert die Verteilung der Fachbereichs-Variablen:

Fachbereich

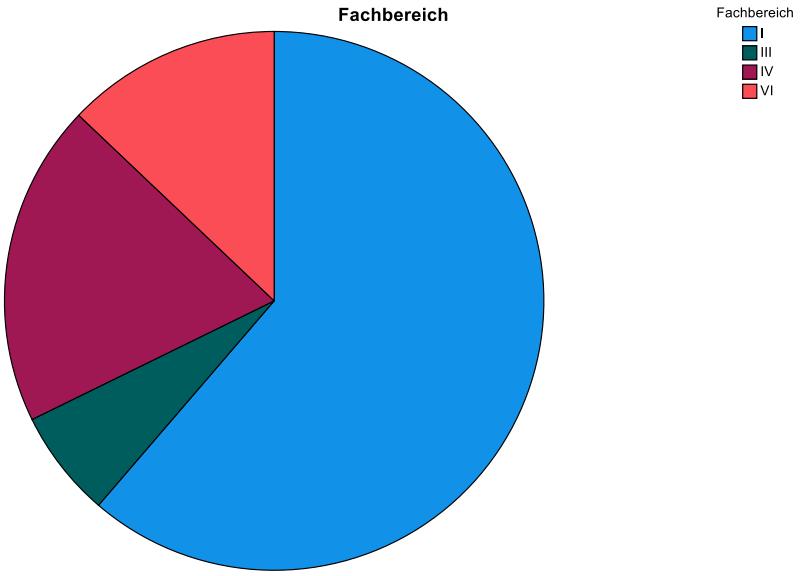
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	I	19	61,3	61,3	61,3
	III	2	6,5	6,5	67,7
	IV	6	19,4	19,4	87,1
	VI	4	12,9	12,9	100,0
	Gesamt	31	100,0	100,0	



Wir sehen, dass in der Manuskriptstichprobe der Fachbereich I sehr stark vertreten ist, was mit dem Kurstermin zusammenhängen mag. Im Fachbereich I der Universität Trier (Fächer: Philosophie, Pädagogik, Psychologie) ist der Frauenanteil relativ hoch (vgl. Abschnitt 14.2).

Ein Balkendiagramm zeigt für jeden aufgetretenen Wert seine absolute oder relative Häufigkeit durch einen Balken entsprechender Höhe. Man kann die Balkenhöhen, also die (relativen) Häufigkeiten der Werte, gut miteinander vergleichen, während der Vergleich mit dem Ganzen (alle Balken übereinander gestapelt) weniger gut gelingt.

Der Kreis ist wohl das ideale Sinnbild für das Ganze, und dementsprechend lässt sich die Zerlegung (Verteilung) des Ganzen auf die Werte einer Variablen sehr gut durch Kreissegmente mit passendem Winkel veranschaulichen, z. B. bei der Variablen FB:

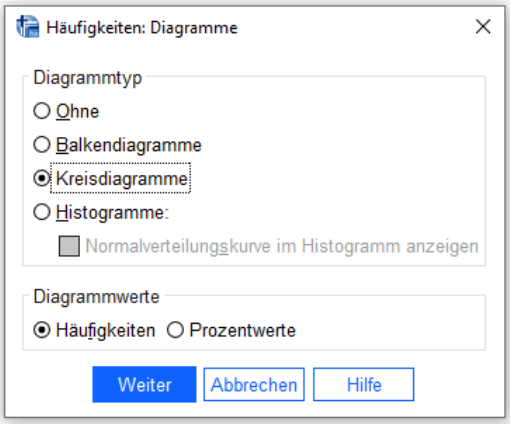


Für den Wert x mit der absoluten Häufigkeit $h(x)$ wird der Winkel $w(x)$ so ermittelt, dass gilt (mit N als Bezeichnung für die Anzahl der gültigen Werte):

$$\frac{h(x)}{N} = \frac{w(x)}{360}$$

Man kann für jeden Wert seinen Anteil gut beurteilen, während der Vergleich mit anderen, nicht direkt benachbarten Werten gelegentlich schwerer fällt.

Um Kreisdiagramme statt Balkendiagrammen zu erstellen, müssen Sie lediglich Ihre Wahl im Subdialog **Diagramme** der Häufigkeitsprozedur ändern:

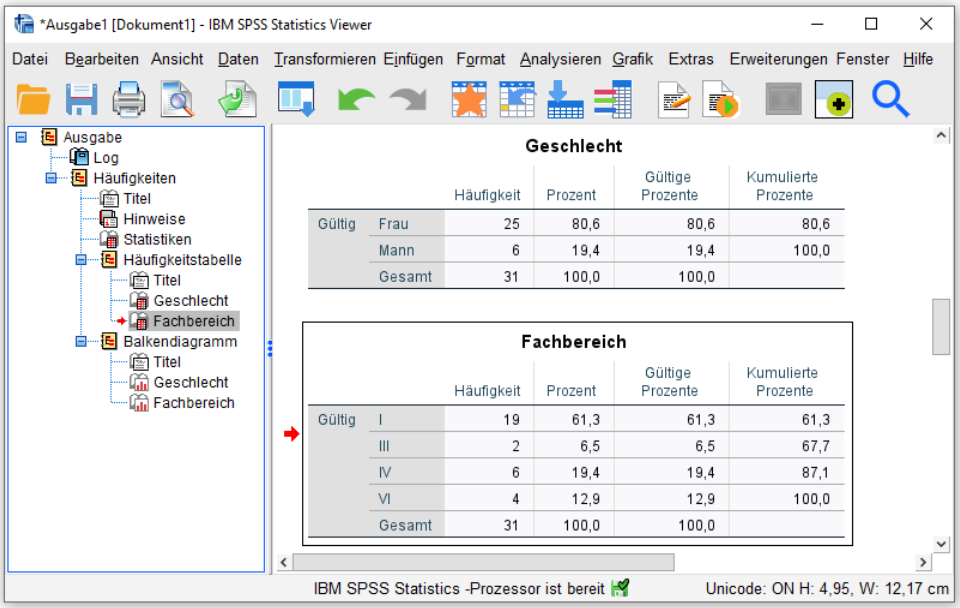


Der aktuelle Abschnitt sollte nur einen ersten Eindruck von den SPSS-Grafikoptionen vermittelt. Wir haben eine integrierte Option der Dialogbox zur Häufigkeitsanalyse benutzt und auf eine Nachbearbeitung per Diagrammeditor verzichtet (siehe Abschnitt 11.2). Die meisten grafischen Darstellungsmöglichkeiten bietet SPSS über das Hauptmenü **Grafik** an, mit dem wir uns in Kapitel 11 befassen werden.

Die obigen SPSS-Ausgaben wurden übrigens über die Windows-Zwischenablage in das Textverarbeitungsprogramm Microsoft Word© übertragen. Mit dieser Form des Datenaustauschs und mit anderen Optionen beim Arbeiten mit dem Ausgabefenster beschäftigen wir uns im nächsten Abschnitt.

5.4 Arbeiten mit dem Ausgabefenster (Teil I)

In seiner voreingestellten Variante ist das SPSS-Ausgabefenster (betitelt mit: **IBM SPSS Statistics Viewer**) zweigeteilt in den Navigationsbereich (die Gliederungsansicht) am linken Rand und den Inhaltsbereich:



So wird ein schnelles Navigieren zwischen den Ausgabebestandteilen ermöglicht.

Um die Platzverteilung auf die beiden Ausgabefenstersegmente zu ändern, klickt man auf die Trennlinie und verschiebt sie bei gedrückter Maustaste.

Wesentliche Bestandteile des Inhaltsbereichs sind Tabellen und Diagramme. Zu ihrer Nachbearbeitung steht jeweils ein spezieller Editor zur Verfügung, der per Doppelklick auf ein Objekt gestartet wird. Außerdem können in einem Ausgabefenster noch SPSS-Kommandos, Textausgaben älterer Prozeduren, Warnungen, Anmerkungen und Titelzeilen auftreten.

5.4.1 Arbeiten im Navigationsbereich

Die meisten der anschließend beschriebenen Aktionen im Navigationsbereich wirken sich synchron auch auf den Inhaltsbereich aus.

5.4.1.1 Fokus positionieren

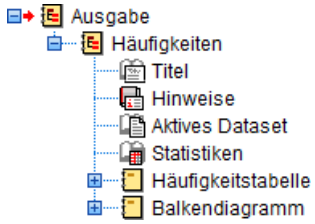
Ein kleiner roter Pfeil zeigt im Gliederungsbereich auf die Bezeichnung derjenigen Ausgabe, die gerade im Inhaltsbereich privilegiert dargestellt wird. Per Mausklick auf eine andere Ausgabenbeschriftung kann dieser Fokus verschoben werden.

5.4.1.2 Ausgabeblöcke bzw. Teilausgaben aus- oder einblenden

Ein *Block* mit zusammengehörigen Ausgaben (in der Regel entstanden aus einer Analyseanforderung) wird ...

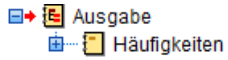
- ausgeblendet: per Mausklick auf das Minus-Zeichen ⊟ neben dem Symbol für einen expandierten Block oder per Doppelklick auf das Block-Symbol.

  Beispiel:



- eingeblendet: per Mausklick auf das Plus-Zeichen ⊞ neben Symbol für einen reduzierten Block oder per Doppelklick auf das Block-Symbol.

  Beispiel:



Eine *Teil*ausgabe innerhalb eines Blocks wird per Doppelklick auf das zugehörige Buchsymbol aus- bzw. eingeblendet. Das Buchsymbol erscheint dementsprechend zugeklappt (im Beispiel: **Hinweise**) oder aufgeklappt (im Beispiel: **Statistiken**).

5.4.1.3 Ausgabeblöcke oder Teilausgaben markieren

Im Navigationsbereich kann man Ausgabeblöcke und/oder Teilausgaben (z. B. zum anschließenden Löschen) so markieren:

- Einen Ausgabeblock: Per Mausklick auf das Blocksymbol
- Eine Teilausgabe: Per Mausklick auf das Buchsymbol
- Mehrere Blöcke bzw. Teile: Durch Mausklicks in Kombination mit der ⇧ - bzw. Strg - Taste

5.4.1.4 Ausgabeblöcke oder Teilausgaben löschen oder verschieben

Markierte Blöcke bzw. Teilausgaben kann man ...

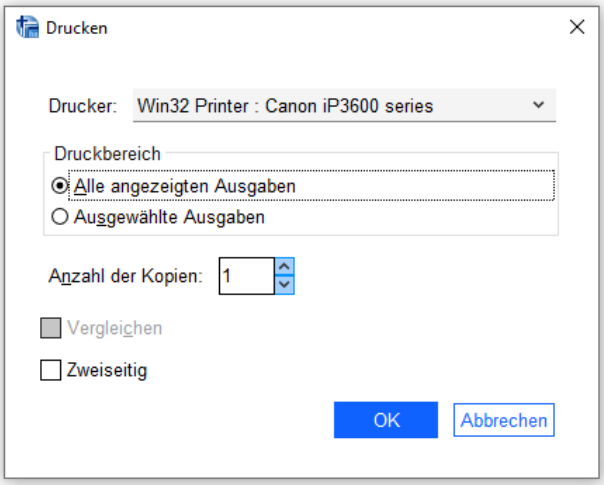
- löschen: mit der **Entf**-Taste oder mit dem Menübefehl **Bearbeiten > Löschen**
- per Maus verschieben: Ziehen und Ablegen

5.4.2 Ausgabebestandteile drucken

Über den Menübefehl

Datei > Drucken

kann man alle angezeigten oder alle markierten Ausgabebestandteile drucken, z. B.:



5.4.3 Ausgaben sichern und öffnen

Zum Speichern eines Ausgabefensters dienen die Menübefehle

Datei > Speichern unter bzw. **Datei > Speichern**

Dabei entstehen Viewer-Dateien mit der Namenserweiterung **.spv**. SPSS-Ausgaben sollten z. B. dann gespeichert werden, wenn sie (auszugsweise) in Dokumente anderer Programme eingegangen sind. Mit SPSS ist eine nachträgliche Modifikation der Ausgaben einfacher vorzunehmen (z. B. per Pivot- oder Diagrammeditor, siehe unten) als mit den Fremdprogrammen.

Das Öffnen einer Viewer-Datei gelingt per Menübefehl

Datei > Öffnen > Ausgabe bzw. **Datei > Zuletzt verwendete Dateien**

oder durch eine Mausaktion (Doppelklick bzw. Drag & Drop)

5.4.4 Objekte via Zwischenablage in andere Anwendungen übertragen

Mit der Tastenkombination **Strg+C**, mit dem Kontextmenü-Item **Kopieren** oder mit dem Menübefehl

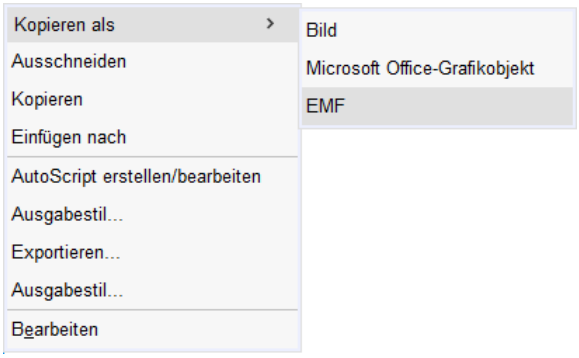
Bearbeiten > Kopieren

fordert man SPSS auf, die markierten Ausgabeobjekte (z. B. Tabellen und/oder Diagramme) in die Windows-Zwischenablage zu befördern. Dabei werden meist mehrere Varianten der voreingestellten Formatfamilie verwendet (z. B. mehrere Bitmap-Formate bei Diagrammen).

Um SPSS zu veranlassen, *ein* markiertes Ausgabeobjekt *in einer bestimmten Formatfamilie* in die Zwischenablage zu befördern, wählt man das Kontextmenü-Item **Kopieren als** oder den Menübefehl

Bearbeiten > Kopieren als

Die daraufhin bei einem Diagramm angebotenen Optionen



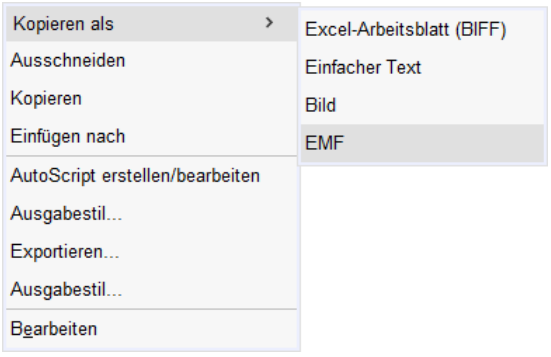
sind so zu verstehen:

- **Bild**
  In der Zwischenablage werden Bitmap-Daten abgelegt (in den Formaten PNG und JPG), sodass im Zielprogramm nur eine Pixelgrafik ankommt. Spätestens nach einer Größenänderung im Zielprogramm ist die Qualität suboptimal.
- **Microsoft Office-Grafikobjekt**
  Es sollte ein Grafikobjekt entstehen, das sich in Microsoft Office uneingeschränkt bearbeiten lässt. Je nach Office-Version klappt das mehr oder weniger gut.
- **EMF**
  In der Zwischenablage landet das *Enhanced Metafile Format* (EMF). Hier verwendet SPSS einen Vektor-basierten Grafikaufbau, sodass im Zielprogramm Größenänderungen ohne Qualitätsverlust möglich sind. Es kommt allerdings gelegentlich zu Übertragungsfehlern, die beim **Bild**-Format praktisch nie auftreten.

Bei einer *Tabelle* führt der (z. B. über das Kontextmenü-Item **Kopieren**) veranlasste Standardtransfer meist zu einem erwünschten Resultat, wenn ...

- LibreOffice Writer oder Microsoft Word das Zielprogramm ist,
- und in Writer bzw. Word beim Einfügen des Zwischenablageninhalts der voreingestellte Modus verwendet wird.

Dann erhält man eine Tabelle, die mit den Mitteln des Zielprogramms beliebig modifiziert werden kann. Von den somit weniger relevanten Optionen im **Kopieren als** - Kontextmenü zu einer *Tabelle* im SPSS-Ausgabefenster



soll hier nur **EMF** erwähnt werden. Sie sorgt dafür, dass eine Tabelle auch als vektor-basierte Grafik in die Zwischenablage gelangt.

Sind Sie bei der Entnahme eines SPSS-Objekts aus der Zwischenablage mit dem vom Zielprogramm bevorzugten Format zufrieden, dann taugt bei traditioneller Menübedienung der Befehl

Bearbeiten > Einfügen

Bei Microsoft Word© ab Version 2007 ist aus der Gruppe **Zwischenablage** im **Start**-Menüband das **Einfügen**-Symbol zu verwenden (siehe Abbildung unten). Mit der Tastenkombination **Strg+V** sollte das Einfügen im voreingestellten Format bei allen Programmen klappen.

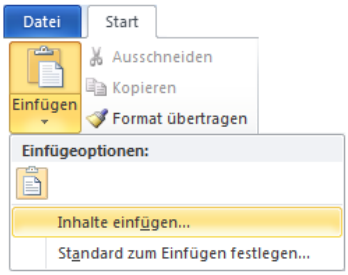
In der Regel befördert SPSS eine Tabelle oder Grafik in verschiedenen Formaten in die Windows-Zwischenablage. LibreOffice Writer und Microsoft Word bevorzugen in der Rolle des Zielprogramms ...

- von einer SPSS-Tabelle in der Windows-Zwischenablage das RTF-Format, sodass man schließlich eine Word-Tabelle erhält, die sich uneingeschränkt mit Word-Techniken editieren lässt.
- von einem SPSS-Diagramm in der Windows-Zwischenablage das Bitmap-Format, sodass man zwar von Übertragungsfehlern verschont bleibt, aber eine suboptimale Darstellungsqualität akzeptieren muss.

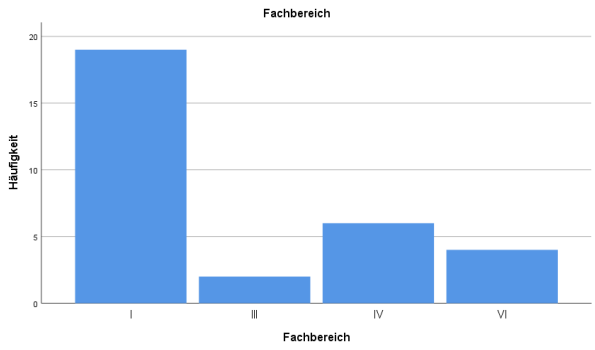
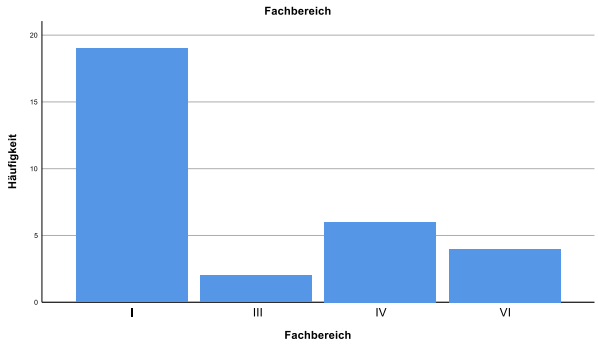
Speziell bei Diagrammen kann es sinnvoll sein, auf das vom Zielprogramm entnommene Format Einfluss zu nehmen, was traditionell über den Menübefehl

Bearbeiten > Inhalte einfügen

möglich ist und auch bei Programmen mit Menüband-Bedienung gelingt:

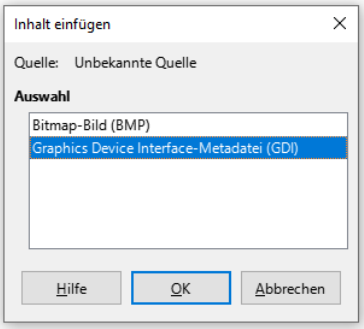


Wenn Sie beim Einfügen eines *Diagramms* das Format **Bild (Erweiterte Metadatei)** wählen, erhalten Sie in Microsoft Word eine Vektorgrafik, die im Vergleich zu der per Voreinstellung übernommenen Bitmap-Grafik eine bessere Qualität (Schärfe) besitzt, z. B.:

Übernahme im Bitmap-Format	Übernahme im Metadatei-Format
	

Zudem hat eine Vektorgrafik den Vorteil, dass ihre Qualität bei einer Größenänderung unverändert bleibt, während eine Bitmap-Grafik speziell bei einer Vergrößerung deutlich an Qualität verliert.

In LibreOffice Writer wählt man den Zwischenablageninhalt im Vektorgrafikformat über die Option **Graphics Device Interface-Metadatei (GDI)**:



Wenn sich SPSS weigert, ein Diagramm im **EMF**-Format in die Zwischenablage zu befördern, hilft es, zunächst einen Transfer als **Bild** und unmittelbar danach den Transfer als **EMF** anzufordern.

Wenn Sie beim Einfügen einer als **EMF** in die Zwischenablage kopierten SPSS-Tabelle in Word das Format **Bild (Erweiterte Metadatei)** bzw. in Writer das Format **Graphics Device Interface-Metadatei (GDI)** wählen, dann erhalten Sie in der Zielanwendung ein Grafikimplantat mit dem Originaldesign aus dem SPSS-Ausgabefenster, das Größenänderungen ohne Qualitätsverlust erlaubt.

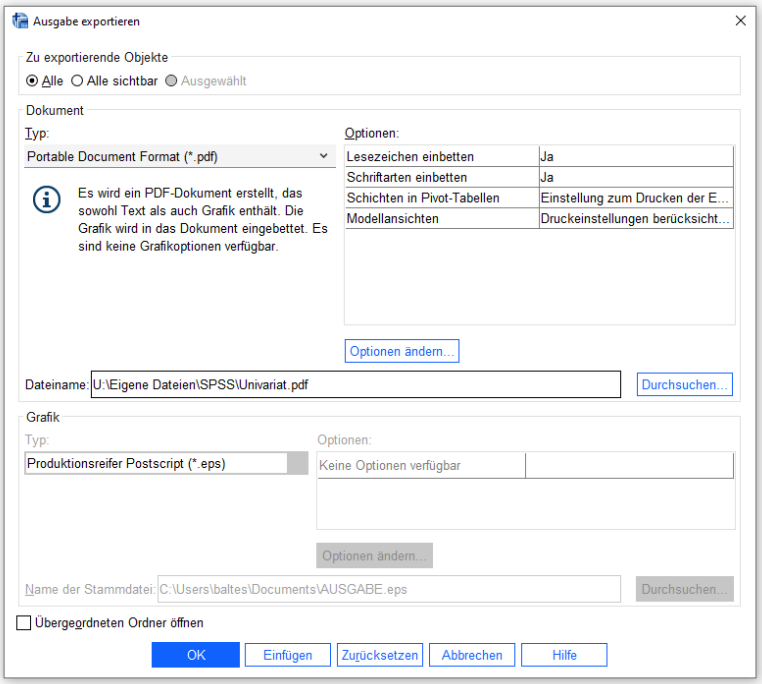
Wenn sich SPSS weigert, eine Tabelle im **EMF**-Format in die Zwischenablage zu befördern, hilft es, zunächst einen Transfer als **einfacher Text** und unmittelbar danach den Transfer als **EMF** anzufordern.

5.4.5 Ausgaben exportieren

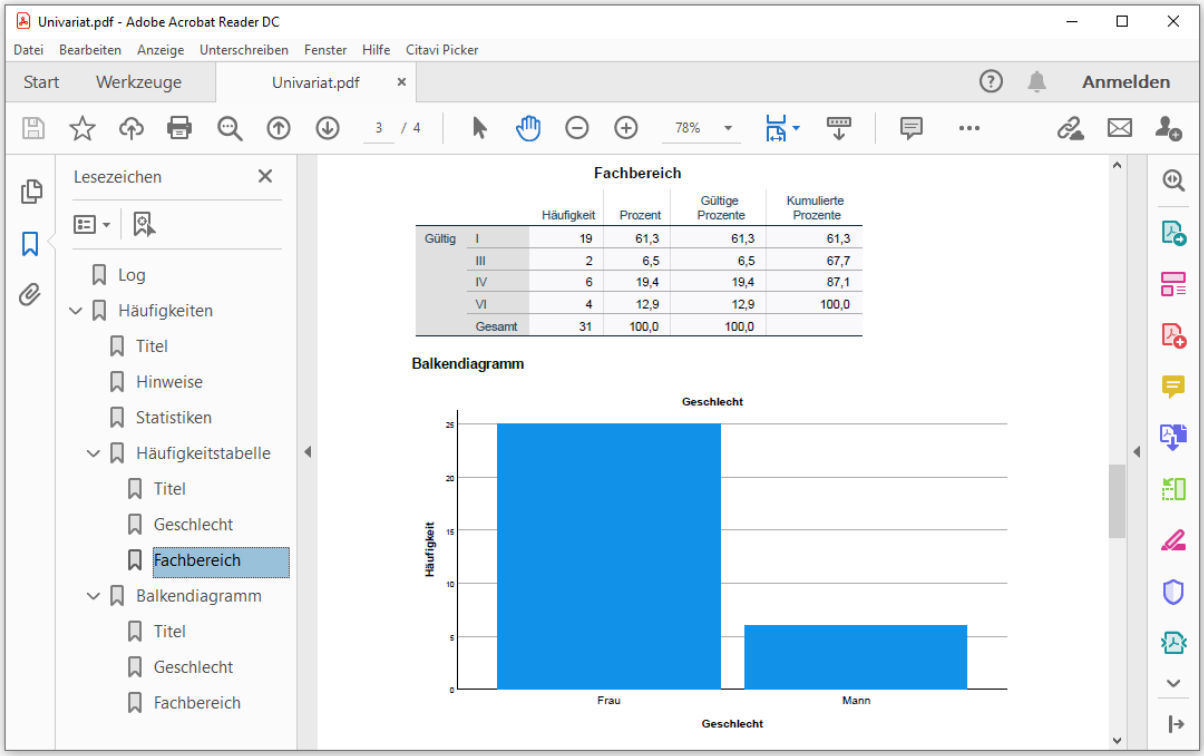
Über den Menübefehl

Datei > Exportieren...

lassen sich Ausgabebestandteile (alle, die sichtbaren oder die ausgewählten) in diversen Formaten exportieren (z. B. HTML, PDF, Microsoft Word, Microsoft PowerPoint, Text). Mit folgender Dialogbox wird z. B. das gesamte Ausgabefenster im PDF-Format exportiert:



Markiert man das Kontrollkästchen **Übergeordneten Ordner öffnen**, dann wird nach erfolgter Exportproduktion der Ordner mit dem Ergebnis vom Windows-Explorer angezeigt. Bei der PDF-Produktion entstehen Lesezeichen aus den Elementen der Ausgabefenster-Gliederungsansicht, z. B.:

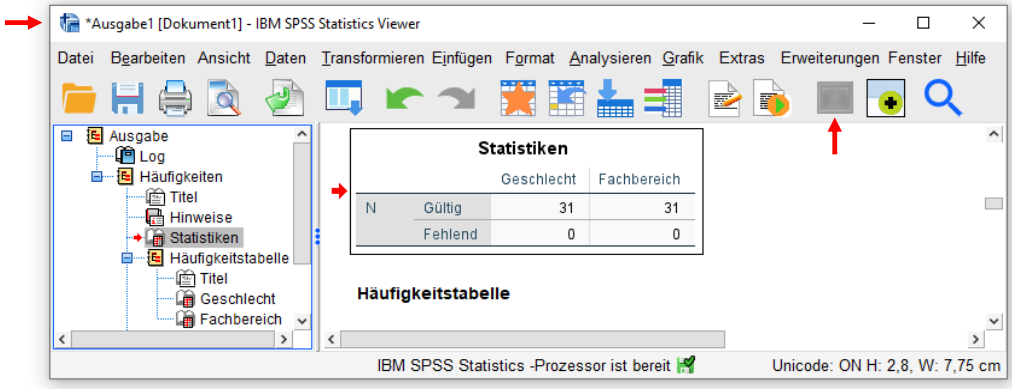


5.4.6 Mehrere Ausgabefenster verwenden

Bislang war immer von *dem* Ausgabefenster die Rede. Im Verlauf einer längeren Auswertungsarbeit kann es aber der Übersichtlichkeit dienen, zusätzliche Ausgabefenster anzufordern. Dazu taugt der Menübefehl:

Datei > Neu > Ausgabe

Sind mehrere Ausgabefenster vorhanden, muss geregelt sein, in welchem Fenster zukünftige Ausgaben landen sollen. Daher ist stets ein *Haupt*ausgabefenster festgelegt. Es ist an einem Pluszeichen im Symbol zum Systemmenü (siehe linken Rand der Titelzeile) sowie an einem *passiven* Hauptfenster-Schalter in seiner Symbolleiste zu erkennen, z. B.:



Dieser Schalter dient nämlich im aktiven Zustand  dazu, ein Ausgabefenster zum *Haupt*ausgabefenster zu ernennen.

Um ein bestimmtes Ausgabefenster in den Vordergrund zu holen, können Sie es anklicken oder über das **Fenster**-Menü eines beliebigen SPSS-Fensters auswählen. Es wird dabei (abweichend vom Verhalten eines Datenfensters) *nicht* automatisch zum Hauptausgabefenster. Jedes Ausgabefenster kann auf windows-übliche Weise geschlossen werden, z. B. über den Menübefehl:

Datei > Schließen

5.4.7 Übung

1) Markieren Sie im Ausgabefenster den zuerst angeforderten Ausgabeblock (Häufigkeitsanalysen für GESCHL und FB mit Balkendiagrammen), und löschen Sie ihn per **Entf**-Taste.
2) Öffnen Sie erneut die Dialogbox zur Häufigkeitsanalyse. Statt den zugehörigen Menübefehl zu wiederholen, können Sie mit dem Symbol  eine Liste der zuletzt benutzten Dialogboxen aufrufen und daraus per Mausklick den Eintrag **Häufigkeiten** wählen. Die Dialogbox ist noch im selben Zustand, den Sie eben verlassen haben. Dies gilt generell in SPSS, sodass Sie bei der sukzessiven Modifikation einer Anforderung innerhalb einer Sitzung jeweils auf dem letzten Stand weitermachen können. SPSS verwaltet ggf. sogar für mehrere Datenblätter jeweils eine separate Liste mit Dialogfeldkonfigurationen.
3) Wählen Sie in der Subdialogbox **Diagramme** durch **Prozentwerte** beschriftete Balkendiagramme.

5.5 Verteilungsanalysen für metrische Merkmale

Bei metrischen Merkmalen mit hoher Messgenauigkeit und infolgedessen annähernd stetiger Verteilung treten in der Stichprobe zahlreiche verschiedene Werte auf, und eine Häufigkeitstabelle ist zur Beschreibung der Verteilung wenig geeignet. Im Extremfall haben alle aufgetretenen Werte die Häufigkeit eins, sodass eine große, unübersichtliche und wenig informative Tabelle entstehen würde. Allerdings kann ein metrisches Merkmal auch eine *diskrete* Verteilung mit relativ wenigen Werten besitzen, sodass eine Häufigkeitstabelle angemessen ist. Das passiert insbesondere bei *Zählvariablen* (z. B. Anzahl der im letzten Jahr gelesenen Bücher, Anzahl der Punkte in der Flensburger Verkehrssünderkartei).[1]

Durch statistische Kennwerte für die zentrale Tendenz, die Dispersion und die Gestalt können bei metrischen Merkmalen die wichtigsten Aspekte der Verteilung kompakt beschrieben werden.

Zur grafischen Darstellung verwendet man bei approximativ stetiger Verteilung ein Histogramm, das für *Intervalle* geeigneter Breite die Häufigkeit durch einen Balken darstellt, wobei die Balken lückenlos aneinander grenzen. Ist ein metrisches Merkmal diskret verteilt, kommt eher ein Balkendiagramm in Frage, das für jeden aufgetretenen Wert die (absolute oder relative) Häufigkeit durch einen Balken darstellt.

---

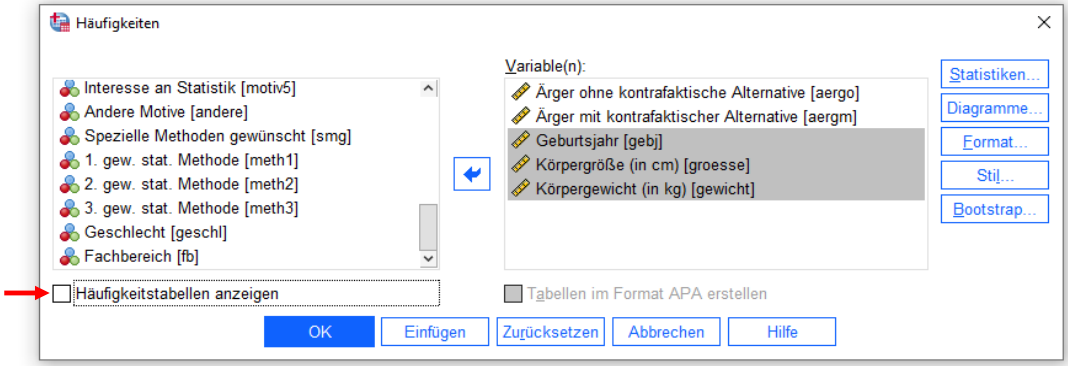
[1] Für kontinuierliche Variablen konstruierte Analyseverfahren arbeiten auch mit diskret verteilten Variablen akzeptabel, wenn die Anzahl der verschiedenen Werte nicht zu klein (≥ 5) und die Verteilung der Residuen nicht zu schief ist. Man sollte Modelle, die normalverteile Residuen benötigen, aber nicht überstrapazieren, sondern ggf. auf generalisierte lineare Modelle (z. B. für Zählvariablen) ausweichen (siehe z. B. Agresti 2007; Baltes-Götz 2016b).

Die mit dem Menübefehl

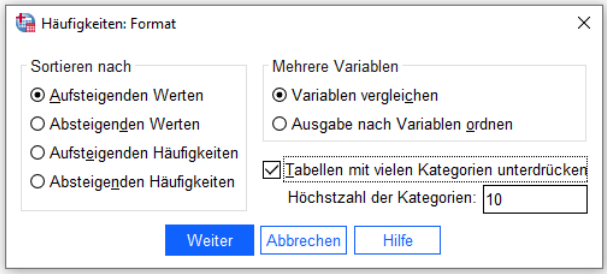
Analysieren > Deskriptive Statistiken > Häufigkeiten

zu startende und schon in Abschnitt 5.3 verwendete Prozedur für univariate Verteilungsanalysen ist auch für metrische Variablen geeignet. Um die Prozedur davon abzuhalten, (übergroße) Häufigkeitstabellen zu erstellen, haben Sie zwei Möglichkeiten:

- Entfernen Sie die per Voreinstellung vorhandene Markierung des für Häufigkeitstabellen zuständigen Kontrollkästchens in der Dialogbox **Häufigkeiten**:



- Im Subdialog **Format** lassen sich Häufigkeitstabellen unterdrücken, die eine vorgegebene Anzahl unterschiedlicher Werte überschreiten:



5.5.1 Zentrale Tendenz

Ein statistischer Kennwert für die Lage (die zentrale Tendenz) einer Verteilung soll die *typische Ausprägung* liefern. Die SPSS-Prozedur zur Häufigkeitsanalyse kann folgende Kennwerte für die zentrale Tendenz berechnen:

- **Modus**
  Den Wert mit der größten Häufigkeit zu bestimmen, ist bei *jedem* Messniveau möglich, also auch bei metrischen Variablen mit diskreter Verteilung. Tritt die maximale Häufigkeit bei *mehreren* Werten auf, nennt SPSS den kleinsten Wert aus dieser Menge.

- **Median**
  Wenn ein Merkmal mindestens ordinales Messniveau besitzt (also insbesondere bei einem metrischen Merkmal), dann ist der als *Median* (abgekürzt: *Mdn*) bezeichnete Wert von Interesse, den 50% der Beobachtungswerte *nicht* übertreffen. Alternative Bezeichnungen sind:
  - 50. Perzentil (P_{50})
  - 2. Quartil (Q_2)

  Wenn mit

$$x_{(1)}, x_{(2)}, x_{(3)}, .., x_{(n)}$$

die geordneten Beobachtungswerte der Variablen X in einer Stichprobe mit dem Umfang *n* bezeichnet werden, dann ist die Stichprobenschätzung des Medians folgendermaßen definiert:

$$\text{Mdn} := \begin{cases} x_{((n+1)/2)} & \text{für ungerades } n \\ \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{für gerades } n \end{cases}$$

Wenn die Variablenausprägungen als Intervallmittelpunkte zu interpretieren sind (gruppierte Daten), dann ist eine aufwändigere Berechnung des Medians anzuwenden (siehe Abschnitt 5.5.5). Die Interpretationen der Variablenausprägungen als Intervallmittelpunkte ist dann plausibel, wenn ein kontinuierliches Merkmal durch eine Variable mit wenigen Ausprägungen erfasst wird.

- **Arithmetisches Mittel**
  Das arithmetische Mittel von Stichprobenbeobachtungen x_1, x_2, x_3, ..., x_n mit der Definition

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

  hat folgende Eigenschaften:
  - Es minimiert die Summe der quadrierten Abweichungen von allen Werten x_i:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \min$$

  - Die Summe seiner Abweichungen von allen Werten x_i ist gleich 0:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

  Der Grenzwert des arithmetischen Mittels bei wachsender Stichprobengröße ist der Erwartungswert der Populationsverteilung. Für den Erwartungswert existieren viele mathematische Gesetze, und das häufig verwendete lineare Modell strebt nach einer guten Erklärung für den bedingten Erwartungswert gegeben eine Prädiktorwertekombination.

Empfehlungen zur Beurteilung der zentralen Tendenz (Lage) bei metrischen Merkmalen:

- Bei metrischen Merkmalen mit symmetrischer Verteilung sind der Erwartungswert und der Populationsmedian identisch, und das arithmetische Mittel ist als beste Stichprobenschätzung für den Erwartungswert zu bevorzugen.
- Bei metrischen Merkmalen mit schiefer (asymmetrischer) Verteilung ist der Populationsmedian als Kennwert für die zentrale Tendenz (als typischer Wert) besser geeignet als der Erwartungswert. In dieser Situation sollte man aus den Stichprobendaten das arithmetische Mittel *und* eine Schätzung für den Median bestimmen.

Auf Ausreißer (extreme, ungewöhnliche Werte, siehe Abschnitt 9.1) reagiert das arithmetische Mittel empfindlich, während der Stichprobenmedian relativ robust ist.

Beim Mittelwert und beim Median sind auch die *Vertrauensintervalle* von Interesse, die leider von der SPSS-Prozedur FREQUENCIES zur Häufigkeitsanalyse *nicht* geliefert werden. In Kapitel 9 lernen Sie die SPSS-Prozedur EXAMINE kennen, die auf die Verteilungsanalyse für metrische Variablen spezialisiert ist und auch das Vertrauensintervall für den Mittelwert anbietet (siehe Abschnitt 9.6.1). Per Bootstrapping lässt sich auch für den Median ein Vertrauensintervall bestimmen (siehe Abschnitt 9.6.2).

5.5.2 Streuung

Um das Streuungsverhalten eines metrischen Merkmals zu quantifizieren, kommen die folgenden Statistiken in Frage:

- **Spannweite**
  Dies ist schlicht die Differenz zwischen dem größten und kleinsten Wert in der Stichprobe. Weil nur wenig Information eingeht, ist die Spannweite selten ein gutes Streuungsmaß.

- **Interquartilsabstand**
  Der Interquartilsabstand (Interquartilsbereich, engl.: *interquartile range*, abgekürzt: IQR) ist definiert durch die Differenz zwischen dem 3. und dem 1. Quartil bzw. zwischen dem 75. und dem 25. Perzentil:[1]

  $$IQR := Q_3 - Q_1 = P_{75} - P_{25}$$

  Während der Median (das zweite Quartil) 50% aller Fälle links von sich lässt, sind es beim ersten bzw. beim dritten Quartil 25% bzw. 75%. Die Schätzung von Q_1 und Q_3 erfolgt analog zur Schätzung des Medians (siehe oben). Bei dem in Kapitel 9 behandelten Boxplot zur Verteilungsbeschreibung und zur Diagnose von Ausreißern wird der Interquartilsabstand zur Beurteilung der Dispersion verwendet.

- **Varianz**
  Auf der Populationsebene ist die Varianz einer Variablen *X* definiert durch die erwartete (mittlere) quadrierte Abweichung vom Erwartungswert E(*X*):

  $$Var(X) := E[(X - E(X))^2]$$

  Eine erwartungstreue Schätzung aufgrund von Stichprobendaten gelingt durch die folgende Statistik:

  $$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  Eine Normalverteilung wird durch ihren Erwartungswert (erstes Moment) und ihre Varianz (zweites Moment) vollständig beschrieben, was die Bedeutung der Varianz als Verteilungsparameter unterstreicht. Die Varianz von abhängigen Variablen in Modellen lässt sich in Bestandteile zerlegen (z. B. aufgeklärter Varianzanteil und Residualvarianz).
  Wenn der Erwartungswert einer Verteilung als typischer Vertreter taugt, dann ist die Varianz am besten dazu geeignet, die Dispersion einer Verteilung oder die Erklärungsleistung eines Modells zu quantifizieren.

- **Standardabweichung**
  Die Standardabweichung ist definiert durch die Wurzel aus der Varianz. Man erhält Werte in der Maßeinheit des betrachteten Merkmals, was die Interpretation im Vergleich zur Varianz erleichtert. Wie der Name suggeriert, gibt die Standardabweichung eines Merkmals die typische Abweichung einer zufälligen Realisation vom Erwartungswert an. Bei einer Normalverteilung mit dem Erwartungswert μ und der Standardabweichung σ befinden sich ...

[1] Die Sequenz aus einem Doppelpunkt und einem Gleichheitszeichen (:=) bedeutet, dass keine Identität behauptet, sondern der Begriff IQR durch den Ausdruck auf der rechten Seite definiert wird.

- im Intervall [μ - σ, μ + σ] ca. 68% der Verteilungsmasse
- im Intervall [μ - 2σ, μ + 2σ] ca. 95% der Verteilungsmasse

Empfehlungen zur Beurteilung der Dispersion bei metrischen Merkmalen:

- Bei einer symmetrischen Verteilung ist der Erwartungswert als typischer Vertreter geeignet, und dann charakterisiert man die Dispersion durch die Varianz oder die Standardabweichung. Beide Statistiken enthalten dieselbe Information und werden je nach Fragestellung eingesetzt.
- Bei einer schiefen Verteilung ist der Interquartilsabstand gegenüber der Varianz bzw. Standardabweichung zu bevorzugen.

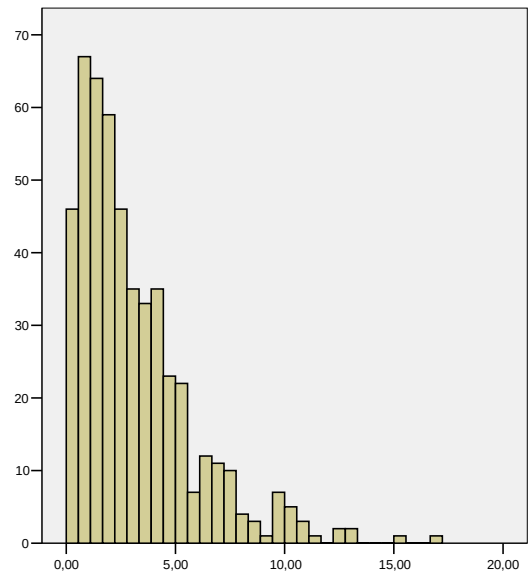
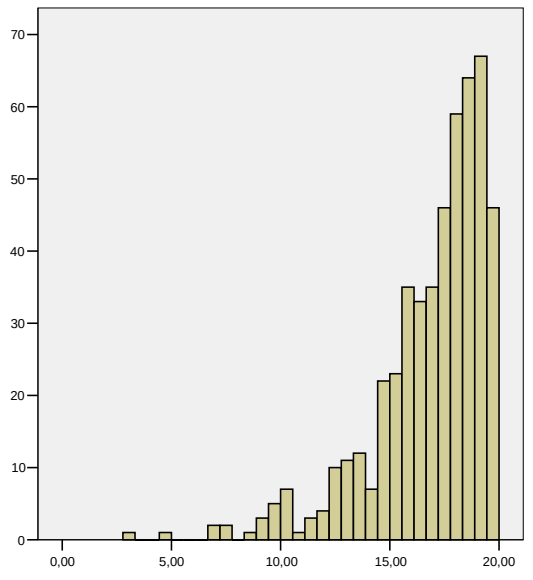
Auf Ausreißer (extreme, ungewöhnliche Werte, siehe Abschnitt 9.1) reagieren Varianz und Standardabweichung sehr empfindlich, während der Interquartilsabstand relativ robust ist.

5.5.3 Schiefe und Wölbung

In diesem Abschnitt werden Statistiken vorgestellt, welche die *Form* einer Verteilung beschreiben. Weil nur bei metrischen Merkmalen die Verteilungsform unter zulässigen Transformationen erhalten bleibt, sind die Statistiken zur Beschreibung der Verteilungsform nur bei metrischen Merkmalen sinnvoll.

5.5.3.1 Schiefe

Bei einem symmetrisch verteilten Merkmal hat die Schiefestatistik den Wert null. Sie wird positiv bei einem linkssteil (rechtsschief) verteilten Merkmal, wenn also die Verteilungsmasse am linken Rand konzentriert ist, und negativ bei einem rechtssteil (linksschief) verteilten Merkmal, z. B.:

Positiv schiefe Verteilung	Negativ schiefe Verteilung
	

Zur Stichprobenschiefe schätzt SPSS auch den zugehörigen Standardfehler, also die Standardabweichung der aus unendlich vielen Stichproben zu gewinnenden Verteilung von Schiefeschätzungen. Mit Hilfe des Standardfehlers kann man Signifikanztests zur Populationsschiefe durchführen, die allerdings nur approximativ (bei großen Stichproben) gültig und folglich bei

kleinen Stichproben mit Vorsicht zu genießen sind. Ihr Vorzug gegenüber den später vorzustellenden Omnibus-Normalverteilungs-Anpassungstests (siehe Kapitel 9) besteht darin, dass sie gezielt auf Verletzungen der Verteilungs*symmetrie* ansprechen.

Bei einem α-Fehlerrisiko von 5% ist die *ungerichtete* Nullhypothese, dass die Schiefe in der Population gleich null sei, zu verwerfen, wenn sich in der Stichprobe für den Quotienten aus dem Betrag und dem Standardfehler der Schiefe ergibt:

$$\frac{|\text{Schiefe}|}{\text{SF(Schiefe)}} > 1{,}96$$

Beim Wert 1,96 handelt es sich um das 97,5 - Perzentil der Standardnormalverteilung.[1]

Der Test zum *gerichteten* Hypothesenpaar:

$$H_0: \text{Schiefe} \geq 0 \quad \text{versus} \quad H_1: \text{Schiefe} < 0$$

entscheidet sich beim selben α-Fehlerrisiko gegen seine Nullhypothese, wenn der Quotient aus der Schiefe und ihrem Standardfehler das 5. Perzentil der Standardnormalverteilung unterbietet:

$$\frac{\text{Schiefe}}{\text{SF(Schiefe)}} < -1{,}65$$

Analog lässt sich auch die gerichtete Hypothese mit umgekehrtem Vorzeichen prüfen.[2]

5.5.3.2 Wölbung (Kurtosis)

Die Wölbung (synonym: *Kurtosis*) ist bei einer Normalverteilung gleich 0.[3] Sie wird ...

- *positiv*, wenn sich im Vergleich zur Normalverteilung mit demselben Erwartungswert und derselben Varianz *mehr* Verteilungsmasse an den Rändern befindet. Damit dieselbe Varianz resultiert, müssen gleichzeitig in der Verteilungsmitte viele *kleine* Varianzbeiträge auftreten. Das „Gedränge um den Mittelwert“ sorgt für einen schmalgipfligen Eindruck.
- *negativ*, wenn sich im Vergleich zur Normalverteilung mit demselben Erwartungswert und derselben Varianz *weniger* Verteilungsmasse an den Rändern befindet. Damit dieselbe Varianz resultiert, dürfen gleichzeitig nur wenige kleine Varianzbeiträge auftreten, sodass ein breitgipfliger Eindruck entsteht.

Mit Hilfe des zugehörigen Standardfehlers können analog zum Vorgehen bei der Schiefe-Statistik (siehe Abschnitt 5.5.3.1) approximativ (bei großen Stichproben) gültige Signifikanztests zur Wölbung in der Population durchgeführt werden.

---

[1] Als *Standardnormalverteilung* bezeichnet man die Normalverteilung mit dem Erwartungswert 0 und der Varianz 1.

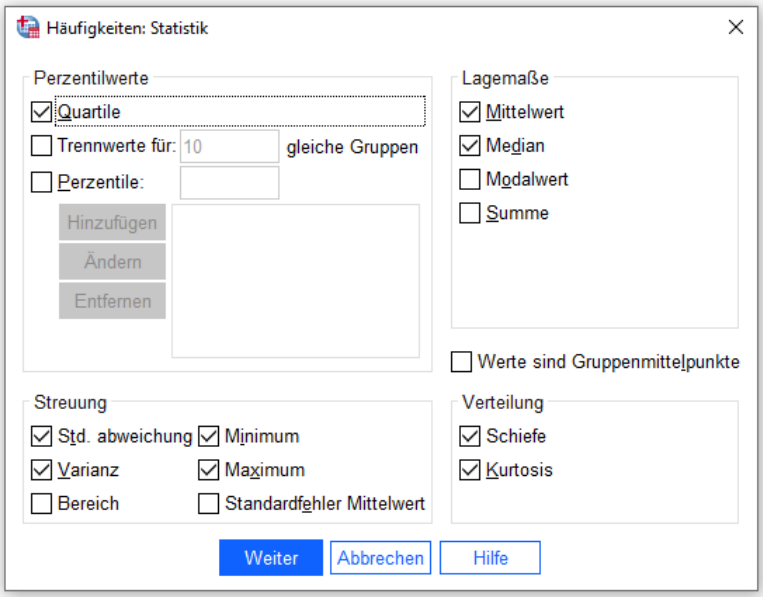
[2] Wer in seinem Gedächtnis nicht mehr genügend Kenntnisse zur Inferenzstatistik reaktivieren konnte, der sei auf den Abschnitt 8.1 vertröstet.

[3] Andere Statistikprogramme und auch Lehrbücher verwenden eine alternative Definition der Wölbung, die zum Wert 3 für die Normalverteilung führt. Bei Aussagen zur Wölbung (in der Statistikliteratur oder in einer Programmausgabe) muss also die zugrunde liegende Definition berücksichtigt werden.

5.5.3.3 Übung

Im Demonstrationsprojekt sind die beiden Ärgermessungen sowie die Merkmale Geburtsjahr, Größe und Gewicht (approximativ) intervallskaliert. Öffnen Sie erneut die Dialogbox zur Häufigkeitsanalyse. Statt den zugehörigen Menübefehl zu wiederholen, können Sie mit dem Symbol  eine Liste der zuletzt benutzten Dialogboxen aufrufen und daraus per Mausklick den Eintrag **Häufigkeiten** wählen. Lassen Sie sich für die genannten Variablen ausgeben:

- Histogramme mit eingezeichneter Normalverteilungsdichte
  Um diesen Wunsch zu artikulieren, müssen Sie den Subdialog **Diagramme** öffnen.
- Keine Häufigkeitstabellen
  Das für Häufigkeitstabellen zuständige Kontrollkästchen in der Dialogbox **Häufigkeiten** ist per Voreinstellung markiert. Sie müssen also die Markierung entfernen.
- Folgende Statistiken: die Quartile (also das 25., das 50. und das 75. Perzentil), Mittelwert, Median, Standardabweichung, Varianz, Minimum, Maximum, Schiefe, Kurtosis
  Zur Auswahl der gewünschten **Statistiken** müssen Sie die zuständige Subdialogbox per Mausklick öffnen:



Die Vergleiche mit der Normalverteilung über das Histogramm sowie die Statistiken Schiefe und Kurtosis erfolgen hier aus purem Interesse an den Verteilungen der betrachteten Variablen, ohne dabei an die Verteilungs*voraussetzungen* irgendwelcher Hypothesentests zu denken. Diese Voraussetzungen beziehen sich in der Regel nicht auf die momentan von uns analysierten *unbedingten* Verteilungen von Variablen, sondern auf die Verteilungen der *Residuen* eines bestimmten statistischen Modells. Genauere Aussagen sind nur im Zusammenhang mit konkreten Testverfahren möglich.

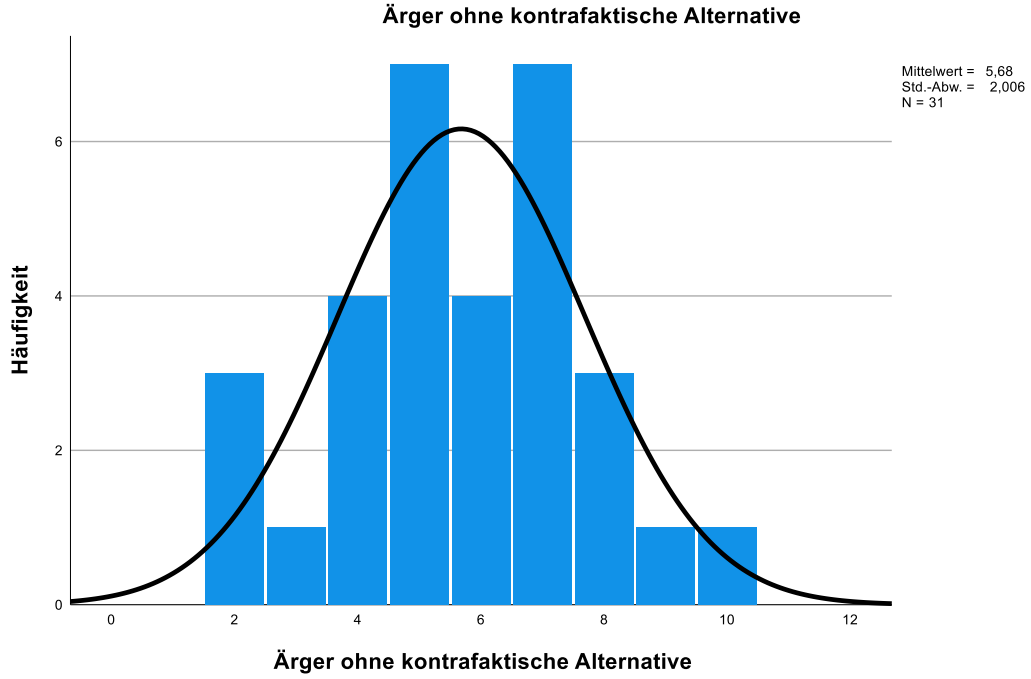
5.5.4 Diskussion ausgewählter Ergebnisse

Wie die Statistiken Minimum und Maximum zeigen, liegen keine irregulären Werte vor:

Statistiken

		Ärger ohne kontrafaktische Alternative	Ärger mit kontrafaktischer Alternative	Geburtsjahr	Körpergröße (in cm)	Körpergewicht (in kg)
N	Gültig	31	31	31	31	31
	Fehlend	0	0	0	0	0
Mittelwert		5,68	7,68	1968,94	172,81	63,484
Median		6,00	8,00	1969,00	174,00	60,000
Std.-Abweichung		2,006	2,271	3,214	8,288	10,4940
Varianz		4,026	5,159	10,329	68,695	110,125
Schiefe		-,080	-1,451	,017	,448	1,265
Standardfehler der Schiefe		,421	,421	,421	,421	,421
Kurtosis		-,277	2,013	,241	-,166	1,889
Standardfehler der Kurtosis		,821	,821	,821	,821	,821
Minimum		2	1	1961	158	50,0
Maximum		10	10	1975	192	96,0
Perzentile	25	4,00	6,00	1967,00	165,00	55,000
	50	6,00	8,00	1969,00	174,00	60,000
	75	7,00	9,00	1970,00	178,00	70,000

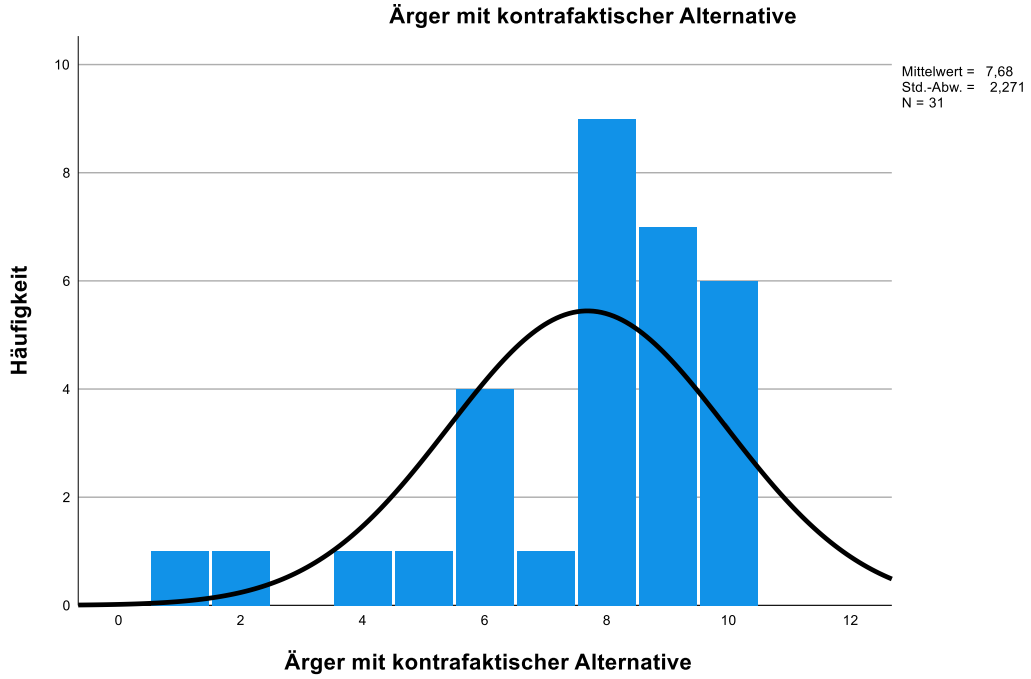
Die Verteilung der Ärgermessung in der Situation *ohne* kontrafaktische Alternative (AERGO) macht einen ziemlich „normalen“ Eindruck:



Die Verteilungskennwerte Schiefe (= -0,08) und Kurtosis (= -0,277) sind nach den oben angegebenen Tests nicht signifikant von null verschieden.

Wir sind nun sehr gespannt auf die Verteilung der Ärgermessung in der Situation *mit* kontrafaktischer Alternative (AERGM), weil sich ein KFA-Effekt hier deutlich abzeichnen sollte. Es ist generell zu empfehlen, sich anhand von Diagrammen und deskriptiven Statistiken ein präzises Bild von der Effektlage zu verschaffen, statt einem Signifikanztest blind zu vertrauen, der eventuell durch technische Fehler verfälscht ist. Im Vergleich zur relativ symmetrischen Verteilung

von AERGO um den Mittelwert 5,68 ist die AERGM-Verteilung deutlich nach rechts verschoben (Mittelwert 7,68) und „deformiert“:



Wir sehen einen Anstieg des Ärgermittelwerts um ca. 20° (bei Rückübersetzung in die Celsius-Skala des Fragebogens). Außerdem ist die AERGM-Verteilung am rechten Rand konzentriert und deutlich verschieden von einer Normalverteilung, was sich auch in signifikanten Ergebnissen der Tests zu Schiefe und Kurtosis zeigt:

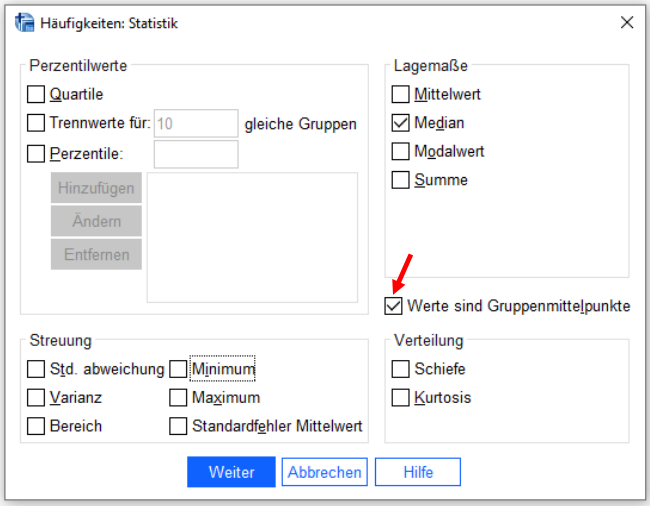
$$\frac{|\,\text{Schiefe}\,|}{\text{SF(Schiefe)}} = \frac{1{,}45}{0{,}42} = 3{,}45 > 1{,}96$$

$$\frac{|\text{Kurtosis}|}{\text{SF(Kurtosis)}} = \frac{2{,}01}{0{,}82} = 2{,}45 > 1{,}96$$

Hier sind *zwei*seitige Tests durchzuführen, weil vor Beginn der Datenerhebung keine gerichteten Hypothesen vorlagen. Wir haben zwar eine explizite Hypothese über die Richtung des KFA-Effekts (vgl. Abschnitt 2.3.2), doch muss die Verschiebung einer Verteilung nach rechts nicht zwangsläufig zu einer negativen Schiefe führen (siehe Abbildung in Abschnitt 2.3.2). Offenbar war aber der KFA-Effekt so stark, dass er die Ärgerverteilung an die „Decke“ geschoben und damit rechtssteil (negativ schief) gemacht hat.

5.5.5 Median und andere Perzentile aus gruppierten Daten

Man kann SPSS im **Statistiken**-Dialog der Häufigkeitsanalyse anweisen, von *gruppierten* Daten auszugehen, also die für eine Variable vorhandenen Werte als Mittelpunkte von Intervallen aufzufassen:



Diese Situation liegt z. B. vor, wenn für alle Fälle mit einem Alter zwischen 30 und 40 Jahren der Wert 35 eingetragen wurde. In dieser Lage ist ein relativ aufwändiges Verfahren angemessen, um den Median und andere Perzentile für die *ungruppierten* Daten zu schätzen. Den etwas anstrengenden Rest des Abschnitts sollte nur lesen, wer die Berechnungsvorschrift für gruppierte Daten mit Intervallmittelpunkten als Werten in nächster Zeit anwenden möchte.

Für die k der Größe nach aufsteigend geordneten Werte (Intervallmittelpunkte) $x_1 < x_2 < ... < x_k$ in der Stichprobe mit den zugehörigen Häufigkeiten $n_1, n_2, ... n_k$ werden kumulative Häufigkeiten c_j folgendermaßen definiert:

$$c_j := \sum_{\upsilon=1}^{j-1} n_\upsilon + \frac{n_j}{2}$$

Hinter dieser Formel steckt die Annahme, dass von den n_j Fällen mit dem Wert (Intervallmittelpunkt) x_j gerade 50% einen wahren Wert kleiner oder gleich x_j und 50% einen wahren Wert größer als x_j haben. Das folgende Zahlenbeispiel illustriert die Definition:

x_j	n_j	c_j
1	4	2 = 0+4/2
2	6	7 = 4 + 6/2
3	2	11 = 10 + 2/2

Aus den Punkten (x_j, c_j) wird durch lineare Interpolationen eine Funktion mit geschätzten kumulativen Häufigkeiten für beliebige Werte konstruiert, und der Schnittpunkt dieser Funktion mit der zur X-Achse parallelen Geraden in der Höhe $\frac{n}{2}$ ist der geschätzte Median.[1]

Zur Berechnung des so definierten Stichprobenmedians ist der Beobachtungswert (der Intervallmittelpunkt) x_m zu bestimmen mit:

$$c_{m-1} \leq \frac{n}{2} < c_m$$

Um den Median zu ermitteln, ist zu x_{m-1} noch ein Reststück r aus dem Intervall $(x_m - x_{m-1})$ zu addieren:

[1] http://www-01.ibm.com/support/docview.wss?uid=swg21475408

$$\text{Mdn} = x_{m-1} + r$$

Für *r* gilt wegen der linearen Interpolation:

$$\frac{\frac{n}{2} - c_{m-1}}{c_m - c_{m-1}} = \frac{r}{x_m - x_{m-1}}$$

Eine Termumformung ergibt die folgende Bestimmungsgleichung für *r*:

$$r = \frac{\frac{n}{2} - c_{m-1}}{c_m - c_{m-1}} (x_m - x_{m-1})$$

Im Zahlenbeispiel resultieren $x_m = 2$ und:

$$r = \frac{\frac{12}{2} - 2}{7 - 2} (2 - 1) = 0{,}8$$

$$\text{Mdn} = 1 + 0{,}8 = 1{,}8$$

Genau diese Median-Schätzung liefert SPSS für die als gruppiert deklarierten Beispieldaten,

Statistiken

x

N	Gültig	12
	Fehlend	0
Median		1,8000 [a]

a. Aus gruppierten Daten berechnet

während die in Abschnitt 5.5.1 mitgeteilte Definition bzw. Berechnungsvorschrift zum Wert 2 führt.

5.6 Übung

Führen Sie die restlichen Verteilungs- bzw. Fehleranalysen zu unserem Forschungsprojekt durch. Die mehrfach benötigte **Häufigkeiten**-Dialogbox sollte jeweils über den Schalter **Zurücksetzen** von alten Einstellungen (auch in den Subdialogboxen) befreit werden.

1) Lassen Sie sich für die LOT-Variablen ausgeben:
   - Häufigkeitstabellen
   - keine Diagramme
   - folgende Statistiken: Mittelwert, Median, Modalwert, Standardabweichung, Varianz, Minimum, Maximum

2) Lassen Sie sich für die Variablen MOTIV1 bis MOTIV5, ANDERE, SMG und METH1 bis METH3 ausgeben:
   - Häufigkeitstabellen
   - keine Diagramme
   - keine Statistiken

Bei den Variablen MOTIV1 bis MOTIV5, ANDERE, SMG und METH1 bis METH3 ist zu beachten, dass die Behandlung fehlender Werte durch Techniken der Datentransformation noch aussteht (vgl. Abschnitt 7.5).

3) Kontrollieren Sie bei allen Variablen, ob unzulässige Werte vorliegen.

5.7 Suche nach Daten

Die Fehleranalyse liefert nur einen „Treffer“. In der Häufigkeitstabelle zur Variablen LOT10 entdecken wir den verbotenen Wert 0:

lot10

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0	1	3,2	3,2	3,2
	1	4	12,9	12,9	16,1
	2	10	32,3	32,3	48,4
	3	9	29,0	29,0	77,4
	4	7	22,6	22,6	100,0
	Gesamt	31	100,0	100,0	

Diese Fehlerquote kann als erfreulich niedrig eingestuft werden.

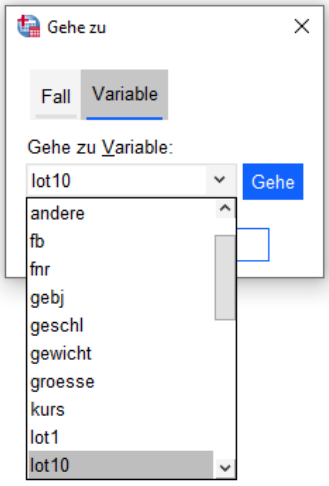
Nun möchten wir natürlich wissen, bei welchem Fall dieser Wert auftritt, um eine Korrektur vornehmen zu können. Der betroffene Fall ist leicht zu ermitteln:

- Holen Sie nötigenfalls das Datenfenster mit der Arbeitsdatei in den Vordergrund.
- Markieren Sie in der **Datenansicht** die Variable LOT10 durch einen Mausklick auf ihren Namen in der Spaltenbeschriftungszone.

In unserem kleinen Datensatz ist eine einzelne Variable leicht zu lokalisieren. SPSS eignet sich aber auch für Projekte mit Tausenden von Variablen und stellt nach dem Menübefehl

Bearbeiten > Gehe zu Variable...

oder einem Klick auf den Symbolleistenschalter  mit der Dialogbox **Gehe zu** eine nützliche Navigationshilfe zur Verfügung, z. B.:



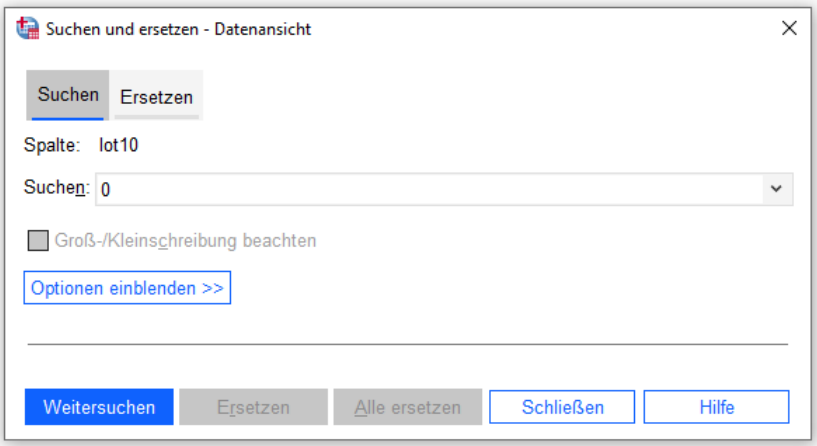
Passend zum bereits eingetippten Namensanfang wird eine Liste mit allen Variablennamen geöffnet und die erste kompatible Zeile markiert. Nach dem Quittieren einer Auswahl mit dem Schalter **Gehe** ist im Datenfenster die zugehörige Variable markiert.

Im Datenfenster mit der markierten Variablen LOT10 findet man leicht zu den Fällen mit einem interessierenden Wert:

- Klicken Sie auf das Symbol  , oder wählen Sie den Menübefehl:

  Bearbeiten > Suchen...

  Dann erscheint die folgende Dialogbox:



- Tragen Sie den zu suchenden Wert ein, und klicken Sie auf den Schalter **Weitersuchen**. Für die Suche nach SYSMIS ist ein Punkt einzutragen.
- Daraufhin markiert SPSS die erste Trefferzelle, und Sie kennen den Fall mit fehlerhaftem LOT10-Wert: Es ist zufällig der erste Fall (FNR = 1), dessen ausgefüllter Fragebogen im Manuskript wiedergegeben ist (siehe Seite 59), sodass Sie den korrekten Wert 3 ablesen und im Datenfenster eintragen können. Nach dieser Datenkorrektur sollten Sie die Arbeitsdatei sichern und damit die SPSS-Datendatei **kfar.sav** auf den neuen Stand bringen.

Ist mangels entsprechender Dokumente keine Korrektur möglich, muss ein irregulärer Wert neutralisiert werden. Man kann ihn löschen oder als MD-Indikator deklarieren. Die zweite Methode ist zu bevorzugen, wenn viele Fälle betroffen sind, weil nur *eine* Änderung im Datenlexikon erforderlich ist.

5.8 Populationsanteil einer Kategorie und Vertrauensintervall

Liegt eine Zufallsstichprobe aus einer Population vor, dann beschränkt man sich nicht darauf, die Stichprobenverteilungen der untersuchten Merkmale zu beschreiben, sondern man strebt nach Aussagen über die Verteilungen in der *Population*. Bei einem nominalskalierten Merkmal interessieren als Populationsparameter die Anteile (Wahrscheinlichkeiten) der einzelnen Kategorien. Aus einer Stichprobe mit Führungskräften möchte man z. B. bei einer Analyse des Merkmals Geschlecht möglichst präzise Informationen über den Frauenanteil in der Population der Führungskräfte gewinnen.

Wir untersuchen anschließend die Variable GESCHL aus unserem Kursprojekt, deren Häufigkeitstabelle wir schon in Abschnitt 5.3 angefordert und diskutiert haben:

Geschlecht

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Frau	25	80,6	80,6	80,6
	Mann	6	19,4	19,4	100,0
	Gesamt	31	100,0	100,0	

Wir gehen davon aus, dass eine Zufallsstichprobe aus der Population der Interessenten für die ZIMK-Lehrveranstaltung *Statistisches Praktikum mit SPSS* vorliegt. Für die **Wahrscheinlichkeit**, beim zufälligen Ziehen aus dieser Population eine Frau anzutreffen, liefert die Stichprobe mit der relativen Häufigkeit die folgende **Punktschätzung**:

$$\frac{25}{31} = 0{,}806 = \hat{p}$$

Als Symbol für die Wahrscheinlichkeit verwenden wir den Kleinbuchstaben *p* (Abkürzung für *probability*), und für die geschätzte Wahrscheinlichkeit setzen wir ein Dach darüber ($\hat{p}$).

Vermutlich rechnen Sie nicht damit, in einer anderen Stichprobe aus derselben Population denselben Schätzwert zu erhalten. Jede Stichprobe liefert einen anderen Frauenanteil, also eine andere Schätzung für die interessierende Wahrscheinlichkeit. Immerhin haben alle Schätzungen die interessierende Wahrscheinlichkeit als gemeinsamen Erwartungswert, d. h. sie pendeln um diesen Wert herum.

Der Auftraggeber Ihrer statistischen Analyse ist an der **Präzision** der gelieferten Anteilsschätzung interessiert. Konkret will er wissen, in welchem Intervall um die Anteilsschätzung aus der einzig vorhandenen Stichprobe die Populationswahrscheinlichkeit mit einer gewünschten Sicherheit von z. B. 95% anzunehmen ist. Bei einer Wählerbefragung möchte das beauftragende Presseorgan (und auch der Medienkonsument) wissen, in welchem Intervall um den Schätzwert der wahre Stimmenanteil für eine Partei mit einer gewünschten Sicherheit von z. B. 95% liegt. Folglich müssen Sie auch eine **Intervallschätzung** abliefern. Deren Ergebnis bezeichnet man als **Konfidenz-** oder **Vertrauensintervall**.

Bei der statistischen Sicherheit zu einem Vertrauensintervall (mit typischen Werten von 95% oder 99%) spricht man vom **Konfidenzniveau**, und das Gegenstück nennt man **Irrtumswahrscheinlichkeit** (mit typischen Werten von 5% oder 1%).

Der Begriff *Vertrauensintervall* ist im Kurs schon mehrfach ohne Erläuterung benutzt worden in der Annahme, dass Grundbegriffe in einem statistischen Praktikum als bekannt vorausgesetzt werden dürfen. Allerdings liegt erfahrungsgemäß bei manchen Kursteilnehmern die statistische Ausbildung schon länger zurück, sodass bei besonders wichtigen Begriffen neben der praktischen Anwendung auch eine wiederholende Erläuterung angemessen ist. Der aktuelle Abschnitt 5.8 soll daher Ihr Wissen über die Intervallschätzung auffrischen, wobei wir als Beispiel die Intervallschätzung zu einer Proportion betrachten.

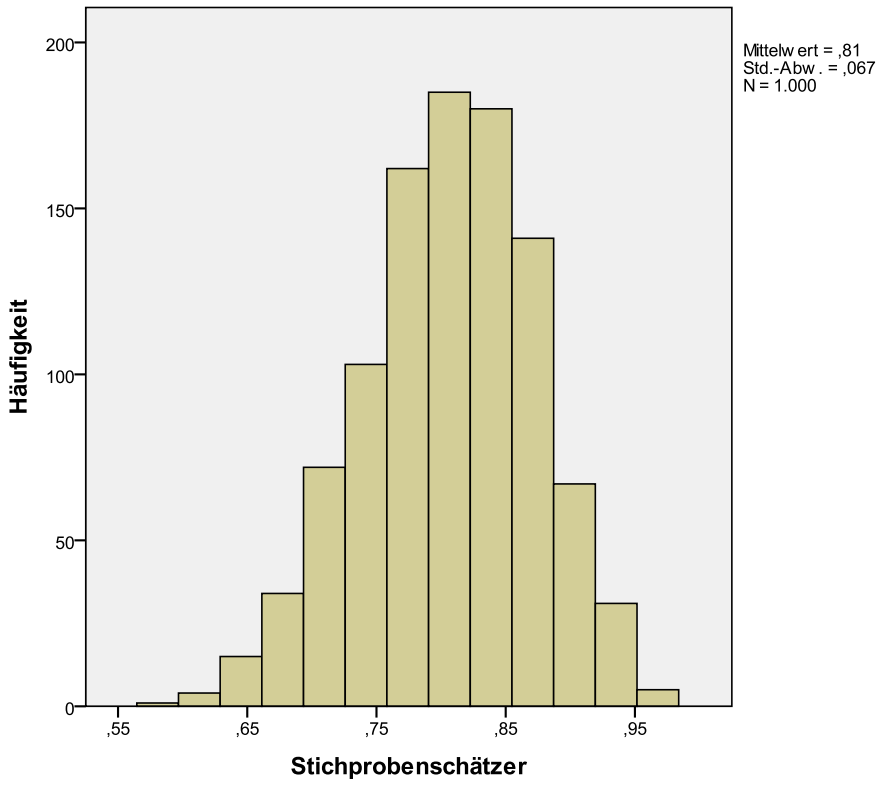
5.8.1 Vertrauensintervall verstehen und näherungsweise berechnen

Wir werden später das Vertrauensintervall für eine Anteilsschätzung von SPSS berechnen lassen, wobei ein komplexer Algorithmus zum Einsatz kommt. Vorab lernen Sie ein Verfahren kennen, um ein approximatives, aber in vielen Fällen ausreichend genaues Vertrauensintervall auf sehr

einfache Weise manuell zu ermitteln. Für die Beschäftigung mit dem manuellen Verfahren gibt es mindestens zwei gute Gründe:

- Sie gewinnen ein Verständnis für die statistischen Grundlagen von Vertrauensintervallen.
- Sie lernen ein einfaches Verfahren kennen, um bei der Planung einer Studie für eine vorgegebene Vertrauensintervallbreite und ein gefordertes Konfidenzniveau den benötigten Stichprobenumfang zu berechnen (siehe Abschnitt 5.8.3). Weil SPSS und auch G*Power vergleichbare Kalkulationen *nicht* anbieten, hat das Do-It-Yourself - Verfahren einen hohen praktischen Nutzen.

Wir stellen uns vor, unendlich viele Zufallsstichproben mit demselben Umfang aus einer Population zu ziehen und jeweils die Anteilsschätzung zu berechnen. Zunächst geht es um die Frage, von welcher Gestalt die resultierende Verteilung der Anteilsschätzungen ist. Sie können es glauben, nach wenigen Jahren Mathematikstudium beweisen (Stichwort: zentraler Grenzwertsatz) oder mit Hilfe einer Computer-Simulation (z. B. mit SPSS) ausprobieren: Es resultiert die als **Gauß**- oder **Normalverteilung** bezeichnet Glockengestalt! In einem Zufallsexperiment mit SPSS-Hilfe habe ich aus einer Population mit der wahren Wahrscheinlichkeit 0,806 für ein Ereignis 1000 Stichproben der Größe 31 gezogen und jeweils die relative Häufigkeit für das Ereignis ermittelt. Wie das folgende Histogramm zeigt, kommt das Ergebnis trotz des relativ kleinen Stichprobenumfangs der Glockenform schon ziemlich nahe, wenngleich noch eine (aus dem hohen Populationsanteil von 0,806 resultierende) Tendenz zur negativen Schiefe erkennbar ist (vgl. Abschnitt 5.5.3.1):



Was Sie hier sehen, ist eine simulierte Stichprobenverteilung für den Anteilsschätzer aus Stichproben der Größe $n = 31$ bei einer Populationswahrscheinlichkeit von $p = 0{,}806$.

Nach Rumsey (2008, S. 181) müssen folgende Anforderungen an die Stichprobengröße erfüllt sein, um eine akzeptable Normalverteilungsanpassung für die Verteilung der Anteilsschätzungen zu erreichen:[1]

$$n \cdot p \geq 5$$
$$n \cdot (1-p) \geq 5$$

Ist die Wahrscheinlichkeit *p* unbekannt, darf man zur Prüfung der beiden Voraussetzungen auch den Anteilsschätzer $\hat{p}$ aus der Stichprobe verwenden. Im Beispiel sind die Bedingungen erfüllt:

$$31 \cdot 0{,}806 = 24{,}986$$
$$31 \cdot (1-0{,}806) = 6{,}014$$

Zwar hat die Stichprobenverteilung als Mittel- bzw. Erwartungswert gerade die gesuchte Populationswahrscheinlichkeit *p*, doch zeigt sich keinesfalls in *jeder* Stichprobe genau dieser Wert. Bei der Konstruktion eines Konfidenzintervalls für die Schätzung aus der *einen* vorhandenen Stichprobe spielt die Standardabweichung der Stichprobenverteilung eine entscheidende Rolle. Weil man diese Standardabweichung als den typischen Fehler bei der Schätzung des Populationsanteils durch eine Stichprobe interpretieren kann, spricht man vom **Standardfehler** des Anteilsschätzers $\hat{p}$. Als Symbol für den Standardfehler verwenden wir anschließend $\sigma(\hat{p})$.

Für konkrete Werte von *n* und *p* lässt sich der Standardfehler zum Anteilsschätzer $\hat{p}$ mit der folgenden Formel berechnen:

$$\sigma(\hat{p}) = \sqrt{\frac{p \cdot (1-p)}{n}}$$

In der Praxis steht statt der Populationswahrscheinlichkeit *p* nur eine Schätzung $\hat{p}$ aus der Stichprobe zur Verfügung, sodass man sich beim Standardfehler mit der Schätzung $\hat{\sigma}(\hat{p})$ zufrieden geben muss:

$$\hat{\sigma}(\hat{p}) = \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

Für eine Normalverteilung mit dem Mittelwert μ und der Standardabweichung σ weiß man generell z. B., dass sich 95% der Verteilungsmasse im folgenden Intervall befinden:

$$[\mu - 1{,}96 \cdot \sigma;\ \mu + 1{,}96 \cdot \sigma]$$

Beim Wert 1,96 handelt es sich um das 97,5 - Perzentil der Standardnormalverteilung (mit dem Erwartungswert 0 und der Varianz 1).

Bei einer ausreichenden Stichprobengröße ist die Stichprobenverteilung des Anteilsschätzers $\hat{p}$ approximativ normal und die Stichprobenschätzung des Standardfehlers hinreichend präzise, sodass 95% aller Stichproben eine Schätzung $\hat{p}$ liefern mit:

$$\hat{p} \in \left[p - 1{,}96 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}};\ p + 1{,}96 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \right]$$

---

[1] Andere Autoren sind skeptisch bzgl. der Normalverteilungsapproximation bei kleinen Stichproben (siehe z. B. Agresti 2007, S. 9; Brown et al. 2001 und 2002). Diese Debatte ist wenig relevant für unser Ziel, das Vertrauensintervall am Beispiel der Anteilsschätzung zu erläutern. Für die Forschungspraxis wird in Abschnitt 5.8.2 die Verwendung eines aufwändiger (von SPSS) berechneten Vertrauensintervalls empfohlen.

Eine einfache Termumformung ergibt, dass in all diesen Fällen für den Populationsanteil *p* erfüllt ist:

$$p \in \left[\hat{p} - 1{,}96 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}};\ \hat{p} + 1{,}96 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \right]$$

Folglich ist

$$\left[\hat{p} - 1{,}96 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}};\ \hat{p} + 1{,}96 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \right]$$

ein 95% - Vertrauensintervall für den Populationsparameter *p*.

Für den Frauenanteil im *statistischen Praktikum mit SPSS* ergibt sich so aus der Manuskriptstichprobe das folgende 95% - Vertrauensintervall:

$$\left[0{,}806 - 1{,}96 \cdot \sqrt{\frac{0{,}806 \cdot (1-0{,}806)}{31}};\ 0{,}806 + 1{,}96 \cdot \sqrt{\frac{0{,}806 \cdot (1-0{,}806)}{31}} \right] = [0{,}667;\ 0{,}945]$$

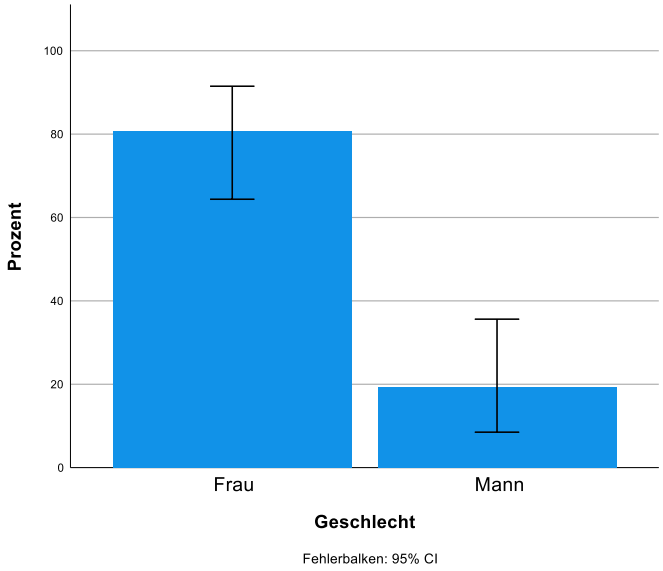
Wenn Sie ein 95% - Vertrauensintervall regelkonform ermitteln (also eine echte Zufallsstichprobe ziehen und die Rechnung korrekt ausführen), dann wird es mit einer Wahrscheinlichkeit von 0,95 den korrekten Populationsparameter enthalten. Immerhin mit einer Wahrscheinlichkeit von 0,05 werden Sie hingegen ein Konfidenzintervall erhalten, das den fraglichen Parameter *nicht* enthält. Für das einzige, in einer Studie tatsächlich berechnete Vertrauensintervall kann man *nicht* sagen, dass es den Parameter mit 95% - iger Wahrscheinlichkeit enthält. Die statistische Sicherheit betrifft nicht ein konkretes Konfidenzintervall, sondern das *Verfahren* zur Ermittlung von Konfidenzintervallen. In der Praxis sind die streng genommen fehlerhaften Wahrscheinlichkeitsaussagen über konkrete Vertrauensintervalle häufig anzutreffen. Allerdings weicht die übliche Redeweise nur in harmloser Weise von der Logik der Vertrauensintervalle ab.

Nicht ganz so harmlos sind beim beschriebenen approximativen Berechnungsverfahren die Abweichungen zwischen der nominellen und der tatsächlichen Überdeckungswahrscheinlichkeit des Vertrauensintervalls bei unzureichender Normalverteilungsapproximation. Bei einer Stichprobengröße von $n = 31$ hat die Stichprobenverteilung der Anteilsschätzungen zur Populationswahrscheinlichkeit von $p = 0{,}806$ noch keine akzeptable Normalverteilungsapproximation erreicht. Daher liefert das oben beschriebene approximative Verfahren leider nicht mit der Wahrscheinlichkeit 0,95 ein Vertrauensintervall, das den wahren Populationsparameter enthält. Im gleich folgenden Abschnitt 5.8.2 wird beschrieben, wie man von SPSS ein Vertrauensintervall berechnen lässt, das die versprochene Überdeckungswahrscheinlichkeit auch bei kleinen Stichproben einhält.

5.8.2 Jeffreys-Vertrauensintervall von SPSS berechnen lassen

Das in Abschnitt 5.8.1 vorgestellte approximative Verfahren war aufgrund seiner Einfachheit gut geeignet, das Konzept des Vertrauensintervalls am Beispiel der Anteilsschätzung zu erläutern. Allerdings ist die verwendete Normalverteilungsapproximation bei kleinen Stichproben (auch bei Beachtung der Abschnitt 5.8.1 angegebenen Regeln) zu ungenau (siehe z. B. Agresti 2007, S. 9). Im Resultat erhält man *nicht* mit der Wahrscheinlichkeit 0,95 ein Vertrauensintervall, das den wahren Parameter überdeckt. Die tatsächliche Überdeckungswahrscheinlichkeit des Verfahrens zur Berechnung von approximativen Vertrauensintervallen zur Anteilsschätzung ist kleiner als 0,95.

SPSS 28 beherrscht neun verschiedene Methoden zur Berechnung des Vertrauensintervalls zu einem Anteil, sodass wir die Qual der Wahl haben. Wird in SPSS per Diagrammerstellung (siehe Kapitel 11) für eine kategoriale (nominale oder ordinale) Variable ein Balkendiagramm mit Fehlerindikatoren angefordert, dann verwendet SPSS per Voreinstellung das 95% - ige Jeffreys-Vertrauensintervall, z. B.:



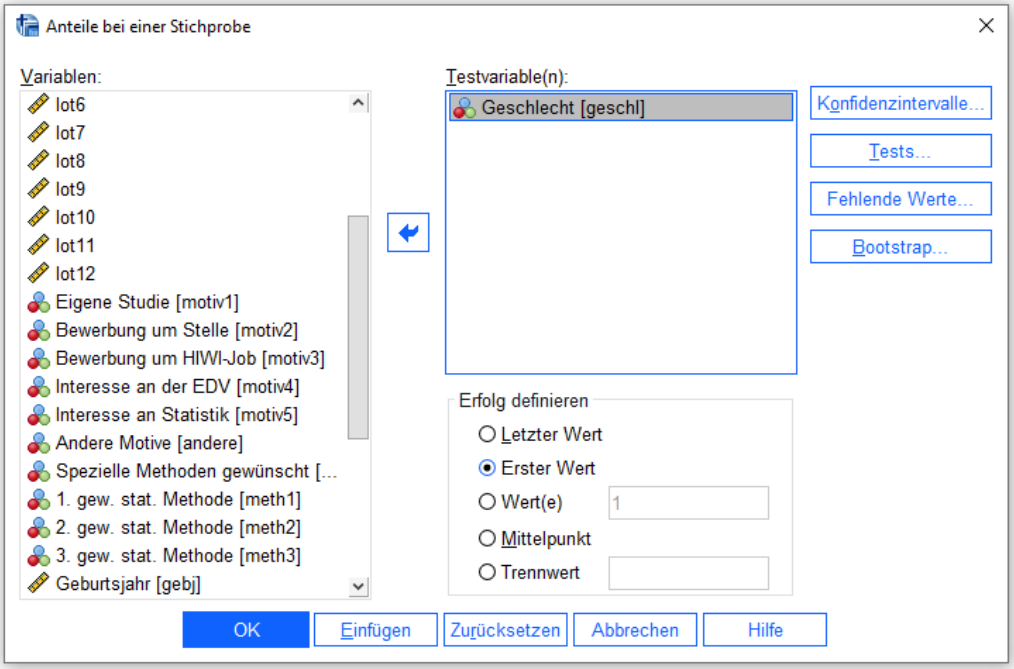
Nach Brown et al. (2002, S. 186) hält das Jeffreys-Vertrauensintervall die versprochene Überdeckungswahrscheinlichkeit mit akzeptabler Genauigkeit ein und fällt dabei relativ klein (präzise) aus. Daher entscheiden wir uns für den Jeffreys-Vorschlag zur Berechnung des Vertrauensintervalls zum Anteil.[1]

Nach dem Menübefehl

Analysieren > Mittelwerte vergleichen > Anteile bei einer Stichprobe

wählen wir im folgenden Dialog GESCHL als Testvariable und eine **Erfolgsdefinition** über den **ersten Wert**:

[1] Eine für Statistikeinsteiger geheimnisvoll klingende Beschreibung, mit der einige Vorbelastete vielleicht etwas anfangen können: Das Jeffreys-Intervall wird nach den Prinzipien der Bayes-Statistik mit einer nichtinformativen a-priori - Verteilung berechnet.



Wir erhalten das Jeffreys-Vertrauensintervall [0,644; 0,915]:

Konfidenzintervalle der Anteile bei einer Stichprobe

		Beobachtet			Asymptotischer Standardfehler	95% Konfidenzintervall	
	Intervalltyp	Erfolge	Versuche	Anteil		Unterer Wert	Oberer Wert
Geschlecht = Frau	Agresti-Coull	25	31	,806	,071	,633	,912
	Jeffreys	25	31	,806	,071	,644	,915
	Wilson-Score	25	31	,806	,071	,637	,908

Es weicht deutlich vom approximativen Vertrauensintervall [0,667; 0,945] ab, das wir im Abschnitt 5.8.1 bestimmt haben. Es ist etwas kleiner und *nicht* symmetrisch um die Punktschätzung. Bei einer Stichprobe der Größe $n = 31$ und einer Populationswahrscheinlichkeit von 0,806 ist das im zentralen Grenzwertsatz beschriebene Streben der Stichprobenverteilung der Anteilsschätzung in Richtung Normalität offenbar noch nicht abgeschlossen (siehe Histogramm in Abschnitt 5.8.1).

5.8.3 Stichprobenumfang für eine gewünschte Präzision berechnen

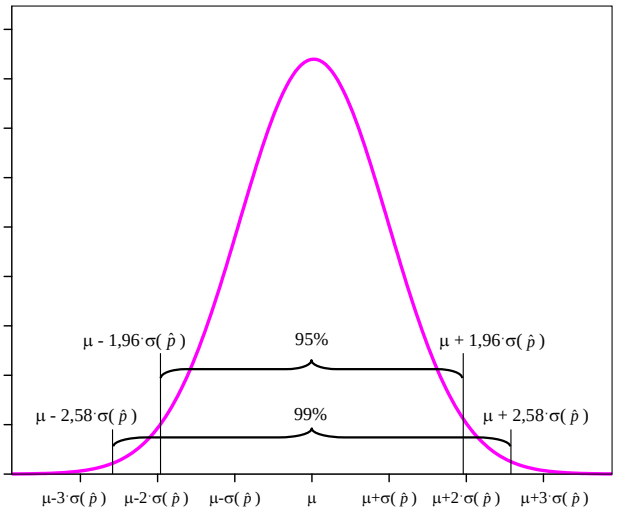
Bei der *Präzision* einer Intervallschätzung sind zwei Forderungen im Spiel:

- Das Konfidenzniveau (die statistische Sicherheit) soll möglichst hoch sein.
- Das Konfidenzintervall soll möglichst klein sein.

Dummerweise verhalten sich die beiden Forderungen konträr: Je höher das Konfidenzniveau (je kleiner die tolerierte Irrtumswahrscheinlichkeit), desto breiter wird das Konfidenzintervall. In der Planungsphase können Sie aber doch beide Forderungen unter einen Hut bringen, indem Sie die Stichprobe so groß wählen, dass bei einer festgelegten Irrtumswahrscheinlichkeit ein Konfidenzintervall der gewünschten Breite resultiert.

Um einzusehen, warum die Steigerung des geforderten Konfidenzniveaus zu einem breiteren Konfidenzintervall führt, betrachten wir die Stichprobenverteilung des Anteils, von der wir mittlerweile wissen, dass sie zumindest bei großen Stichproben annähernd normal ist. Wird etwa

eine Sicherheit von 99% statt 95% verlangt, dann muss ein breiteres Segment der Stichprobenverteilung einbezogen werden, um entsprechend viel Verteilungsmasse einzusammeln:



In der Formel zur näherungsweisen Berechnung des Vertrauensintervalls ist im Vergleich zur 95% - Version

$$\left[\hat{p} - 1{,}96 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}};\ \hat{p} + 1{,}96 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}\right]$$

der Faktor zu ändern, mit dem der geschätzte Standardfehler $\hat{\sigma}(\hat{p})$ der Stichprobenverteilung zu multiplizieren ist. Wei Sie aus dem Abschnitt 5.8.1 wissen, stammt dieser Faktor aus der Standardnormalverteilung (mit dem Erwartungswert 0 und der Streuung 1). Zum Konfidenzniveau 95% (also zur Irrtumswahrscheinlichkeit 5%) gehört der Wert 1,96, weil dieser bei einer Standardnormalverteilung genau 2,5% der Verteilungsmasse am rechten Rand abschneidet. Mit anderen Worten: Bei der Ziehung einer Zufallsgröße mit Standardnormalverteilung wird der Wert 1,96 mit einer Wahrscheinlichkeit von 0,025 überschritten. Der zu einer beliebigen Irrtumswahrscheinlichkeit α gehörige Wert, der am rechten Rand der Standardnormalverteilung die Verteilungsmasse α/2 abschneidet, soll anschließend als $z_{1-\alpha/2}$ bezeichnet werden.

Mit dem Konfidenzniveau wächst der Faktor $z_{1-\alpha/2}$ und damit auch die Breite des Vertrauensintervalls. In der folgenden Tabelle sind häufig verwendete Werte angegeben:

Konfidenzniveau	$\mathbf{z_{1-\alpha/2}}$
80%	1,282
90%	1,645
95%	1,960
99%	2,576

Mit dem zu einem Konfidenzniveau gehörigen Wert $z_{1-\alpha/2}$ ermittelt man folgendermaßen das Konfidenzintervall zur Anteilsschätzung $\hat{p}$ aus einer Stichprobe mit *n* Fällen:

$$\left[\hat{p} - z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}};\ \hat{p} + z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}\right]$$

Im Beispiel mit dem Frauenanteil im statistischen Praktikum erhalten wir das approximative 99% - Vertrauensintervall $[0{,}623\,;\,0{,}989]$. Es ist erwartungsgemäß um den Faktor $\frac{2{,}58}{1{,}96} = 1{,}32$ breiter als das 95% - Gegenstück.

Wer ein schmales Vertrauensintervall angeben möchte, muss bei konstanter Stichprobengröße den Faktor $z_{1-\alpha/2}$ reduzieren, was nach obiger Tabelle zu Lasten der statistischen Sicherheit geht.

Wenn Sie rechtzeitig (vor der Datenerhebung) planen und beim Stichprobenumfang nicht an Grenzen stoßen, können Sie beide Aspekte der Präzision kontrollieren:

- Legen Sie das Konfidenzniveau fest.
- Wählen Sie die gewünschte Intervallbreite.
  Üblicherweise interessiert man sich für die *einseitige* Fehlerbreite *b* und berichtet das Vertrauensintervall in der Notation $\hat{p} \pm b$.
- Berechnen Sie die erforderliche Stichprobengröße.

Nur für den Fall, dass Ihnen die Formel zum dritten Punkt nicht spontan einfällt, gebe ich sie samt der recht simplen Herleitung an. Die folgende Formel für die Fehlerbreite *b*

$$b = z_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

muss nach der Stichprobengröße *n* aufgelöst werden. Zur Berechnung des Standardfehlers wird hier kein Anteilsschätzer verwendet, sondern die angenommene Populationswahrscheinlichkeit, weil wir uns geistig in die Planungsphase einer Studie versetzt haben, also noch keinen Schätzwert zur Verfügung haben. Wir dividieren beide Seiten der Gleichung durch $z_{1-\alpha/2}$ und quadrieren:

$$\frac{b^2}{z_{1-\alpha/2}^2} = \frac{p \cdot (1-p)}{n}$$

Der Rest ist simpel:

$$n = z_{1-\alpha/2}^2 \frac{p \cdot (1-p)}{b^2}$$

Ist bei einer angenommenen Populationswahrscheinlichkeit von 0,3 ein Konfidenzniveau von 95% (nach obiger Tabelle $z_{1-\alpha/2}$ = 1,96) sowie eine Fehlerbreite (gleich halbe Konfidenzintervallbreite) von 0,02 gewünscht, dann resultiert als minimal erforderlicher Stichprobenumfang:

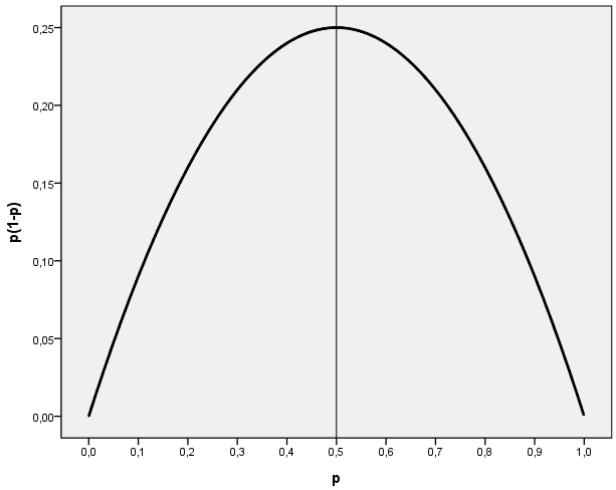
$$n = 1{,}96^2 \frac{0{,}3 \cdot (1-0{,}3)}{0{,}02^2} \approx 3{,}84 \frac{0{,}21}{0{,}0004} = 2016$$

Um also z. B. einen Stimmenanteil von 30% mit einer Genauigkeit von ± 2% schätzen zu können, muss man mindestens 2016 zufällig ermittelte Wahlberechtigte befragen.

Dummerweise muss der unbekannte, in der Stichprobe zu schätzende Populationsparameter für die Berechnung des benötigten Stichprobenumfangs bereits zur Verfügung stehen. Wenn noch nicht einmal eine grobe Vorstellung über den Populationsparameter existiert, geht man vom ungünstigsten Fall *p* = 0,5 aus, der den größten Standardfehler und damit den größten Stichprobenbedarf zur Folge hat. Wenn Sie dieser Behauptung über das Maximum der Funktion

$$p \cdot (1-p) = p - p^2$$

nicht trauen, werfen Sie doch einfach einen Blick auf den Graphen der Funktion:



Mit $p = 0{,}5$ statt $p = 0{,}3$ erhöht sich der zuletzt berechnete Stichprobenbedarf auf 2400:

$$n = 1{,}96^2 \frac{0{,}5 \cdot (1-0{,}5)}{0{,}02^2} \approx 3{,}84 \frac{0{,}25}{0{,}0004} = 2400$$

Meinungsforschungsinstitute sollten also zur Erfassung der Parteipräferenzen bei

- einem Konfidenzniveau von 95%
- und einer Fehlerbreite (gleich halbe Konfidenzintervallbreite) von 0,02

eine Stichprobengröße von ca. 2400 Fällen verwenden.[1]

Man rechnet leicht nach, dass die in Abschnitt 2.1.2.1 zitieren Genauigkeitsversprechen der Forschungsgruppe Wahlen zum Politbarometer im März 2022

> Die Interviews wurden in der Zeit vom 8. bis 10. März 2022 bei 1.345 zufällig ausgewählten Wahlberechtigten telefonisch erhoben. ... Der Fehlerbereich beträgt bei einem Anteilswert von 40 Prozent rund +/- drei Prozentpunkte und bei einem Anteilswert von 10 Prozent rund +/-zwei Prozentpunkte.

korrekt sind und sogar noch eine Sicherheitsreserve enthalten, z. B.:

$$1{,}96^2 \frac{0{,}4 \cdot (1-0{,}4)}{0{,}03^2} \approx 1024$$

$$1{,}96^2 \frac{0{,}1 \cdot (1-0{,}1)}{0{,}02^2} \approx 864$$

Bei den in diesem Abschnitt ermittelten Stichprobengrößen, die sehr weit über dem in Abschnitt 5.8.1 im Zusammenhang mit der näherungsweisen Berechnung des Vertrauensintervalls betrachteten Wert 31 liegen, ist übrigens die Normalverteilungsapproximation der Stichprobenverteilung des Anteils nahezu perfekt. Die in Abschnitt 5.8.2 beim Jeffreys-Vertrauensintervall für den Anteilsschätzer aus einer Stichprobe mit 31 Fällen beobachtete Asymmetrie tritt dann praktisch nicht mehr auf, und die Notation $\hat{p} \pm b$ ist gerechtfertigt.

---

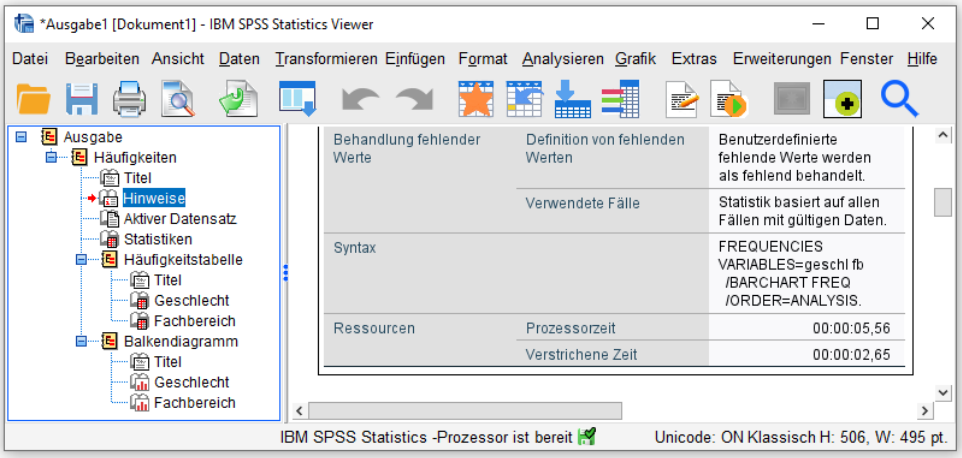
[1] Auf der folgenden Webseite finden Sie die Stichprobenumfänge von zahlreichen Forsa-Umfragen: http://www.wahlrecht.de/umfragen/forsa.htm

6 Speichern der SPSS-Kommandos zu wichtigen Anweisungsfolgen

6.1 Zur Motivation

Eventuell erhalten Sie nach Abschluss der Fehlerkontrolle noch weitere ausgefüllte Fragebögen. Sie freuen sich natürlich über die Stichprobenerweiterung und erfassen sofort die neuen Fälle. Dann allerdings fällt Ihnen ein, dass nun alle Kontrollanalysen wiederholt werden müssen.

Um solchen Frust zu vermeiden, brauchen wir eine Möglichkeit, aufwändige und potentiell mehrfach benötigte Anweisungssequenzen zur späteren Wiederverwendung abzuspeichern. In SPSS eignen sich dazu die **Kommandos**, die den einzelnen Dialogboxen zugrunde liegen, und die von SPSS stets im Hintergrund erzeugt und ausgeführt werden, wenn wir eine ausgefüllte Dialogbox mit **OK** abschicken. SPSS protokolliert zu jeder Analyseanforderung in der zunächst zugeklappten Teilausgabe **Hinweise** u. a. die zugrunde liegende Syntax, z. B. bei der Häufigkeitsanalyse für die Variablen GESCHL und FB:



In diesem Zusammenhang lohnt sich ein kurzer Blick auf die Architektur des SPSS-Systems, das aus den beiden folgenden Komponenten besteht:

- **Bedienoberfläche**
  Wir interagieren mit der Bedienoberfläche, die unsere Anweisungen entgegennimmt und die Ergebnisse präsentiert. Wir können der Bedienoberfläche unsere Anweisungen durch Dialogboxen oder durch SPSS-Kommandos übergeben.
- **SPSS-Prozessor**
  Die Bedienoberfläche gibt unsere Anweisungen in jedem Fall in Form von SPSS-Kommandos an den Prozessor weiter, der im Hintergrund arbeitet. Wir erfahren übrigens in der Statuszeile der SPSS-Fenster, was der Prozessor gerade treibt. Da wir den Prozessor bislang nur minimal belastet haben, war in der Statuszeile meist zu lesen:

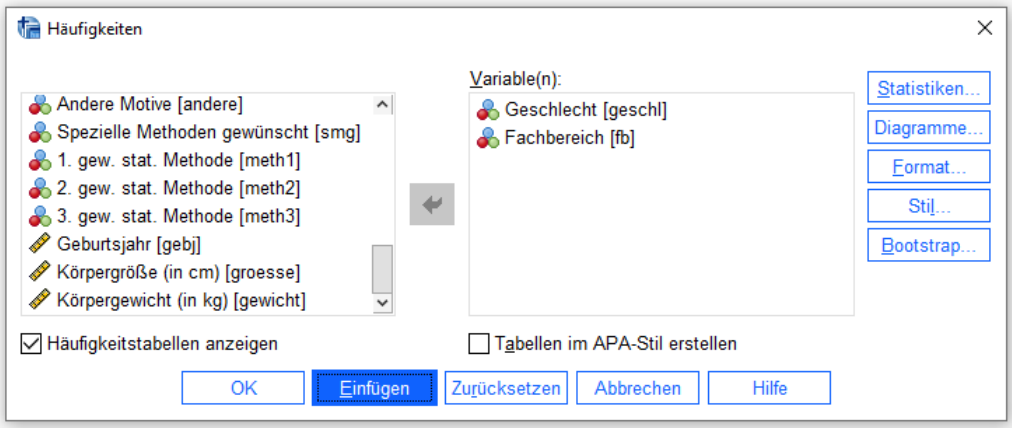
  IBM SPSS Statistics - Prozessor ist bereit

  Während der Prozessor arbeitet, wird in der Statuszeile angezeigt, mit welchem SPSS-Kommando er gerade beschäftigt ist. Nach dem Abschicken einer Häufigkeitsdialogbox erscheint z. B. (bei unserem kleinen Datensatz allerdings nur sehr kurz):

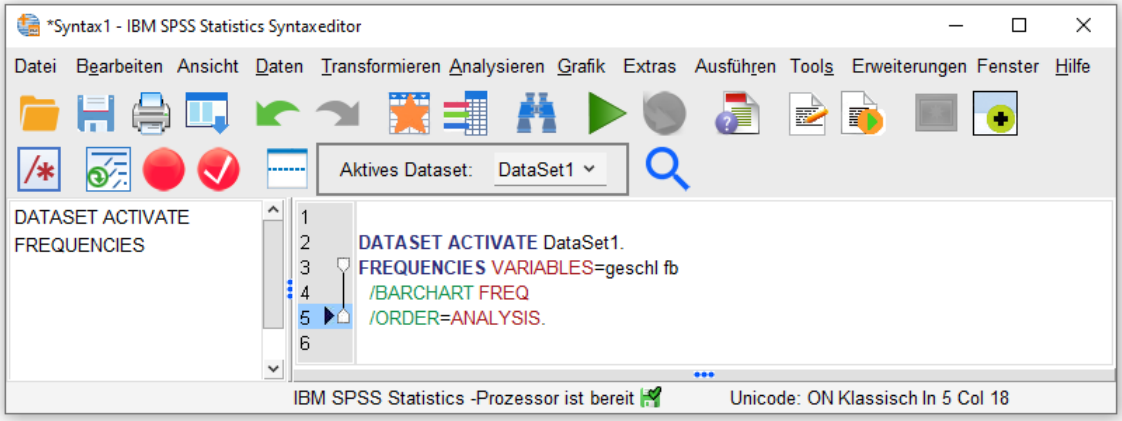
  Ausführen: FREQUENCIES

Wenn wir eine ausgefüllte Häufigkeitsdialogbox mit **OK** quittieren, dann führt der SPSS-Prozessor also im Hintergrund das korrespondierende FREQUENCIES-Kommando aus.

In fast allen SPSS-Dialogboxen kann man über die Schaltfläche **Einfügen**



die zugrunde liegenden SPSS-Kommandos produzieren lassen. Diese werden dann *nicht* ausgeführt, sondern in ein sogenanntes **Syntaxfenster** übertragen, das die Bearbeitung von Kommandos u. a. durch eine Navigationszone am linken Rand, die farbliche Unterscheidung verschiedener Syntaxbestandteile und eine Syntaxvervollständigung unterstützt:



Hier kann man zusammengehörige Kommandos zu einer Sequenz ansammeln, nach Bedarf editieren, einzeln oder gemeinsam ausführen lassen und schließlich in einer Datei abspeichern. Später kann man diese Datei wieder öffnen und die Kommandos - eventuell nach manueller Überarbeitung - erneut ausführen lassen. Das genaue Vorgehen wird in Abschnitt 6.2 an einem konkreten Beispiel geübt.

Eine Folge von SPSS-Kommandos kann man (leicht hochstaplerisch) als **SPSS-Programm** bezeichnen. In fast jedem Projekt sollte es mindestens *ein* SPSS-Programm geben, nämlich das bereits in Abschnitt 4.2.4 vorgeschlagene **Transformationsprogramm**, das aus der Rohdatendatei durch diverse Transformationen die Fertigdatendatei des Projekts erstellt. Wir werden für unser KFA-Projekt ein solches Programm in Kapitel 7 erstellen.

Ob sich bei einer konkreten Anweisungssequenz das Abspeichern als SPSS-Programm lohnt, muss von Fall zu Fall entschieden werden. Bei kurzen, simplen Sequenzen mit geringer Wiederholungswahrscheinlichkeit ist ein Konservieren unrentabel.

Es soll nicht verschwiegen werden, dass die Ausführung einer Anweisungssequenz mit dem Umweg über ein Syntaxfenster geringfügig mehr SPSS-Kenntnisse erfordert als die direkte Ausführung durch Quittieren von Dialogboxen mit **OK**. Wer sich beim Umgang mit Kommandos unsicher fühlt, bei seinem relativ kleinen Projekt eventuell erforderliche Wiederholungen von Dialogbox-Sequenzen nicht scheut und das Risiko inkonsistenter Datenzustände durch Sorgfalt kontrolliert, der kann auf das Erzeugen und Abspeichern von SPSS-Kommandos verzichten.

Für ambitionierte SPSS-Anwender(innen) muss noch klargestellt werden, dass die Erstellung, Überarbeitung und Ausführung von Programmen in einem Syntaxfenster eine eigenständige Methode der SPSS-Benutzung darstellt, über die fast alle Leistungen des Programms erreichbar sind. Viele SPSS-Optionen stehen sogar *ausschließlich* über die Syntax zur Verfügung, z. B.:

- die in der Marktforschung eingesetzte Conjoint-Analyse
- Wiederholungsanweisungen (Schleifen), mit denen man Datentransformationen auf effiziente Weise durchführen kann (DO REPEAT - Schleife, LOOP-Schleife)
- die MATRIX-Programmiersprache, mit der man eigene Statistikprozeduren erstellen kann

Im aktuellen Kapitel 6 werden nur sehr elementare Hinweise zur Kommandosprache gegeben. Diese sollten genügen für Anwender, die nicht frei programmieren, sondern nur gelegentlich ein per SPSS-Dialogbox erzeugtes Kommando modifizieren wollen. Der Anhang zu diesem Manuskript enthält eine ausführlichere Beschreibung der Kommandosprache. Eine vollständige Dokumentation auf ca. 2450 Seiten finden Sie als PDF-Dokument im Hilfesystem von SPSS über

Hilfe > Befehlssyntaxreferenz

Dieses Dokument ist nicht nur ein gutes Nachschlagewerk mit Details zu allen Kommandos, sondern bietet im Kapitel *Universals* auch eine Einführung in generelle Themen im Zusammenhang mit Kommandos, Dateien, Variablen und Transformationen. Wertvolle Informationen bietet auch das kostenlos verfügbare Buch *Programming and Data Management for IBM SPSS Statistics 24* (IMB Corp. 2016).

Wie schon erwähnt, sind die Dialogboxen beim Erstellen eines SPSS-Programms sehr nützlich. Mit Hilfe der bislang ignorierten Schaltfläche **Einfügen** kann man die zu einer Dialogbox-Bearbeitung äquivalente Kommandofolge in ein Syntaxfenster übertragen. Man muss sich also nicht zwischen zwei unabhängigen SPSS-Bediensystemen entscheiden, sondern kann eine rationelle Kombination der beiden Techniken verwenden.

6.2 Dialogunterstützte Erstellung von SPSS-Programmen

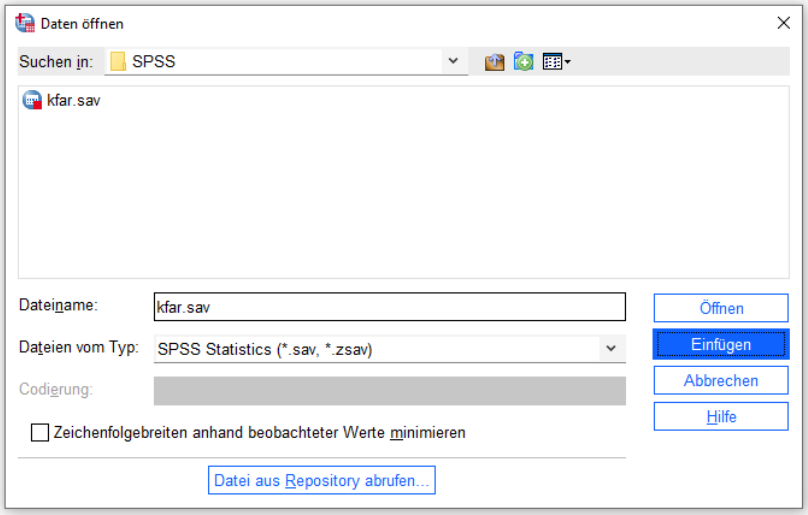
Das folgende SPSS-Programm führt für unser KFA-Projekt die Verteilungs- und Fehleranalysen für die Variablen GESCHL, FB, AERGO, AERGM, GEBJ, GROESSE und GEWICHT durch (vgl. Kapitel 5):

```
GET
  FILE='U:\Eigene Dateien\SPSS\kfar.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.

DATASET ACTIVATE DataSet1.
FREQUENCIES VARIABLES=geschl fb
  /BARCHART FREQ
  /ORDER=ANALYSIS.

FREQUENCIES VARIABLES=aergo aergm gebj groesse gewicht
  /FORMAT=NOTABLE
  /NTILES=4
  /STATISTICS=STDDEV VARIANCE MINIMUM MAXIMUM MEAN MEDIAN
   SKEWNESS SESKEW KURTOSIS SEKURT
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.
```

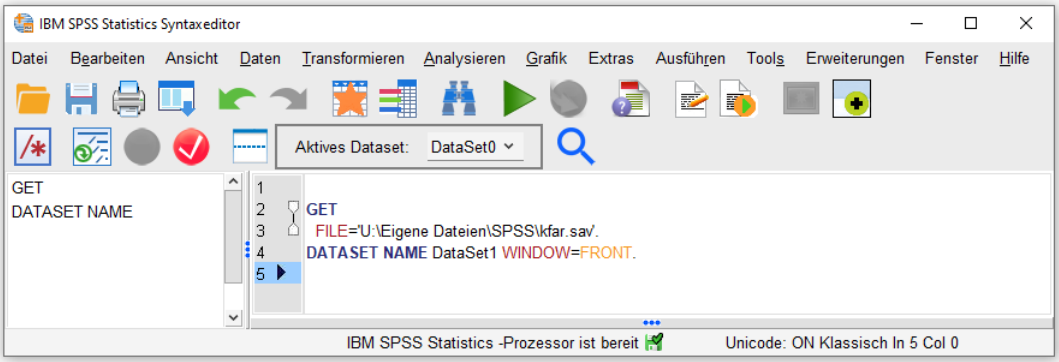
Wir werden dieses Programm gleich mit drei Mausklicks auf **Einfügen**-Schalter produzieren und dabei seine Bestandteile kurz beschreiben. Als Ausgangssituation für die anschließenden Erläuterungen wird eine neue SPSS-Sitzung mit einem leeren Datenfenster angenommen. Verzichten Sie also z. B. beim SPSS-Start auf das Öffnen einer Datendatei per Begrüßungsdialog. Dabei erhalten Sie ein leeres Datenfenster mit dem Namen **DataSet0**. Rufen Sie die Dialogbox zum Öffnen einer Datendatei mit dem folgenden Menübefehl auf:

Datei > Öffnen > Daten

Navigieren Sie zum Ordner mit Ihrer Rohdatendatei, schreiben oder klicken Sie deren Namen in das Feld **Dateiname**, und betätigen Sie dann den Schalter **Einfügen**.



Daraufhin beginnt SPSS *nicht* damit, aus der angegebenen Datendatei ein neues Datenblatt zu erstellen, sondern schreibt das für diese Aktionen zuständige GET-Kommando in ein Syntaxfenster:



Das GET-Kommando ist sehr einfach aufgebaut:

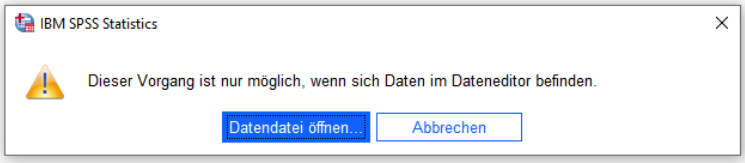
- Es beginnt mit dem Kommandonamen GET.
- Im FILE-Subkommando wird die zu öffnende Datendatei angegeben.
- Am Ende steht bei jedem SPSS-Kommando ein **Punkt**.

Das zusätzlich angelegte Kommando DATASET NAME gibt dem per GET erzeugten, noch unbenannten neuen Datenblatt den Namen **DataSet1** und holt es in den Vordergrund, macht es also zur Arbeitsdatei. Auch das Kommando DATASET NAME endet mit einem Punkt.

Noch sind die beiden Kommandos nicht ausgeführt worden, und das leere Datenblatt mit dem Namen **DataSet0** ist immer noch die Arbeitsdatei. Der Versuch, zur Produktion des ersten FREQUENCIES-Kommandos über dem Menübefehl

Analysieren > Deskriptive Statistiken > Häufigkeiten

eine Häufigkeits-Dialogbox zu öffnen, führt daher zur Meldung:



Brechen Sie diesen Dialog ab, und lassen Sie die im Syntaxfenster befindlichen Kommandos GET und DATASET NAME ausführen, um die Daten einzulesen. Wählen Sie dazu im Syntaxfenster den Menübefehl

Ausführen > Alle

Daraufhin werden folgende Aktionen ausgeführt:

- Es wird ein neues, noch unbenanntes Datenblatt angelegt.
- Dorthin werden die Daten aus der Rohdatendatei **kfar.sav** kopiert.
- Das neue Datenblatt ist mit der Rohdatendatei **kfar.sav** verbunden.
- Das beim Programmstart vorhandene, leere Datenblatt mit dem Namen **DataSet0** wird geschlossen.
- Das neue Datenblatt erhält den Namen **DataSet1** und gelangt in den Vordergrund, wird also zur Arbeitsdatei

Spezifizieren Sie jetzt mit Hilfe der zuständigen Dialogbox dieselbe Häufigkeitsanalyse zu den Variablen GESCHL und FB wie in Abschnitt 5.3. Verlassen Sie die Dialogbox jedoch nicht mit

OK, sondern mit **Einfügen**. Daraufhin erscheint am Ende des Syntaxfensters ein FREQUENCIES-Kommando (siehe oben):

- Es beginnt mit dem Kommandonamen FREQUENCIES.
- Im VARIABLES-Subkommando wird angegeben, welche Variablen analysiert werden sollen.
- Das BARCHART-Subkommando sorgt dafür, dass Balkendiagramme mit einer Häufigkeitsbeschriftung erscheinen.
- Das ORDER-Subkommando entscheidet bei der Analyse mehrerer Variablen darüber, ob die Statistiken für jede Variable in einer eigenen Tabelle oder für alle Variablen in einer gemeinsamen Tabelle erscheinen sollen. Um diese Entscheidung in der **Häufigkeiten**-Dialogbox zu treffen, müssen Sie übrigens die **Format**-Subdialogbox öffnen und im Rahmen **Mehrere Variablen** die passende Option wählen.
- Das FREQUENCIES-Kommando wird wie jedes SPSS-Kommando durch einen **Punkt** abgeschlossen.

Das zusammen mit der Häufigkeitsanalyse automatisch erstellte Kommando

```
DATASET ACTIVATE DataSet1.
```

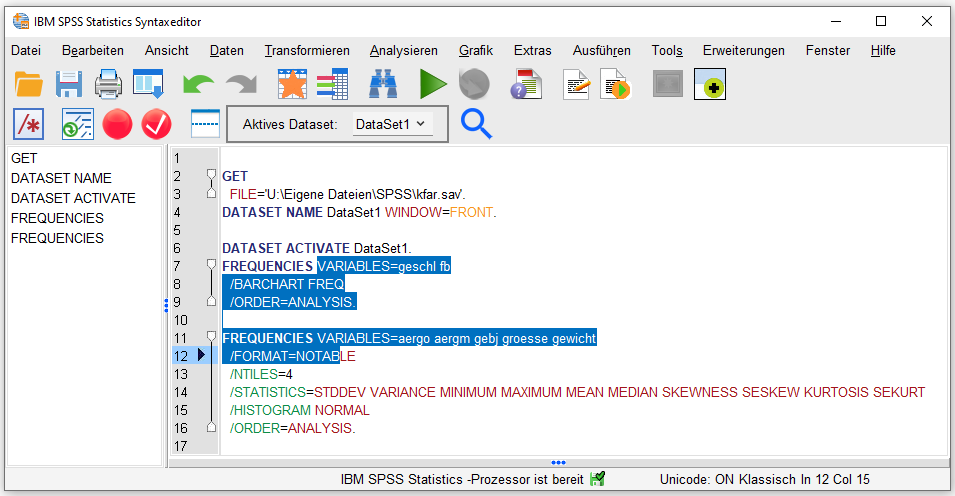
aktiviert vorsichtshalber das mit der Rohdatendatei **kfar.sav** verbundene **DataSet1**. Das Kommando ist momentan überflüssig, weil gar kein anderes Datenblatt existiert.

Produzieren Sie als Nächstes die Syntax zu der in Abschnitt 5.5 durchgeführten Häufigkeitsanalyse für die Variablen AERGO, AERGM, GEBJ, GROESSE und GEWICHT.

Nun sollte Ihr Syntaxfenster den zu Beginn des Abschnitts wiedergegebenen Inhalt haben. Die Kommandos GET und DATASET NAME sind schon gelaufen. Das Kommando DATASET ACTIVATE ist momentan überflüssig, doch würde seine Ausführung keinen Schaden anrichten.

Um die Häufigkeitsanalysen zu erhalten, müssen Sie die beiden FREQUENCIES-Kommandos ausführen lassen. Weil es sich um eine Teilmenge der im Syntaxfenster vorhandenen Kommandos handelt, müssen Sie folgendermaßen vorgehen:

- Markieren Sie zunächst per Maus *die beiden* auszuführenden Kommandos, wobei von jedem Kommando wenigstens ein Zeichen in die Markierung einbezogen werden muss, z. B.:



- Klicken Sie dann auf den Symbolleistenschalter ▶, oder drücken Sie die Tastenkombination **Strg+R**. Daraufhin werden alle Kommandos im Syntaxfenster ausgeführt, die (zumindest teilweise) markiert sind.

SPSS protokolliert im Ausgabefenster zu jeder Analyseanforderung in der zunächst zugeklappten Teilausgabe **Hinweise** u. a. die zugrundeliegende Syntax, z. B.:

Ob zusätzlich die verarbeiteten Kommandos in **Log**-Teilausgaben protokolliert werden sollen, kann im **Optionen**-Dialog auf der Registerkarte **Viewer** eingestellt werden (siehe Abschnitt 4.2.3). Während diese Option bis zur SPSS-Version 27 per Voreinstellung aktiv war, ist sie in SPSS 28 per Voreinstellung passiv.

Damit sich durch spätere Wiederverwendung der SPSS-Kommandos der gewünschte Rationalisierungseffekt einstellen kann, müssen Sie Ihr SPSS-Programm sichern. Wechseln Sie dazu nötigenfalls zum Syntaxfenster, und wählen Sie den Menübefehl:

Datei > Speichen unter...

Verwenden Sie im Dateinamen die vorgeschlagene Erweiterung **.sps**, indem Sie *keine* Erweiterung angeben.

Wenn Sie später dieselbe Auswertung nochmals benötigen, müssen Sie lediglich das vorhandene Programm mit dem Menübefehl:

Datei > Öffnen > Syntax

öffnen und ausführen lassen.

Um die Ausführung *sämtlicher* Kommandos in einem Syntaxfenster anzuordnen, haben Sie folgende Möglichkeiten:

- Menübefehl **Ausführen > Alle**
- Alle Kommandos markieren (z. B. mit **Strg+A**) und die Ausführung anfordern (z. B. per Mausklick auf das Symbol ▶ oder mit der Tastenkombination **Strg+R**)

Lässt man obiges Programm innerhalb einer SPSS-Sitzung erneut komplett ausführen, dann erscheinen die folgenden Warnungen im Ausgabefenster:

```
Warnungsnummer 67.  Befehlsname: GET FILE
Das Dokument wird bereits von einem anderen Benutzer oder Prozess verwendet.
Wenn Sie Änderungen an dem Dokument vornehmen, können diese gegebenenfalls
Änderungen anderer Benutzer überschreiben bzw.  Ihre Änderungen werden von
anderen Benutzern überschrieben.
Geöffnete Datei: U:\Eigene Dateien\SPSS\kfar.sav
```

Datasetname

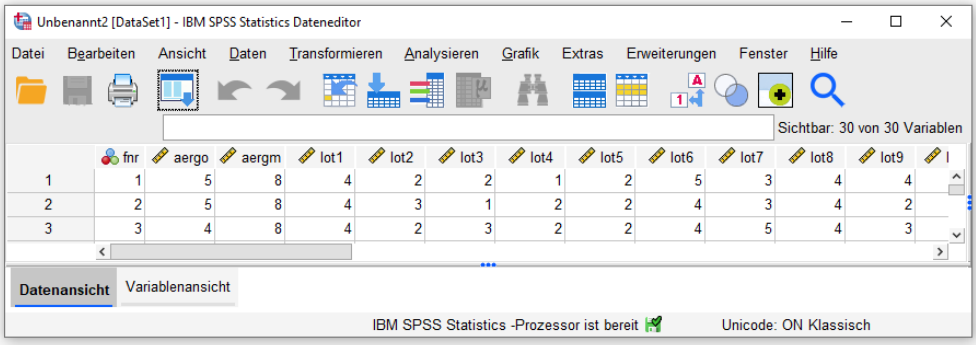
Warnungen

Das aktive Daten-Set ersetzt das vorhandene Daten-Set mit dem Namen DataSet1.

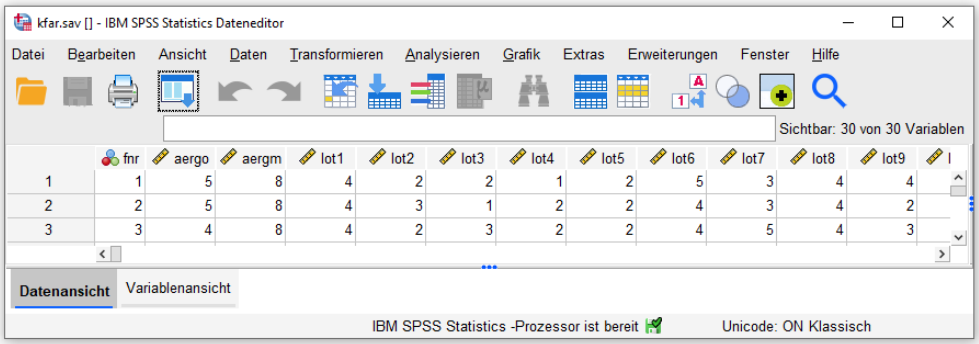
Leider sorgt die generell begrüßenswerte Möglichkeit, in einer SPSS-Sitzung *mehrere* Datenblätter zu verwenden, aktuell für eine Komplikation. Damit daraus keine Konfusion wird, müssen wir das Geschehen im Detail verfolgen:

- Vor dem erneuten Ausführen des Programms existiert ein Datenblatt mit dem Namen **DataSet1**, das mit der Rohdatendatei **kfar.sav** verbunden ist.
- Das erneut ausgeführte GET-Kommando erzeugt ein neues (noch anonymes) Datenblatt und kopiert den Inhalt der Rohdatendatei dorthin. In der ersten Warnung wird darüber informiert, dass zwei Datenblätter mit den Rohdaten als Inhalt existieren.
- Die Rohdatendatei **kfar.sav** bleibt mit dem älteren Datenblatt verbunden, das noch den Namen **DataSet1** trägt.
- Das erneut ausgeführte Kommando DATASET NAME gibt dem neuen Datenblatt den bereits in Verwendung befindlichen Namen **DataSet1**, woraufhin das alte Datenblatt mit diesem Namen geschlossen wird (siehe zweite Warnung).

Insgesamt führt die erneute Ausführung des Programms dazu, dass ein Datenblatt namens **DataSet1** mit dem Inhalt der Rohdatendatei existiert, das aber *nicht* mit der Rohdatendatei verbunden ist:



Wenn dieses Verhalten stört, kann man z. B. die Kommandos DATASET NAME und DATASET ACTIVATE löschen. Dann bleibt das per GET befüllte und mit der Rohdatendatei verbundene Datenblatt unbenannt und wird bei jeder Ausführung des Programms überschrieben(!):



6.3 Arbeiten mit dem Syntax-Fenster

In einem Syntaxfenster lassen sich automatisch erstellte SPSS-Kommandos leicht modifizieren, um z. B. die in einer Statistikprozedur zu analysierenden Variablen auszutauschen.

Man kann ein neues Syntaxfenster auch unabhängig vom **Einfügen**-Schalter einer Dialogbox direkt anfordern mit:

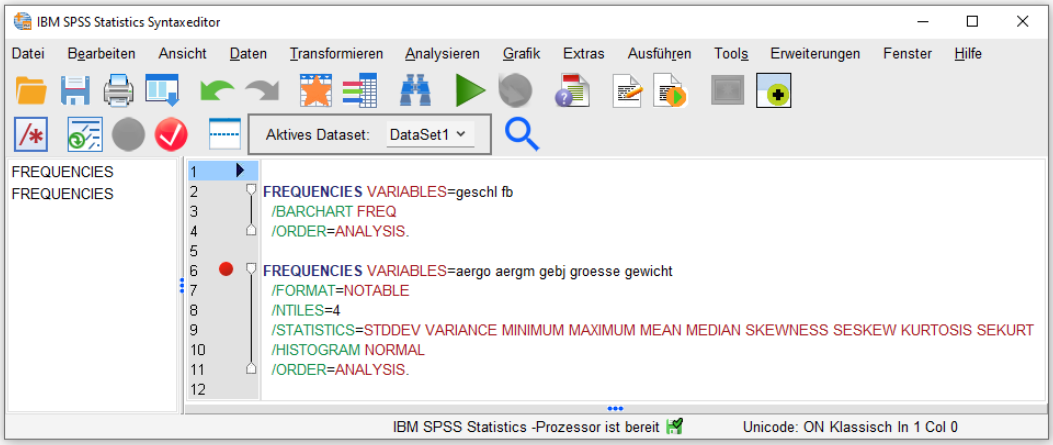
Datei > Neu > Syntax

Wenn *mehrere* Syntaxfenster vorhanden sind, dann muss geregelt werden, in welches Fenster SPSS die per **Einfügen**-Schalter automatisch erzeugten Kommandos übertragen soll. Dies geschieht genauso wie bei den Ausgabefenstern: Ein Mausklick auf den aktiven Schalter  in seiner Symbolleiste macht ein Syntaxfenster zum **Hauptfenster** in seiner Kategorie. Es ist an einem Pluszeichen im Symbol zum Systemmenü  zu erkennen (siehe linken Rand der Titelzeile).

Um ein bestimmtes Syntaxfenster in den Vordergrund zu holen, können Sie es anklicken oder über das **Fenster**-Menü eines beliebigen SPSS-Fensters wählen. Es wird dabei (abweichend vom Verhalten eines Datenfensters) *nicht* automatisch zum Hauptsyntaxfenster. Jedes Syntaxfenster kann auf windows-übliche Weise geschlossen werden, z. B. über den Menübefehl:

Datei > Schließen

Bei der Ausführung von Kommandos ist oft das aktive Datenblatt relevant. Wenn mehrere Datenblätter geöffnet sind, muss eventuell über die in einem Syntaxfenster vorhandene Drop-Down - Liste **Aktives Dataset** für die korrekte Einstellung gesorgt werden:



Allerdings kann ein SPSS-Programm über das Kommando DATASET ACTIVATE auch selbst dafür sorgen, das für die folgenden Kommandos benötigte Datenblatt zu aktivieren.

Wenn SPSS bei der Ausführung eines Programms unerwartet und ohne Fehlermeldung stoppt, dann haben Sie vermutlich per Mausklick im Bereich der Zeilennummerierung einen **Haltepunkt** gesetzt. Im obigen Syntaxfenster ist ein Haltepunkt neben der Startzeile des zweiten FREQUENCIES-Kommandos zu sehen. Nach dem Starten über den Menübefehl

Ausführen > Alle

wird das Programm durch den Haltepunkt gestoppt und kann über den Symbolschalter  fortgesetzt werden. Um einen Haltepunkt zu beseitigen, setzt man einen Mausklick darauf. Über den Menübefehl

Tools > Alle Haltepunkte löschen

wird man alle Haltepunkte wieder los.

Wenn Sie längere Zeit mit SPSS arbeiten, wird sich vermutlich Ihr Umgang mit SPSS-Syntax in folgenden Stufen weiterentwickeln:

- Kommandos automatisch erzeugen lassen und später unverändert wiederverwenden
  Bei dieser Arbeitsweise müssen Sie nur wissen, wie man SPSS-Kommandos per Dialogbox in ein Syntaxfenster befördert, und wie man überflüssige Kommandos löscht.
- Automatisch erzeugte Kommandos modifizieren
  Es zeigt sich, dass SPSS-Kommandos meist leicht zu durchschauen und zu modifizieren sind.
- Freies Programmieren

6.4 Elementare Regeln zur SPSS-Syntax

Für den im Kurs vorgeschlagenen Einsatz von SPSS-Kommandos sollte die Kenntnis der folgenden Regeln genügen:

- Ein Kommando besteht aus seinem Namen und den Spezifikationen, die sich aus Schlüsselwörtern (z. B. VARIABLES, STATISTICS), Variablennamen usw. zusammensetzen, z. B.:

  Kommandoname → `FREQUENCIES`

  Spezifikationen
  ```
  VARIABLES=geschl fb
  /BARCHART FREQ
  /ORDER=ANALYSIS.
  ```

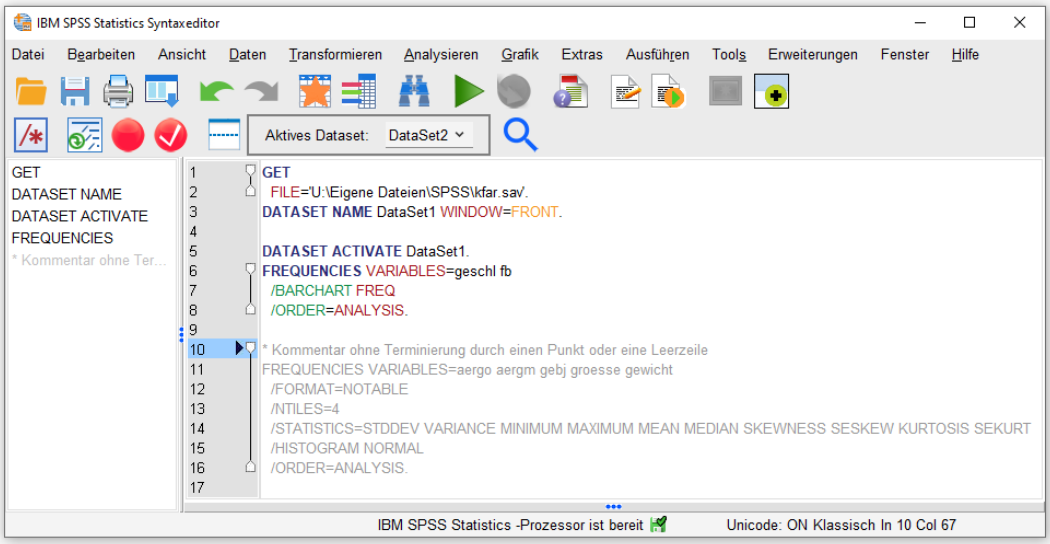
- Bei den Schlüsselwörtern der SPSS-Kommandosprache ist die Groß-/Kleinschreibung irrelevant. Bei der automatisch (per **Einfügen**-Schalter) erzeugten Syntax schreibt SPSS die Schlüsselwörter in Großbuchstaben.
- Zwei Elemente der Kommandosprache sind durch mindestens ein Leerzeichen oder durch einen Zeilenwechsel voneinander zu trennen. Manche Zeichen mit spezieller Bedeutung wie z. B. "=", "/", "(", "+", ">" sind allerdings selbstbegrenzend, d. h. davor und danach sind keine Leerzeichen nötig (aber erlaubt).
- Ein Kommando kann sich über beliebig viele Fortsetzungszeilen erstrecken, dabei dürfen aber *innerhalb* des Kommandos keine Leerzeilen auftreten. Diese signalisieren nämlich (wie ein Punkt) das Ende des Kommandos.
- *Zwischen* zwei Kommandos dürfen Leerzeilen stehen, was eine übersichtliche Gestaltung von SPSS-Programmen erlaubt.
- Jedes Kommando muss in einer neuen Zeile beginnen.
- **Es ist sehr zu empfehlen, jedes Kommando explizit mit einem Punkt zu beenden.**

Gut kommentierte Programme sind später leichter zu verstehen. Die SPSS-Syntax bietet zum Kommentieren das Kommando COMMENT, dessen Name durch ein Sternchen ersetzt werden darf, z. B.:

```
* Mit diesem Programm wird die Rohdatendatei KFAR.SAV
  auf Erfassungsfehler untersucht.
GET
  FILE='U:\Eigene Dateien\SPSS\KFAR.SAV'.
   .   .   .
```

Beachten Sie bitte beim Kommentar-Kommando die folgenden Hinweise:

- Es darf sich ebenfalls über beliebig viele Fortsetzungszeilen erstrecken.
- **Jedes Kommentar-Kommando sollte unbedingt mit einem Punkt abgeschlossen werden.** Wenn Sie den Punkt am Ende vergessen, dann betrachtet SPSS den folgenden Programmtext bis zum nächsten Punkt (oder zur nächsten Leerzeile) als Teil des Kommentars! Dank der farblichen Unterscheidung von Syntaxbestandteilen (mit grauer Schrift für die Kommentare) fällt dieser Fehler allerdings sofort auf, z. B.:



- Endet eine Kommentarzeile mit einem Punkt, so betrachtet SPSS das Kommentar-Kommando als abgeschlossen. Wenn Sie einen Punkt als *Satzzeichen* ans Ende einer Kommentarzeile gesetzt haben, müssen Sie die nächste Kommentarzeile also wieder mit COMMENT oder ***** einleiten.
- Punkte innerhalb einer Kommentarzeile sind kein Problem.

7 Datentransformation

Die zur Untersuchung unserer differentialpsychologischen Hypothese benötigte Optimismus-Variable existiert noch nicht, sondern muss erst aus den 12 LOT-Variablen berechnet werden. Vor dieser Berechnung müssen allerdings die aus messtechnischen Gründen umgepolten (negativ formulierten) LOT-Fragen recodiert werden (z. B. Frage 3).[1] Es ist typisch für empirische Studien, dass vor dem eigentlichen Start der Datenanalyse aus den Rohvariablen mit zahlreichen Datentransformationen neue oder modifizierte Fertigvariablen erstellt werden müssen. Dabei geht es sowohl um sorgfältig zu absolvierende Pflichtübungen (z. B. LOT-Berechnung nach den Vorschriften der Testautoren) als auch um kreative Begriffsbildungen mit dem Ziel, durch geschickte Kombination vorhandener Informationen einen begrifflichen Mehrwert zu schaffen. Wir werden z. B. aus den „einfachen" Begriffen *Gewicht* und *Größe* den für unsere ernährungswissenschaftliche Studien relevanten *Body Mass Index* nach der folgenden Formel berechnen:

$$\frac{Gewicht\ (in\ kg)}{Größe^2\ (in\ m)}$$

Im aktuellen Kapitel werden Sie häufig benötigte SPSS-Kommandos zur Datentransformation kennenlernen. Diese wirken sich auf ein Datenblatt aus, wo entweder neue Variablen aufgenommen oder vorhandene Variablen verändert werden.

Per Voreinstellung werden dabei *alle Fälle gleichermaßen* behandelt. Man kann die Ausführung einer Datentransformation aber auch von einer Bedingung abhängig machen, sodass nicht mehr alle Fälle davon betroffen sind. Diese Möglichkeit werden wir dazu verwenden, die MD-Behandlung bei den Motiv-Variablen in Ordnung zu bringen, indem wir für die Fälle mit

MOTIV1 = MOTIV2 = ... = ANDERE = 0

bei allen Motiv-Variablen den Wert 0 durch den MD-Indikator SYSMIS ersetzen.

SPSS unterstützt Transformationen für Variablen beliebigen Typs. Wir beschränken uns jedoch auf die besonders wichtigen numerischen Variablen.

7.1 Vorbemerkungen

7.1.1 Transformationsprogramm

In Abschnitt 4.2.4 wurde vorgeschlagen, zu jedem Projekt ein SPSS-Transformationsprogramm zu erstellen, dessen Aufgabe darin besteht, ausgehend von der Rohdatendatei alle Fertigvariablen zu entwickeln, die im weiteren Verlauf wiederholt benötigt werden. *Alle* potentiell relevanten Variablen (roh oder fertig) sollen in eine erweiterte Datendatei gesichert werden, die sich für alle Auswertungsarbeiten eignet.[2] Mit Rücksicht auf diese Idee haben wir die bislang existierende

---

[1] Das Recodieren ist keine zwingende Voraussetzung für die Berechnung des Optimismus-Schätzwerts, hat aber erhebliche Vorteile, indem es z. B. die Berechnung des Optimismus-Schätzwerts vereinfacht und die Möglichkeit zu einer Skalenanalyse (mit Berechnung der internen Konsistenz) eröffnet.

[2] Unter gewissen, am ehesten in großen Projekten anzutreffenden Umständen kann es sinnvoll bzw. notwendig sein, die auszuwertenden Daten in *mehreren* Dateien bereitzuhalten. Werden die Variablen oder Fälle einer Tabelle auf mehrere Dateien verteilt, kann es leicht zu dem Problem kommen, dass sich die in einer Analyse zu vergleichenden Fälle oder Variablen in verschiedenen Dateien befinden. Treten in einem Projekt mehrere Entitäten auf (z. B. Kunden und Mitarbeiter), werden natürlich entsprechend viele Datendateien benötigt.

Datendatei mit **kfar.sav** (*r* für *roh*) bezeichnet. Im Namen der Fertigdatendatei werden wir das **r** weglassen.

Wir werden im Verlauf des aktuellen Kapitels das SPSS-Transformationsprogramm zu unserem KFA-Projekt erstellen, indem wir passend konfigurierte Dialogboxen mit dem Schalter **Einfügen** quittieren, um die äquivalenten SPSS-Kommandos in einem Syntaxfenster zu sammeln (vgl. Kapitel 6). Dabei ist eine hohe Sorgfalt erforderlich, weil Fehler im Transformationsprogramm schwerwiegende Konsequenzen für die weitere Projektarbeit haben können.

Das fertige Transformationsprogramm wird anschließend ausgeführt, wobei die Fertigdatendatei entsteht. Außerdem wird das Transformationsprogramm in einer Datei gespeichert, damit es z. B. nach einer Stichprobenerweiterung erneut ausgeführt werden kann. Als Dateinamen werden wir **kfat.sps** verwenden.

Man kann alle erforderlichen Transformationen auch durch direkte Ausführung der zuständigen Dialogboxen erledigen (Schalter **OK**). Diese Arbeitsweise ist zweifellos für Anfänger leichter zu handhaben als die programmorientierte Methode, hat aber folgende Nachteile:

- Beim sukzessiven manuellen Modifizieren der Datendatei geht bei größeren Projekten leicht der Überblick verloren. Z. B. weiß irgendwann von einer abgeleiteten Variablen niemand mehr, in welchen Zwischenschritten sie aus welchen Rohvariablen berechnet worden ist. Spätestens nach dem Auftreten unplausibler Ergebnisse muss die *tatsächlich* angewendete Berechnungsvorschrift als mögliche Fehlerquelle überprüft werden. Bei der Verwendung eines Transformationsprogramms ist die Entstehung der abgeleiteten Variablen verlässlich **dokumentiert**.
- Sind **Wiederholungen von Datentransformationen** erforderlich, dann müssen diese komplett neu spezifiziert werden. Solche Wiederholungen sind z. B. nach einer Datenkorrektur fällig, weil SPSS abgeleitete Variablen **nicht** automatisch anpasst, wenn sich Werte der Ursprungsvariablen ändern. Nach einer Korrektur bei einer Rohvariablen müssen Sie also alle Datentransformationen wiederholen, bei denen diese Rohvariable direkt oder indirekt beteiligt ist. Ein weiterer potentieller Anlass für die Wiederholungen von Datentransformationen ist die Erweiterung der Stichprobe.

Die für ein Projekt erforderlichen Datentransformationen in Form von SPSS-Kommandos zu konservieren, lohnt sich meistens, denn:

- Die einzelnen Anweisungen sind relativ komplex und damit ebenso fehleranfällig wie **zeitaufwändig**.
- Es ist relativ wahrscheinlich, dass die gesamte Anweisungsfolge **wiederholt** durchgeführt werden muss (z. B. bei entdeckten Fehlern in den Rohvariablen oder bei einer Stichprobenerweiterung).
- Die Anweisungen zur Datentransformation sind **dokumentationspflichtig**.

7.1.2 Datensicherheit

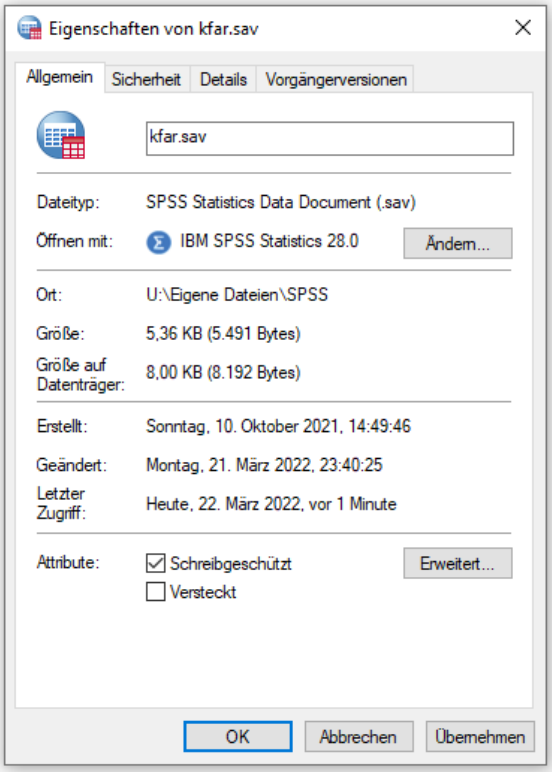
Die Rohdaten können nach der sorgfältigen Datenerfassung und -prüfung als korrekt gelten. Sichern Sie den erreichten Stand, indem Sie die Rohdaten in mindestens **zwei** Dateien speichern (möglichst auf verschiedenen Datenträgern). Bei besonders wichtigen Daten sollte sogar die **3-2-1 – Regel** für ein professionelles Backup beachtet werden:

- Es sind mindestens **drei** Kopien vorhanden.
- Diese befinden sich auf mindestens **zwei** verschiedenen Datenträgern (z. B. lokale Festplatte und Cloud-Speicher).
- **Eine** Backup-Kopie befindet sich an einem externen Ort, um vor Elementarschäden (z. B. durch Feuer oder Wasser) am primären Speicherort geschützt zu sein (z. B. Cloud-Speicher).

Für eine zum Backup gehörende Datendatei ist es sinnvoll, das Schreibschutzattribut zu setzen, was für eine in SPSS geöffnete Datendatei über den folgenden Menübefehl des Datenfensters geschehen kann:

Datei > Datei als schreibgeschützt markieren

Alternativ lässt sich der Schreibschutz per Windows-Explorer im Eigenschaftsdialog einer betroffenen Datei aktivieren, z. B.:



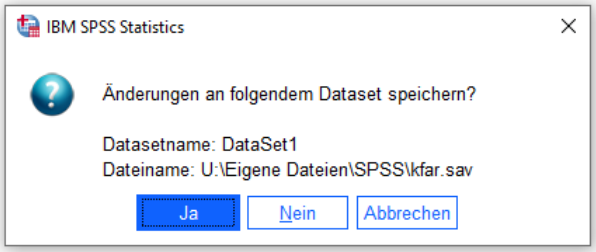
Vor der geplanten Änderung einer Datei muss der Schreibschutz wieder aufgehoben werden.

Mit dem SPSS-Kommando PERMISSIONS lässt sich der Schreibschutz auch per Syntax (de)aktivieren. Man muss das Kommando manuell erstellen, was aber aufgrund seines einfachen Aufbaus keine Schwierigkeiten macht, z. B.:

```
PERMISSIONS FILE = 'U:\Eigene Dateien\SPSS\kfar.sav' /PERMISSIONS READONLY.
PERMISSIONS FILE = 'U:\Eigene Dateien\SPSS\kfar.sav' /PERMISSIONS WRITEABLE.
```

Nach seiner Fertigstellung verdient auch das Transformationsprogramm eine sorgfältige Aufbewahrung.

Wenn Sie beim Verlassen von SPSS gefragt werden, ob Sie ein Daten- oder Syntaxfenster sichern wollen, sollten Sie sorgfältig prüfen, ob beim betroffenen Dokument während der Sitzung tatsächlich nur geplante Veränderungen stattgefunden haben.



Antworten Sie im Zweifelsfall mit **Nein**. Möglicherweise haben Sie unbeabsichtigte Veränderungen vorgenommen. Diese Fehler sollten dann auf keinen Fall auf einen permanenten Datenträger (z B. auf eine Festplatte) geschrieben werden.

7.1.3 Initialisierung neuer numerischer Variablen

Wenn Sie in einer Datentransformationsanweisung die Erstellung einer *neuen* numerischen Variablen anfordern, dann wird die (Fälle × Variablen) - Datenmatrix im aktiven Datenblatt um eine Spalte erweitert (am rechten Rand). SPSS initialisiert zunächst die neue Variable, indem alle Fälle den MD-Indikator SYSMIS als vorläufigen Wert erhalten. Gelingt anschließend für einen Fall die Ermittlung der neuen Variablenausprägung, dann wird der Initialwert entsprechend ersetzt. Anderenfalls bleibt SYSMIS stehen, sodass der betroffene Fall bei allen Berechnungen mit der neuen Variablen ausgeschlossen wird.

7.2 Alte Werte einer Variablen auf neue abbilden (Umcodieren)

Mit den Befehlen zum **Umcodieren** aus dem Menü **Transformieren** bzw. mit dem äquivalenten RECODE-Kommando können die Werte einer bestehenden Variablen in neue Werte überführt werden. Man kann die Ausgangsvariable verändern oder eine neue Variable mit einem recodierten Wertevektor erstellen.[1]

7.2.1 Das praktische Vorgehen am Beispiel einer Gruppenbildung

Da wir im Kapitel 7 das KFA-Transformationsprogramm sukzessive aufbauen wollen, benötigen wir eine Arbeitsdatei mit unseren Rohdaten. Öffnen Sie daher nötigenfalls über den Menübefehl

Datei > Öffnen > Daten

die Rohdatendatei **kfar.sav**, wobei ein benanntes, mit der Rohdatendatei verbundenes Datenblatt entsteht.

Um das Umcodieren zu üben, wählen wir ein mäßig sinnvolles Beispiel aus unserer Studie: Wir konstruieren unter dem Namen DEKADE eine vergröberte Variante der Jahrgangsvariablen, bei der alle in den 60'er Jahren geborenen Personen den Wert 1 und alle in den 70'er Jahren geborenen Personen den Wert 2 erhalten sollen. Wie man sich anhand der Häufigkeitstabelle zur Variablen GEBJ

[1] Der Vollständigkeit halben soll noch eine dritte, seltener benötigte Option erwähnt werden: Man kann auch eine vorhandene Variable mit dem recodierten Wertevektor der Ausgangsvariablen überschreiben.

Geburtsjahr

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1961	1	3,2	3,2	3,2
	1964	1	3,2	3,2	6,5
	1965	1	3,2	3,2	9,7
	1966	2	6,5	6,5	16,1
	1967	7	22,6	22,6	38,7
	1968	3	9,7	9,7	48,4
	1969	2	6,5	6,5	54,8
	1970	7	22,6	22,6	77,4
	1972	3	9,7	9,7	87,1
	1974	2	6,5	6,5	93,5
	1975	2	6,5	6,5	100,0
	Gesamt	31	100,0	100,0	

überzeugen kann, ist damit für alle Fälle in unserer Stichprobe ein DEKADE-Wert definiert. Mit Hilfe der neuen Variablen kann man z. B. den Einfluss des Geburtsjahrzehnts auf diverse abhängige Variablen untersuchen, wobei man sich von der Informationsreduktion (im Vergleich zu GEBJ) keinen allzu großen Nutzen versprechen sollte.

Es hat sich mittlerweile herumgesprochen, dass eine informations-reduzierende und willkürliche Gruppenbildung selten die Wissenschaft voranbringt. Von diesen Problemen ist insbesondere die beliebte Median-Dichotomisierung betroffen (siehe z. B. MacCallum et al. 2002).

Bei der geplanten Recodierung wird die (Fälle × Variablen) - Datenmatrix der Arbeitsdatei um eine neue Variable erweitert, die folgendermaßen aus der vorhandenen Variablen GEBJ entsteht:

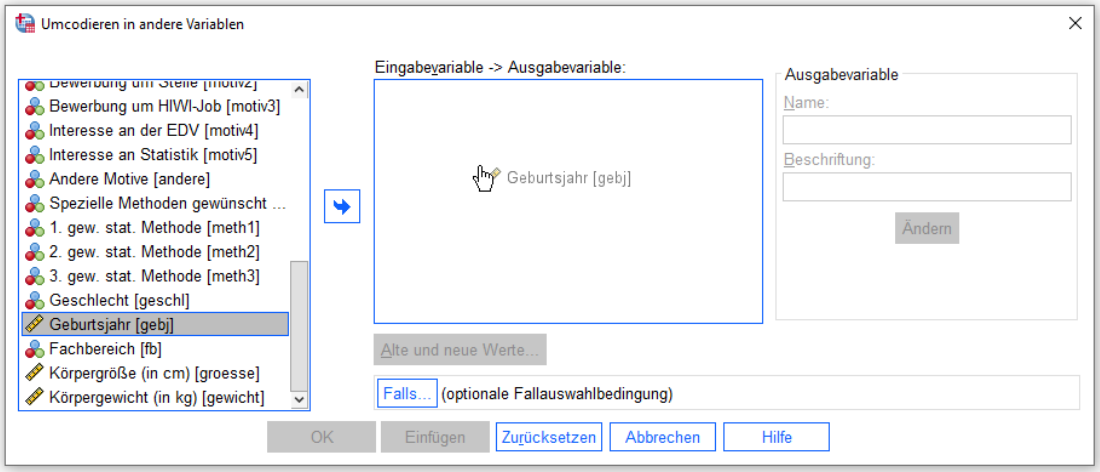
GEBJ		DEKADE
1969	→	1
1970	→	2
1969	→	1
1967	→	1
.		.
.		.
.		.
1972		2
1968	→	1
1967	→	1
1967	→	1

Wählen Sie den Menübefehl

Transformieren > Umcodieren in andere Variablen

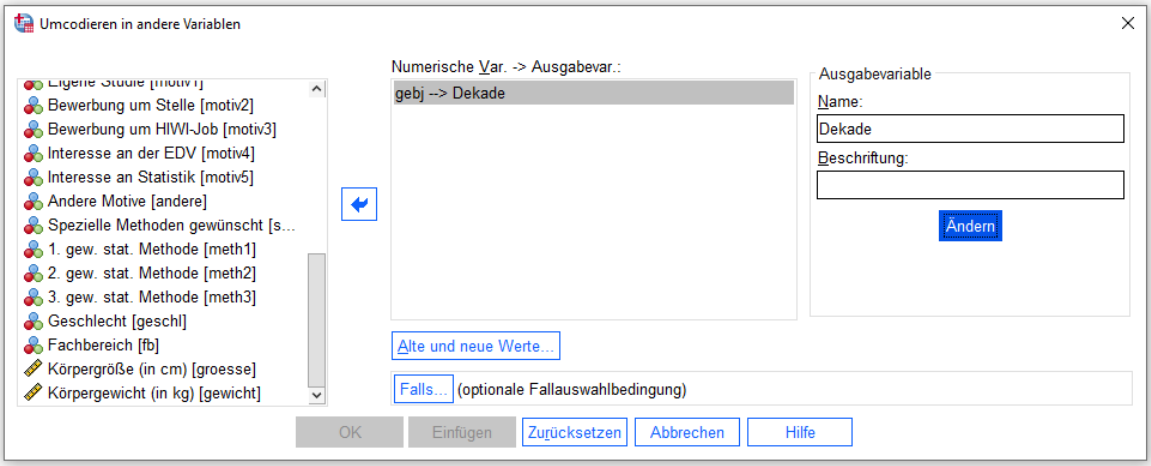
und machen Sie folgendermaßen weiter:

- Befördern Sie in der Dialogbox **Umcodieren in andere Variablen** die Variable GEBJ in das Feld **Eingabevariable -> Ausgabevariable**. Statt den Schalter zu benutzen, können Sie in SPSS solche Transportaufgaben auch per Drag & Drop (Ziehen und Ablegen) erledigen:



- Tragen Sie im Bereich **Ausgabevariable** den gewünschten **Namen** der neu zu erzeugenden Variablen ein.
- Optional kann eine **Beschriftung** ergänzt werden. Wir verzichten darauf, sodass der Variablenname *Dekade* auch zur Beschriftung der Ausgabe verwendet wird. In dieser Situation sollte man beim Variablennamen auf die korrekte Schreibweise achten.
- Klicken Sie auf **Ändern**.

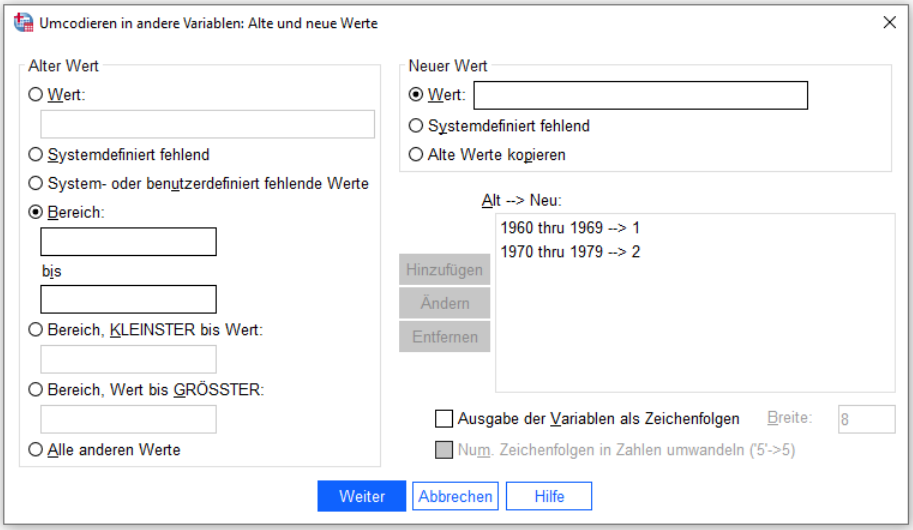
Danach müsste Ihre Dialogbox ungefähr so aussehen:



Legen Sie nun die Abbildungsregeln fest:

- Aktivieren Sie mit dem Schalter **Alte und neue Werte** die Subdialogbox **Umcodieren in andere Variablen: Alte und neue Werte**.
- Geben Sie im Rahmen **Alter Wert** den **Bereich** von 1960 bis 1969 an, und wählen Sie als zugehörigen **Neuen Wert** die 1.
- Komplettieren Sie die Definition der ersten Abbildungsvorschrift mit **Hinzufügen**.
- Vereinbaren Sie analog die Zuordnungsvorschrift: „1970 bis 1979 → 2“.

Jetzt müssten Sie dieses Bild sehen:

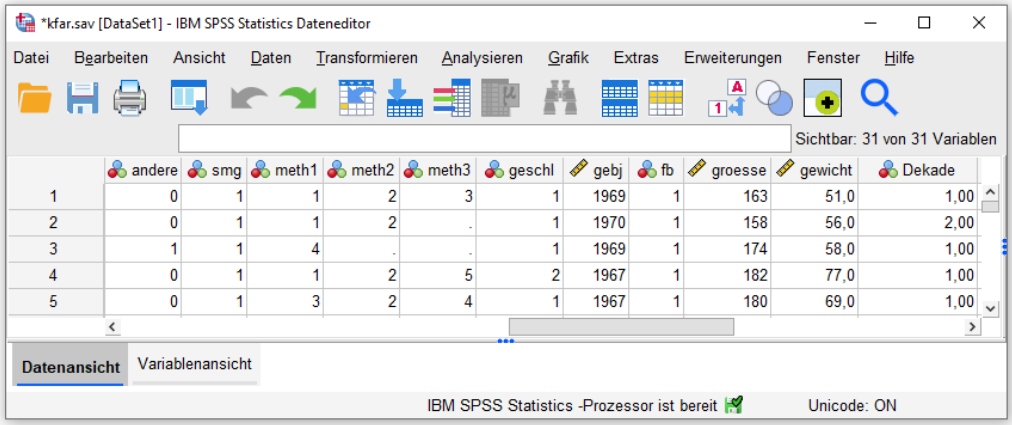


Damit ist die geplante Recodierung vollständig festgelegt. Quittieren Sie die Subdialogbox mit **Weiter**. Da wir das KFA-Transformationsprogramm sukzessive aufbauen wollen, müssen Sie nun in der Dialogbox **Umcodieren in andere Variablen** auf den Schalter **Einfügen** klicken, um die implizit definierten Kommandos zu produzieren. Wir erhalten ein Syntaxfenster mit dem folgenden Inhalt:

```
DATASET ACTIVATE DataSet1.
RECODE gebj (1960 thru 1969=1) (1970 thru 1979=2) INTO Dekade.
EXECUTE.
```

Das erste Kommando macht das **DataSet1** zur Arbeitsdatei und soll verhindern, dass die nachfolgenden Kommandos auf ein ungeeignetes Datenblatt treffen. Sein Nutzen ist aber fragwürdig, weil bei einer späteren Wiederverwendung des Programms nicht unbedingt ein Datenblatt mit dem Namen **DataSet1** und einem passenden Inhalt vorhanden ist. Hinter das RECODE-Kommando, das die Umcodierung bewirkt, hat SPSS noch ein EXECUTE gesetzt, dessen Rolle in Abschnitt 7.3 erläutert wird.

Unabhängig von den guten Argumenten für die Erstellung eines Transformationsprogramms gibt es in Ihrer aktuellen Lernphase einen Grund, die obige **Umcodieren**-Dialogbox per **OK**-Schalter zu quittieren oder die zugehörigen Kommandos jetzt schon ausführen zu lassen: Sie können den Effekt auf die Arbeitsdatei sofort beobachten, statt bis zum Ausführen des kompletten Transformationsprogramms warten zu müssen. Weil keine Konflikte mit unserer langfristigen Strategie zu befürchten sind, kehren wir (z. B. über den Symbolschalter ) zur **Umcodieren**-Dialogbox zurück und quittieren sie mit **OK**. Anschließend befindet sich am rechten Rand der Arbeitsdatei die neue Variable DEKADE:



Um die Attribute der neuen Variablen DEKADE (z. B. Messniveau, Anzahl der Dezimalstellen) kümmern wir uns später.

7.2.2 Technische Details

Obwohl das Umcodieren eine simple Datentransformation ist, sind bei der praktischen Anwendung doch einige technische Details zu beachten:

- Über den Dialog **Umcodieren in andere Variablen** erhält man neue Variablen als recodierte Varianten der Ausgangsvariablen. Über den Dialog **Umcodieren in dieselben Variablen** werden vorhandene Variablen verändert.
- Bei einem Einsatz eines Umcodieren-Dialogs kann man *beliebig viele* Variablen gleichzeitig recodieren.
- Bei der Spezifikation der alten Werte, die auf einen neuen Wert abgebildet werden sollen, kann man angeben:
  - Einen einzelnen **Wert**
  - **Systemdefiniert fehlend**
    So ist es also möglich, den systemseitigen MD-Indikator SYSMIS durch einen anderen Wert zu ersetzen.
  - **System- oder benutzerdefiniert fehlende Werte**
    Alle MD-Indikatoren werden ersetzt.
  - Den **Bereich** von einem ersten Wert bis zu einem zweiten Wert (inklusive Grenzwerte)
    Bei allen Bereichen (auch den anschließend behandelten halboffenen Bereichen) ist zu beachten, dass im Bereich befindliche *benutzerdefinierte* MD-Indikatoren einbezogen werden. Dies lässt sich z. B. mit der *davor positionierten* Ersetzungsvorschrift MISSING = SYSMIS verhindern. Um diese Vorschrift per Dialogbox zu erzeugen, wählt man als alten Wert **System- oder benutzerdefiniert fehlende Werte** und als neuen Wert **Systemdefiniert fehlend** (siehe unten). SPSS platziert Ersetzungsvorschriften mit einem Bereich alter Werte automatisch *hinter* alle Ersetzungsvorschriften mit …
    - einem einzelnen alten Wert
    - oder **Systemdefiniert fehlend**
    - oder **System- oder benutzerdefiniert fehlende Werte**.
  - Den **Bereich** vom kleinsten Wert in der Stichprobe bis zu einem bestimmten Wert (inklusive Grenzwert)

  - Den **Bereich** von einem bestimmten Wert bis zum größten Wert in der Stichprobe (inklusive Grenzwert)
  - **Alle anderen Werte**
    Damit sind alle in keiner anderen Ersetzungsvorschrift genannten Werte angesprochen (inklusive MD-Indikatoren, auch SYMIS). Um zu verhindern, dass auch MD-Indikatoren einbezogen werden, muss man diese Werte zuvor in einer speziellen Ersetzungsvorschrift behandeln, z. B. MISSING = SYSMIS (siehe Erläuterung zum **Bereich**). **Alle anderen Werte** kann nur in *einer* Ersetzungsvorschrift angegeben werden. Diese wird von SPSS in der Liste aller Ersetzungsvorschriften automatisch an die letzte Stelle gesetzt und damit bei der Ausführung zuletzt abgearbeitet.

- Als neuen Wert, auf den die alten Werte einer Ersetzungsvorschrift abgebildet werden sollen, können Sie angeben:
  - Einen konkreten **Wert**
  - **Systemdefiniert fehlend**
    Dann werden alle zugehörigen alten Werte durch SYSMIS ersetzt.
  - **Alte Werte kopieren**
    Diese Möglichkeit steht nur beim Umcodieren in *andere* Variablen zur Verfügung und bewirkt für die zugehörigen alten Werte eine unveränderte Übernahme. Dies ist besonders nützlich, wenn die alten Werte mit **Alle anderen Werte** spezifiziert worden sind.
- Man kann beliebig viele Ersetzungsvorschriften definieren.
- Bei jedem Fall wird nur die *erste zutreffende* Ersetzungsregel angewendet. Eine Regel trifft bei einem Fall zu, wenn ihre Menge alter Werte den aktuellen Wert des Falles enthält. Alle weiteren (eventuell ebenfalls zutreffenden) Ersetzungsregeln werden bei einem bereits behandelten Fall ignoriert. Die Definitionsreihenfolge der Ersetzungsregeln ist also relevant.
- Haben die Ausprägungen einer Variablen potentiell viele Dezimalstellen (z. B. 3,1415926), dann kann man auf einfache Weise *lückenlos* aufeinander folgende Intervalle definieren, indem man dieselbe Zahl zweimal als Intervallgrenze verwendet, z. B.

  ```
  (1 thru 5 = 1) (5 thru 10 = 2)
  ```

  Im Beispiel resultiert die Abbildung:

  $[1, 5] \rightarrow 1, (5, 10] \rightarrow 2$

  Bei umgekehrter Anordnung derselben Ersetzungsregeln

  ```
  (5 thru 10 = 2) (1 thru 5 = 1)
  ```

  erhält man hingegen die Abbildung:

  $[1, 5) \rightarrow 1, [5, 10] \rightarrow 2$
- Wenn beim Umcodieren in andere Variablen eine *neue* Variable entsteht, dann wird diese zunächst mit dem Wert SYSMIS initialisiert (vgl. Abschnitt 7.1.3). Wird der alte Wert eines Falles in keiner Übersetzungsregel angesprochen, dann bleibt bei der neuen Variablen der Initialisierungswert SYSMIS stehen. Dies würde in Beispiel aus Abschnitt 7.2.1 etwa einem 1980 geborenen Untersuchungsteilnehmer passieren.
- Benutzerdefinierte MD-Indikatoren werden wie gültige Werte behandelt! Ist z. B. beim **Umcodieren in dieselben Variablen** für eine Variable der Wert 99 ein benutzerdefinierter MD-Indikator, und wird die 99 recodiert zur 98, dann **bleibt** die 99 ein MD-Indikator der Variablen, und die 98 wird **nicht** zum MD-Indikator. Eventuell muss also nach der Recodierung die Variablendeklaration angepasst werden. Oben wurde schon erläutert, wie man bei *Berei-*

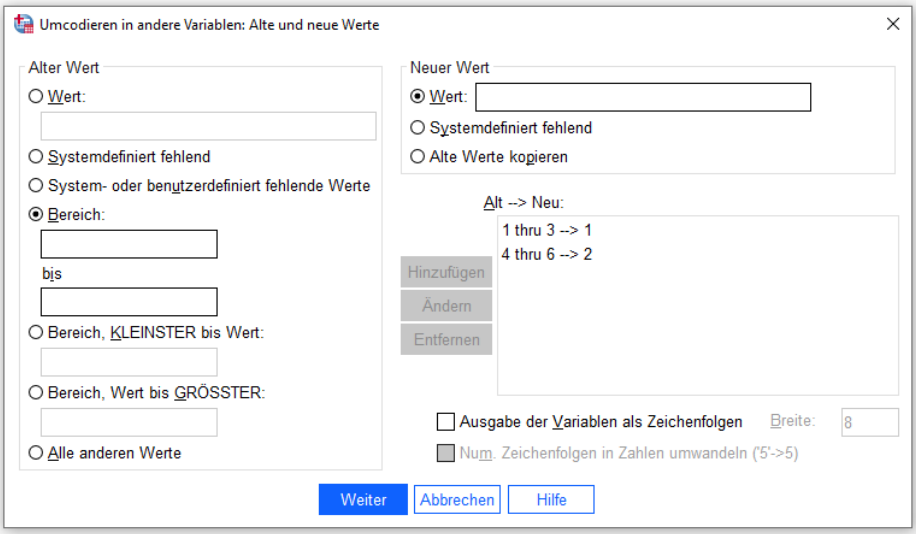
chen alter Werte bzw. bei der Argumentspezifikation **Alle anderen Werte** die unerwünschte Mitbehandlung von (benutzerdefinierten) MD-Indikatoren verhindert.

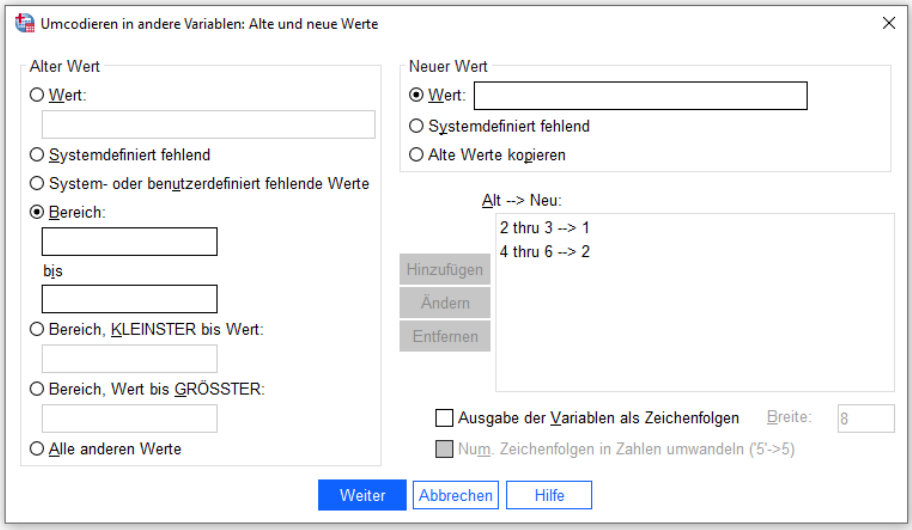
Gelegentlich kommt es vor, dass numerische Merkmalsausprägungen versehentlich in einer Zeichenfolgenvariablen landen und infolgedessen eine zusätzliche Variable mit den numerischen Werten erstellt werden muss. Dies kann per RECODE-Kommando über das Schlüsselwort CONVERT erfolgen, das nur in Verbindung mit dem Schlüsselwort INTO zulässig ist, z. B.:

```
recode astr (convert) into anum.
```

7.2.3 Übungen

1) In den beiden folgenden Dialogboxen, die wir in unserem Projekt *nicht* ausführen wollen, wird jeweils eine Umcodierung der Fachbereichsvariablen (FB) in eine andere, neue Variable spezifiziert. Hätten die beiden Dialogboxen denselben Effekt?





2) Bei unserem LOT-Fragebogen wurden die Fragen 3, 4, 5, und 12 aus messtechnischen Gründen umgepolt (negativ formuliert). Indem eine optimistische Antwort abwechselnd durch Zustimmung oder Ablehnung zum Ausdruck kommt, wird verhindert, dass systematische Ja- oder Neinsager einen extremen Optimismuswert erhalten. Bevor wir aus den LOT-Items durch Mittelwertsbildung einen Optimismus-Schätzwert berechnen können, müssen die negativ gepolten Variablen folgendermaßen umcodiert werden:[1]

$$\begin{aligned} 5 &\rightarrow 1 \\ 4 &\rightarrow 2 \\ 2 &\rightarrow 4 \\ 1 &\rightarrow 5 \end{aligned}$$

Wählen Sie den Menübefehl:

Transformieren > Umcodieren in dieselben Variablen

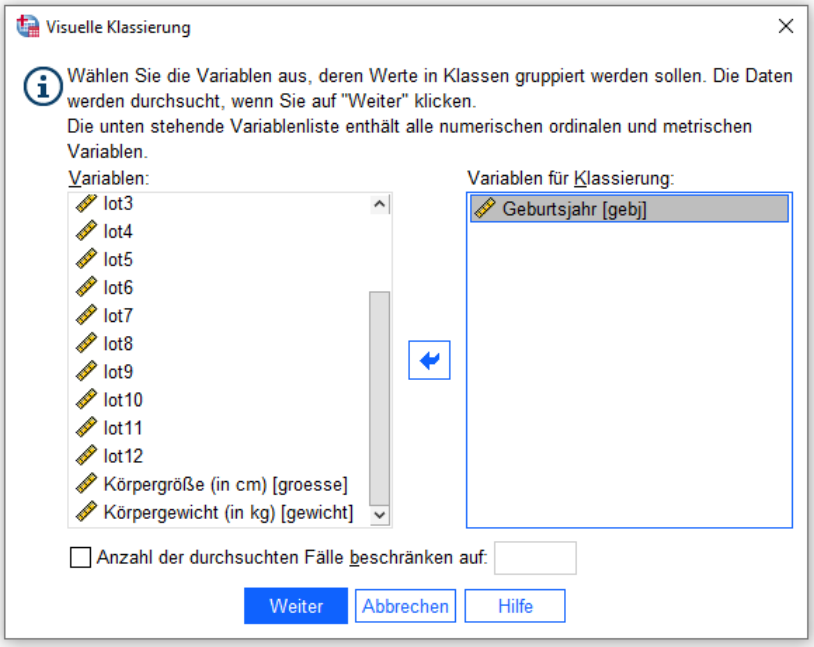
Quittieren Sie die bearbeitete Dialogbox **Umcodieren in dieselben Variablen** nicht mit **OK**, sondern mit **Einfügen**, damit das zugehörige RECODE-Kommando in das Syntaxfenster eingetragen wird, in dem wir gerade unser Transformationsprogramm aufbauen. Machen Sie sich klar, warum die Abbildungsvorschrift „3 → 3“ beim Umcodieren **in dieselben Variablen** überflüssig ist, beim Umcodieren in andere, neue Variablen aber unbedingt erforderlich wäre.

7.2.4 Visuelles Klassieren

Über den Menübefehl

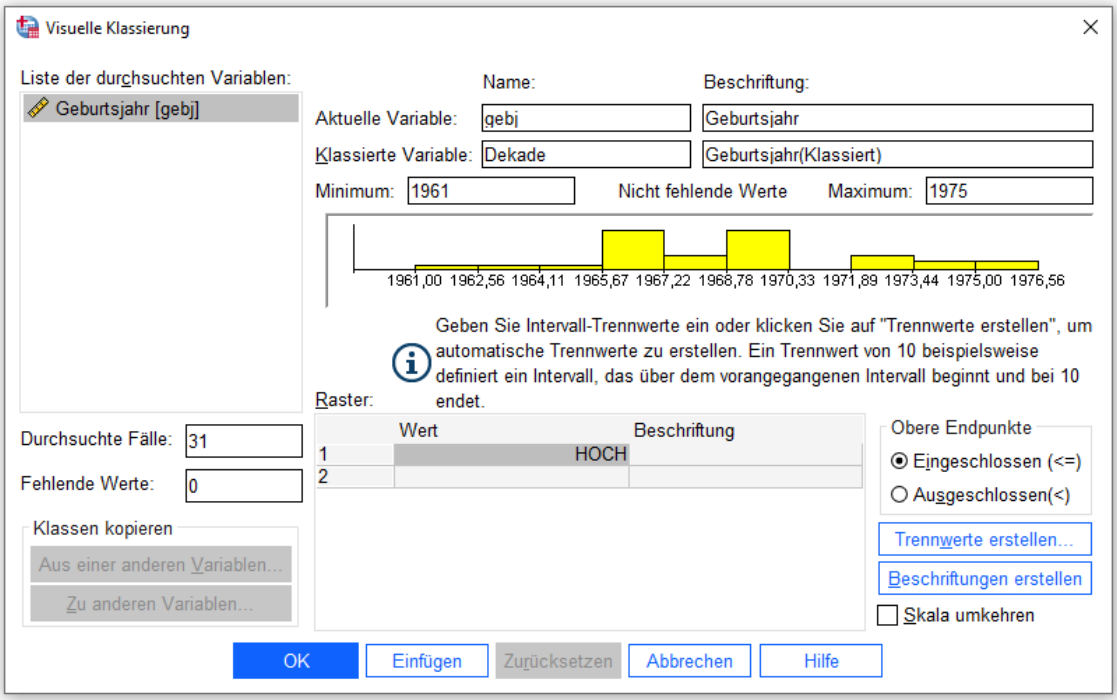
Transformieren > Visuelle Klassierung

wird ein Assistent zur Unterstützung der Klassenbildung gestartet. Im ersten Schritt wählt man die Ausgangsvariable, z. B.:



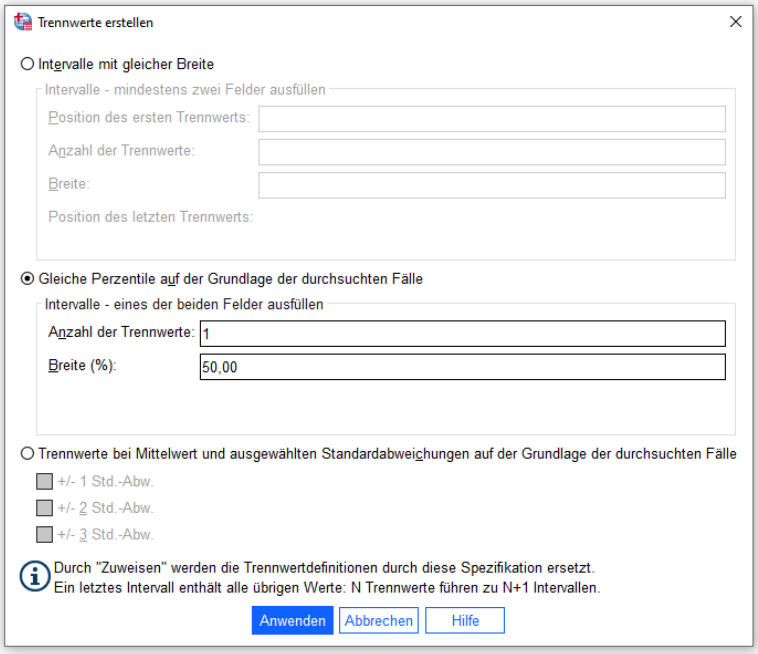
[1] Das Recodieren ist keine zwingende Voraussetzung für die Berechnung des Optimismus-Schätzwerts, hat aber erhebliche Vorteile, indem es z. B. die Berechnung des Optimismus-Schätzwerts vereinfacht und die Möglichkeit zu einer Skalenanalyse (mit Berechnung der internen Konsistenz) eröffnet.

Im nächsten Dialog gibt man den **Namen** und (optional) eine **Beschriftung** für die Zielvariable an:

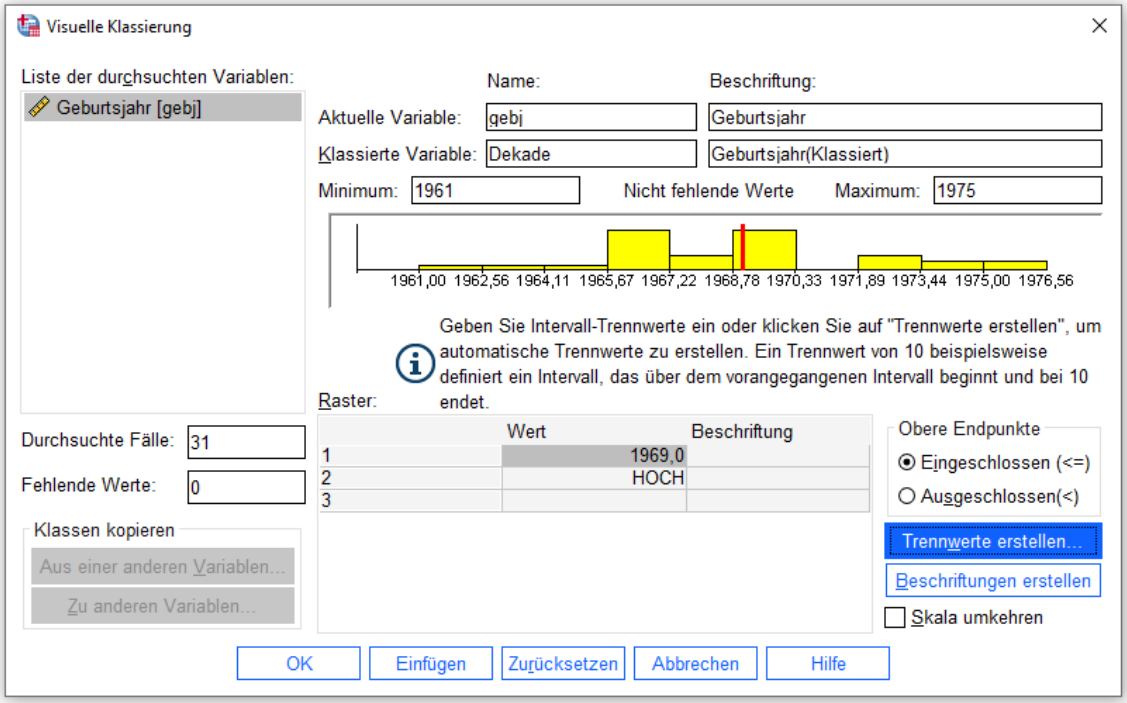


Ein Histogramm gibt eventuell Anregungen zur Aufteilung, und mit dem Kontrollkästchen **Skala umkehren** könnte man im Beispiel dafür sorgen, dass die Klasse mit den niedrigsten Geburtsjahren den höchsten Wert bei der Zielvariablen erhält.

Nach einem Klick auf den Schalter **Trennwerte erstellen** kann man im folgenden Dialog z. B. die Bildung von zwei annähernd gleich stark besetzten Klassen veranlassen:



Im Hauptdialog wird nun der Trennwert angezeigt, z. B.:



Über den Schalter **Einfügen** erhält man u. a. das vom Assistenten erstellte RECODE-Kommando, z. B.:

```
RECODE  gebj (MISSING=COPY) (LO THRU 1969=1) (LO THRU HI=2) (ELSE=SYSMIS) INTO Dekade.
```

Es führt im Beispiel zum selben Ergebnis wie unsere eigene Syntax (siehe Abschnitt 7.2.1) und demonstriert, wie man durch die geschickte Anordnung von Abbildungsvorschriften mit überlappenden Intervallen alter Werte dafür sorgt, dass *alle* alten Werte versorgt werden.

7.3 Zur Rolle des EXECUTE-Kommandos

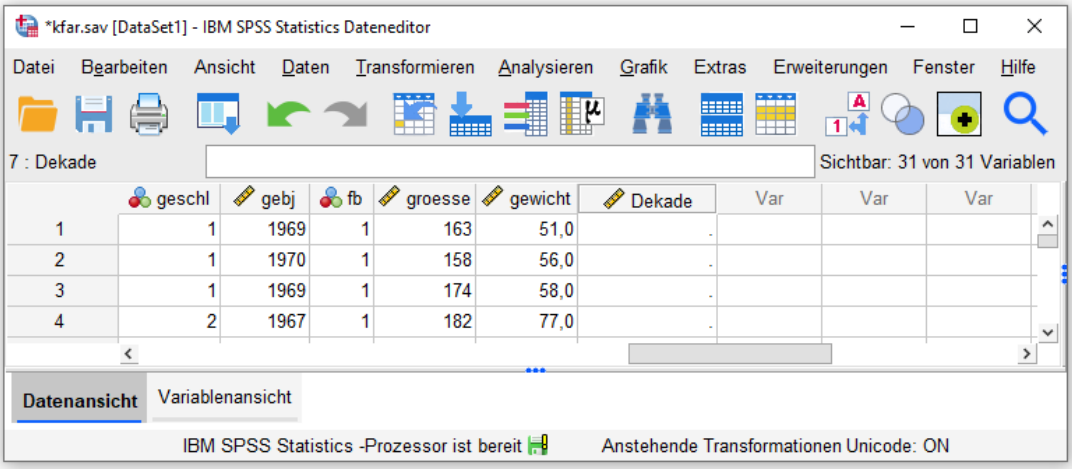
Wenn Sie eine **Umcodieren**-Dialogbox mit **OK** quittieren, dann führt SPSS per Voreinstellung die angeforderte Recodierung sofort in der Arbeitsdatei aus. Obwohl dieses Verhalten sehr naheliegend erscheint, gibt es doch eine erwägenswerte Alternative. Zum Recodieren muss SPSS nämlich die Arbeitsdatei vollständig durchlaufen, was bei einer großen Stichprobe durchaus einige Zeit in Anspruch nehmen kann. Bei einer nächsten und übernächsten Transformationsanweisung (z. B. Recodierung oder Neuberechnung) ist jeweils ein weiterer Durchlauf fällig. Dabei könnte SPSS zeitsparend *alle* Transformationen in einer *einzigen* Datenpassage erledigen. Diese könnte so lange aufgeschoben werden, bis wegen der Anforderung einer Statistikprozedur das Durchackern der Daten unvermeidlich wäre. Genau in dem zuletzt beschriebenen, zeitökonomischen Sinn funktionieren seit jeher in SPSS die Transformationskommandos, die Änderungen an der Datenmatrix bewirken: Sie werden vorgemerkt und erst bei der nächsten Prozedur gemeinsam ausgeführt. Allerdings kann dieses zeitoptimierte Verhalten SPSS-Neulinge verwirren. Daher setzt die SPSS-Bedienoberfläche hinter jedes per Dialogbox produzierte Transformationskommando per Voreinstellung automatisch ein EXECUTE-Kommando, das die *sofortige Ausführung* aller offenen Transformationen erzwingt. Wenn wir z. B. eine **Umcodieren**-Dialogbox mit **OK** quittieren, verarbeitet der SPSS-Prozessor im Hintergrund ein RECODE- *und* ein EXECUTE-Kommando. Das erste bewirkt nur eine Arbeitsvorbereitung, das zweite erzwingt die

Ausführung der vorbereiteten Arbeit. Quittieren wir dieselbe Dialogbox mit **Einfügen**, dann erscheinen die beiden Kommandos in einem Syntaxfenster, z. B.:[1]

```
RECODE gebj (1960 thru 1969=1) (1970 thru 1979=2) INTO Dekade.
EXECUTE.
```

Im gerade entstehenden Transformationsprogramm sind die von SPSS produzierten EXECUTE-Kommandos in der Regel überflüssig. Aufgrund der heutzutage verfügbaren Rechenleistung lohnt es sich allerdings nur bei einer sehr großen Datei, die überflüssigen EXECUTE-Kommandos aus einem automatisch produzierten Programm zu entfernen. „Sehr groß“ ist z. B. eine Datei mit 200.000 Fällen. In dieser Situation wurde für ein Testprogramm mit 10 Transformationen und 9 überflüssigen EXECUTE-Anweisungen auf einem PC mit der Intel-CPU Core i7 860 eine Laufzeit von ca. 10 Sekunden gemessen, die sich durch Entfernen der überflüssigen EXECUTE-Kommandos auf ca. 1 Sekunde reduzieren ließ. Bei einer typischen Datendatei (mit ≤ 500 Fällen) kann man die überflüssigen EXECUTE-Kommandos ignorieren.

Beim Arbeiten mit einem Syntaxfenster kann es zu dem folgenden, frustrierenden Erlebnis kommen: Sie lassen wohlgeformte Transformationskommandos ausführen, doch in der Arbeitsdatei stellt sich nur ein partieller Erfolg ein. Zwar erscheinen die neu anzulegenden Variablen, doch haben alle Fälle den Wert SYSMIS, z. B.:



Die Ursache ist dann meist: Sie haben nach den Transformationskommandos noch kein Prozedur- oder EXECUTE-Kommando ausführen lassen, sodass SPSS zwar die neue Variablen initialisiert, aber noch keine Werte ermittelt hat. In dieser Situation wird in der Statuszeile angezeigt, dass **anstehende Transformationen** vorhanden sind. Sie können deren Ausführung erzwingen, indem Sie in einem Syntaxfenster ein EXECUTE-Kommando abschicken oder den folgenden Menübefehl wählen:

---

[1] Man kann nach

Bearbeiten > Optionen > Daten

im Rahmen **Optionen für Transformieren und Zusammenfügen** mit der Option **Werte vor Verwendung berechnen** die voreingestellte EXECUTE-Inflation abstellen. Dann zeigt SPSS das oben beschriebene zeitoptimierte Verhalten, führt also z. B. nach dem Quittieren einer **Umcodieren**-Dialogbox mit **OK** das zugrundeliegende RECODE-Kommando zunächst noch *nicht* aus, sondern reiht es in die Warteschlange der offenen Transformationen ein. Diese werden vom SPSS-Prozessor erst dann ausgeführt, wenn er ein Prozedur- oder ein EXECUTE-Kommando erhält.

Transformieren > Anstehende Transformationen ausführen

Es soll nicht verschwiegen werden, dass hier für SPSS-Neulinge Schwierigkeiten auftauchen, die bei rein interaktiver Dialogboxnutzung und voreingestelltem EXECUTE-Einsatz nicht entstehen können.

Für angehende SPSS-Profis soll noch erwähnt werden, dass man aus einer Sequenz von Transformationskommandos nicht immer schadlos alle EXECUTE-Kommandos mit Ausnahme des letzten entfernen darf. In dem folgenden (manuell erstellten) Beispiel wird mit Hilfe des Transformationskommandos SELECT IF jeder zweite Fall aus der Arbeitsdatei entfernt:

```
compute nr = $casenum.
execute.
select if (mod(nr,2) = 1).
execute.
```

Lässt man das erste EXECUTE-Kommando weg, werden jedoch *alle* Fälle mit Ausnahme des ersten entfernt.

7.4 Berechnung von Variablen nach mathematischen Formeln

In der Dialogbox **Variable berechnen** bzw. im äquivalenten COMPUTE-Kommando wird ein numerischer Ausdruck (z. B. GROESSE - 100) definiert und einer Ergebnisvariablen zugewiesen. Dabei kann man eine *neue* Variable erzeugen oder eine vorhandene verändern.

7.4.1 Beispiel

Wir werden später anhand unserer Stichprobe untersuchen, ob die Trierer Studierenden im Mittel wenigstens das folgende Idealgewicht auf die Waage bringen (Nullhypothese)

$$\textit{Gewicht (in kg)} \stackrel{!}{=} \textit{Größe (in cm) - 100}$$

oder ob sie relativ zu dieser Formel zu leicht sind (Alternativhypothese). Mit den Symbolen μ_R für den Realgewichtsmittelwert und μ_I für den Idealgewichtsmittelwert kann man die beiden konkurrierenden Hypothesen so notieren:

$$H_0 : \mu_R \geq \mu_I \quad \text{versus} \quad H_1 : \mu_R < \mu_I$$

Zur Klärung der Frage durch einen t-Test für verbundene Stichproben muss die Arbeitsdatei um eine neue Variable, z. B. IDGEW genannt, erweitert werden, deren Werte nach der Formel

GROESSE - 100

aus der Körpergröße zu berechnen sind. Anschließend enthält die (Fälle × Variablen)-Datenmatrix in der Arbeitsdatei u. a. die beiden folgenden Variablen:

GROESSE	IDGEW
163	63
158	58
174	74
182	82
.	.
.	.
.	.
176	76
176	76
170	70
169	69

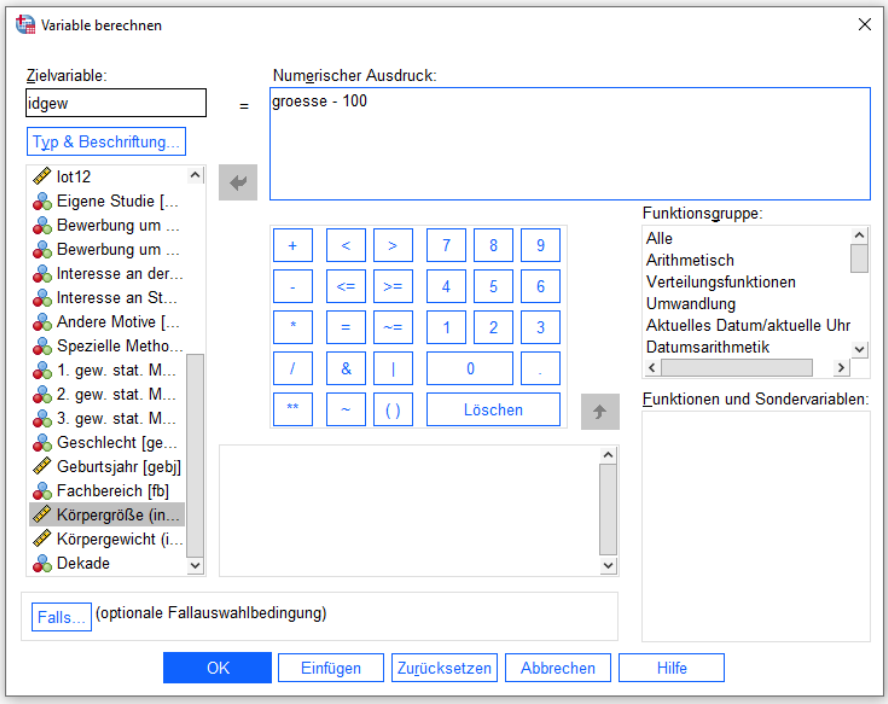
Starten Sie zum Definieren der neuen Variablen die Dialogbox **Variable berechnen** mit:

Transformieren > Variable berechnen

Tragen Sie zunächst im Feld **Zielvariable** den Namen für die zu erstellende Variable ein (IDGEW), und schreiben Sie dann in das Feld **Numerischer Ausdruck** die Definitionsvorschrift (GROESSE - 100), wobei einige Schreibhilfen zur Verfügung stehen:

- Den Variablennamen GROESSE kann man aus einer Liste per Transportschalter  , Drag & Drop oder Doppelklick übernehmen.
- Mit Hilfe einer virtuellen Tastatur kann man das Minuszeichen und die Zahl 100 auch per Maus eingeben.

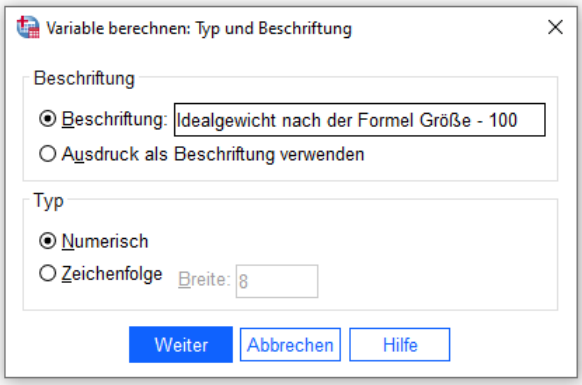
Anschließend sollte die Dialogbox **Variable berechnen** ungefähr so aussehen:



Die Dialogbox bietet über unsere momentanen Bedürfnisse hinausgehend auch die in SPSS verfügbaren Funktionen (vgl. Abschnitt 7.4.2.1) in **Funktionsgruppen** geordnet zum Transport in das Feld **Numerischer Ausdruck** an, sodass man Tippfehler vermeiden und Zeit sparen kann.

Am Anfang der Liste mit **allen** Funktionen befinden sich zudem spezielle Systemvariablen (z. B. **$Casenum** mit einer bei 1 startenden Nummerierung der Fälle in der Arbeitsdatei).

Rufen Sie nun die Subdialogbox **Typ und Beschriftung** mit dem gleichnamigen Schalter auf, und tragen Sie dort zur Variablen IDGEW die **Beschriftung** *Idealgewicht nach der Formel Größe - 100* ein:



Quittieren Sie die Subdialogbox mit **Weiter** und die Hauptdialogbox mit **Einfügen**. Daraufhin erhalten Sie im Syntaxfenster ein COMPUTE - und ein VARIABLE LABELS - Kommando:

```
COMPUTE idgew = groesse - 100.
VARIABLE LABELS idgew 'Idealgewicht nach der Formel Größe - 100'.
EXECUTE.
```

Außerdem erscheint ein EXECUTE-Kommando, dessen (meist überflüssige) Rolle in Abschnitt 7.3 erläutert wurde.

7.4.2 Technische Details

Im mehrzeiligen Texteingabefeld **Numerischer Ausdruck** der Dialogbox **Variable berechnen** sind wir trotz der SPSS-Schreibhilfen mit der Aufgabe konfrontiert, „auf einem leeren Blatt“ einen formalsprachlichen Ausdruck nach gewissen Regeln zu erstellen. Zum Glück sind uns numerische Ausdrücke aus der Schule wohlbekannt.[1]

Ein numerischer Ausdruck im Sinne von SPSS darf folgende Bestandteile enthalten:

- bereits definierte Variablen
- Zahlen
- arithmetische Operatoren:
  - Addition (+)
  - Subtraktion (-)
  - Multiplikation (*)
  - Division (/)
  - Potenzieren (**)
- Klammern
- Funktionen

[1] Zwar gibt es gewisse Unterschiede zwischen mathematischen *Gleichungen* (z. B. $y = a + b \cdot x$) und IT-sprachlichen *Zuweisungen* (z. B. compute x = x + 2.), doch sind die Regeln für die numerischen Ausdrücke auf der *rechten* Seite weitgehend identisch.

7.4.2.1 Numerische Funktionen

In numerischen Ausdrücken können Sie zahlreiche Funktionen verwenden, die numerische Variablen oder Zahlen als Argumente (in den folgenden Syntaxdarstellungen vertreten durch den Platzhalter *arg*) verarbeiten.[1] Diese Funktionen sind in mehrere Gruppen aufgeteilt, aus denen anschließend jeweils einige wichtige Vertreter genannt werden:

- **Arithmetische Funktionen**, z. B.:
  - ABS(*arg*) Absoluter Wert
  - EXP(*arg*) Exponentialfunktion
  - LN(*arg*) Natürlicher Logarithmus
  - MOD(*arg1*, *arg2*) Rest aus der Division von *arg1* durch *arg2*
  - RND(*arg*) Auf die nächst gelegene ganze Zahl gerundeter Wert
  - SQRT(*arg*) Quadratwurzel

  Beispiel:
  ```
  compute lnsal = ln(sal).
  ```
  Für jeden Fall wird von der vorhandenen Variablen SAL der Logarithmus berechnet und in der neuen Variablen LNSAL gespeichert. Durch die logarithmische Transformation kann oft die Linearität eines Regressionsmodells verbessert werden (siehe Baltes-Götz 2019).

- **Statistische Funktionen**, z. B.:
  - MEAN[.*m*](*arg1*, *arg2*[, ...]) Arithmetisches Mittel
  - MEDIAN[.*m*](*arg1*, *arg2*[, ...]) Median
  - MAX[.*m*](*arg1*, *arg2*[, ...]) Maximum
  - MIN[.*m*](*arg1*, *arg2*[, ...]) Minimum
  - SD[.*m*](*arg1*, *arg2*[, ...]) Standardabweichung
  - SUM[.*m*](*arg1*, *arg2*[, ...]) Summe

  Regeln:
  - Die eckigen Klammern schließen optionale Angaben ein.
  - Mit „[, ...]“ wird zum Ausdruck gebracht, dass die Liste der Argumente optional beliebig verlängert werden darf.
  - Der optionale Funktionsparameter *m* hat folgende Bedeutung: Wenn bei einem Fall mindestens *m* valide Argumente vorliegen, wird der Funktionswert berechnet. Ansonsten wird dem Fall der Wert SYSMIS zugewiesen. Wird *m* nicht angegeben, gelten die sehr liberalen Voreinstellungen 1 (z. B. bei MEAN) oder 2 (z. B. bei SD).

    Zwei häufige Fehler beim Einsatz des Minimalanforderungsparameters *m* sind:
    - Punkt zwischen dem Funktionsnamen und *m* vergessen

      Dieser Funktionsaufruf
      ```
      mean2(hund, katze, maus)
      ```
      hat (ohne Fehlermeldung!) denselben Effekt wie der Aufruf
      ```
      mean(hund, katze, maus)
      ```

[1] SPSS kennt auch zahlreiche Funktionen für String- und Datumsvariablen, die aber aus Zeitgründen in diesem Kurs nicht behandelt werden. Informieren Sie sich bei Bedarf im Hilfesystem, z. B. über eine Suche nach dem Stichwort *Funktionen*.

- Leerzeichen zwischen dem Funktionsnamen und dem Punkt gesetzt
  Dieser Funktionsaufruf
  ```
  mean .2(hund, katze, maus)
  ```
  führt zu einer Fehlermeldung.

- Man kann eine Serie von Variablen, *die im Datenblatt hintereinander stehen*, über das Schlüsselwort TO bequem in einer Argumentenliste angeben:

  erste TO *letzte*

Beispiel:
```
compute mfrei = mean.45(sport to angeln).
```
Wenn für einen Fall bei den Variablen SPORT bis ANGELN, die im Datenblatt hintereinander stehen, mindestens 45 valide Argumente vorliegen, wird deren Mittelwert der Variablen MFREI zugewiesen. Anderenfalls wird der MD-Indikator SYSMIS zugewiesen.

Beachten Sie den Unterschied zwischen den gerade beschriebenen statistischen *Funktionen* und den Statistik*prozeduren*, mit denen wir z. B. die univariaten Verteilungsanalysen durchgeführt haben:

- Wenn wir in der Dialogbox **Häufigkeiten** (erreichbar über **Analysieren > Deskriptive Statistiken > Häufigkeiten**) z. B. den Mittelwert der Variablen GEWICHT anfordern, dann werden die (validen) Gewichtsangaben aller Fälle in der Stichprobe gemittelt. Es werden also die Ausprägungen *einer Variablen* über *alle Fälle* gemittelt. SPSS arbeitet sich *senkrecht* durch eine komplette Variable bzw. Spalte der Arbeitsdatei. Es resultiert ein einziger Stichprobenkennwert, der im Ausgabefenster erscheint.
- Mit der statistischen Funktion MEAN können wir für *jede einzelne Person* z. B. den Mittelwert über *mehrere LOT-Variablen* berechnen lassen. SPSS geht *waagerecht* vor, wobei dasselbe Verfahren *auf jeden Fall, d. h. auf jede Zeile* der Datenmatrix angewendet wird. Es wird eine Variable, d. h. eine Spalte im Datenblatt, erzeugt oder modifiziert, in die für jeden Fall sein Berechnungsergebnis eingetragen wird.

- **Funktionen für fehlende Werte**, z. B.:
  - NMISS(*arg1*[, ...]) — Anzahl fehlender Werte bei den aufgelisteten Variablen
  - NVALID(*arg1*[, ...]) — Anzahl gültiger Werte bei den aufgelisteten Variablen
  - VALUE(*arg*) — Es wird der Wert der Variablen *arg* geliefert, wobei *benutzerdefinierte* MD-Deklarationen ignoriert werden.

Regeln:
- Mit „[, ...]“ wird zum Ausdruck gebracht, dass die Liste der zu untersuchenden Variablen optional beliebig verlängert werden darf.
- Mit dem Schlüsselwort TO kann eine Serie von Variablen angegeben werden (siehe obige Erläuterungen bei den statistischen Funktionen).

Beispiel:
```
compute nmfrei = nmiss(sport to angeln).
```
Der numerische Ausdruck liefert die Anzahl der fehlenden Werte (SYMIS oder benutzerdefiniert) bei den Variablen SPORT bis ANGELN, die im Datenblatt hintereinander stehen.

- **Pseudozufallszahlgeneratoren**, z. B.:
  - NORMAL(*arg*) Die Funktion liefert normalverteilte Pseudozufallszahlen mit dem Mittelwert 0 und der Standardabweichung *arg*.
  - UNIFORM(*arg*) Die Funktion liefert gleichverteilte Pseudozufallszahlen im Intervall von 0 bis *arg*.

Beispiel:

```
COMPUTE av = NORMAL(1).
EXECUTE.
T-TEST
  GROUPS=geschl(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=av
  /CRITERIA=CIN(.95).
```

Die Kommandos in diesem Beispiel wurden mit Hilfe von Dialogboxen erzeugt (Schalter **Einfügen**). Im COMPUTE-Kommando wird die standardnormalverteilte Zufallsvariable AV erstellt. Es ist klar, dass Frauen und Männer bei AV denselben Erwartungswert (Populationsmittelwert) 0 haben. Damit können wir ausprobieren, wie sich der t-Test für unabhängige Stichproben bei Gültigkeit der Nullhypothese verhält. Die Dialogbox zu diesem t-Test erhält man mit **Analysieren > Mittelwerte vergleichen > t-Test bei unabhängigen Stichproben**.

Wenn Ihnen die Erläuterungen zu diesem Beispiel „spanisch“ vorkommen, hilft Ihnen vielleicht der Abschnitt 8.1 weiter, wo die Grundprinzipien der Inferenzstatistik erläutert werden. Mit dem t-Test für unabhängige Stichproben beschäftigen wir uns „offiziell“ in Kapitel 12.

Hinweis: Bei NORMAL und UNIFORM wird ein Pseudozufallszahlengenerator verwendet, der per Voreinstellung mit dem festen Wert 2.000.000 startet und damit stets dieselben Zahlen liefert. Bei der Verwendung von Pseudozufallszahlen bietet SPSS die folgenden Optionen:

- Der per Voreinstellung verwendete Pseudozufallszahlengenerator mit MC-Technik (*Multiplicative Congruential*) kann durch eine modernere Alternative mit MT-Technik (*Mersenne Twister*) ersetzt werden.
- Man kann als Startwert für den Pseudozufallszahlengenerator eine bestimmte Zahl festlegen oder einen zufälligen Startwert verlangen.

Die Konfiguration des Pseudozufallszahlengenerators kann erfolgen per …

- Dialog, nach dem Menübefehl:
  Transformieren > Zufallszahlengeneratoren
- SPSS-Kommando:
  a) Traditioneller Zufallszahlengenerator:

  SET RNG=MC SEED={*start* | RANDOM}**.**

  b) Mersenne-Twister:

  SET RNG=MT MTINDEX={*start* | RANDOM}**.**

7.4.2.2 Regeln für die Bildung numerischer Ausdrücke

Auch bei Verwendung der Dialogbox **Variable berechnen** müssen wir den numerischen Ausdruck im Wesentlichen selbst formulieren. Dabei sind folgende Regeln zu beachten:

- Sind mehrere Operatoren vorhanden, ist die **Auswertungsreihenfolge** relevant. Diese hängt von den Prioritäten der Operatoren ab. Es gilt folgende Rangordnung:

  Priorität 1: Funktionen
  Priorität 2: Potenzieren (**)
  Priorität 3: Multiplikation (*), Division (/)
  Priorität 4: Addition (+), Subtraktion (-)

  Bei gleicher Priorität erfolgt die Auswertung von links nach rechts. Eine alternative Auswertungsreihenfolge kann durch Klammern erzwungen werden: Klammerausdrücke werden zuerst ausgewertet. Bei geschachtelten Klammern erfolgt die Auswertung von innen nach außen.

- Bei Funktionen mit mehreren Argumenten müssen die einzelnen Argumente **durch jeweils genau ein Komma** (optional ergänzt durch Leerzeichen) getrennt werden.

  Beispiel: `compute mabc = mean.2(a,b, c).`

- Obwohl SPSS *im Daten- und im Ausgabefenster* das ländertypische Dezimaltrennzeichen benutzt, bei uns also das Komma, ist in numerischen Ausdrücken der Punkt als Dezimaltrennzeichen zu verwenden:

  Richtig: `2.75`
  Falsch: `2,75`

  Dies gilt sowohl für das Feld **Numerischer Ausdruck** der Dialogbox **Variable berechnen** als auch für das COMPUTE-Kommando in einem Syntaxfenster. Es kann also durchaus passieren, dass Sie ein und dieselbe Zahl im Datenfenster (als Wert eines Falles für eine bestimmte Variable) mit Dezimal*komma* und in der Dialogbox **Variable berechnen** (als Konstante in einem numerischen Ausdruck) mit Dezimal*punkt* schreiben müssen.

- Bei den meisten Funktionen sind auch numerische Ausdrücke als Argumente zugelassen.

  Beispiel: `compute albmax = max(a, ln(b)).`

7.4.2.3 Sonstige Hinweise

a) SYSMIS als Ergebnis eines numerischen Ausdrucks

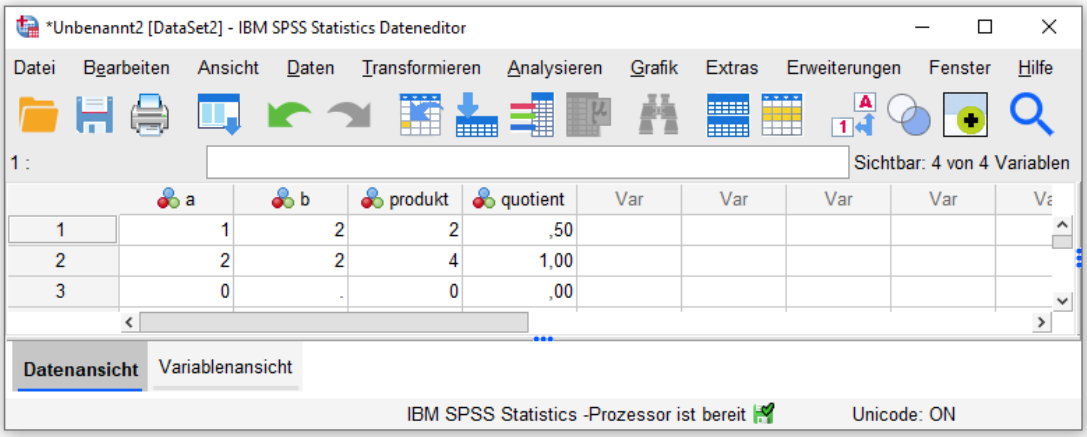
Durch eine Berechnungsanweisung wird der Wert des numerischen Ausdrucks auch dann der Zielvariablen zugewiesen, wenn dieser Wert gleich SYSMIS ist (z. B. bei fehlenden Argumenten). Ist die Zielvariable bereits *vorhanden*, bleibt bei missglückter Berechnung des numerischen Ausdrucks keinesfalls der alte Wert bestehen, sondern dieser wird durch SYSMIS ersetzt.

b) Rechnen mit fehlenden Werten

Fehlt bei einem Fall zur Berechnung eines numerischen Ausdrucks eine Argumentvariable, dann erhält die Ergebnisvariable den Wert SYSMIS. Ausnahmen sind die folgenden Regeln für das „Rechnen“ mit fehlenden Werten:

- 0 * unbekannt = 0
  Diese Regel ist schlau, denn für beliebige reelle Zahlen *x* gilt:
  $$0 \cdot x = 0$$
- 0 / unbekannt = 0
  Diese Regel ist kritisierbar, denn:
  $$\frac{0}{x} = \begin{cases} 0 & x \neq 0 \\ \text{undefiniert} & x = 0 \end{cases} \quad \text{für}$$

Im folgenden Datenfenster erhält der dritte Fall (mit dem Wert 0 bei der Variablen A und einem fehlenden B-Wert) für das Produkt A * B *und* für den Quotienten A / B von SPSS den Ergebniswert null:



7.4.3 Übungen

1) Welche Werte haben die folgenden numerischen Ausdrücke?

```
(3 + 4) / 2
 3 + 4 / 2
(3**2 / 2) + 4
 3**2 / 2  + 4
```

2) Erstellen Sie im KFA-Projekt die Variablen, auf die sich unsere zentralen Hypothesen beziehen (vgl. Abschnitt 2.3):

- Berechnen Sie die Variable LOT als arithmetisches Mittel der (nötigenfalls recodierten!) LOT-Items 1, 3, 4, 5, 8, 9, 11 und 12. Die restlichen Items dienen nicht zur Messung von Optimismus, sondern sollen verhindern, dass der Zweck des Fragebogens deutlich wird. Dies könnte das Antwortverhalten verzerren. Tolerieren Sie bei der Berechnung des Mittelwerts bis zu *zwei* fehlende Werte. Bei der Schätzung einer latenten Variablen durch den Mittelwert aus manifesten Indikatoren einige fehlende Items zu tolerieren und den Mittelwert aus den vorhandenen Items zu berechnen, ist übrigens eine spezielle Technik zur Behandlung des Problems fehlender Werte (siehe Baltes-Götz 2013, S. 21).

- Berechnen Sie die Variable AERGAM als arithmetisches Mittel der beiden Ärgervariablen und die Variable AERGZ als Ärgerzuwachs auf Grund der kontrafaktischen Alternative. AERGAM benötigen wir zum Testen der differentialpsychologischen Hypothese. Beim geplanten t-Test für abhängige Stichproben zum Vergleich der Mittelwerte von AERGO und AERGM wird letztlich mit einem Einstichproben - t-Test geprüft, ob der Erwartungswert der Differenzvariablen (AERGZ = AERGM - AERGO) signifikant größer 0 ist. Dabei wird vorausgesetzt, dass die *Differenzvariable* in der Population normalverteilt ist (vgl. Abschnitt 8.1). Für die Durchführung des t-Tests mit SPSS ist es nicht erforderlich, die Differenzvariable per Datentransformation zu erstellen. Allerdings bietet die t-Test-Prozedur keine Möglichkeit, die Normalverteilungsvoraussetzung zu prüfen. Daher berechnen wir die Variable AERGZ explizit und prüfen ihre Verteilung mit der Prozedur zur explorativen Datenanalyse auf Normalität (siehe Kapitel 9).

Rufen Sie jeweils mit dem Menübefehl:

Transformieren > Variable berechnen

die zuständige Dialogbox auf. Quittieren Sie Ihre Eintragungen nicht mit **OK**, sondern mit **Einfügen**, damit die zugehörigen COMPUTE-Kommandos in das Syntaxfenster eingetragen werden, in dem gerade das Transformationsprogramm zum KFA-Projekt entsteht.

Weil SPSS eine Folge von mehreren Kommandos stets in der natürlichen Reihenfolge abarbeitet, wird beim späteren Ablauf unseres Transformationsprogramms z. B. die für einige Items angeordnete Recodierung (vgl. Abschnitt 7.2.3) bereits erledigt sein, wenn das COMPUTE-Kommando zur LOT-Berechnung ausgeführt wird.

3) Erstellen Sie eine Variable namens BMI mit dem aus Körpergröße und Körpergewicht nach folgender Formel

$$\frac{Gewicht\ (in\ kg)}{Größe^2\ (in\ m)}$$

berechneten **Body Mass Index**. Wir werden später im Rahmen unserer ernährungsphysiologischen Begleitstudie der Frage nachgehen, ob beim BMI Geschlechtsunterschiede bestehen (siehe Kapitel 12).

4) Berechnen Sie aus dem Geburtsjahr der Untersuchungsteilnehmer das Alter.[1] Wir haben bei der Datenerhebung nach dem Geburtsjahr gefragt, weil manche Auskunftspersonen diese Information leichter und genauer liefern können als das Alter. Bei der Forschungsarbeit und in Ergebnisberichten ist das Alter jedoch anschaulicher. Außerdem ist zu befürchten, dass mit dem Wissen um den Erhebungszeitpunkt einer Studie irgendwann das Wissen um das Alter der Befragten verlorengeht.

[1] Bei Verwendung der Manuskriptstichprobe muss berücksichtigt werden, dass diese aus dem Jahr 1999 stammt.

7.5 Bedingte Datentransformation

Gelegentlich ist es erforderlich, eine Datenmodifikation auf diejenigen Fälle zu beschränken, die eine Bedingung erfüllen. Wir benötigen z. B. im KFA-Projekt eine solche Möglichkeit, um für die Motivations- und Methodenvariablen die Behandlung fehlender Werte zu komplettieren (siehe Abschnitt 2.4.3.2.5).

Bei manchen Transformationen ist eine **Fallunterscheidung** erforderlich. Z. B. könnte im Rahmen einer ernährungsphysiologischen Studie ein Idealgewichtsbegriff zum Einsatz kommen, der bei Frauen und Männern unterschiedliche Formeln vorschreibt.

In den Transformations-Dialogboxen von SPSS erreichen Sie über den Schalter **Falls** eine Subdialogbox zur Definition einer Bedingung, unter der die Transformation ausgeführt werden soll. Sie können z. B. eine bedingte Umcodierung (vgl. Abschnitt 7.2), Berechnung (vgl. Abschnitt 7.4) oder Werteauszählung (vgl. Abschnitt 7.6) vornehmen.

Wenn unter ein und derselben Bedingung gleich *mehrere* Transformationen vorgenommen werden sollen, dann muss diese Bedingung in allen benötigten Transformations-Dialogboxen wiederholt werden. Ähnlich umständlich ist die Realisation von Fallunterscheidungen mit Hilfe der Transformations-Dialogboxen. Bei solchen Aufgaben ist es oft einfacher, in einem Syntaxfenster eine DO IF - ELSE - END IF – Konstruktion zu erstellen, z. B.:

```
do if (geschl = 1).
  compute  idgew = (groesse - 100) * 0.85.
  compute  rle = 83.5 - alter.
else.
  compute  idgew = (groesse - 100) * 0.9.
  compute  rle = 78.5 - alter.
end if.
```

7.5.1 Beispiel

In diesem Abschnitt soll endlich die Behandlung fehlender Werte für die Motivationsvariablen abgeschlossen werden. Wir haben bei den Variablen MOTIV1 bis MOTIV5 und ANDERE die markierten Kästchen mit 1 und die leeren Kästchen mit 0 codiert. Ein Fall mit Nullen bei MOTIV1 bis MOTIV5 *und* ANDERE hat aber offenbar den Fragebogenteil 3a komplett ausgelassen, denn:

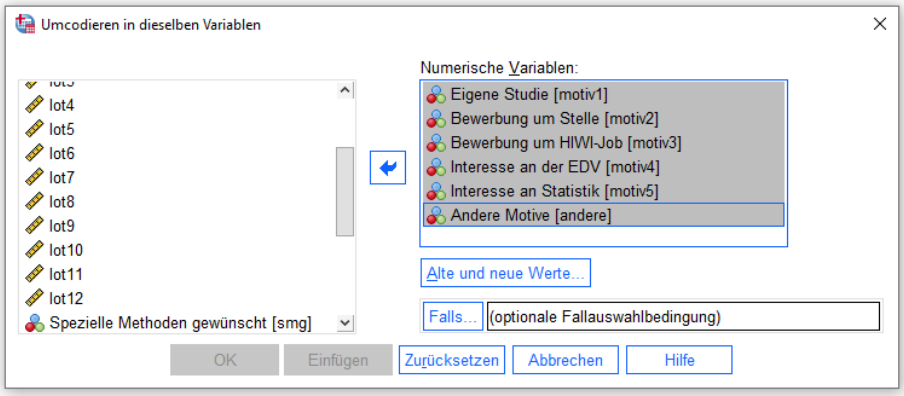
- In der Stichprobe befinden sich ausschließlich Kursteilnehmer.
- Aufgrund der Restkategorie (Variable ANDERE) sind alle möglichen Motive zur Kursteilnahme berücksichtigt.

Daher müssen für genau diese Fälle die Nullen bei den Variablen MOTIV1 bis MOTIV5 und ANDERE in SYSMIS (oder einen anderen MD-Indikator) umcodiert werden. Gehen Sie folgendermaßen vor:

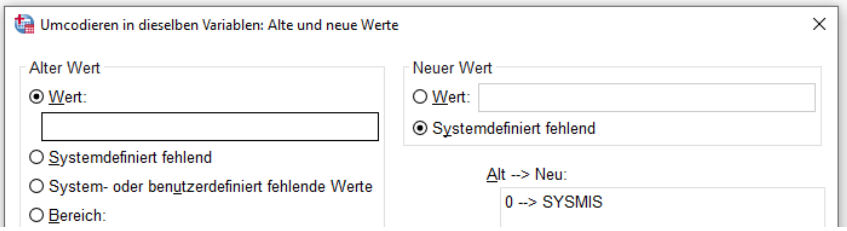
- Wählen Sie den Menübefehl:

  Transformieren > Umcodieren in dieselben Variablen

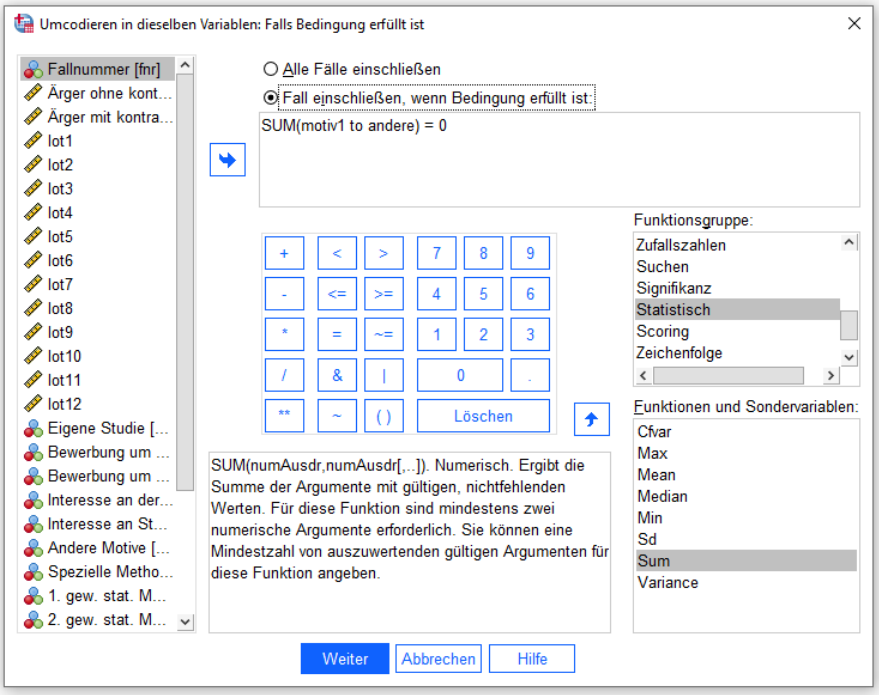
- Transportieren Sie die Variablennamen MOTIV1 bis MOTIV5 und ANDERE in die Teilnehmerliste der **Umcodieren**-Dialogbox.



- Legen Sie in der Subdialogbox **Alte und neue Werte** die benötigte Abbildungsvorschrift fest:



- Öffnen Sie die **Falls**-Subdialogbox, markieren Sie die Option **Fall einschließen, wenn Bedingung erfüllt ist**, und tragen Sie in das darunterliegende Textfeld eine geeignete Bedingung ein, z. B.:



Aufgrund unserer Datenüberprüfung (siehe Kapitel 5) können wir uns darauf verlassen, dass aktuell bei den Variablen MOTIV1 bis MOTIV5 und ANDERE ausschließlich die Werte 0 und 1 vorkommen. Daher ist die Summe dieser Variablen genau dann gleich 0, wenn jede einzelne Variable gleich 0 ist. Somit lässt sich die erforderliche Bedingung mit Hilfe der statistischen Funktion SUM (vgl. Abschnitt 7.4.2.1) besonders einfach formulieren.

Die obige Eintragung im Bedingungsfeld (siehe Bildschirmfoto) kann „semiautomatisch“ z. B. so erzeugt werden:

- Wählen Sie die Funktionsgruppe **Statistisch**, markieren Sie die Funktion **Sum**, und klicken Sie auf den Transportschalter ↑, sodass im Bedingungsfeld eine Vorlage für einen SUM() - Funktionsaufruf erscheint:

  SUM(?,?)

- Transportieren Sie aus der Variablenliste am linken Rand der Dialogbox per Transportschalter →, Drag & Drop oder Doppelklick die Variable MOTIV1 in das Bedingungsfeld, wobei in der Vorlage das markierte Fragezeichen automatisch durch den Variablennamen ersetzt wird.
- Ersetzen Sie das Komma in der Vorlage durch das Schlüsselwort TO (mit begrenzenden Leerzeichen), und komplettieren Sie die Liste durch den Variablennamen ANDERE, den Sie wiederum aus der Variablenliste in das Bedingungsfeld transportieren können.
- Erstellen Sie aus dem bis jetzt vorhandenen SUM() - Funktionsaufruf eine Bedingung, indem Sie ein Gleichheitszeichen und den Wert 0 ergänzen.

- Machen Sie **weiter**, und quittieren Sie die **Umcodieren**-Dialogbox mit **Einfügen**.

Daraufhin wird Ihr Transformationsprogramm um die folgende Kommandosequenz mit einer DO IF - END IF - Kontrollstruktur erweitert:

```
DO IF (SUM(motiv1 to andere) = 0).
RECODE motiv1 motiv2 motiv3 motiv4 motiv5 andere (0=SYSMIS).
END IF.
EXECUTE.
```

Wenn Sie diese Kommandos ausführen lassen, gleichgültig ob direkt per **OK** in der **Umcodieren**-Dialogbox oder indirekt via Syntaxfenster, passiert bei jedem einzelnen Fall in der Stichprobe folgendes:

- SPSS überprüft die Bedingung, die auch als **logischer Ausdruck** bezeichnet werden kann.
- Ist bei einem Fall die Bedingung erfüllt, dann wird umcodiert, anderenfalls passiert nichts.

Weil die Variablen MOTIV1 bis MOTIV5 und ANDERE vor dem Umcodieren garantiert nur Nullen oder Einsen als Werte aufweisen, hat unser logischer Ausdruck die Eigenschaft, in jedem Fall entweder wahr oder falsch zu sein. Das ist in der empirischen Forschung z. B. wegen des nahezu allgegenwärtigen Problems fehlender Werte keineswegs der Normalfall. Generell kann z. B. der logische Ausdruck „GESCHL = 1“ folgende Wahrheitswerte annehmen:

- wahr ⇔ Der GESCHL-Wert ist gleich 1.
- falsch ⇔ Der GESCHL-Wert ist eine von 1 verschiedene Zahl.
- unbestimmt ⇔ GESCHL hat einen MD-Indikator als Wert.

Komplettiert um Regeln für unbestimmte logische Ausdrücke ist das Verhalten von SPSS bei bedingten Transformationen so zu beschreiben:

- Ist der logische Ausdruck **wahr**, dann wird die Transformation ausgeführt.
  Im Fall einer bedingten Berechnung (COMPUTE-Kommando) wird der Ergebnisvariablen also der Wert des numerischen Ausdrucks zugewiesen. Die Zuweisung erfolgt auch dann, wenn der numerische Ausdruck den Wert SYSMIS hat (z. B. wegen einer Division durch null).
- Ist der logische Ausdruck **falsch oder unbestimmt**, dann passiert **nichts**, d. h.:
  - Bei einer bereits vorhandenen Ergebnisvariablen behält der betroffene Fall seinen bisherigen Wert.
  - Bei einer neu definierten Variablen behält der betroffene Fall den Initialisierungswert SYSMIS.

7.5.2 Bedingungen formulieren

Der im obigen Beispiel aufgetretene logische Ausdruck war recht einfach aufgebaut, weil er aus einem einzigen Vergleich bestand. Obwohl Ihnen auch komplexere Exemplare vertraut sein dürften, soll der Begriff *logischer Ausdruck* zur Klärung einiger Detailfragen exakt definiert werden. Zunächst wird der einfachere Begriff *Vergleich* eingeführt, wobei wir uns auf numerische Variablen bzw. Ausdrücke beschränken.

7.5.2.1 Vergleich

Ein Vergleich besteht aus zwei numerischen Ausdrücken und einem Vergleichsoperator:

numerischer_ausdruck vergleichsoperator numerischer_ausdruck

Die **Vergleichsoperatoren** können in SPSS alternativ durch IT-Varianten der mathematischen Symbole oder durch Schlüsselwörter dargestellt werden:

Bedeutung	Symbol	Schlüsselwort
gleich	=	EQ
ungleich	<> oder ~=	NE
kleiner als	<	LT
kleiner oder gleich	<=	LE
größer als	>	GT
größer oder gleich	>=	GE

Beispiele:
```
beruf >= 4
beruf ge 4
```

7.5.2.2 Logischer Ausdruck

Ausgehend vom einfachen Begriff *Vergleich* wird nun durch eine rekursive Definition der komplexere Begriff *logischer Ausdruck* konstruiert:

i) Jeder Vergleich ist ein logischer Ausdruck.

ii) Durch Anwendung des logischen Operators **NOT** auf einen logischen Ausdruck oder durch Anwendung der logischen Operatoren **AND** bzw. **OR** auf zwei logische Ausdrücke entsteht ein neuer logischer Ausdruck:

NOT *logischer_ausdruck*

logischer_ausdruck_1 AND *logischer_ausdruck_2*

logischer_ausdruck_1 OR *logischer_ausdruck_2*

Den Wahrheitswert eines zusammengesetzten logischen Ausdrucks erhält man aus den Wahrheitswerten der Argumente nach den Regeln für logische Operatoren, die in den sogenannten *Wahrheitstafeln* festgelegt sind (siehe unten).

Es lassen sich sukzessiv beliebig komplexe logische Ausdrücke aufbauen, die für jeden konkreten Fall die Wahrheitswerte *wahr*, *falsch* oder *unbestimmt* haben können.

Beispiel: `(beruf >= 4) and (schule <> 7)`

Konjunktion und Adjunktion können per Schlüsselwort oder Symbol ausgedrückt werden:

Schlüsselwort	**Symbol**
AND	&
OR	|

Mit unbestimmten Wahrheitswerten in logischen Ausdrücken verfährt SPSS analog zum Rechnen mit fehlenden Werten in numerischen Ausdrücken (siehe Abschnitt 7.4.2.3). Die folgenden Wahrheitstafeln sind gegenüber der klassischen Aussagenlogik um den Wahrheitswert *unbestimmt* erweitert (*la1* und *la2* seien logische Ausdrücke):

la1	**NOT *la1***
wahr	falsch
falsch	wahr
unbestimmt	unbestimmt

la1	*la2*	*la1* **AND** *la2*	*la1* **OR** *la2*
wahr	wahr	wahr	wahr
wahr	falsch	falsch	wahr
wahr	unbestimmt	unbestimmt	wahr
falsch	wahr	falsch	wahr
falsch	falsch	falsch	falsch
falsch	unbestimmt	falsch	unbestimmt
unbestimmt	wahr	unbestimmt	wahr
unbestimmt	falsch	falsch	unbestimmt
unbestimmt	unbestimmt	unbestimmt	unbestimmt

7.5.2.3 Regeln für die Auswertung logischer Ausdrücke

Bei der Auswertung von logischen Ausdrücken gelten in SPSS folgende Regeln:

- Die Auswertungsreihenfolge hängt von den Prioritäten der Operatoren ab. Es gilt folgende Rangordnung:

  Priorität 1: Funktionen
  Priorität 2: Potenzieren (**)
  Priorität 3: Multiplikation (*), Division (/)
  Priorität 4: Addition (+), Subtraktion (-)
  Priorität 5: Vergleichsoperatoren
  Priorität 6: NOT
  Priorität 7: AND
  Priorität 8: OR

- Bei gleicher Priorität: Auswertung von links nach rechts.
- Eine alternative Auswertungsreihenfolge kann durch Klammern erzwungen werden.

Das Beispiel für einen zusammengesetzten logischen Ausdruck aus Abschnitt 7.5.2.2 kann wegen der voreingestellten Auswertungsreihenfolge auch kürzer geschrieben werden:

```
beruf >= 4  and  schule <> 7
```

Die aus Computer-Sicht überflüssigen Klammern verbessern allerdings die Lesbarkeit des Ausdrucks für Menschen und reduzieren so das Fehlerrisiko.

7.5.3 Übung

Bei den Variablen METH1 bis METH3 haben wir zur Vereinfachung der Erfassung im Codierplan festgelegt, dass „unbenutzte“ Variablen einfach leer bleiben sollen. Nun wollen wir aber bei Fällen mit regulärem Antwortmuster die SYSMIS - Werte durch Nullen ersetzen. Die 0 soll z. B. bei der Variablen METH2 bedeuten: Die Option, einen zweiten Methodenwunsch zu äußern, wurde nicht genutzt.

Die folgende Tabelle, die wir in Abschnitt 2.4.3.2.5 vereinbart haben, legt im Einzelnen fest, was unter den möglichen Bedingungskonstellationen geschehen soll:

		Mindestens eine speziell interessierende Methode angegeben?	
		Ja	**Nein**
SMG	**1**	METH1 ... METH3: SYSMIS → 0 Bem.: Korrektes Antwortverhalten. Variablen zu nicht benutzten Optionen (gem. Codierplan bisher auf SYSMIS) werden auf 0 gesetzt.	SMG: 1 → SYSMIS Bem.: Irreguläres Antwortverhalten. METH1 bis METH3 behalten SYMIS. SMG wird ebenfalls auf SYMIS gesetzt.
	0	SMG: 0 → 1 METH1 ... METH3: SYSMIS → 0 Bem.: Leicht irreguläres Antwortverhalten. Wir sind großzügig und setzen SMG auf 1 sowie die Variablen zu nicht benutzten Optionen auf 0.	METH1 ... METH3: SYSMIS → 0 Bem.: Korrektes Antwortverhalten. Die Variablen zu allen Optionen (gem. Codierplan bisher auf SYSMIS) werden auf 0 gesetzt.
	SYSMIS	SMG: SYSMIS → 1 METH1 ... METH3: SYSMIS → 0 Bem.: Leicht irreguläres Antwortverhalten. Wir sind großzügig und setzen SMG auf 1 sowie die Variablen zu nicht benutzten Optionen auf 0.	Bem.: Irreguläres Antwortverhalten. Alle Variablen behalten den Wert SYSMIS.

In den beiden obersten Zeilen jeder Zelle sind die erforderlichen Korrekturen bei SMG bzw. METH1 bis METH3 angegeben. Erweitern Sie Ihr Programm **kfat.sps** um passende Transformationsanweisungen.

7.6 Häufigkeit bestimmter Werte bei einem Fall ermitteln

Mit dem Befehl **Werte in Fällen zählen** aus dem Menü **Transformieren** bzw. mit dem zugrunde liegenden COUNT-Kommando kann man für …

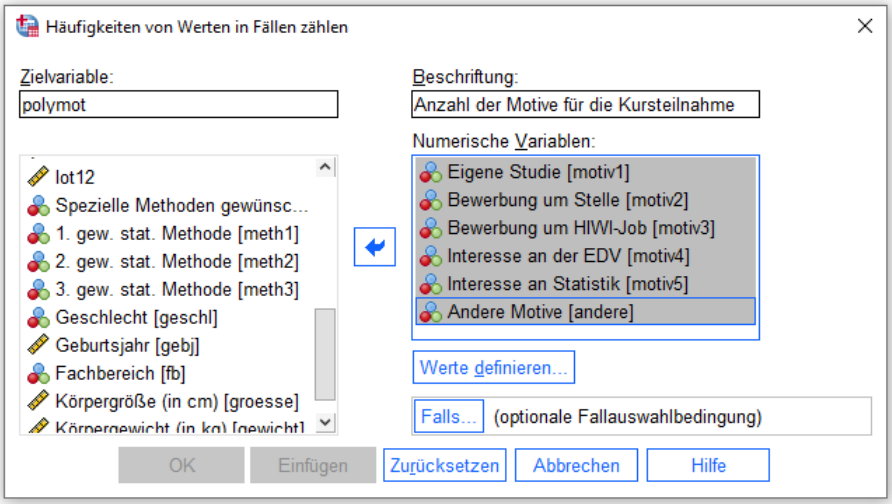
- eine Liste von *k* Ausgangsvariablen
- sowie eine Liste mit kritischen (relevanten) Werten

eine Zählvariable erstellen lassen, die für jeden Fall festhält, wie viele von den Ausgangsvariablen einen kritischen Wert haben. Das minimale Zählergebnis ist 0 (keine Ausgangsvariable hat einen kritischen Wert), und das maximale Ergebnis ist *k* (jede Ausgangsvariable hat einen kritischen Wert).

Wir wollen eine neue Variable namens POLYMOT berechnen lassen, die für jede Person festhält, wie viele Motive zur Kursteilnahme sie im Fragebogenteil 3a angegeben hat. Aktivieren Sie die Dialogbox **Häufigkeiten von Werten in Fällen zählen** mit

Transformieren > Werte in Fällen zählen

Vereinbaren Sie für die **Zielvariable** den Namen POLYMOT sowie die **Beschriftung** *Anzahl der Motive für die Kursteilnahme*, und transportieren Sie die Variablen MOTIV1 bis ANDERE in die Liste der **numerischen Variablen**. Danach müsste Ihre Dialogbox ungefähr so aussehen:



Wechseln Sie jetzt mit dem Schalter **Werte definieren** in die Subdialogbox **Werte in Fällen zählen: Welche Werte**, tragen Sie dort den kritischen **Wert** 1 ein, und klicken Sie auf **Hinzufügen**:

Die in dieser Subdialogbox angebotenen sonstigen Möglichkeiten zur Festlegung der Trefferwerte kennen Sie schon aus der Subdialogbox **Umcodieren: Alte und neue Werte** (siehe Abschnitt 7.2).

Da SPSS eine Folge von mehreren Kommandos stets in der natürlichen Reihenfolge abarbeitet, wird beim späteren Ablauf unseres Transformationsprogramms die MD-Behandlung für die Variablen MOTIV1 bis ANDERE bereits erledigt sein, wenn die **Zählen**-Anweisung an die Reihe kommt. Bei Personen, die den Fragebogenteil 3a *nicht* korrekt bearbeitet haben, wird also gelten:

MOTIV1 = MOTIV2 = ... = ANDERE = SYSMIS

Wir müssen die folgende Eigenschaft des COUNT-Kommandos beachten: Die Ergebnisvariable hat *stets* einen validen Wert größer oder gleich 0. Wenn ein Fall z. B. bei allen Ausgangsvariablen den Wert SYSMIS hat, dann resultiert das valide Ergebnis 0! In dieser Situation wissen wir aber *nichts* von den Motiven der Person und dürfen ihr keine Motivationslosigkeit (POLYMOT = 0) unterstellen.

Weil im konkreten Beispiel das Zählergebnis 0 generell als irregulär einzustufen ist, könnten wir durch ein gewöhnliches (unbedingtes) Umcodieren

$$0 \rightarrow \text{SYSMIS}$$

dafür sorgen, dass ein Fall bei POLYMOT den Wert SYSMIS erhält, wenn er den Fragebogenteil 3a nicht korrekt bearbeitet hat. Im Allgemeinen kann das Zählergebnis 0 jedoch auch auf reguläre Weise zustande kommen, und auch ein von 0 verschiedenes Zählergebnis kann MD-belastet sein. Daher ist es meist erforderlich, durch eine *bedingte* Datentransformation MD-belastete Zählergebnisse zu verhindern.

Wir verwenden das generelle Verfahren der Übung halber auch im aktuellen Beispiel und sorgen durch eine bedingte Datentransformation dafür, ...

- dass bei Personen mit gültigen Werten für *alle* Motivvariablen die Anzahl der Motivvariablen mit dem Wert 1 ermittelt und der Ergebnisvariablen POLYMOT zugewiesen wird,
- dass bei anderen Fällen die Variable POLYMOT den Initialisierungswert SYSMIS behält.

Das Zählergebnis soll nur dann ermittelt und der Ergebnisvariablen zugewiesen werden, wenn die folgende, mit Hilfe der in Abschnitt 7.4.2.1 beschriebenen Funktion NMISS formulierte Bedingung erfüllt ist:

NMISS(MOTIV1 TO ANDERE) = 0

Die Bedingung ist genau dann wahr, wenn ein Fall bei den Variablen MOTIV1, ..., MOTIV5, ANDERE sechs gültige (nicht MD-deklarierte) Werte hat.

Klicken Sie in der Dialogbox **Häufigkeiten von Werten in Fällen zählen** auf den **Falls**-Schalter, und tragen Sie die vorgeschlagene Bedingung ein. Wenn Sie dann **weiter**machen und die Hauptdialogbox mit **Einfügen** quittieren, erhalten Sie im Syntaxfenster die folgenden Kommandos:

```
DO IF  (NMISS(motiv1 to andere) = 0).
COUNT polymot=motiv1 motiv2 motiv3 motiv4 motiv5 andere(1).
VARIABLE LABELS  polymot 'Anzahl der Motive für die Kursteilnahme'.
END IF.
EXECUTE.
```

Was hier zählt, ist offenbar das COUNT-Kommando. Es enthält im Wesentlichen eine Liste der zu untersuchenden Variablen, gefolgt von einer eingeklammerten Liste der kritischen (relevanten) Werte. Das Zählergebnis wird nur dann ermittelt und der neuen Variablen POLYMOT als Wert zugewiesen, wenn die Bedingung im DO IF - Kommando erfüllt ist. Anderenfalls behält POLYMOT den Initialisierungswert SYSMIS.

Das VARIABLE LABELS - Kommando sorgt für eine Variablenbeschriftung und geht auf unserer Eintragung im **Beschriftungs**-Textfeld des Dialogs **Häufigkeiten von Werten in Fällen zählen** zurück.

Weil bei den Variablen MOTIV1, MOTIV2, MOTIV3, MOTIV4, MOTIV5 und ANDERE nur die Werte 0, 1 und SYSMIS auftreten, kann man die Variable POLYMOT auch per COMPUTE-Kommando erstellen:

```
COMPUTE polymot = motiv1 + motiv2 + motiv3 + motiv4 + motiv5 + andere.
```

Fälle mit einem MD-Indikator bei mindestens einer Variablen erhalten automatisch den Ergebniswert SYSMIS, sodass im Vergleich zur obigen COUNT-Lösung Aufwand gespart wird. Bei

anderen Zählaufträgen ist COUNT aber unverzichtbar, z. B.: Wie viele Variablen aus einer Liste haben den Wert 1, 2, 3 oder 7?

7.7 Erstellung der Fertigdatendatei mit dem Transformationsprogramm

Aufgrund von nachvollzogenen Demonstrationsbeispielen oder Übungsaufgaben in den Abschnitten 7.2 (Erstellung von DEKADE durch das Kategorisieren von GEBJ, Umpolen der negativ formulierten LOT-Fragen), 7.4 (Berechnung von IDGEW, LOT, AERGAM, AERGZ, BMI und ALTER), 7.5 (MD-Behandlung für die Motiv- und die Methoden-Variablen) und 7.6 (Zählen der Kursmotive) sollten jetzt alle vorläufig im KFA-Projekt benötigten Transformationskommandos in einem Syntaxfenster stehen.

7.7.1 Transformationsprogramm vervollständigen

Um daraus ein komfortables SPSS-Programm zu machen, das die Rohdatendatei **kfar.sav** selbständig in ein neues Datenblatt einließt, alle Transformationen ausführt und schließlich das Ergebnis in die Fertigdatendatei **kfa.sav** sichert, müssen wir an den Anfang des Syntaxfensters noch ein GET-Kommando zum Lesen aus **kfar.sav** und ans Ende noch ein SAVE-Kommando zum Sichern in **kfa.sav** setzen. Wie Sie das GET-Kommando produzieren können, haben Sie schon in Abschnitt 6.2 erfahren. Wenn Sie das Kommando jetzt erzeugen lassen, erscheint es am Ende des Syntaxfensters, und Sie müssen es an den Anfang verschieben.

Im automatisch erzeugten, in der Regel hinter dem GET-Kommando positionierte DATASET NAME - Kommando wählen wir einen passenden Namen:[1]

```
DATASET NAME KfaFertigdaten WINDOW=FRONT.
DATASET ACTIVATE KfaFertigdaten.
```

Für einen Datenblattnamen gelten ähnliche Regeln wie für einen Variablennamen. Das ebenfalls automatisch erstellte DATASET ACTIVATE - Kommando wird passend aktualisiert oder gestrichen. Dem per GET-Kommando erstellten und aktivierten Datenblatt wird bei der kompletten Ausführung des Transformationsprogramms kaum ein anderes Datenblatt die Rolle der Arbeitsdatei streitig machen.

Um das SAVE-Kommando zu generieren, wechseln wir ins Datenfenster und aktivieren mit **Datei > Speichern unter** die zugehörige Dialogbox. Dann tragen wir den gewünschten Dateinamen **kfa.sav** ein und erzeugen mit **Einfügen** das benötigte SAVE-Kommando.

Um Komplikationen aus dem Weg zu gehen, sollte das Transformationsprogramm stets *komplett* ausgeführt werden (siehe Abschnitt 7.7.2). Damit ist sichergestellt, ...

- dass sich alle Kommandos auf dasselbe (unbenannte) Datenblatt beziehen, das erstellt, mit den Rohdaten befüllt, transformiert und schließlich in die Fertigdatendatei gesichert wird,
- dass jedes Kommando genau einmal ausgeführt wird.

Hinweise zur Ausgabedatei eines Transformationsprogramms:

---

[1] Die in früheren Versionen des Skripts empfohlene Verwendung des *unbenannten* Datenblatts kollidiert mit den in SPSS 27 eingeführten Restorepunkten. Bei der Erstellung eines Restorepunkts (per Voreinstellung alle 10 Minuten), erhält das unbenannte Datenblatt automatisch einen Namen.

- Verwenden Sie niemals dieselbe Datei als Quelle und Ziel des Transformationsprogramms. Schreiben Sie also keinesfalls mit Ihrem Transformationsprogramm in die Rohdatendatei. Wenn Sie der Empfehlung in Abschnitt 7.1.2 folgend für die Rohdatendatei das Schreibschutzattribut gesetzt haben, kann dieser Fehler auch nicht versehentlich passieren.
- Bei der Ausführung des Transformationsprogramms darf für seine Ausgabedatei, also für die Fertigdatendatei, das Schreibschutzattribut natürlich nicht gesetzt sein.

Abgesehen von einigen Ergänzungen, die gleich erläutert werden, sollte das von Ihnen erstellte Transformationsprogramm so aussehen:

```
GET
  FILE='U:\Eigene Dateien\SPSS\kfar.sav'.

DATASET NAME KfaFertigdaten WINDOW=FRONT.

* DEKADE.
RECODE gebj (1960 thru 1969=1) (1970 thru 1979=2) INTO Dekade.
value labels  dekade  1 '60er'  2 '70er'.
EXECUTE.

* LOT-Fragen Umcodieren.
RECODE lot3 lot4 lot5 lot12 (5=1) (4=2) (2=4) (1=5).
EXECUTE.

* IDGEW berechnen.
COMPUTE idgew = groesse - 100.
VARIABLE LABELS idgew 'Idealgewicht nach der Formel Größe - 100'.
EXECUTE.

* LOT berechnen.
COMPUTE lot = MEAN.6(lot1,lot3,lot4,lot5,lot8,lot9,lot11,lot12).
VARIABLE LABELS lot 'LOT-Optimismus'.
EXECUTE.

* AERGAM berechnen.
COMPUTE aergam = (aergo + aergm)/2.
VARIABLE LABELS aergam 'Mittel der Ärgervariablen'.
EXECUTE.

* AERGZ berechnen.
COMPUTE aergz =  aergm - aergo.
VARIABLE LABELS aergz 'Ärgerzuwachs durch die KFA'.
EXECUTE.

* BMI berechnen.
COMPUTE bmi = gewicht / (groesse/100)**2.
VARIABLE LABELS bmi 'Body Mass Index'.
EXECUTE.

* Alter berechnen.
COMPUTE Alter = 1999 - gebj.
EXECUTE.
```

```
* MD-Behandlung für die Motiv-Variablen.
DO IF (SUM(motiv1 to andere) = 0).
RECODE motiv1 motiv2 motiv3 motiv4 motiv5 andere (0=SYSMIS).
END IF.
EXECUTE.

* MD-Behandlung für die Methoden-Variablen, Zelle (1,1) der Tabelle.
DO IF (smg=1 and nmiss(meth1 to meth3) < 3).
RECODE meth1 meth2 meth3 (SYSMIS=0).
END IF.
EXECUTE.
* MD-Behandlung für die Methoden-Variablen, Zelle (1,2) der Tabelle.
DO IF (smg=1 and nmiss(meth1 to meth3) = 3).
RECODE smg  (1=SYSMIS).
END IF.
EXECUTE.

* MD-Behandlung für die Methoden-Variablen, Zelle (2,1) der Tabelle.
DO IF ((smg = 0)  and  (nmiss(meth1 to meth3) < 3)).
RECODE smg  (0=1).
END IF.
EXECUTE .
DO IF ((smg = 0)  and  (nmiss(meth1 to meth3) < 3)).
RECODE meth1 meth2 meth3 (SYSMIS=0).
END IF.
EXECUTE.

* MD-Behandlung für die Methoden-Variablen, Zelle (2,2) der Tabelle.
DO IF (smg=0 and nmiss(meth1 to meth3) = 3).
RECODE meth1 meth2 meth3 (SYSMIS=0).
END IF.
EXECUTE.

* MD-Behandlung für die Methoden-Variablen, Zelle (3,1) der Tabelle.
DO IF ((nmiss(smg) = 1)  and  (nmiss(meth1 to meth3) < 3)).
RECODE smg  (SYSMIS=1).
END IF.
EXECUTE.
DO IF ((nmiss(smg) = 1)  and  (nmiss(meth1 to meth3) < 3)).
RECODE meth1 meth2 meth3 (SYSMIS=0).
END IF.
EXECUTE.

* POLYMOT berechnen.
DO IF (NMISS(motiv1 to andere) = 0).
COUNT polymot=motiv1 motiv2 motiv3 motiv4 motiv5 andere(1).
VARIABLE LABELS polymot 'Anzahl der Motive für die Kursteilnahme'.
END IF.
EXECUTE.
```

```
* Variablenattribute setzen.
formats dekade idgew aergz alter polymot (f8.0) aergam (f8.1) lot bmi (f8.2).
variable width  dekade (9) idgew (7) lot (5) aergam (9)
                aergz to Alter (7) polymot (9).
variable level dekade (nominal) / idgew to polymot (scale).
variable role
  /input dekade idgew lot alter
  /target aergam aergz bmi
  /both polymot.

SAVE OUTFILE='U:\Eigene Dateien\SPSS\kfa.sav'
   /COMPRESSED.
```

In diesen Lösungsvorschlag ist etwas Handarbeit eingeflossen:

- Zwischen manchen Kommandos sind der Übersichtlichkeit halber Leerzeilen eingefügt worden. Man darf aber auf keinen Fall *innerhalb* eines Kommandos eine Leerzeile einfügen (vgl. Abschnitt 6.4).
- Die mit einem Sternchen (*) eingeleiteten Zeilen beinhalten *Kommentare*, die nachträglich eingefügt wurden, um die spätere Orientierung im Programm zu erleichtern (vgl. Abschnitt 6.4).

  Wichtig: Ein Kommentar hat ebenfalls Kommandostatus und muss daher unbedingt mit einem Punkt abgeschlossen werden. Sonst erstreckt sich der Kommentar bis zur nächsten Zeile, die entweder komplett leer ist oder mit einem Punkt endet.
- Für die nominalskalierte Variable DEKADE sollten Wertelabels definiert werden, was über das folgende VALUE LABELS – Kommando geschehen kann:

  ```
  value labels  dekade  1 '60er'  2 '70er'.
  ```

  Fügen Sie das Kommando hinter dem RECODE-Kommando ein, das die Variable DEKADE erstellt. Wie in Abschnitt 6.4 erläutert, ist bei den Schlüsselwörtern der SPSS-Kommandosprache (hier: Kommandoname VALUE LABELS) die Groß-/Kleinschreibung syntaktisch irrelevant. Bei selbst verfassten Kommandos muss man sich also nicht an die von SPSS bei automatisch erstellten Kommandos meist verwendete Großschreibung halten. Zur Groß-/Kleinschreibung beim Variablennamen DEKADE soll der Klarheit halber noch einmal an die folgenden Regeln erinnert werden:
  - Bzgl. der Identifikation von Variablen ist die Groß-/Kleinschreibung irrelevant.
  - Wird bei einer Variablen auf die optionale Beschriftung verzichtet, dann erscheint in Ergebnistabellen und Diagrammen der Name. In dieser Situation sollte an folgenden Stellen auf eine „publikationsreife“ Schreibweise des Namens geachtet werden:
    - Bei der Variablendeklaration im Dateneditor (Registerkarte **Variablenansicht**).
    - In Transformationskommandos, die eine neue Variable erzeugen, z. B.:

      ```
      RECODE gebj (1960 thru 1969=1) (1970 thru 1979=2) INTO Dekade.
      ```

    Wird in einem Kommando eine bereits *vorhandene* Variable angesprochen, ist die Groß-/Kleinschreibung irrelevant.
  - In Manuskript werden SPSS-Variablennamen zur Hervorhebung in Großbuchstaben geschrieben.

- Eventuell legen Sie Wert darauf, dass auch die neu berechneten Variablen mit einer optimalen Anzahl von Dezimalstellen angezeigt werden. Eine manuelle Einstellung per Dateneditor (vgl. Abschnitt 4.2.2) ist wenig attraktiv, weil unser Transformationsprogramm ja mit einiger Wahrscheinlichkeit mehrfach ausgeführt werden muss. Die bessere Alternative besteht darin, das Programm um ein FORMATS-Kommando zu erweitern, das die Anzahl der Dezimalstellen festlegt:

  ```
  formats  dekade idgew aergz alter polymot (f8.0) aergam (f8.1) lot bmi (f8.2).
  ```

  In der für numerische Variablen geeigneten Formatdefinition „(f*b*.*d*)“ legt man mit *b* die Gesamtbreite der Wertausgabe (entspricht dem Dateneditorattribut **Breite**) und mit *d* die Anzahl der Dezimalstellen fest. Weil bei numerischen Variablen die Gesamtbreite für uns irrelevant ist, haben wir bei den Rohvariablen auf eine Anpassung der Voreinstellung 8 verzichtet. So verfahren wir der Einheitlichkeit halber auch bei den abgeleiteten Variablen.

  Fügen Sie das FORMATS-Kommando am Ende des Transformationsprogramms ein (unmittelbar vor dem SAVE-Kommando).

- Mit den folgenden Kommandos

  ```
  variable width  dekade (9) idgew (7) lot (5) aergam (9)
                  aergz to Alter (7) polymot (9).
  variable level  dekade (nominal) / idgew to polymot (scale).
  variable role
    /input  dekade idgew lot alter
    /target  aergam aergz bmi
    /both  polymot.
  ```

  werden für die neuen Variablen eingestellt:

  - Breite der Datenfensterspalte (entspricht dem Dateneditorattribut **Spalten**)
  - Messniveau
  - Rolle

  Fügen Sie die Kommandos am Ende des Transformationsprogramms ein (hinter dem FORMATS-Kommando).

- Auf die Kommandos FORMATS, VARIABLE WIDTH, VARIABLE LEVEL und VARIABLE ROLE muss *kein* EXECUTE-Kommando folgen (vgl. Abschnitt 7.3). Weil sich diese Kommandos nicht auf die Datenmatrix beziehen, sondern auf das Datenlexikon, werden sie auf jeden Fall sofort ausgeführt.

Beachten Sie im Zusammenhang mit dem Thema Datensicherheit:

- Wenn zum Zeitpunkt der Programmausführung ein *unbenanntes* Datenblatt existiert, wird dieses vom GET-Kommando ohne Nachfrage überschrieben.
- Das SAVE-Kommando überschreibt eine eventuell vorhandene Datei **kfa.sav** ohne Nachfrage, was jedoch bei der im Manuskript vorgeschlagenen Arbeitsweise (vgl. Abschnitt 7.1.1) unproblematisch ist.

Damit ist das Transformationsprogramm zum KFA-Projekt fertig. Falls noch nicht geschehen, müssen Sie es unbedingt sichern, z. B. in das Verzeichnis **U:\Eigene Dateien\SPSS** unter dem oben vorgeschlagenen Dateinamen **kfat.sps**.

7.7.2 Transformationsprogramm ausführen

Um die spätere Erfolgskontrolle zu vereinfachen, sollte vor dem Start des Transformationsprogramms das Hauptausgabefenster leer sein, oder es sollte gar kein Ausgabefenster vorhanden sein. Auf jeden Fall sollten ältere Warnungen bzw. Fehlermeldungen aus dem Hauptausgabefenster gelöscht werden, um Unklarheiten zu vermeiden.

Lassen Sie das Transformationsprogramm ausführen, z. B. mit

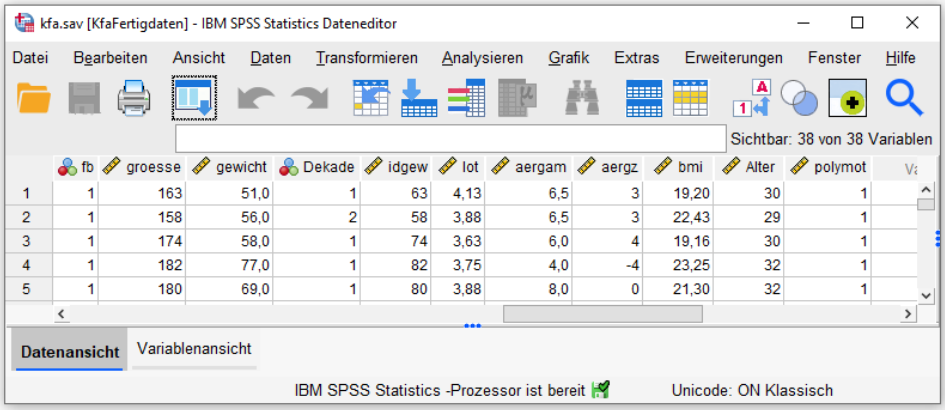
Ausführen > Alle

Bei der in Kapitel 7 beschriebenen Vorgehensweise ist beim Start des Transformationsprogramms die Rohdatendatei bereits geöffnet, und infolgedessen erscheint im Ausgabefenster die folgende Warnung, die ignoriert werden darf:

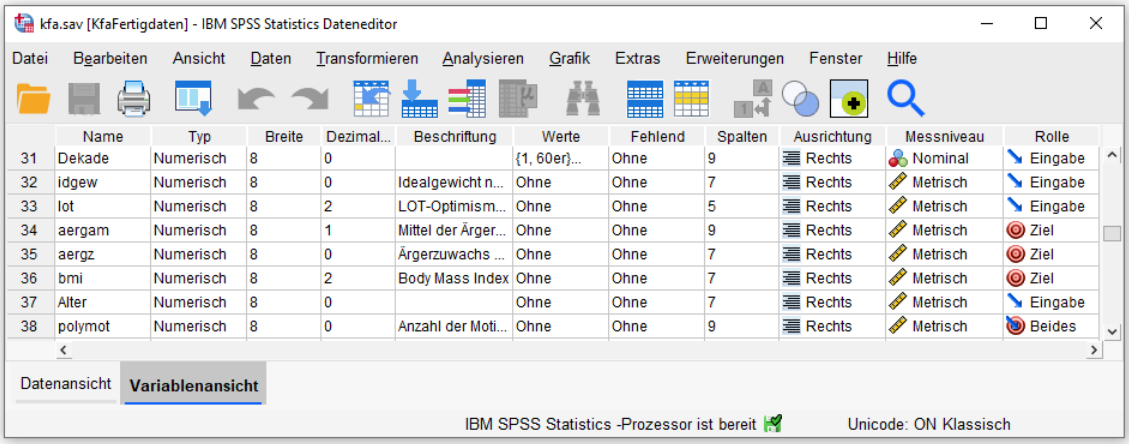
```
Warnungsnummer 67.  Befehlsname: GET FILE
Das Dokument wird bereits von einem anderen Benutzer oder Prozess verwendet.
Wenn Sie Änderungen an dem Dokument vornehmen, können diese gegebenenfalls
Änderungen anderer Benutzer überschreiben bzw.  Ihre Änderungen werden von
anderen Benutzern überschrieben.
Geöffnete Datei: U:\Eigene Dateien\SPSS\kfar.sav
```

Ansonsten enthält das Ausgabefenster nach der erfolgreichen Ausführung des Transformationsprogramms noch die ausgeführten Kommandos, sofern das in SPSS 28 per Voreinstellung deaktivierte Protokollieren der Kommandos per **Optionen**-Dialog eingeschaltet worden ist (vgl. Abschnitt 6.2).

Nach einem gelungenen Lauf unseres Transformationsprogramms existiert ein Datenblatt mit dem Namen `KfaFertigdaten`, das mit der Datei **kfa.sav** verbunden ist. Am rechten Rand der Datenmatrix sind die neuen Variablen zu finden, z. B.:



Außerdem sind die Attribute der neuen Variablen korrekt gesetzt:



Das seit Beginn unserer Arbeit am Transformationsprogramm vorhandene, mit der Rohdatendatei verbundene Datenblatt ändert sich durch die Programmausführung *nicht*, denn:

- Beim Öffnen der Rohdatendatei per Dialogbox hat das Datenblatt einen Namen erhalten (z. B. **DataSet1**).
- Das GET-Kommando des Transformationsprogramms befördert die Rohdaten in das anonyme Datenblatt, das anschließend den Namen `KfaFertigdaten` erhält.

Sie müssen vor der Erfolgskontrolle das tatsächlich relevante Datenfenster ansteuern (z. B. per **Fenster**-Menü).

Sie dürfen Ihre Erfolgskontrolle keinesfalls auf das Datenfenster beschränken, sondern müssen unbedingt das Ausgabefenster auf Fehlermeldungen und Warnungen überprüfen. SPSS stoppt nämlich die Programmausführung **nicht** beim Auftreten des ersten fehlerhaften Kommandos, sondern ignoriert das fehlerhafte Kommando und macht unverdrossen mit den nächsten Kommandos weiter. Diese arbeiten aber in der Regel aufgrund des vorangegangenen Fehlers mit falschen Zwischenergebnissen und produzieren Unsinn. Es kann also passieren, dass nach einem fehlerhaften Lauf des Transformationsprogramms alle erwarteten neuen Variablen vorhanden sind, jedoch teilweise falsche Werte enthalten.

Zwar kann es durchaus sinnvoll sein, *einzelne* Transformationskommandos ausführen zu lassen, doch sollten SPSS-Einsteiger vorläufig darauf verzichten. Damit eine korrekte Fertigdatendatei entsteht, müssen alle Kommandos

- genau **einmal**
- und in der **korrekten Reihenfolge**

ausgeführt werden. Bei der separaten Ausführung einzelner Kommandos kann man diese Regeln verletzen. Wiederholt man z. B. das Umcodieren der LOT-Variablen, so entsteht wieder der Ausgangszustand! Von solchen Tücken bleibt verschont, wer das Transformationsprogramm stets komplett ausführen lässt.

8 Hypothesentests

Damit die bald anstehende Prüfung der zentralen Hypothesen unserer Kursstudie (emotionaler Effekt kontrafaktischer Alternativen, Ärgerdämpfung durch Optimismus) fehlerfrei über die Bühne geht, beschäftigen wir uns in diesem Kapitel mit den Grundprinzipien und den Voraussetzungen von statistischen Hypothesentests. Obwohl die im Rahmen der Kursstudie konkret geplanten Verfahren (t-Test für abhängige Stichproben, bivariate lineare Regression) als Beispiele im Vordergrund stehen, lassen sich die Erläuterungen gut auf andere Testverfahren übertragen. Wer sich bei den im bisherigen Kursverlauf (z. B. in Abschnitt 5.5) gelegentlich eingesetzten statistischen Hypothesenprüfungen unsicher gefühlt hat, wird hoffentlich nun ein solides Verständnis (wieder-)erlangen.

8.1 Grundprinzipien am Beispiel des Einstichproben - t-Tests

In diesem Abschnitt werden die Grundprinzipien der Inferenzstatistik nach dem binären Entscheidungskonzept von Neyman und Pearson am Beispiel unserer KFA-Hypothese demonstriert. Dabei handelt es sich *nicht* um eine systematische Behandlung des Themas, die erheblich mehr Platz beanspruchen würde (siehe z. B. Eid et al. 2017, Kap. 8; Lehmann 1993). Im Wesentlichen sollen die statistischen Entscheidungsregeln so präsentiert werden, dass sie mit Hilfe der SPSS-Ausgaben unmittelbar umgesetzt werden können. Zumindest in älteren Statistikbüchern findet man nämlich Formulierungen mit wenig Bezug zu den heute üblichen Ausgaben von Statistikprogrammen.

8.1.1 Gerichtete Hypothese über den KFA-Effekt

Wenn mit μ_O der Erwartungswert (Populationsmittelwert) des Merkmals AERGO und mit μ_M der Erwartungswert des Merkmals AERGM bezeichnet wird, dann lautet unser KFA-Testproblem:

$$H_0 : \mu_M \leq \mu_O \quad \text{vs.} \quad H_1 : \mu_M > \mu_O$$

Mit Hilfe der Differenz- bzw. Zuwachsvariablen AERGZ := AERGM - AERGO, deren Erwartungswert mit μ_Z bezeichnet werden soll, lässt sich das Testproblem äquivalent noch kompakter formulieren:

$$H_0 : \mu_Z \leq 0 \quad \text{vs.} \quad H_1 : \mu_Z > 0$$

Bei der Reformulierung wird die folgende Identität ausgenutzt (Linearität des Erwartungswerts):

$$\mu_Z = \mu_M - \mu_O$$

Durch die Betrachtung der Differenzvariablen AERGZ lässt sich der t-Test für abhängige Stichproben auf den **Einstichproben - t-Test** zurückführen, der nun beschrieben werden soll.

8.1.2 Voraussetzungen für den Einstichproben - t-Tests

Wir setzen voraus, dass die Differenzvariable AERGZ in der Population normalverteilt ist mit dem Erwartungswert μ_Z und der Varianz σ_Z^2:

$$\text{AERGZ} \sim N(\mu_Z, \sigma_Z^2)$$

Für die *n* AERGZ-Beobachtungen in der Stichprobe nehmen wir an, dass sie durch **unabhängiges** Ziehen aus der eben beschriebenen Population entstanden sind.

Wir haben also ...

- eine Voraussetzung über die Verteilung von AERGZ in der Population
- und eine Voraussetzung über die Stichprobenziehung.

8.1.3 Teststatistik

Ein inferenzstatistisches Entscheidungsverfahren verwendet eine sogenannte **Teststatistik *T*** (synonym: **Prüfgröße**), die aus den Stichprobendaten berechnet werden kann.

8.1.3.1 Anforderungen

Für eine Teststatistik muss gelten:

- Sie ist **indikativ für Abweichungen der wahren Populationsverteilung von der Nullhypothesenbehauptung**, indem ihr Betrag stochastisch mit der Effektstärke wächst. Die Teststatistik quantifiziert also, wie gut bzw. schlecht die Nullhypothese mit den Stichprobendaten vereinbar ist.
  Wenn wir uns auf das konkrete Hypothesenpaar
  $$H_0 : \mu_Z \leq 0 \quad \text{vs.} \quad H_1 : \mu_Z > 0$$
  beschränken, dann führt eine Abweichung der wahren Populationsverteilung von der Nullhypothesenverteilung zu einem *positiven* Wert der Prüfgröße *T*.
- Es ist bekannt, welcher **Verteilung die Teststatistik *T* bei gültiger Nullhypothese** folgt. Damit lässt sich für den konkreten Wert T_{emp} der Teststatistik in einer bestimmten Stichprobe berechnen, mit welcher Wahrscheinlichkeit eine Nullhypothesenpopulation Zufallsstichproben mit einer Teststatistikausprägung $\geq T_{emp}$ liefert. Ist diese sogenannte **Überschreitungswahrscheinlichkeit** sehr klein, dann liegt der Schluss nahe, dass die konkret vorliegende Stichprobe *nicht* aus einer Nullhypothesenpopulation stammt, dass also in der zur Stichprobe gehörenden Population die Nullhypothese *nicht* gilt.

Von *der* Überschreitungswahrscheinlichkeit zu sprechen, ist allerdings ungenau, weil die Nullhypothese i.A. (jedenfalls bei einer *gerichteten* Fragestellung) als *Menge von Verteilungen* zu beschreiben ist. Diese Menge enthält viele (meist unendlich viele) Verteilungen, die *unterschiedliche* Überschreitungswahrscheinlichkeiten liefern. In unserem Beispiel enthält die Nullhypothese alle Normalverteilungen $N(\mu_Z, \sigma_Z^2)$ mit $\mu_Z \leq 0$.

Um das Problem mit einer (unendlich) großen Familie von Nullhypothesenverteilungen zu lösen, stellt man eine „Worst Case" - Analyse an und betrachtet die Verteilung der Teststatistik unter derjenigen konkreten Nullhypothese, welche die *maximale* Überschreitungswahrscheinlichkeit liefert. Ist die maximale Überschreitungswahrscheinlichkeit klein genug, dann sind es alle, und wir können schließen, dass in der zur Stichprobe gehörigen Population die Nullhypothese *nicht* gilt.

In unserem konkreten Fall ist die *Worst-Case* - Nullhypothese, bzw. die Menge der *Worst-Case* - Nullhypothesen leicht zu ermitteln: Es ist die Familie der Normalverteilungen $N(0, \sigma_Z^2)$ mit dem Erwartungswert $\mu_Z = 0$ und einer beliebigen Varianz σ_Z^2. Nur für diese Nullhypothesenverteilungen muss bekannt sein, welcher Verteilung die Teststatistik *T* folgt, um die entscheidungsrelevante Überschreitungswahrscheinlichkeit berechnen zu können. Wir haben es immer noch mit einer unendlich großen Menge von Nullhypothesenverteilungen zu tun, doch wie sich gleich

zeigen wird, führen alle Elemente dieser Menge zur selben Verteilung der Teststatistik und damit zur selben Überschreitungswahrscheinlichkeit.

8.1.3.2 Die Prüfgröße zum Einstichproben - t-Test

Zur Prüfung von Hypothesen über den Erwartungswert der Variablen AERGZ eignet sich die folgende Teststatistik T_Z:

$$T_Z := \frac{\bar{Z}}{S_Z}\sqrt{n} \quad \text{mit} \quad \bar{Z} := \frac{1}{n}\sum_{i=1}^{n} Z_i \quad \text{und} \quad S_Z := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2}$$

Dabei ist $\bar{Z}$ das arithmetische Stichprobenmittel und S_Z die Wurzel aus dem erwartungstreuen Schätzer der Populationsvarianz σ_Z^2.

T_Z erfüllt die Abschnitt 8.1.3.1 formulierten Anforderungen an eine Teststatistik:

- Die ergebnisabhängige T_Z - Komponente $\frac{\bar{Z}}{S_Z}$ ist offenbar ein Schätzer für die Effektstärke d_z in der Population, die folgendermaßen definiert ist (vgl. Abschnitt 2.3.2):

  $$d_z := \frac{\mu_z}{\sigma_z}$$

  Folglich wächst T_Z stochastisch bei zunehmender Effektstärke im Sinne unserer Alternativhypothese ($\mu_Z > 0$).

- Für $\mu_Z = 0$ besitzt T_Z (bei beliebigem Nebenparameter σ_Z^2) unter den in Abschnitt 8.1.2 genannten Voraussetzungen eine t-Verteilung mit *n* - 1 Freiheitsgraden. Damit kennen wir das Verhalten der Teststatistik *am Rand* der Nullhypothese. Man kann zeigen, dass die Wahrscheinlichkeit, einen T_Z - Wert größer oder gleich T_{emp} zu erhalten, bei $\mu_Z = 0$ maximal wird. Um beurteilen zu können, ob *alle* Nullhypothesen-Überschreitungswahrscheinlichkeiten klein genug sind, muss man also nur den Fall $\mu_Z = 0$ betrachten (die Worst-Case - Nullhypothese).

Bevor wir uns der Testentscheidung unter Verwendung der T_Z - Prüfgröße zuwenden, beobachten wir noch, dass diese Statistik ein typisches Konstruktionsprinzip für Hypothesentests über Populationsparameter realisiert. Für das Stichprobenmittel $\bar{Z}$ (als Zufallsvariable aufgefasst) ergibt sich die Varianz

$$\mathrm{Var}(\bar{Z}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} Z_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Z_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Z^2 = \frac{n\sigma_Z^2}{n^2} = \frac{\sigma_Z^2}{n}$$

und damit die Standardabweichung (der sogenannte *Standardfehler*)

$$\sqrt{\mathrm{Var}(\bar{Z})} = \frac{\sigma_Z}{\sqrt{n}}.$$

Folglich schätzt $\frac{S_Z}{\sqrt{n}}$ den Standardfehler des Stichprobenmittelwerts, und T_Z ist gerade der **Quotient aus dem Stichprobenmittelwert und seinem geschätzten Standardfehler**:

$$T_Z = \frac{\bar{Z}}{S_Z}\sqrt{n} = \frac{\bar{Z}}{\frac{S_Z}{\sqrt{n}}}$$

Teststatistiken von analoger Bauart sind uns schon bei den approximativen Tests zur Schiefe bzw. Wölbung einer univariaten Verteilung in Abschnitt 5.5 begegnet.

8.1.4 Entscheidungsregel

Die maximale Überschreitungswahrscheinlichkeit (ab jetzt gelegentlich der Kürze halber als *p-Wert* bezeichnet) unter der Nullhypothese, die ab jetzt mit $P_{H_0}(T_Z \geq T_{emp})$ bezeichnet werden soll, tritt nach obigen Überlegungen bei $\mu_Z = 0$ (mit beliebiger Varianz σ_Z^2) auf und kann (von einer Statistik-Software) leicht berechnet werden.

Bei einem akzeptierten **Fehlerrisiko erster Art** von $\alpha = 0{,}05$ verwendet man die folgende **Entscheidungsregel**:

$$P_{H_0}(T_Z \geq T_{emp}) \begin{cases} \geq 0{,}05 & \Rightarrow \ H_0 \text{ beibehalten} \\ < 0{,}05 & \Rightarrow \ H_0 \text{ verwerfen (Entscheidung für } H_1) \end{cases} \qquad (8\text{-}1)$$

Die Nullhypothese wird also abgelehnt, wenn die Teststatistik in der beobachteten Stichprobe einen Wert besitzt, der bei Zufallsstichproben aus einer Nullhypothesenpopulation (genauer: aus einer Population mit $\mu_Z = 0$), nur relativ selten (mit einer Wahrscheinlichkeit kleiner 0,05) erreicht oder übertroffen wird.

8.1.5 Kritischer Wert und Ablehnungsbereich

In Statistiklehrbüchern wird oft für den gerichteten Einstichproben - t-Test (mit der Alternativhypothese $\mu_Z > 0$) zum Niveau $\alpha = 0{,}05$ ein **kritischer Wert** T_{krit} so bestimmt, dass gilt:

$$P_{H_0}(T_Z \geq T_{krit}) = 0{,}05$$

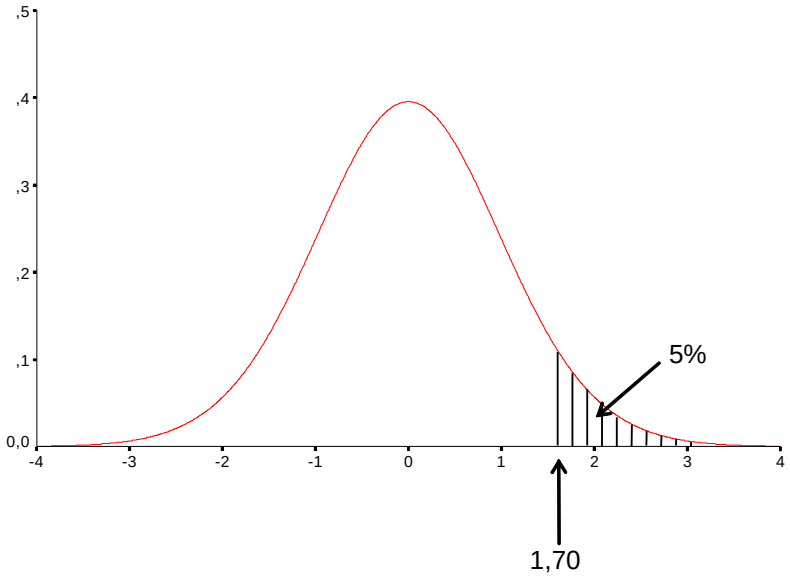
Damit kann obige Entscheidungsregel äquivalent folgendermaßen formuliert werden:

$$T_{emp} \begin{cases} \leq T_{krit} \Rightarrow & H_0 \text{ beibehalten} \\ > T_{krit} \Rightarrow & H_0 \text{ verwerfen (Entscheidung für } H_1) \end{cases} \qquad (8\text{-}2)$$

In unserer Situation ist T_{krit} gerade das 95. Perzentil der t-Verteilung mit n - 1 Freiheitsgraden. Bei der Stichprobengröße $n = 31$ erhalten wir $T_{krit} = 1{,}70$.

Wir haben bei den approximativen Tests zur Schiefe und Wölbung einer univariaten Verteilung (siehe Abschnitt 5.5) die Testentscheidung anhand von kritischen Werten vorgenommen. Dort waren wir ausnahmsweise in der Lage, keine p-Werte zu kennen, aber die kritischen Werte der Teststatistiken (als Perzentile der Standardnormalverteilung) leicht ermitteln zu können. Weil SPSS und vergleichbare Statistikprogramme in der Regel p-Werte liefern, werden die im Anhang vieler Statistiklehrbücher tabellierten kritischen Werte wichtiger Prüfverteilungen (z. B. N, t, F, χ^2) nur noch selten benötigt.

Die folgende Abbildung zeigt die Wahrscheinlichkeitsdichte der t-Verteilung mit 30 Freiheitsgraden und den **H_0-Ablehnungsbereich** (|||) bei einseitiger Fragestellung im Sinne unserer KFA-Hypothese:



Diese Dichte beschreibt das Verteilungsverhalten einer Zufallsgröße, zu der eine einzelne Realisation folgendermaßen zu ermitteln ist: Ziehe aus einer Population mit

$$\text{AERGZ} \sim N(0,\ \sigma_Z^2)$$

eine Zufallsstichprobe der Größe $n = 31$, ermittle die AERGZ-Werte und berechne T_Z.

Damit beschreibt die Abbildung das Verhalten der T_Z - Werte von Stichproben aus einer für unseren Test besonders ungünstigen Nullhypothesenpopulation (mit $\mu_Z = 0$). Wir kommen zu einer Testentscheidung, indem wir unser Stichprobenergebnis T_{emp} vor dem Hintergrund dieses Erwartungshorizonts beurteilen. Wir lehnen die Nullhypothese (eine Menge von Verteilungen) ab, wenn selbst der extreme Vertreter dieser Menge (die Worst-Case - Nullhypothese) nur selten (mit einer Wahrscheinlichkeit < 0,05) ein Stichprobenergebnis mit einem T_Z - Wert größer oder gleich T_{emp} liefert.

8.1.6 Akzeptiertes Risiko erster Art

Wenn wir aus einer Nullhypothesenpopulation mit $\mu_Z = 0$ eine Zufallsstichprobe der Größe $n = 31$ ziehen und T_{emp} ermitteln, werden wir mit der Wahrscheinlichkeit 0,05 einen Wert größer $T_{krit} = 1{,}70$ erhalten und *falsch* gegen die H_0 entscheiden, also einen **Fehler erster Art** begehen. Für die extremen Nullhypothesenpopulationen mit $\mu_Z = 0$ (und einer beliebigen Varianz σ_Z^2) ist das Risiko erster Art also gleich 0,05. Für alle anderen Elemente aus der Nullhypothesen-Verteilungsfamilie ist es kleiner.

Das akzeptierte Risiko erster Art sollte umso niedriger angesetzt werden, je gravierender (schädlicher, teurer) das irrtümliche Ablehnen einer gültigen Nullhypothese ist.

8.1.7 Faktoren mit Einfluss auf das Risiko zweiter Art

Das Risiko, bei Gültigkeit der *Alternativhypothese* falsch zu entscheiden (**Fehler zweiter Art**, **β-Fehler**), hängt von folgenden Faktoren ab:

- **Effektstärke**

  Bei unserem KFA-Testproblem ist die Effektstärke durch $\frac{\mu_Z}{\sigma_Z}$ definiert. Weil die Prüfstatistik im Kern ein Stichprobenschätzer für die Effektstärke ist, gilt offenbar: Je größer die Effektstärke, desto wahrscheinlicher ist ein T_{emp} - Wert mit $P_{H_0}(T_Z \geq T_{\text{emp}}) < 0{,}05$ bzw. $T_{emp} > T_{\text{krit}}$, also eine korrekte Entscheidung gegen die H_0.

- **Akzeptiertes Fehlerrisiko erster Art**

  Reduziert man den akzeptierten α-Fehler, dann steigt der kritische Wert T_{krit}. Folglich steigt auch das Risiko dafür, dass der T_{emp} - Wert einer Stichprobe trotz gültiger Alternativhypothese den kritischen Wert T_{krit} *nicht* übertrifft. In diesem Fall kommt es zu einem β-Fehler.

- **Korrekte Anwendung der ein- oder zweiseitigen Testung**

  Wer sich auf die Richtung des Effekts festlegt und einseitig testet, wird im Fall einer korrekten Annahme mit einer höheren Power (Entdeckungswahrscheinlichkeit) im Vergleich zu der in Statistikprogrammen oft voreingestellten zweiseitigen Testung belohnt (siehe Abschnitt 8.1.8).

- **Sensibilität des verwendeten Signifikanztests**

  Die Wahrscheinlichkeit dafür, dass ein bestimmter Populationseffekt in einer Stichprobe zu einem signifikanten Testergebnis führt, wächst generell mit der **Stichprobengröße**. Wir betrachten exemplarisch, was beim Einstichproben - t-Test passiert, wenn bei gültiger Alternativhypothese ($\mu_Z > 0$) die Stichprobengröße *n* steigt:

  - Der Quotient $\frac{\bar{Z}}{S_Z}$ konvergiert stochastisch gegen $\frac{\mu_Z}{\sigma_Z}$, also gegen eine feste Zahl > 0.
  - Die Prüfgröße

    $$T_Z := \frac{\bar{Z}}{S_Z}\sqrt{n}$$

    wächst stochastisch gegen Unendlich.
  - T_{krit} konvergiert gegen 1,65 (95. Perzentil der Standardnormalverteilung).

  Folglich steigt mit der Stichprobengröße die Wahrscheinlichkeit, die T_{krit} - Hürde zu nehmen, also die falsche H_0 zu verwerfen. Auch bei einem kleinen Effekt ($\frac{\mu_Z}{\sigma_Z}$ nur geringfügig > 0), kann man die Wahrscheinlichkeit für ein signifikantes Ergebnis (die Power des Testverfahrens) durch eine Steigerung der Stichprobengröße beliebig erhöhen (vgl. Abschnitt 2.3.2 über die Stichprobenumfangsplanung).

  Alternative Testverfahren zum selben Entscheidungsproblem verwenden unterschiedliche Prüfgrößen und unterscheiden sich bei ihren Annahmen über das Messniveau und die Verteilung der beteiligten Variablen. Von zwei mathematisch ausgereiften Verfahren besitzt dasjenige mit den stärkeren Annahmen die bessere Power, *falls seine Annahmen erfüllt sind*. Wir werden zur Prüfung der allgemeinpsychologischen Hypothese den t-Test

für verbundene Stichproben nur dann einsetzen, wenn sich die Variable AERGZ in unserer Stichprobe als annähernd normalverteilt erweist. Sind die Voraussetzungen eines Verfahrens erheblich verletzt, darf es wegen potentiell verfälschter Ergebnisse nicht verwendet werden. In der Regel wäre das Verfahren in dieser Situation auch eine schlechte Wahl. Ob bereits eine erhebliche Verletzung der Voraussetzungen vorliegt, oder noch auf die Robustheit eines Verfahrens vertraut werden kann, ist leider nicht immer auf objektive und automatisierte Weise zu entscheiden.

Wie Sie aus der Stichprobenumfangsplanung in Abschnitt 2.3.2 wissen, kann man zum t-Test für abhängige Stichproben für eine konkret vorgegebene Effektstärke d_z, eine Testausrichtung (ein- oder zweiseitig) und ein α-Fehlerniveau …

- die erforderliche Stichprobengröße zu einer gewünschten Teststärke (z. B. $1 - \beta = 0{,}8$) ermitteln,
- die Teststärke bzw. das β-Fehler-Risiko zu einer festen Stichprobengröße ausrechnen.

8.1.8 Zweiseitiges Testproblem

Passend zu unserer KFA-Hypothese haben wir bislang das *einseitige* Testproblem mit der Alternativhypothese $\mu_Z > 0$ behandelt. Wir wollen noch das folgende **zweiseitige Testproblem** betrachten:

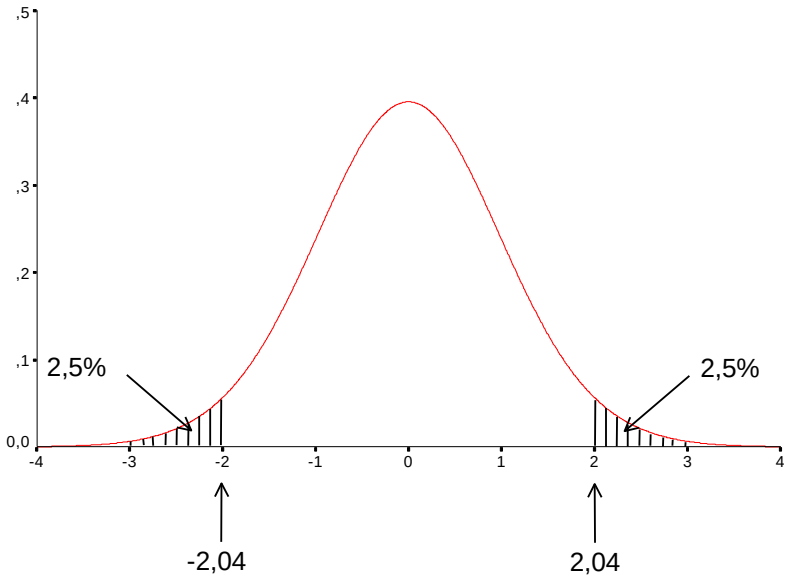
$$H_0 : \mu_M = \mu_O \quad \text{vs.} \quad H_1 : \mu_M \neq \mu_O$$

bzw.

$$H_0 : \mu_Z = 0 \quad \text{vs.} \quad H_1 : \mu_Z \neq 0$$

Die H_0 des zweiseitigen Tests ist identisch mit dem *Rand* der H_0 zum einseitigen Test.

Man verwendet beim zweiseitigen Test dieselbe Teststatistik T_Z wie beim einseitigen Test. Nun sind aber *betragsmäßig* große T_{emp} - Werte (mit *positivem oder negativem* Vorzeichen) indikativ für eine Abweichung von der Nullhypothese. Nach einem generellen Prinzip der Testkonstruktion müssen *alle* Elemente der Alternativhypothese (im zweiseitigen Fall also mit $\mu_Z < 0$ oder $\mu_Z > 0$) eine faire Chance haben, sich in einem signifikanten Ergebnis zu artikulieren. Anderenfalls resultiert ein sogenannter *verfälschter Test*. Daher sind beim zweiseitigen Test *zwei* symmetrisch angeordnete Ablehnungsbereiche zu verwenden, damit T_{emp} - Werte mit einem großen positiven *oder* negativen Abstand vom Wert 0 (T_Z - Erwartungswert unter der Nullhypothese) zum signifikanten Testergebnis führen können:



Damit ein zweiseitiger Test das α-Niveau einhält, muss für seinen kritischen Wert $T_{\text{krit},2}$ gelten:

$$P_{H_0}(T_Z \leq -T_{\text{krit},2}) + P_{H_0}(T_Z \geq T_{\text{krit},2}) = P_{H_0}(|T_Z| \geq |T_{\text{krit},2}|) = \alpha$$

Für $\alpha = 0{,}05$ und $n = 31$ erhält man $T_{\text{krit},2} = \pm 2{,}04$, und dieser Wert liegt betragsmäßig deutlich über dem kritischen Wert für den einseitigen Test ($T_{\text{krit}} = 1{,}70$; siehe Abschnitt 8.1.5).

Statistikprogramme liefern keine kritischen Werte, sondern die Überschreitungswahrscheinlichkeit

$$P_{H_0}(|T_Z| \geq |T_{\text{emp}}|)$$

Damit führt man analog zum einseitigen Fall den Testentscheid zum gewünschten α-Niveau durch, z. B.:

$$P_{H_0}(|T_Z| \geq |T_{\text{emp}}|) \begin{cases} \geq 0{,}05 & \Rightarrow \quad H_0 \text{ beibehalten} \\ < 0{,}05 & \Rightarrow \quad H_0 \text{ verwerfen (Entscheidung für } H_1) \end{cases} \tag{8-3}$$

8.1.9 Beziehung zwischen dem ein- und dem zweiseitigen p-Wert

Weil unsere Teststatistik T_Z bei $\mu_Z = 0$ einer t-Verteilung folgt und somit symmetrisch um den Wert 0 verteilt ist, gilt für $T_{\text{emp}} \geq 0$:

$$P_{H_0}(T_Z \geq T_{\text{emp}}) = \frac{1}{2} \cdot P_{H_0}(|T_Z| \geq |T_{\text{emp}}|) \tag{8-4}$$

Der p-Wert des einseitigen t-Tests ergibt sich also durch Halbieren aus dem p-Wert des zweiseitigen t-Tests (, sofern die Prüfgröße das von der einseitigen H_1 behauptete Vorzeichen besitzt). Während SPSS bis zur Version 27 bei t-Tests nur den *zweiseitigen* p-Wert mitteilt hat, erscheinen seit der Version 28 beide Varianten in der Ausgabe zu den t-Tests.

Sie dürfen den Zusammenhang in Gleichung (8-4) aber nicht auf beliebige Testverfahren generalisieren. Wir werden z. B. im Zusammenhang mit der Kreuztabellenanalyse den exakten Test von Fisher kennenlernen, bei dem eine analoge Gleichung *nicht* gilt (vgl. Abschnitt 14.4.5.1).

8.2 Teststatistik und Annahmen im Modell der bivariaten linearen Regression

Unsere *differentialpsychologische* Hypothese bezieht sich auf den Koeffizienten β_1 in der linearen Regression von AERGAM auf LOT:

$$\text{AERGAM} = \beta_0 + \beta_1 \text{LOT} + \varepsilon, \;\; \varepsilon \sim N(0, \sigma^2)$$

Die Hypothesen des Testproblems lauten:

$$H_0 : \beta_1 \geq 0 \quad \text{vs.} \quad H_1 : \beta_1 < 0$$

8.2.1 Teststatistik

Es kommt eine Teststatistik zum Einsatz, die sich im vorliegenden Fall der *bivariaten* linearen Regression besonders bequem mit Hilfe der Stichprobenkorrelation *r* zwischen dem Kriterium und dem Regressor ausdrücken lässt:

$$T_r := \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Als Effektstärkemaß zur bivariaten linearen Regression haben wir in Abschnitt 2.3.3 kennengelernt:

$$f^2 = \frac{\rho^2}{1-\rho^2}$$

Offenbar ist das Quadrat der Teststatistik

$$T_r^2 = \frac{r^2}{1-r^2}(n-2)$$

abgesehen vom ergebnisunabhängigen Faktor (*n* - 2) ein Schätzer für die Effektstärke. Damit zeigt die Prüfgröße T_r an, wie stark die Daten von der Nullhypothesenbehauptung abweichen.

Außerdem folgt T_r bei einer Nullhypothesenpopulation mit $\beta_1 = 0$ einer t-Verteilung mit *n* - 2 Freiheitsgraden, sofern die Voraussetzungen des Regressionsmodells erfüllt sind (siehe Abschnitt 8.2.2). Man kann sich auf dieses extreme Element der Nullhypothesen-Verteilungsfamilie beschränken, weil es die größte Überschreitungswahrscheinlichkeit für eine Stichprobenrealisation T_{emp} der Prüfgröße T_r liefert (vgl. Abschnitt 8.1.4). Ob als Überschreitungswahrscheinlichkeit $P_{H_0}(T_r \leq T_{emp})$ zu bestimmen ist (wie bei unserer differentialpsychologischen Hypothese) oder $P_{H_0}(T_r \geq T_{emp})$, hängt von dem in der Alternativhypothese behaupteten Vorzeichen des Regressionskoeffizienten β_1 ab.

Für statistisch besonders Interessierte soll noch eine alternative Darstellung für T_r vorgeführt werden. Weil der Stichprobenschätzer b_1 des Steigungskoeffizienten in folgender Beziehung zur Stichprobenkorrelation *r* und den Schätzern $\hat{\sigma}_Y$ und $\hat{\sigma}_X$ für die Standardabweichungen des Kriteriums *Y* und des Regressors *X* steht,

$$b_1 = r\,\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}$$

und der geschätzte Standardfehler zu b_1 gleich

$$\hat{\sigma}(b_1) = \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \frac{\sqrt{1-r^2}}{\sqrt{n-2}}$$

ist (siehe z. B. Cohen et al. 2003, S. 42), kann die Prüfgröße T_r als Quotient aus dem Stichprobenschätzer b_1 und seinem geschätzten Standardfehler geschrieben werden:

$$T_r = b_1 \frac{\hat{\sigma}_X}{\hat{\sigma}_Y} \frac{\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{b_1}{\hat{\sigma}(b_1)}$$

Diese Prüfgrößenkonstruktion ist typisch für Tests über Populationsparameter.

8.2.2 Annahmen

Die Annahmen im Modell der bivariaten linearen Regression werden anschließend der bequemeren Schreibweise halber für ein Kriterium *Y* und einen Regressor *X* beschrieben.

Um verbreiteten Missverständnissen entgegenzuwirken, soll vorab betont werden:

- Es wird *keine* Annahme über die Verteilung des Regressors benötigt.
- Es wird keine Annahme über die Verteilung des Kriteriums (seine sogenannte *Randverteilung*) benötigt.
- Es sind die **Residuen des Modells**, welche die gleich zu erläuternden Verteilungsvoraussetzungen (Erwartungswert 0 für alle Werte des Regressors, Normalität, Varianzhomogenität, Unkorreliertheit) erfüllen müssen.

8.2.2.1 Linearität

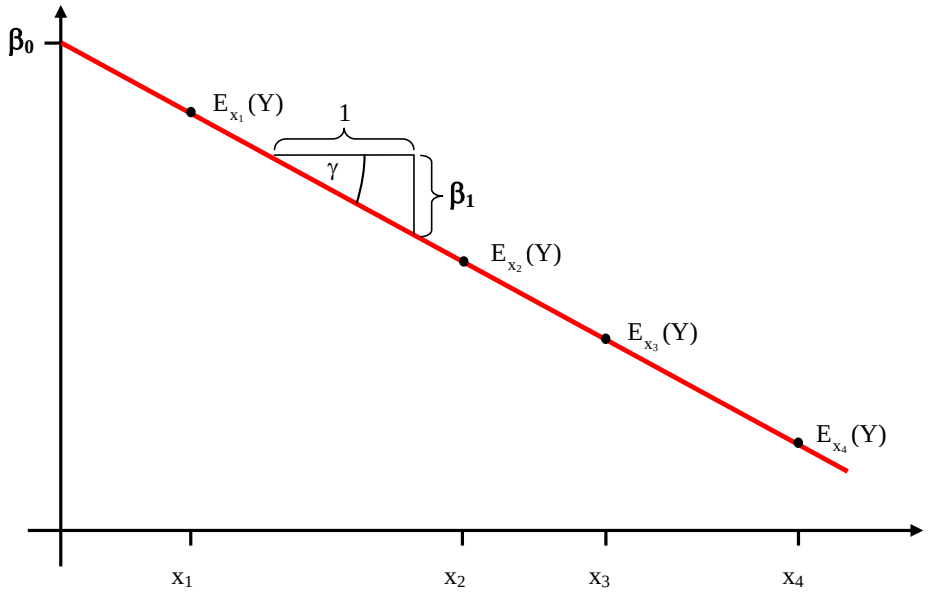
Für die (nicht direkt beobachtbare) Fehler- bzw. Residualvariable ε wird angenommen, dass sie für jeden festen *X*-Wert den Erwartungswert 0 hat ($\varepsilon \sim N(0, \sigma^2)$). Folglich hängt der *X*-bedingte Erwartungswert $E_X(Y)$ des Kriteriums **linear** von *X* ab:

$$E_X(Y) = \beta_0 + \beta_1 X$$

Für jeden festen *X*-Wert liegt der Erwartungswert $E_X(Y)$ der zugehörigen *Y*-Werte auf der **Regressionsgeraden** durch die Punktepaare

$$(X, \beta_0 + \beta_1 X)$$

Dabei ist β_0 der Schnittpunkt der Regressionsgeraden mit der Y-Achse (Ordinatenabschnitt) und β_1 die Steigung der Regressionsgeraden. Unserer differentialpsychologischen Hypothese entspricht eine Regressionsgerade mit negativer Steigung, weil wir eine Ärgerreduktion bei zunehmendem Optimismus erwarten:



Zur Interpretation des Koeffizienten β_1: Erhöht man *X* um eine Einheit, so sinkt modellgemäß der Erwartungswert $E_X(Y)$ um β_1 Einheiten.

Aufgrund der geschätzten Regressionskoeffizienten $\hat{\beta}_0$ und $\hat{\beta}_1$ bestimmt man zu einem beobachteten Werte x_i des Regressors eine Schätzung für den erwarteten Wert

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

und eine Schätzung für das Residuum:

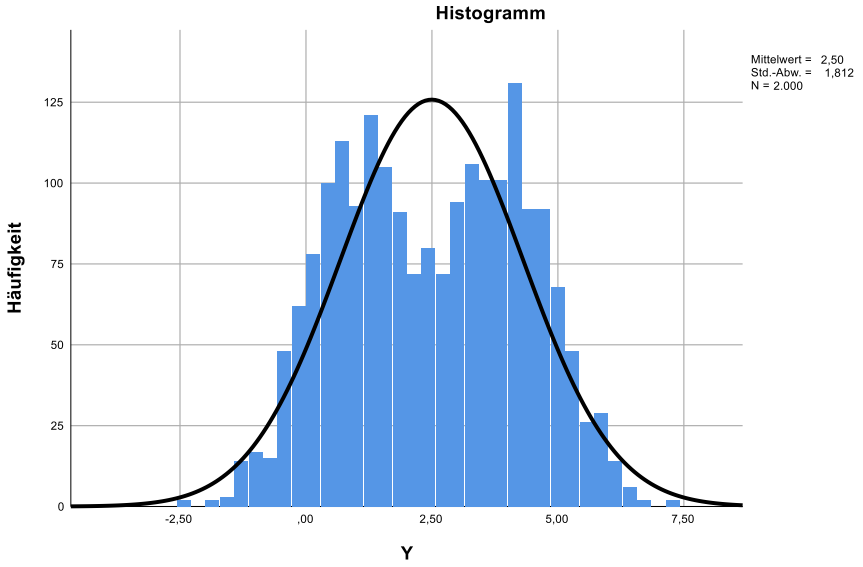
$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Ist die Linearitätsannahme erfüllt, dann haben die *geschätzten* Residuen für beliebige Werte des Regressors den Erwartungswert 0. Folglich lässt sich die Gültigkeit der Linearitätsannahme durch eine Untersuchung der geschätzten Residuen überprüfen.

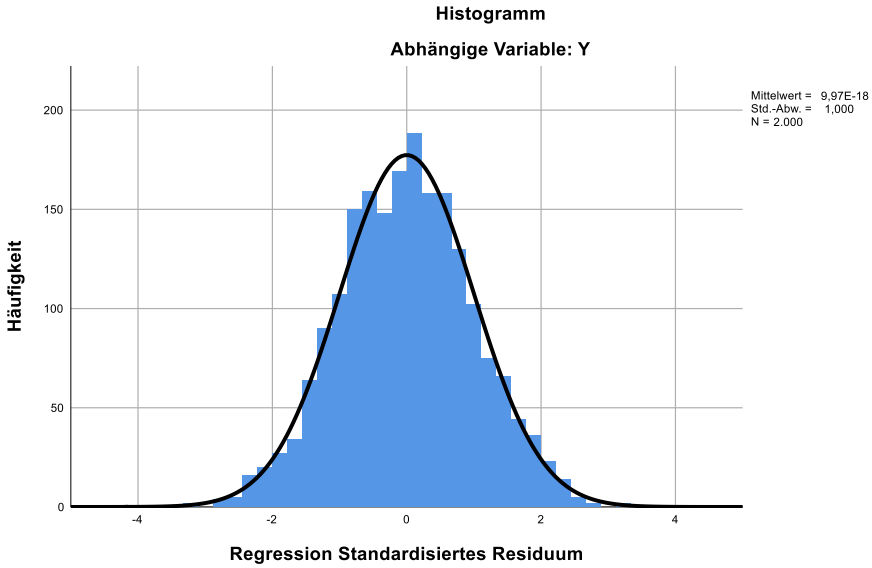
8.2.2.2 Normalität der Residuen

Für die Residualvariable ε wird angenommen, dass sie normalverteilt ist. Sie dürfen sich vorstellen, dass es für jede *X*-Ausprägung eine Normalverteilung potentieller ε-Werte gibt, aus der zufällige Realisationen gezogen werden, die zusammen mit dem konstanten Anteil $\beta_0 + \beta_1 X$ die Realisationen der abhängigen Variablen *Y* ergeben.

In der bivariate Regression mit einem effektstarken dichotom-kategorialen Regressor *X* zeigt die Randverteilung des Kriteriums *Y* wegen des *X*-Effekts eine deutliche Abweichung von der Normalverteilung, z. B.:



Es ist also sinnlos, die *Randverteilung* von *Y* auf Normalität zu untersuchen. Die Forderung nach der Normalität der *Residuen* läuft im Fall eines dichotom-kategorialen Regressors darauf hinaus, dass die Kriteriumsvariable in den beiden Gruppen bzw. Populationen normalverteilt sein muss, was im Beispiel sehr gut erfüllt ist. Hier ist die gemeinsame Verteilung der geschätzten Residuen aus den beiden Teilstichproben zu sehen:



Für jeden Fall wird sein Residuum geschätzt, indem von seinem *Y*-Wert der Gruppen*mittelwert* subtrahiert wird. Das wahre Residuum ist eine nicht beobachtbare Variable, weil zur Berechnung der unbekannte Gruppen*erwartungswert* benötigt wird. Bei der linearen Regression sind generell nur *geschätzte* Residuen verfügbar, weil für die zur Berechnung der Residuen erforderlichen Regressionskoeffizienten nur Schätzungen vorhanden sind.

8.2.2.3 Varianzhomogenität der Residuen

Die Normalverteilungen der ε-Variablen zu den verschiedenen *X*-Ausprägungen haben voraussetzungsgemäß alle dieselbe Varianz σ^2. Statt von *Varianzhomogenität* spricht man auch von *Homoskedastizität*.

8.2.2.4 Unkorreliertheit der Residuen

Für die Residuen zu den einzelnen Beobachtungen (Fällen) in der Stichprobe wird die Unkorreliertheit vorausgesetzt.

In unserer querschnittlich angelegten Studie mit Einzelpersonen als Untersuchungseinheiten sind korrelierte Residuen nahezu ausgeschlossen. Begründete Zweifel an der Unkorreliertheit bestehen in folgenden Situationen:

- Zeitreihen-Stichprobe (z. B. Arbeitsmarktdaten aus 40 aufeinander folgenden Jahren)
  Zu Regressionsmodellen für seriell abhängige Daten siehe z. B. Baltes-Götz (2019).
- Panel-Stichprobe (z. B. 100 Teilnehmer und 5 Untersuchungswellen) oder
  Cluster-Stichprobe (z. B. 300 Schüler aus insgesamt 20 Schulklassen)
  In solchen Fällen kann man eine sogenannte **Mehrebenenanalyse** durchführen (siehe z. B. Baltes-Götz 2020b, 2016a) oder ein **GEE-Modell** (*Generalized Estimating Equation*) anwenden (siehe z. B. Baltes-Götz 2016b).

9 Gründliche Verteilungsanalyse für metrische Variablen

Für die folgenden Schritte wird eine aktive SPSS-Sitzung mit geöffneter Fertigdatendatei **kfa.sav** vorausgesetzt. Ob Sie die SPSS-Kommandos zu den anstehenden Analysen für eine spätere Wiederverwendung konservieren wollen, bleibt Ihnen überlassen.

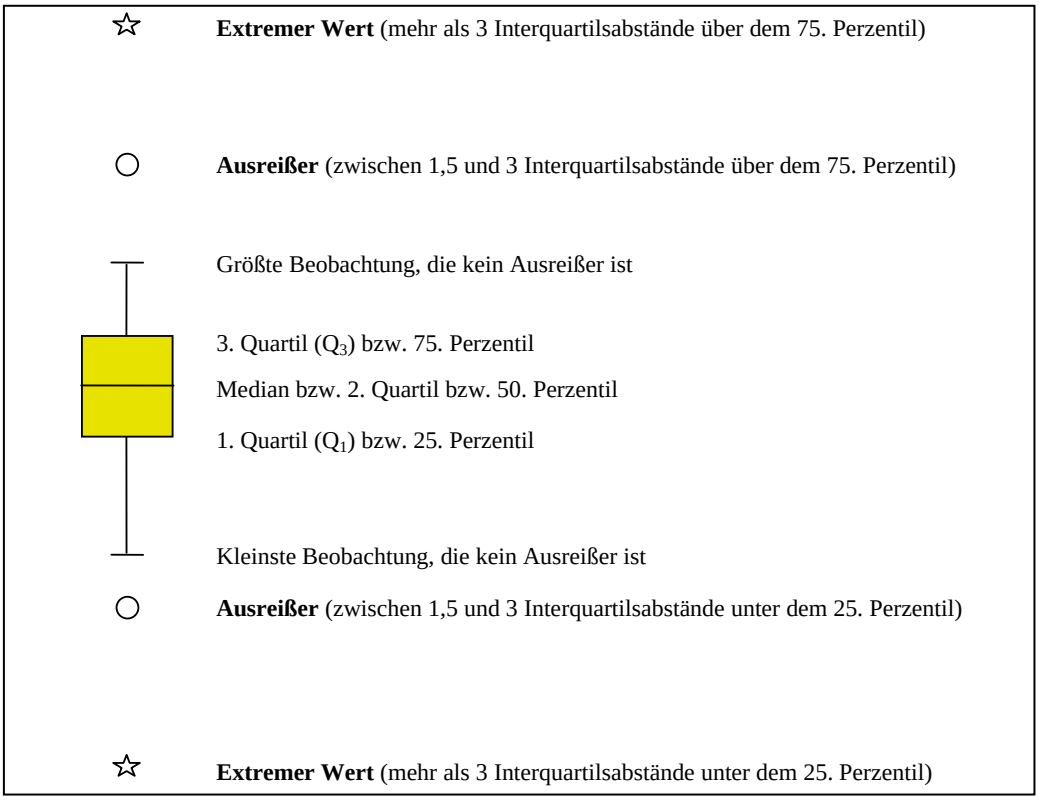
Wir untersuchen die univariaten Verteilungen der abgeleiteten Variablen LOT, AERGAM, AERGZ und BMI. U. a. kümmern wir uns um die gemäß Abschnitt 8.1 beim t-Test für verbundene Stichproben zur Prüfung unserer allgemeinpsychologischen KFA-Hypothese benötigte Normalverteilungsannahme für AERGZ.

Analog zu den Verteilungsanalysen in Kapitel 5, die auch zur Datenprüfung dienten, wollen wir bei den Verteilungen der abgeleiteten Variablen auch auf Anomalien infolge fehlerhafter oder schlecht durchdachter Berechnungsvorschriften achten. Außerdem wollen wir noch eine weitere Gefahrenquelle für unser Forschungsprojekt ins Visier nehmen: Ausreißer.

9.1 Diagnose von Ausreißern

Als **Ausreißer** bezeichnet man Werte, die zwar innerhalb des logisch möglichen Wertebereichs liegen, aber doch mit großer Wahrscheinlichkeit *nicht* aus der interessierenden Population stammen. Diese Werte haben insbesondere auf parametrische Auswertungsverfahren einen starken, verzerrenden Einfluss. Daher wollen wir ab jetzt auch auf Ausreißer achten.

Dazu lassen wir uns für metrische Variablen einen **Boxplot** erstellen. Dieses Instrument der explorativen Datenanalyse zeigt auf prägnante Weise wesentliche Verteilungsinformationen, und ist zur Identifikation von Ausreißern gut geeignet. Die Bestandteile eines Boxplots haben folgende Bedeutung:



Man bezeichnet Q_1 bzw. Q_3 auch als den *unteren* bzw. *oberen Angelpunkt* (engl.: *hinge*) der Verteilung und spricht gelegentlich von den *Tukey-Angelpunkten*, um die Verdienste von John Tu-

key bei der Entwicklung der explorativen Datenanalyse zu würdigen. Die Höhe der Box (Bereich von Q_1 bis Q_3) ist gerade der *Interquartilsabstand* (engl.: *interquartile range*, abgekürzt IQR), den wir in Abschnitt 5.5.2 als robustes Dispersionsmaß kennengelernt haben.

Bei der boxplot-basierten Ausreißerdiagnose differenziert man zwischen zwei Schweregraden: Ist ein Wert von der Box mehr als 1,5 und maximal 3 Interquartilsabstände entfernt, wird er als **Ausreißer** bezeichnet. Ist ein Wert von der Box mehr als 3 Interquartilsabstände entfernt, wird er als **extrem** bezeichnet.

Als Ursachen für Ausreißer und extreme Werte kommen in Frage:

- Erhebungs- bzw. Erfassungsfehler
  Messwerte können falsch ermittelt oder fehlerhaft in die Datenverarbeitung übernommen worden sein.
- Besondere Umstände beim Merkmalsträger
  Bei einer Agrarstudie zum Ertrag verschiedene Getreidesorten kann z. B. der Boden in einer bestimmten Versuchsparzelle durch einen Ölunfall verseucht worden sein.

Eindeutig irreguläre Daten müssen natürlich entfernt werden. Sie können z. B. mit dem Dateneditor in der Rohdatendatei:

- einen Wert löschen, d. h. durch SYSMIS ersetzen
- einen Wert als MD-Indikator deklarieren
- einen kompletten Fall löschen

Natürlich dürfen Sie keine Daten eliminieren, weil sie Ihren Hypothesen widersprechen.

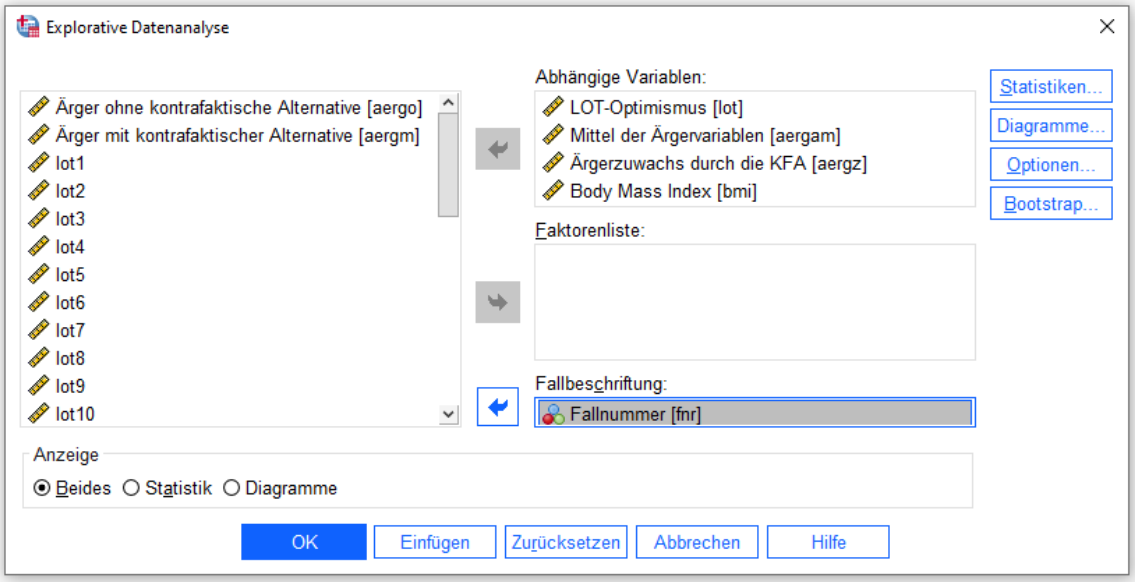
9.2 Die SPSS-Prozedur zur explorativen Datenanalyse

Für die Verteilungsanalyse bei metrischen Variablen (inklusive Ausreißerdiagnose und Normalverteilungsbeurteilung) eignet sich die SPSS-Prozedur zur explorativen Datenanalyse (basierend auf dem Kommando EXAMINE) besser als die in Kapitel 5 der Einfachheit halber bevorzugte Häufigkeitsanalyse. Natürlich sollten Sie in Zukunft auch die Verteilungen von metrischen Rohvariablen mit der leistungsfähigeren explorativen Datenanalyse untersuchen. Die Häufigkeitsanalyse (basierend auf dem Kommando FREQUENCIES) hat aber auch Alleinstellungsmerkmale zu bieten (z.B. die Option zur aufwändigen Berechnung des Medians für gruppierte Daten, Histogramm mit eingezeichneter Normalverteilungsdichte).

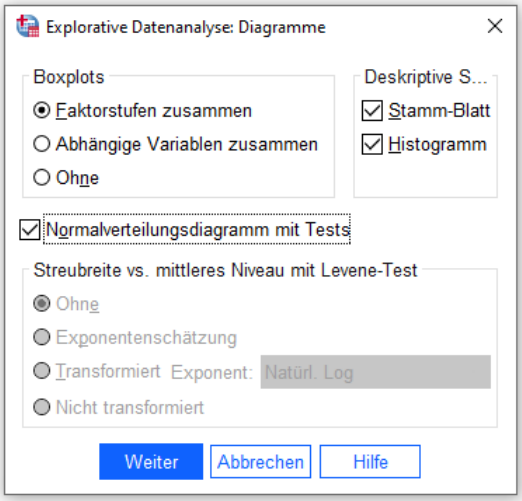
Starten Sie den Dialog der explorativen Datenanalyse mit:

Analysieren > Deskriptive Statistiken > Explorative Datenanalyse

Transportieren Sie die zu untersuchenden Variablen in die Liste der **abhängigen Variablen**, und wählen Sie die Variable FNR zur Fallbeschriftung aus, damit mögliche Ausreißer durch ihre Fallnummer identifiziert werden können:



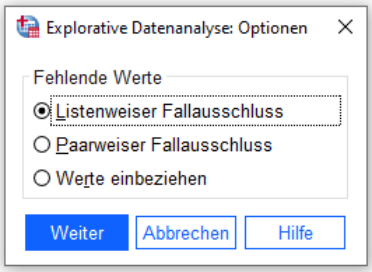
Fordern Sie in der **Diagramme**-Subdialogbox zusätzlich **Histogramme** sowie **Normalverteilungdiagramme mit Tests** an:



Das Kontrollkästchen zum Anfordern von Normalverteilungsanpassungstests (Kolmogorov-Smirnov mit Lilliefors-Korrektur und Shapiro-Wilk) hat SPSS wirklich sehr gut in der **Diagramme**-Subdialogbox der explorativen Datenanalyse versteckt.

Der Klarheit halber soll nochmals betont werden, dass wir nur für die Variable AERGZ eine Normalverteilungsvoraussetzung zu prüfen haben (vgl. Abschnitt 8.1.2). Allerdings sind die teilweise irrelevanten Ausgaben für LOT, AERGAM und BMI kein Grund dafür, zwei verschiedene Analysen anzufordern.

Wie ein Blick in den **Optionen**-Dialog zeigt, arbeitet die explorative Datenanalyse per Voreinstellung mit dem **listenweisen Fallausschluss**, sodass nur Fälle mit gültigen Werten für *alle* abhängigen Variablen in der Analyse verbleiben:

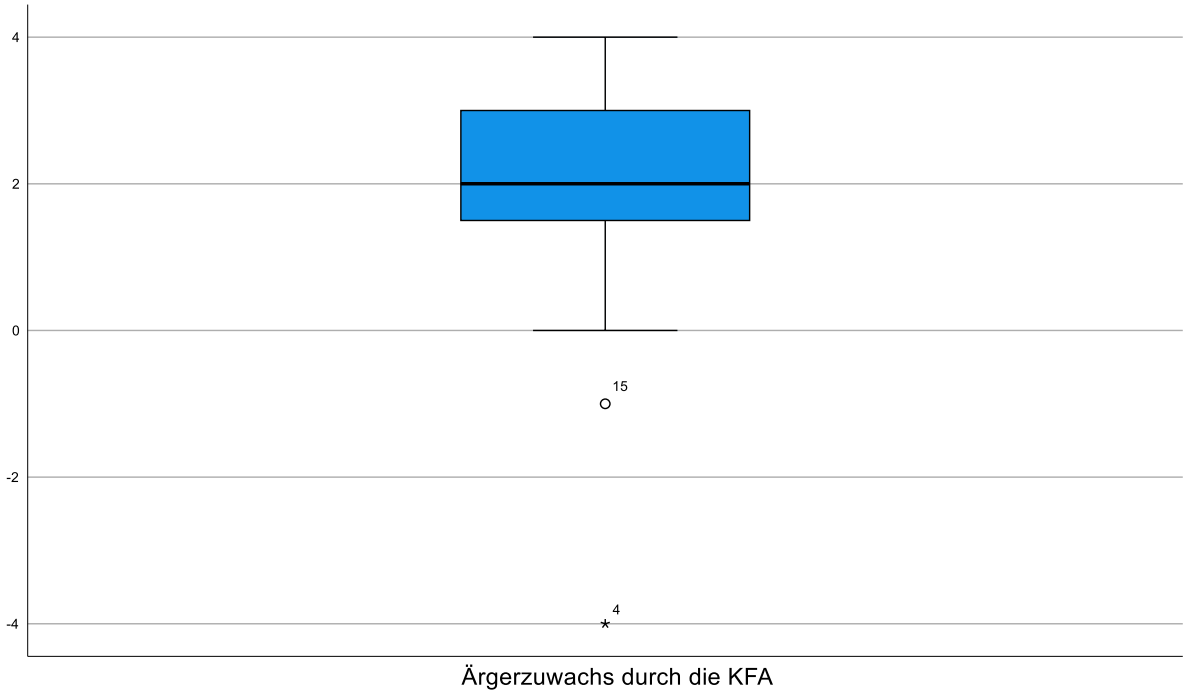


Mit dem alternativen **paarweisen Ausschluss** können Sie z. B. verhindern, dass bei der AERGZ-Verteilungsanalyse Fälle ausgeschlossen werden, weil deren BMI-Wert fehlt. In der Manuskriptstichprobe sind allerdings bei den vier aktuell betrachteten Variablen alle Werte vorhanden.

Wir erhalten im Ausgabefenster u. a. für jede abhängige Variable einen Boxplot.

9.3 Ausreißer- und Normalverteilungsbeurteilung für AERGZ

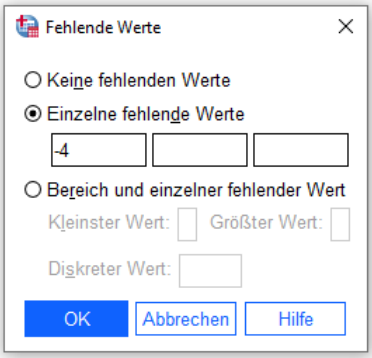
Bei der Ausreißeranalyse gibt es nur einen Problemfall und zwar ausgerechnet bei der Variablen AERGZ, über die unsere zentrale KFA-Hypothese geprüft werden soll. Hier tanzt Fall Nr. 4 aus der Reihe:



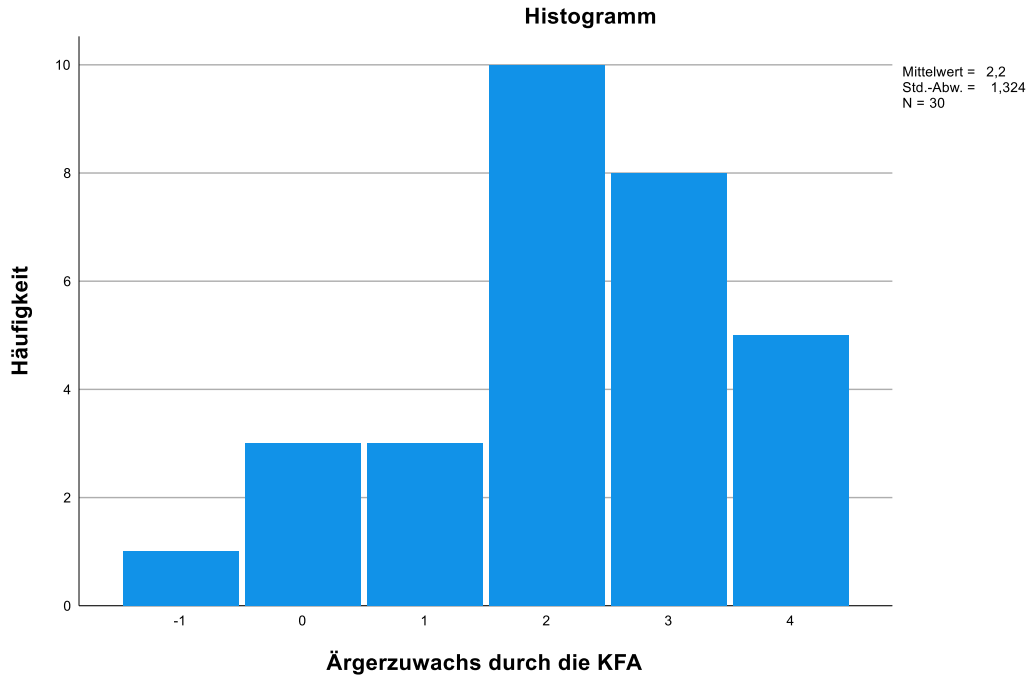
Diese Person hatte *ohne* KFA eine Ärgertemperatur von 60° gemeldet, die sich dann durch die KFA-Komponente auf 20° abkühlte. Zwar darf dieses Muster nicht als verdächtig gelten, weil es unserer Hypothese widerspricht, doch der Boxplot gibt eine klare Empfehlung, den Fall bei dieser Analyse auszuschließen. Allerdings scheut sich ein redlicher Forscher, Daten zu neutralisieren, die der eigenen Hypothese widersprechen.

Vor einer endgültigen Entscheidung wollen wir die Verteilung von AERGZ noch weiter analysieren, da beim geplanten t-Test zur allgemeinpsychologischen KFA-Hypothese vorausgesetzt werden muss, dass AERGZ (in der Population) normalverteilt ist. Damit der extreme AERGZ-Wert von Fall Nr. 4 die weitere Verteilungsanalyse nicht beeinflusst, soll er vorübergehend

neutralisiert werden. Weil wir noch keine Methode kennen, komplette Fälle von einer Analyse fern zu halten (siehe Kapitel 12), deklarieren wir den betroffenen Wert (= -4) kurzerhand als MD-Indikator. Markieren Sie in der Variablenansicht des Datenfensters die **Fehlend** - Zelle zur Variablen AERGZ, klicken Sie auf den Erweiterungsschalter [...], und tragen Sie den Wert –4 als einzelnen MD-Indikator ein:



Das folgende Histogramm zeigt, dass die AERGZ-Verteilung auch *nach* Elimination von Fall Nr. 4 noch deutlich von der Normalität abweicht:



Tatsächlich lehnen auch *nach* der Elimination des Ausreißers die beiden von SPSS angebotenen Normalverteilungstests (Kolmogorov-Smirnov mit Lilliefors-Korrektur und Shapiro-Wilk) die im t-Test benötigte Normalverteilungsannahme ab:

Tests auf Normalverteilung

	Kolmogorov-Smirnov [a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
Ärgerzuwachs durch die KFA	,207	30	,002	,913	30	,018

a. Signifikanzkorrektur nach Lilliefors

Auch diese Testentscheidung folgt der in Abschnitt 8.1 beschriebenen Logik, wobei folgende Hypothesen zur Konkurrenz stehen:

H_0: AERGZ ist normalverteilt versus H_1: AERGZ ist *nicht* normalverteilt

Die von SPSS berechnete Überschreitungswahrscheinlichkeit (**Signifikanz**) ist bei beiden Teststatistiken kleiner als 0,05, sodass beide Tests übereinstimmend die Nullhypothese verwerfen. Dies ist vor allem deshalb ein ernst zu nehmender Befund, weil unsere Stichprobe relativ klein und damit die Power der Tests eher gering ist.

Von den beiden Tests zur Normalitäts-Nullhypothese hat das Verfahren von Shapiro-Wilk den besseren Ruf. Nach Brosius (2013, S. 405) ist es vor allem bei kleineren Stichproben ($N < 50$) zu bevorzugen.

Bei einer *großen* Stichprobe besitzen die Normalitätstests eine hohe Power und decken auch kleine (z. B. für die Validität eines geplanten t-Tests irrelevante) Abweichungen von der Nullhypothese auf. Folglich ist dann ein signifikantes Testergebnis „nicht tragisch". Wenn bei einer *kleinen* Stichprobe ein Normalitätstest „anschlägt", ist jedoch von einer relevanten Verletzung der Normalitätsannahme auszugehen.

9.4 Nichtparametrische Testalternativen

Aufgrund der problematischen Verteilungsverhältnisse bei AERGZ entscheiden wir uns, zur Prüfung der KFA-Hypothese statt des geplanten parametrischen t-Tests für verbundene Stichproben einen nichtparametrischen Lagevergleich mit dem **Vorzeichentest** durchzuführen (siehe z. B. Hartung 1989, S. 242f; Siegel 1976, S. 65ff).

Gegen den häufig als Alternative zum t-Test für abhängige Stichproben verwendeten **Wilcoxon-Test** (auch als *Wilcoxon-Vorzeichen-Rangsummentest* bezeichnet) spricht seine Voraussetzung einer *symmetrischen* Verteilung der Differenzen. Wir beobachten für AERGZ (inkl. Fall Nr. 4) eine negativ-schiefe Verteilung:

Deskriptive Statistik

			Statistik	Standardfehler
Ärgerzuwachs durch die KFA	Mittelwert		2,00	,308
	95% Konfidenzintervall des Mittelwerts	Untergrenze	1,37	
		Obergrenze	2,63	
	5% getrimmtes Mittel		2,16	
	Median		2,00	
	Varianz		2,933	
	Standardabweichung		1,713	
	Minimum		-4	
	Maximum		4	
	Spannweite		8	
	Interquartilbereich		2	
	Schiefe		-1,575	,421
	Kurtosis		3,850	,821

Der in Abschnitt 5.5.3.1 beschriebene zweiseitige Test zur Nullhypothese einer symmetrischen Verteilung wird deutlich signifikant:

$$\frac{|\text{Schiefe}|}{\text{SF(Schiefe)}} = \frac{1{,}575}{0{,}421} = 3{,}741 > 1{,}96$$

Bei erfüllter Symmetriebedingung wäre der Wilcoxon-Test aufgrund seiner besseren Teststärke gegenüber dem Vorzeichentest zu bevorzugen.[1] Er entscheidet sich in seiner gerichteten Variante zwischen den folgenden Hypothesen:

$$H_0: \text{Median}(AERGZ) \leq 0 \quad \text{vs.} \quad H_1: \text{Median}(AERGZ) > 0$$

Mittlerweile wissen wir allerdings, dass in der ARGZ-Verteilung die positiven Werte so stark dominieren, dass die Power des Vorzeichentests mehr als ausreichend sein dürfte. Wir bevorzugen daher den voraussetzungsarmen Vorzeichentest und vermeiden jeden Zweifel an der Korrektheit der Entscheidung.

Die gerichtete Variante des Vorzeichentests entscheidet sich zwischen den folgenden Hypothesen:

H_0: $P(AERGZ > 0) \leq P(AERGZ < 0)$
Die Wahrscheinlichkeit für einen Ärgerzuwachs ist nicht größer als die Wahrscheinlichkeit für eine Ärgerreduktion.

versus

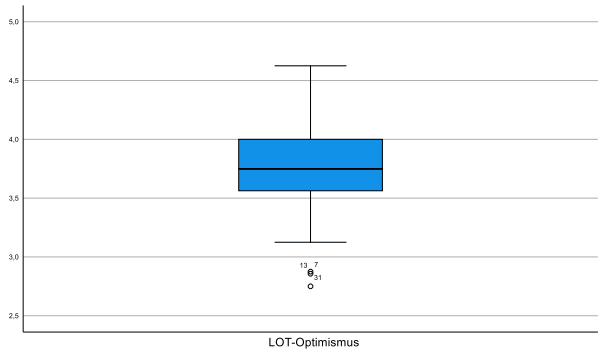
H_1: $P(AERGZ > 0) > P(AERGZ < 0)$
Die Wahrscheinlichkeit für einen Ärgerzuwachs ist größer als die Wahrscheinlichkeit für eine Ärgerreduktion.

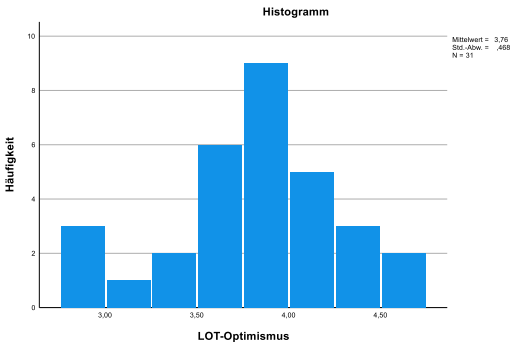
Statt der in Abschnitt 8.1 vorgestellten Teststatistik T_Z verwendet der Vorzeichentest eine Prüfgröße, die im Wesentlichen auf der Anzahl der positiven AERGZ-Ausprägungen in der Stichprobe basiert. Sie wird üblicherweise mit Z bezeichnet, weil sie unter der H_0 (genauer: bei $P(AERGZ > 0) = P(AERGZ < 0)$ approximativ z-verteilt (d. h. standardnormalverteilt) ist.

Man geht davon aus, dass die Verteilungsapproximation ab $n \geq 20$ hinreichend genau ist, sodass wir den Test bei unserer Stichprobe ($n = 31$) in der approximativen Form anwenden dürften. Bei kleineren Stichproben muss die *exakte* Variante des Tests eingesetzt werden, die von SPSS ebenfalls unterstützt wird. Gegen den exakten Test spricht nur der höhere Rechenaufwand, der aber bei $n = 31$ auf einem modernen Rechner vernachlässigbar ist.

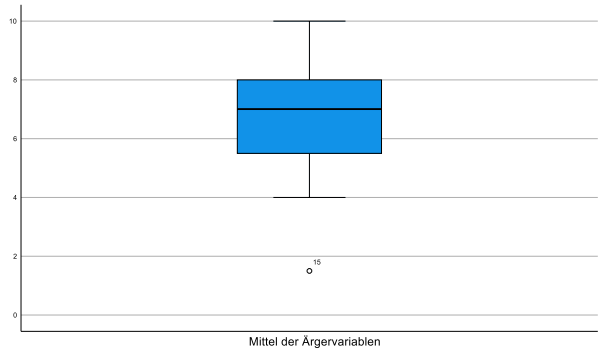
9.5 Ausreißerbeurteilung für LOT, AERGAM und BMI

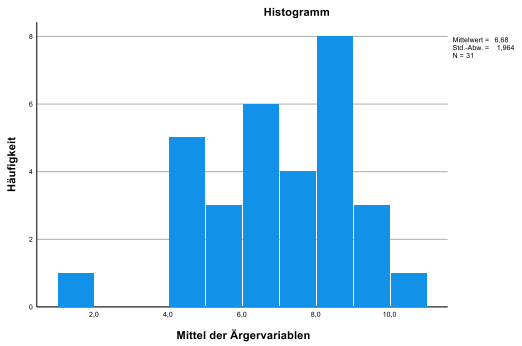
Bei den Variablen LOT, AERGAM und BMI finden sich keine Hinweise auf Fehler in den Berechnungsanweisungen oder auf extreme Ausreißer:

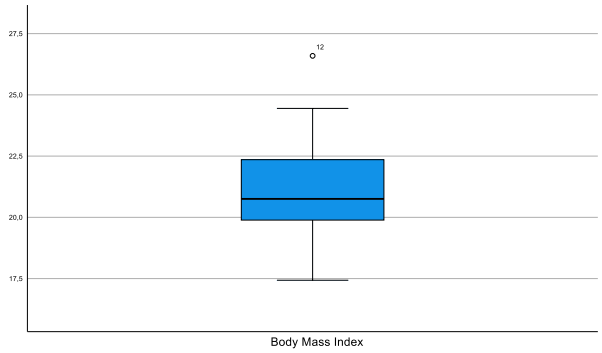


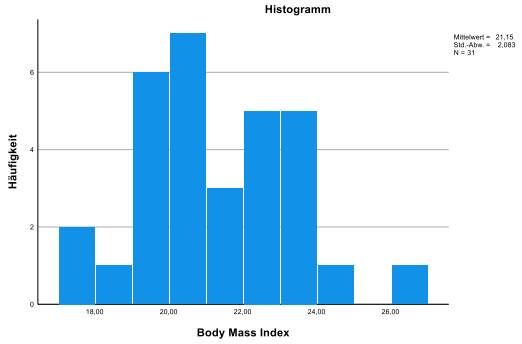


[1] Eine gute Beschreibung des Wilcoxon-Vorzeichen-Rangsummentests ist z. B. hier zu finden: http://www.biostathandbook.com/wilcoxonsignedrank.html









Die in den Boxplots auftauchenden Ausreißer sind nicht extrem (Abstand vom 25. bzw. 75. Perzentil kleiner als drei Interquartilsabstände) und sollten aufgrund einer relativ kleinen Stichprobe, die nur wenig Information über die Populationsverteilungen enthält, *nicht* ausgeschlossen werden.

9.6 Vertrauensintervalle für Lageparameter

Bei der bisherigen Diskussion der AERGZ-Verteilung haben wir uns auf Gefahrenquellen für die Interpretierbarkeit des geplanten KFA-Hypothesentests konzentriert (Ausreißer, Abweichung von der Normalverteilung). Wir sollten aber auch jede Möglichkeit nutzen, anhand von Verteilungsdiagrammen und -statistiken einen Eindruck von der empirischen Bewährung der KFA-Hypothese zu gewinnen. Eine genaue Kenntnis des deskriptiven Ergebnisbilds kann verhindern, dass wir von einem durch technische Defekte verfälschten Testergebnis in die Irre geführt werden. Der Boxplot und das Histogramm zu AERGZ (siehe Seite 193 bzw. 194) sprechen klar für einen starken KFA-Effekt in der erwarteten Richtung. Weil fast alle Fälle in der Bedingung *mit* KFA mehr Ärger erleben, bestehen keine ernsthaften Zweifel an der statistischen Bedeutsamkeit des Effekts.

Neben der Signifikanz und der Richtung des KFA-Effekts interessiert auch seine Stärke, die durch ein Maß der zentralen Tendenz für die Variable AERGZ charakterisiert werden kann. Bezogen auf das „Ärgerthermometer“ aus dem Fragebogen unserer Studie ist von Interesse, um wieviel Grad das Ärgerniveau durch den KFA-Effekt gesteigert wird. Im Unterschied zum Effektstärkenbegriff aus der Stichprobenumfangsplanung (vgl. Abschnitt 2.3.2.4) betrachten wir nun den *unstandardisierten* Ärgerzuwachs.

Wenn die Messung eines Merkmals in einer interpretierbaren Einheit erfolgt (z. B. cm, mmHg, Euro), dann ist die unstandardisierte Effektstärkebeurteilung oft gegenüber der standardisierten Beurteilung zu bevorzugen, weil die unstandardisierte Beurteilung nicht von den Streuungsverhältnissen in einer bestimmten Stichprobe bzw. Population abhängt.

Zu einem Lageparameter ist in der Regel neben der Punktschätzung auch ein Vertrauensintervall gefragt, das den wahren Wert mit einer gewünschten Wahrscheinlichkeit enthält (vgl. Abschnitt 5.8).

9.6.1 Normalverteilungs-Vertrauensintervall für das arithmetische Mittel

Weil die Variable AERGZ auch nach der Elimination von Fall Nr. 4 tendenziell negativ-schief verteilt ist (siehe Abschnitt 9.3), ist das arithmetische Mittel als Maß für die zentrale Tendenz nicht perfekt geeignet (vgl. Abschnitt 5.5.1). In vielen Fällen ist jedoch das Vertrauensintervall für einen *Mittelwert* von Interesse, sodass wir uns jetzt damit beschäftigen wollen, wie ein solches Vertrauensintervall mit SPSS zu ermitteln ist.

Während die SPSS-Prozedur zur Häufigkeitsanalyse, die wir in Kapitel 5 verwendet haben, keine Vertrauensintervalle liefert, erhalten wir von der in Abschnitt 9.2 vorgestellten explorativen Verteilungsanalyse unaufgefordert das Vertrauensintervall für den Mittelwert, z. B. bei AERGZ (aus der Stichprobe *ohne* den als Ausreißer identifizierten Wert -4):

Deskriptive Statistik

			Statistik	Standardfehler
Ärgerzuwachs durch die KFA	Mittelwert		2,20	,242
	95% Konfidenzintervall des Mittelwerts	Untergrenze	1,71	
		Obergrenze	2,69	
	5% getrimmtes Mittel		2,26	
	Median		2,00	
	Varianz		1,752	
	Standardabweichung		1,324	
	Minimum		-1	
	Maximum		4	
	Spannweite		5	
	Interquartilbereich		1	
	Schiefe		-,585	,427
	Kurtosis		-,059	,833

Bei einer akzeptierten Irrtumswahrscheinlichkeit von $\alpha = 0{,}05$ (Konfidenzniveau 95%) berechnet sich das Konfidenzintervall folgendermaßen aus dem geschätzten arithmetischen Mittel $\bar{Z}$, der geschätzten Standardabweichung S_Z und dem $(1-\alpha/2)$ - Quantil der t-Verteilung mit n - 1 Freiheitsgraden, das wie in Abschnitt 8.1.5 als $T_{\text{krit},2}$ bezeichnet werden soll (Bortz & Schuster 2010, S. 119):

$$\left[\bar{Z} - T_{krit,2} \cdot \frac{S_Z}{\sqrt{n}}; \bar{Z} + T_{krit,2} \cdot \frac{S_Z}{\sqrt{n}}\right]$$

Dabei wird die Normalverteilung der betrachteten Variablen vorausgesetzt.

Mit den nach Elimination von Fall 4 noch verbliebenen 30 Fällen resultiert exakt das von der explorativen Datenanalyse berechnete Vertrauensintervall:

$$\left[2{,}2 - 2{,}045 \cdot \frac{1{,}324}{\sqrt{30}}; 2{,}2 + 2{,}045 \cdot \frac{1{,}324}{\sqrt{30}}\right] = [1{,}71; 2{,}69]$$

Es ist allerdings nicht ganz korrekt, weil seine Berechnung auf der Normalverteilungsannahme für AERGZ beruht, von der wir uns bereits distanziert haben. Um den Makel abzuschütteln, kann das in Abschnitt 9.6.2 beschriebene Bootstrapping verwendet werden.

Es besteht ein wichtiger Zusammenhang zwischen dem Normalverteilungs-basierten Vertrauensintervall für den Mittelwert und dem *zweiseitigen* Einstichproben - t-Test. Für den t-Test haben wir in Abschnitt 8.1.8 die folgende Entscheidungsregel notiert:

$$\frac{|\bar{Z}|}{S_Z}\sqrt{n} > T_{\text{krit},2} \Rightarrow \text{H}_0 \text{ verwerfen}$$

Aus der Äquivalenz

$$\frac{|\bar{Z}|}{S_Z}\sqrt{n} > T_{\text{krit},2} \Leftrightarrow |\bar{Z}| - T_{krit,2} \cdot \frac{S_Z}{\sqrt{n}} > 0$$

folgt unmittelbar, dass der zweiseitige t-Test seine Nullhypothese genau dann verwirft, wenn das Vertrauensintervall zum Mittelwert den Wert 0 *nicht* enthält.

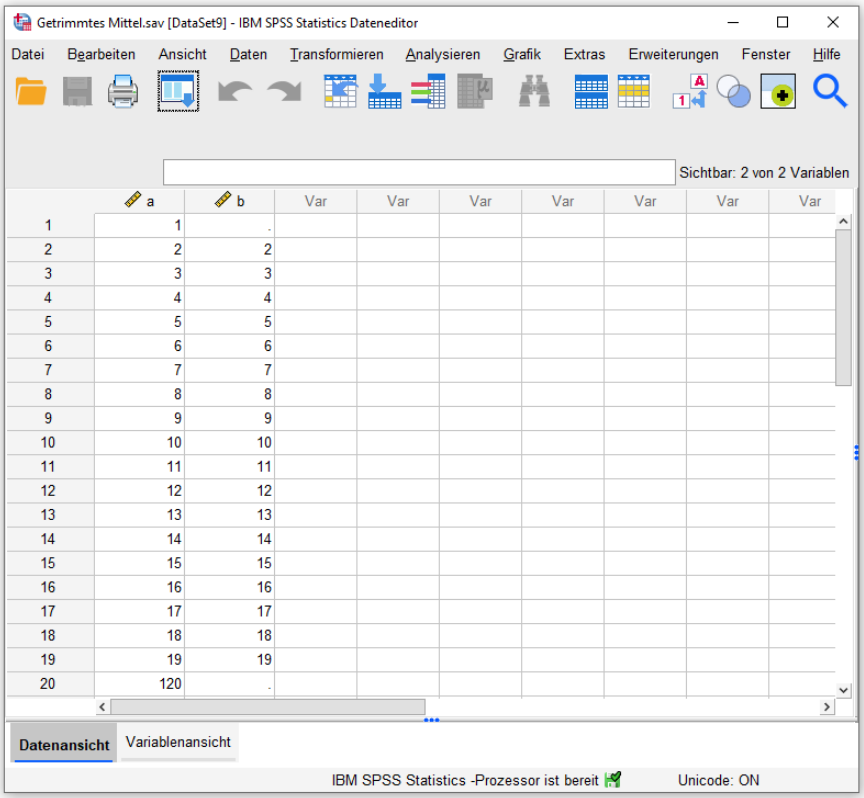
Im Beispiel liegt die 0 weit vom zweiseitigen Vertrauensintervall entfernt, doch hält uns die verletzte Normalverteilungsannahme noch davon ab, die KFA-Nullhypothese zu verwerfen.

Mit dem *einseitigen* Vertrauensintervall zum Mittelwert existiert übrigens unter den Techniken der Intervallschätzung auch eine Entsprechung zum *einseitigen* Einstichproben - t-Test, der in sehr vielen Fällen gegenüber der zweiseitigen (ungerichteten) Variante zu bevorzugen ist. Allerding wird das einseitige Vertrauensintervall zum Mittelwert nur selten benutzt, und SPSS kann es nicht berechnen.

Mit der Berechnung eines Vertrauensintervalls zu einer Parameterschätzung sind also bestimmte Hypothesentests zum Vergleich des Parameters mit einem Referenzwert (z. B. 0) bereits entschieden, wenn die Berechnung des Vertrauensintervalls unter gültigen Voraussetzungen erfolgte.

9.6.2 Bootstrapping-Vertrauensintervall für das getrimmte Mittel

Die explorative Datenanalyse liefert mit dem **5% getrimmten Mittel** einen gegenüber Ausreißern robusten Schätzwert für den Populationserwartungswert. Bei seiner Berechnung werden vor Anwendung der üblichen Formel für das arithmetische Mittel die 5% der Fälle mit den kleinsten sowie die 5% der Fälle mit den größten Werten ausgeschlossen. Im folgenden Beispiel mit 20 Fällen



ist das *getrimmte* Mittel der Variablen A identisch mit dem *arithmetischen* Mittel der Variablen B, weil bei der Berechnung des getrimmten Mittels der Fall mit dem kleinsten Wert und der Fall mit dem größten Wert ausgeschlossen werden:

Deskriptive Statistik

			Statistik	Standardfehler
a	Mittelwert		10,50	1,258
	95% Konfidenzintervall des Mittelwerts	Untergrenze	7,85	
		Obergrenze	13,15	
	5% getrimmtes Mittel		10,50	
	Median		10,50	
	Varianz		28,500	
	Standardabweichung		5,339	
	Minimum		2	
	Maximum		19	
	Spannweite		17	
	Interquartilbereich		10	
	Schiefe		,000	,536
	Kurtosis		-1,200	1,038
b	Mittelwert		10,50	1,258
	95% Konfidenzintervall des Mittelwerts	Untergrenze	7,85	
		Obergrenze	13,15	
	5% getrimmtes Mittel		10,50	
	Median		10,50	
	Varianz		28,500	
	Standardabweichung		5,339	
	Minimum		2	
	Maximum		19	
	Spannweite		17	
	Interquartilbereich		10	
	Schiefe		,000	,536
	Kurtosis		-1,200	1,038

Das getrimmte Mittel ist gegenüber Ausreißern ähnlich robust wie der Median, verwertet aber mehr Informationen. Es eignet sich insbesondere dann, wenn eine Verteilung relativ symmetrisch, aber durch Ausreißer belastet ist. Bei der Variablen AERGZ wird bei der Berechnung des getrimmten Mittels insbesondere der im Boxplot als Ausreißer identifizierte Fall mit dem Wert -4 neutralisiert:

Deskriptive Statistik

			Statistik	Standardfehler
Ärgerzuwachs durch die KFA	Mittelwert		2,00	,308
	95% Konfidenzintervall des Mittelwerts	Untergrenze	1,37	
		Obergrenze	2,63	
	5% getrimmtes Mittel		2,16	
	Median		2,00	
	Varianz		2,933	
	Standardabweichung		1,713	
	Minimum		-4	
	Maximum		4	
	Spannweite		8	
	Interquartilbereich		2	
	Schiefe		-1,575	,421
	Kurtosis		3,850	,821

Auch für das getrimmte Mittel sollte ein Vertrauensintervall angegeben werden, wobei aber die komplexe Definition eine mathematische Herleitung verhindert. In dieser Situation hilft uns eine flexible statistische Technik namens **Bootstrapping** weiter.

In der amerikanischen Variante einer bekannten Münchhausen-Geschichte schafft es der Held, sich an den eigenen Stiefelriemen aus dem Sumpf zu ziehen, und dieses Verfahren liefert den Namen für eine durchaus solide statistische Schätz- und Testmethodologie, die erstmals von Efron (1982) ausformuliert worden ist.

Man behandelt die Stichprobe als *Population*, ermittelt durch Ziehen *mit Zurücklegen* zahlreiche Sekundärstichproben (z. B. 1000) mit derselben Größe wie die Original- bzw. Primärstichprobe, wobei in einer Sekundärstichprobe in der Regel etliche Fälle *mehrfach* vertreten sind. Aus jeder Sekundärstichprobe wird mit den üblichen Methoden ein Schätzer für den interessierenden Parameter gewonnen, sodass man eine empirische Stichprobenverteilung erhält. Diese ersetzt die theoretische Stichprobenverteilung, die z. B. auf der Normalitätsannahme basiert. Aus der empirischen Stichprobenverteilung lassen sich Vertrauensintervalle und Testentscheidungen konstruieren, die nicht von der Normalverteilungsannahme abhängen.

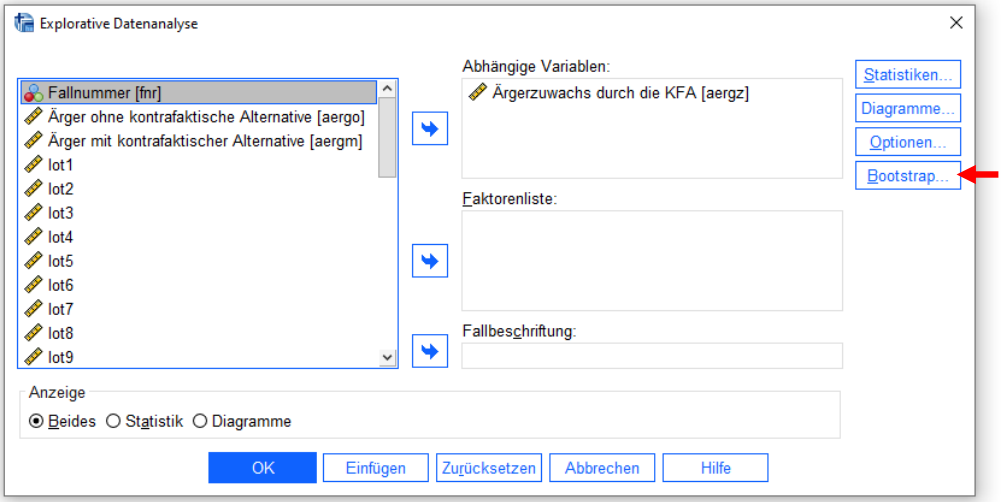
Lange war die benötigte Rechenleistung ein Hindernis für die Weiterentwicklung und Anwendung der Bootstrap-Technologie, doch mittlerweile suchen die CPU-Hersteller nach relevanten Anwendungen für ihre Gigahertz- und Multicore-Boliden.

In den folgenden Situationen ist die sehr generell einsetzbare Ermittlung von Vertrauensintervallen zu Parameterschätzungen durch Bootstrap-Methoden von Interesse:

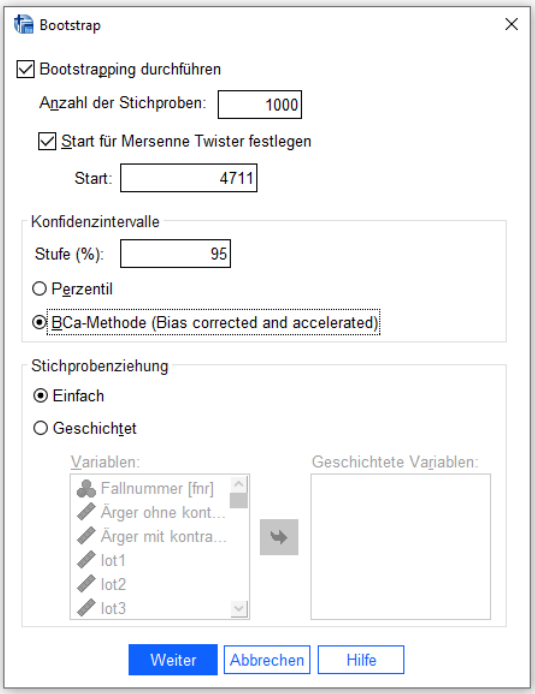
- Für manche Parameter (z. B. getrimmtes Mittel) ist die mathematische Herleitung des Vertrauensintervalls sehr aufwändig oder unmöglich. Per Bootstrap-Technik ersetzt man die theoretische Stichprobenverteilung durch die oben beschriebene empirische Stichprobenverteilung.
- Wo die Berechnung eines Vertrauensintervalls aufgrund eines mathematischen Modells möglich ist, sind in der Regel Voraussetzungen im Spiel (z.B. die Normalverteilungsannahme). Bei grob verletzten Voraussetzungen resultiert ein Vertrauensintervall, das die

erwartete Überdeckungswahrscheinlichkeit *nicht* besitzt (siehe z.B. Abschnitt 5.8.1). Mit Hilfe der Bootstrap-Technik erhält man in vielen Fällen realistischere Ergebnisse.

SPSS kann bei vielen Prozeduren das Erstellen von Bootstrap-Sekundärstichproben und die Zusammenfassung der Ergebnisse automatisieren. Wenn eine Prozedur das Bootstrapping unterstützt, ist ein entsprechender Schalter in ihrer Dialogbox vorhanden, z. B. bei der explorativen Datenanalyse:



Im **Bootstrap**-Dialog



aktiviert man das Bootstrapping und wählt eine Anzahl von Sekundärstichproben (z. B. 1000, 5000 oder 10000). Ein Startwert für den Pseudozufallszahlengenerator (**Mersenne Twister**) macht das Bootstrap-Ergebnis reproduzierbar.

Für die **Vertrauensintervalle** legt man des Konfidenzniveau fest und wählt in der Regel die **Bias corrected and accelerated** - Methode.

Für die Variable AERGZ erhalten wir (mit der kompletten Stichprobe) zum getrimmten Mittel das 95% - Konfidenzintervall [1,69; 2,50].

Deskriptive Statistik

					Bootstrap [a]			
							BCa 95% Konfidenzintervall	
			Statistik	Standardfehler	Verzerrung	Standardfehler	Unterer Wert	Oberer Wert
Ärgerzuwachs durch die KFA	Mittelwert		2,00	,308	-,02	,29	1,44	2,42
	95% Konfidenzintervall des Mittelwerts	Untergrenze	1,37					
		Obergrenze	2,63					
	5% getrimmtes Mittel		2,16		-,06	,28	1,69	2,50
	Median		2,00		,15	,36	2,00	2,00
	Varianz		2,933		-,062	1,167	1,194	5,004
	Standardabweichung		1,713		-,054	,347	1,107	2,202
	Minimum		-4					
	Maximum		4					
	Spannweite		8					
	Interquartilbereich		2		0	1	2	2
	Schiefe		-1,575	,421	,356	,619	-2,705	,623
	Kurtosis		3,850	,821	-1,591	2,325	,481	3,398

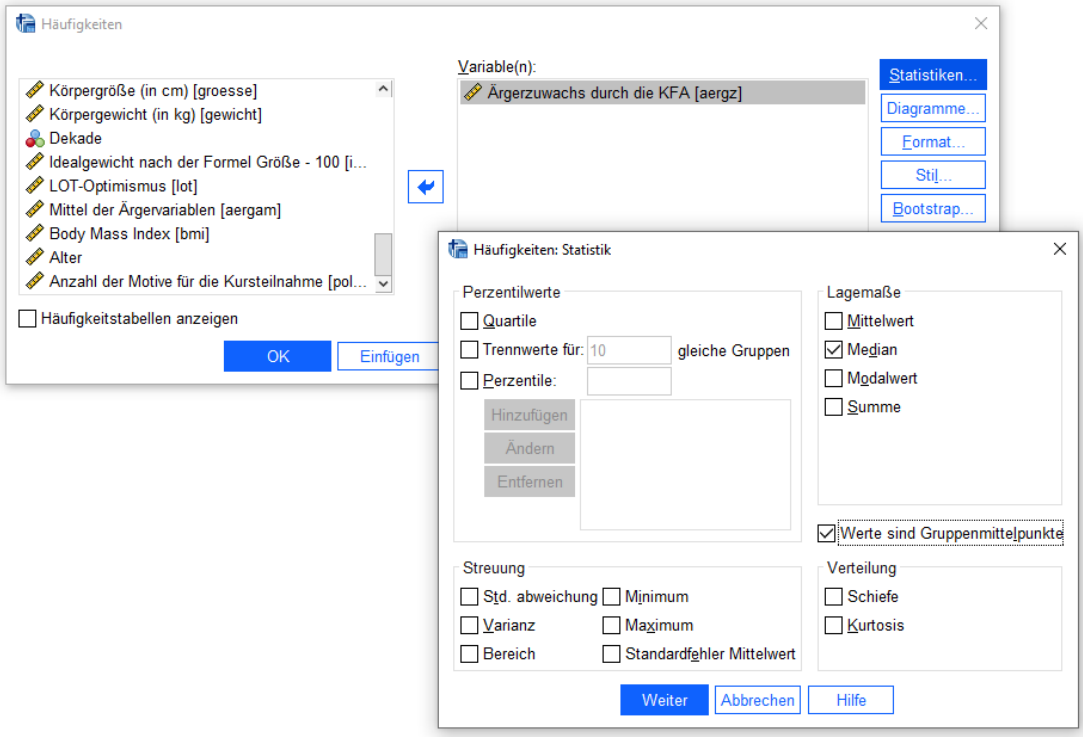
a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

Auch mit einem Bootstrap-basierten Vertrauensintervall lässt sich ein Signifikanztest zum Vergleich eines Parameters mit einem Referenzwert durchführen (vgl. Abschnitt 9.6.1). Im Beispiel liegt der Referenzwert 0 eindeutig links vom Bootstrap-Vertrauensintervall für das getrimmte Mittel. Wir schließen daraus, dass in der Population der Erwartungswert von AERGZ größer als 0 ist. Damit ist die KFA-Nullhypothese im Grunde verworfen. Der in Abschnitt 9.4 geplante Vorzeichentest ist allerdings nicht ganz überflüssig, weil sein Ergebnis weder von der Normalverteilungsannahme, noch von der Bootstrap-Technik abhängig und damit für manche Forschungsinteressenten besonders überzeugend ist.

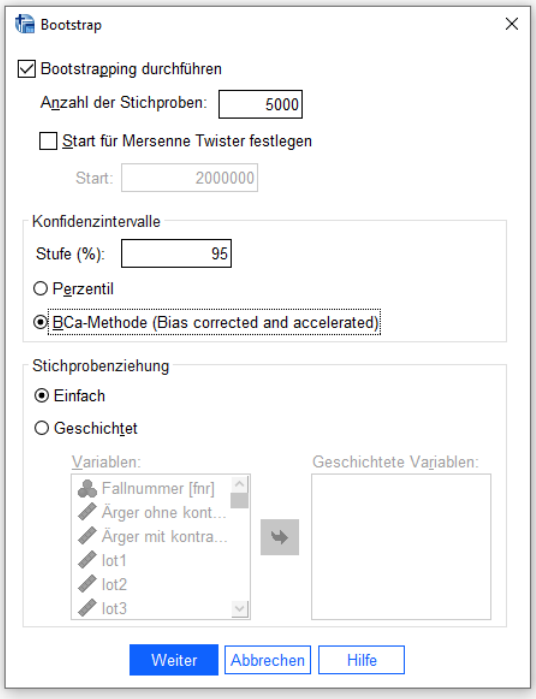
Bei einer ausgeprägt schiefen KFA-Verteilung ist zudem der Median gegenüber dem Erwartungswert als Lagemaß zu bevorzugen, und infolgedessen ist das Vertrauensintervall für den Median nützlicher als das Vertrauensintervall für das (getrimmte) Mittel.

9.6.3 Bootstrapping-Vertrauensintervall für den Median aus gruppierten Daten

Wie die zuletzt (in Abschnitt 9.6.2) präsentierte Tabelle zeigt, resultiert für den gemäß Abschnitt 5.5.1 berechneten AERGZ-Median kein sinnvolles Bootstrap-Vertrauensintervall. Es ist anzunehmen, dass die relativ grobe Messung dazu führt, dass der Median und sein Vertrauensintervall nicht sinnvoll geschätzt werden können. In Abschnitt 5.5.5 wurde eine alternative, von der Annahme gruppierter Daten ausgehende Berechnungsmethode für den Median vorgestellt, die im Rahmen der FREQUENCIES-Prozedur (anzufordern über **Analysieren > Deskriptive Statistiken > Häufigkeiten**) zur Verfügung steht. Wird sie auf AERGZ angewendet



und zusätzlich ein Bootstrap-Vertrauensintervall (z. B. unter Verwendung von 5000 Sekundärstichproben)



angefordert, dann erhält man eine plausible Schätzung (2,28) und ein sinnvolles Vertrauensintervall ([1,73; 2,75]) für den AERGZ-Median:

Statistiken

Ärgerzuwachs durch die KFA

			Bootstrap [b]			
					BCa 95% Konfidenzintervall	
		Statistik	Verzerrung	Standard Fehler	Unterer Wert	Oberer Wert
N	Gültig	31	0	0	.	.
	Fehlend	0	0	0	.	.
Median		2,28 [a]	-,01	,25	1,73	2,75

a. Aus gruppierten Daten berechnet

b. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 5000 Bootstrap-Stichproben

10 Prüfung der zentralen Projekt-Hypothesen

In diesem Kapitel machen wir uns daran, die zentralen Hypothesen der Kursstudie zu prüfen (emotionaler Effekt kontrafaktischer Alternativen, Ärgerdämpfung durch Optimismus).

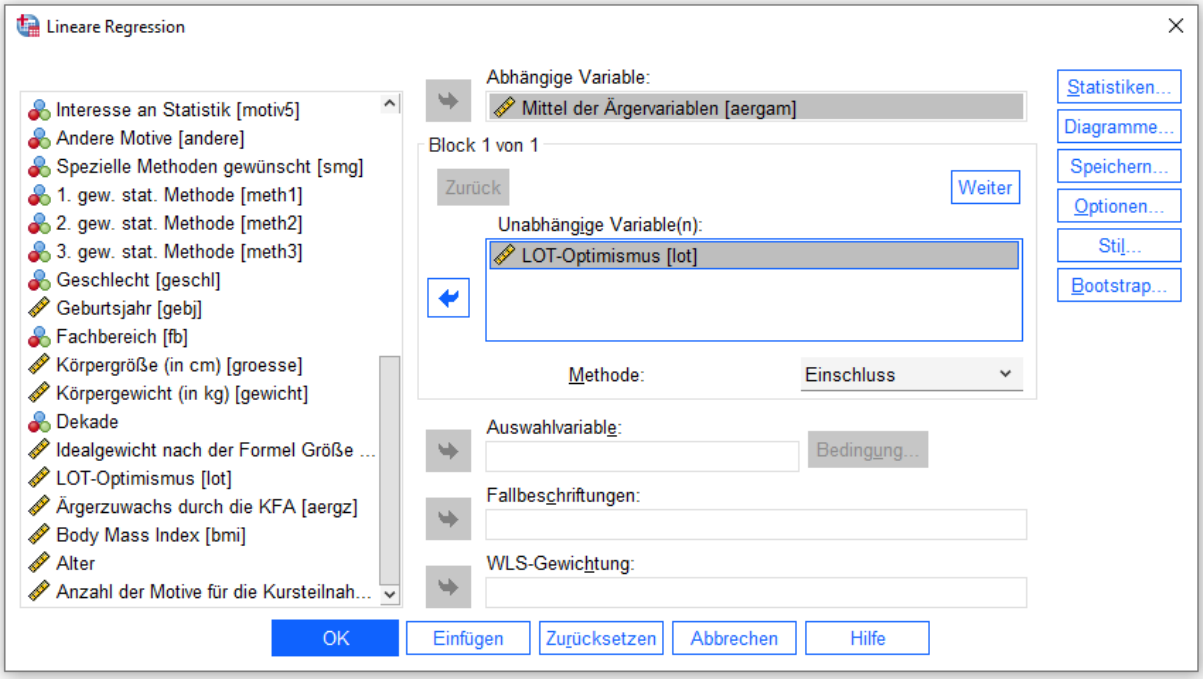
10.1 Prüfung der differentialpsychologischen Hypothese

10.1.1 Regression von AERGAM auf LOT

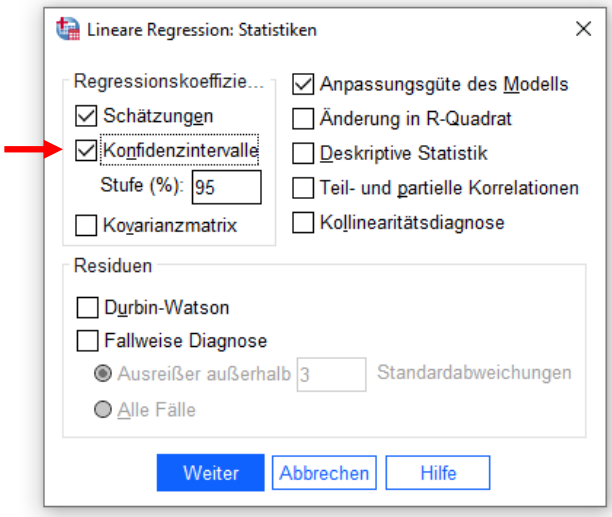
Um die lineare Regression von AERGAM auf LOT zu untersuchen, öffnen wir mit dem Menübefehl

Analysieren > Regression > Linear

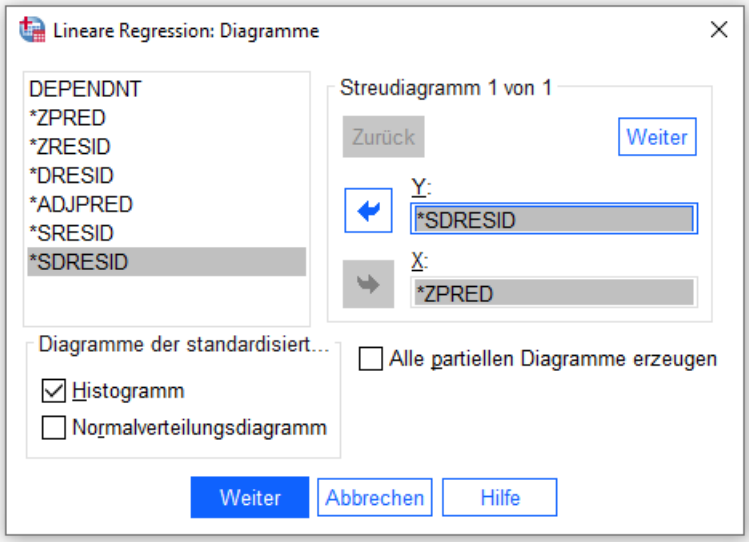
die folgende Dialogbox:



In der **Statistiken** - Subdialogbox verlangen wir über die Voreinstellung hinausgehend die Berechnung von **Konfidenzintervallen** zu den Regressionskoeffizienten:



Zur Prüfung der in Abschnitt 8.2 beschriebenen Voraussetzungen ordern wir in der **Diagramme**-Subdialogbox



folgende Ausgaben:

- Das **Streudiagramm** der ausgelassen-studentisierten Residuen (interne Variable SDRESID) gegen die standardisierte Modellprognose (interne Variable ZPRED)

  Weil die Modellprognose eine lineare Funktion des Regressors ist ($\hat{Y} = b_0 + b_1 X$), erhalten wir letztlich einen Plot der ausgelassen-studentisierten Residuen gegen den Regressor. Damit lassen sich die Linearität und die Varianzhomogenität beurteilen.

- Das **Histogramm** der standardisierten Residuen
  Damit lässt sich die Normalität der Residuen beurteilen. Zwar wäre ein Histogramm der ausgelassen-studentisierten Residuen wünschenswert, doch sind keine relevante Unterschiede zwischen den beiden Residualdiagrammen zu erwarten.

Mit den beiden Diagrammen lassen sich drei Voraussetzungen (Linearität, Varianzhomogenität und Normalität der Residuen) optisch prüfen, die unser Modell

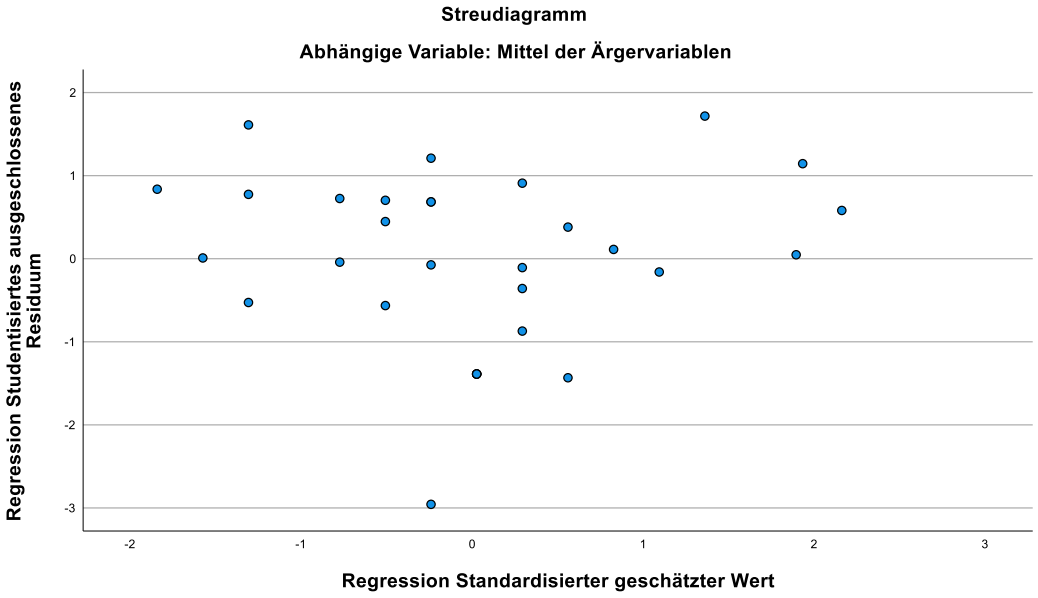
$$\text{AERGAM} = \beta_0 + \beta_1 \text{LOT} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

sehr kompakt formuliert in der Formel

$$\varepsilon \sim N(0, \sigma^2)$$

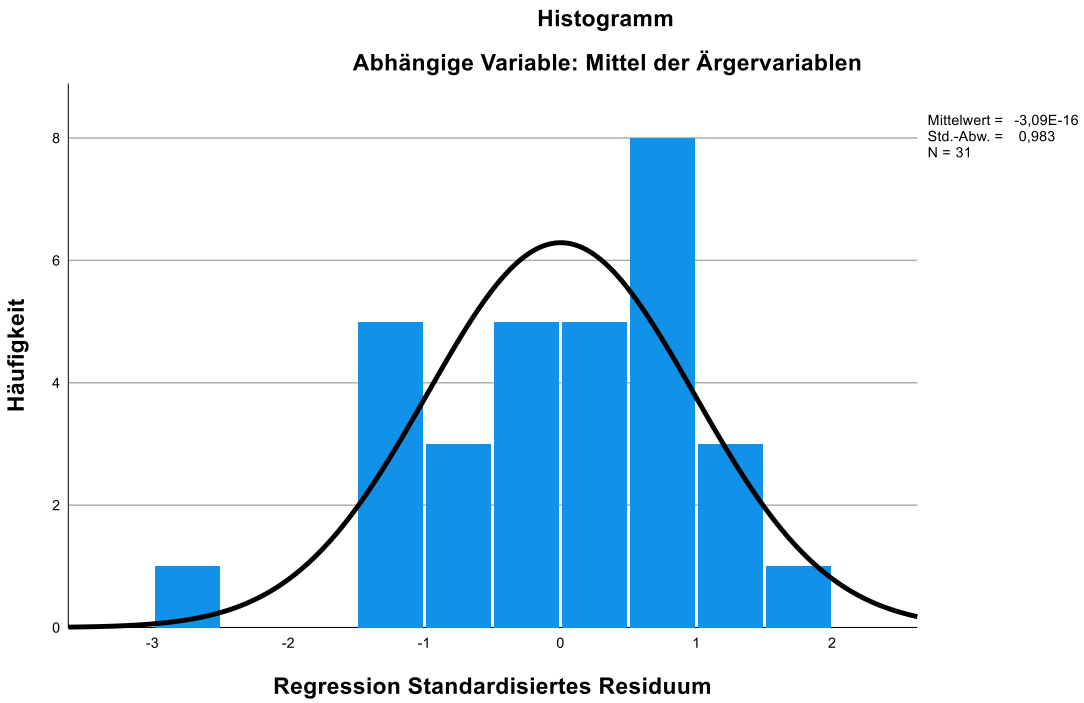
enthält.

Das Streudiagramm bietet wenig Anlass zur Sorge um die Linearität oder um die Varianzhomogenität:

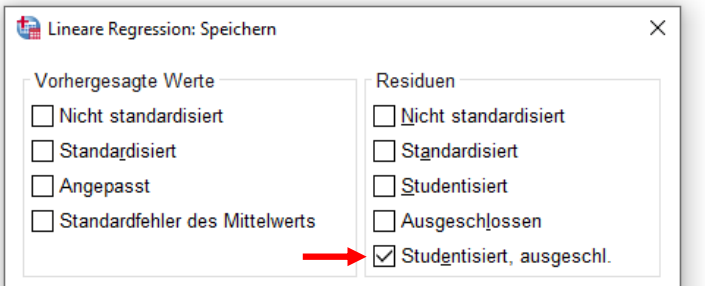


Wir sehen ein „signifikantes“ Residuum (standardisierter Wert betragsmäßig größer 2), was aber bei 31 Fällen mit der Annahme eines gültigen Modells vereinbar ist. Bei Gültigkeit aller Modellvoraussetzungen sind ca. 5 % standardisierte Residuen mit einem Betrag größer 2 zu erwarten.

Das Histogramm der standardisierten Residuen zeigt eine zufriedenstellende Normalverteilungsapproximation:



Mit den per **Speichern**-Subdialog



in eine neue Variable der Arbeitsdatei geschriebenen Residuen lässt sich auch ein formaler Normalverteilungsanpassungstest durchführen (vgl. Abschnitt 9.3). Dabei verwenden wir die zur Normalverteilungsbeurteilung besonders gut geeigneten ausgelassen-studentisierten Residuen.

Generell führen derartige Voraussetzungsprüfungen per Signifikanztest nicht unbedingt auf einfache Weise zu einer guten Entscheidung, denn:

- Bei einer kleinen Stichprobe sind Verletzungen der Normalität ernst zu nehmen, können aber mangels Power der Normalverteilungsprüfung schwer nachgewiesen werden.
- Bei einer großen Stichprobe verliert die Normalitätsannahme an Bedeutung, was sich durch den zentralen Grenzwertsatz der Statistik begründen lässt. Doch haben Tests zur Normalverteilungsprüfung in dieser Situation eine große Power, sodass auch kleine, für die geplante Inferenzstatistik irrelevante Abweichungen von der idealen Glockenform signifikant werden. Nach Bühner & Ziegler (2009, S. 674) ist bei der multiplen linearen Regression ab ca. 100 Fällen eine Verletzung der Normalverteilung der Residuen „weniger problematisch“, wenn keine Ausreißer vorhanden sind. Backhaus et al. (2008, S. 90) rechnen schon ab einem Stichprobenumfang von $N = 40$ mit einer Robustheit der Regressionsanalyse gegenüber Verletzungen der Fehlernormalverteilung.

In unserem Beispiel (mit einer kleinen Stichprobe) übersteht die Annahme (Nullhypothese) normalverteilter Residuen die Signifikanztests nach Kolmogorov-Smirnov und Shapiro-Wilk:

Tests auf Normalverteilung

	Kolmogorov-Smirnov [a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
Studentized Deleted Residual	,118	31	,200 *	,948	31	,133

*. Dies ist eine untere Grenze der echten Signifikanz.

a. Signifikanzkorrektur nach Lilliefors

Nachdem wir die Voraussetzungen als gültig akzeptiert haben, steht einer Inspektion der Regressionsergebnisse nichts mehr im Weg. Wir erhalten zwar, wie erwartet, einen negativen Regressionskoeffizienten, doch ist dieser bei weitem nicht signifikant:

Koeffizienten [a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten			95,0% Konfidenzintervalle für B	
		Regressionskoeffizient B	Standardfehler	Beta	T	Sig.	Untergrenze	Obergrenze
1	(Konstante)	7,669	2,947		2,602	,014	1,641	13,697
	LOT-Optimismus	-,264	,778	-,063	-,339	,737	-1,854	1,327

a. Abhängige Variable: Mittel der Ärgervariablen

SPSS ermittelt eine zweiseitige Überschreitungswahrscheinlichkeit von 0,737, die auch nach der zulässigen Halbierung aufgrund unserer einseitigen Fragestellung von der Signifikanzgrenze 0,05 sehr weit entfernt ist. Der LOT-Optimismus zeigt entgegen unserer Annahme fast keinen Effekt auf den mittleren Ärger in der fiktiven Situation.

Wer sich ausführlicher über die lineare Regressionsanalyse mit SPSS informieren möchte, kann ein ZIMK-Manuskript zu diesem Thema (Baltes-Götz 2019) auf dem Webserver der Universität Trier finden:

https://www.uni-trier.de/?id=22489

10.1.2 Methodische Anmerkungen

10.1.2.1 Explorative Analysen im Anschluss an einen „gescheiterten" Hypothesentest

Auf das „Scheitern" einer konfirmatorischen Forschungsbemühung werden in der Regel explorative Analysen folgen, wobei revidierte bzw. neue Hypothesen entstehen können. Wir werden uns in Abschnitt 11.4 z. B. dafür interessieren, ob eventuell das Geschlecht den Zusammenhang zwischen Optimismus und Ärger moderiert. Allerdings ist es *nicht* möglich, revidierte oder neue Hypothesen anhand *derselben* Stichprobe zu testen. Sie dürfen und sollen aus Ihren Daten etwas lernen, aber ein Test der dabei generierten Hypothesen erfordert eine neue, unabhängige Stichprobe.

Außerdem sollten Sie es nicht unterlassen, das „Scheitern" einer Hypothese zu veröffentlichen. Ansonsten tragen Sie dazu bei, in der Fachliteratur ein systematisch verzerrtes Bild der Wirklichkeit aufzubauen.

10.1.2.2 Post hoc - Poweranalyse

Bei der Interpretation des Testausgangs zur differentialpsychologischen Hypothese ist zu beachten, dass die Power des t-Tests zum Regressionskoeffizienten in unserer relativ kleinen Stichprobe recht bescheiden ist, sodass kleine Effekte leicht übersehen werden können (siehe Power-Analyse in Abschnitt 2.3.3). Unser Testergebnis kann nicht als *Beleg* für die Gültigkeit der Nullhypothese interpretiert werden, doch spricht es gegen die Existenz eines *starken* Effekts. Um zu genaueren Aussagen zu kommen, betrachten wir die Power des t-Tests bei unterschiedlichen Effektstärken.

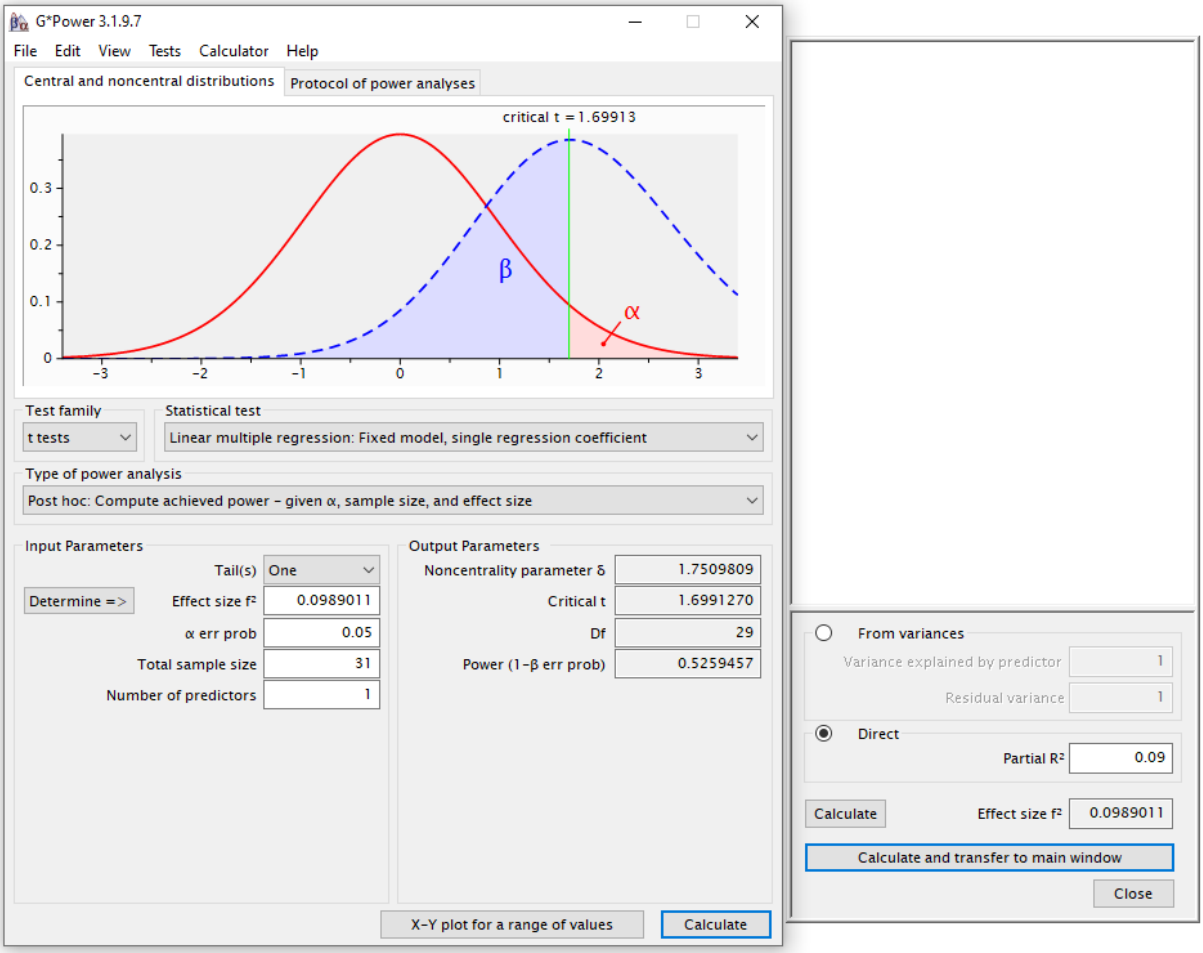
Dabei verwenden wir erneut das Programm **G*Power 3.1**, das schon bei der Stichprobenumfangsplanung in Abschnitt 2.3 zum Einsatz kam. Auf den Pool-PCs der Universität Trier unter dem Betriebssystem Windows ist G*Power 3.1 im Startmenü folgendermaßen zu finden:

Statistik > GPower

Zunächst untersuchen wir in einer Post hoc - Power-Analyse, mit welcher Wahrscheinlichkeit ein Effekt der Stärke 0,1 (ca. 9% Varianzaufklärung durch den Regressor) in einer Zufallsstichprobe mit 31 Fällen zu einem signifikanten Ergebnis führt. Wir wählen:

- **Test family**: **t-Tests**
- **Statistical test**: **Linear Multiple Regression: Fixed model, single regression coefficient**
- **Type of power analysis** **Post hoc**
- **Tail(s)** **One**
- **Effect size f^2** 0.0989011
- **α err prob** 0.05
- **Total sample size** 31
- **Number of predictors** 1

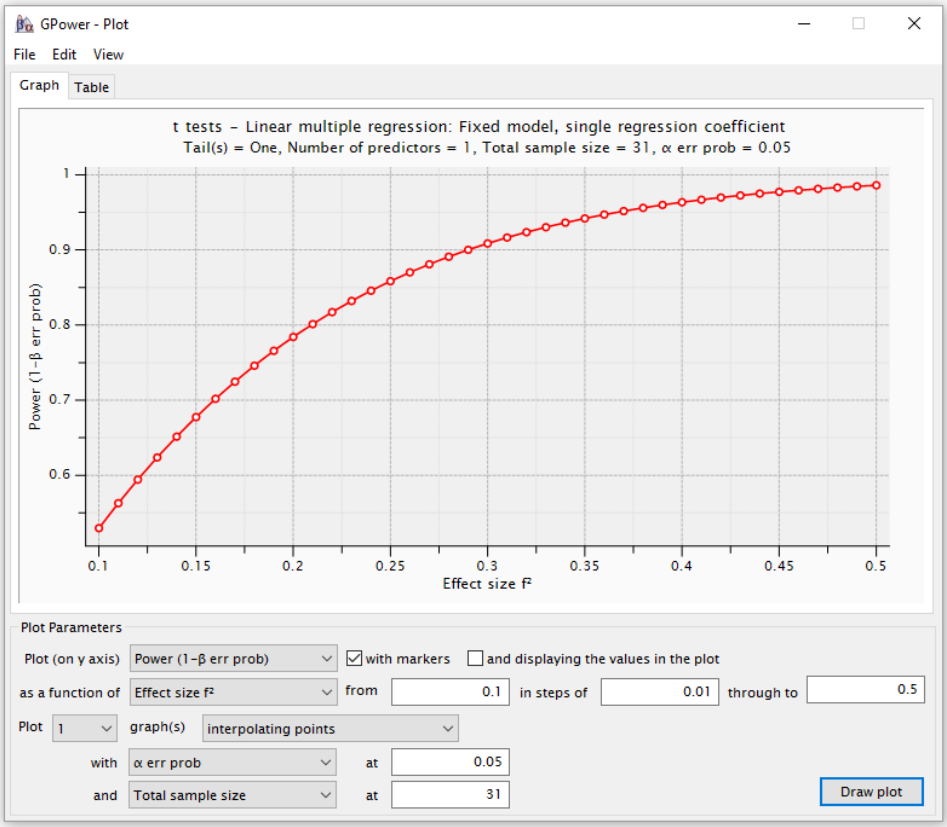
Nach einem Klick auf den Schalter **Calculate** wird für den Test zur differentialpsychologischen Hypothese eine Teststärke (Power) von lediglich 0,53 berechnet:



Um zu einer Darstellung der Power als Funktion der Effektstärke zu gelangen, klicken wir auf den Schalter **X-Y plot for a range of values** und wählen

- **Plot (on y axis)** **Power (1 - β err prob)**
- **as a function of** **Effect size f^2**
- **from** 0.1
- **in steps of** 0.01
- **through to** 0.5
- **Plot** 1

Nach einem Klick auf den Schalter **Draw Plot** zeigt die folgende Abbildung, wie bei fester Stichprobengröße (n = 31) die Power des einseitigen Tests von der Effektstärke abhängt:



Erst ab einer Effektstärke von ca. f^2 = 0,35 (bzw. ρ^2 = 0,26) ist die Power so groß (ca. 0,95), dass man die ausgebliebene Signifikanz als Beleg gegen einen Effekt dieser Stärke werten kann. Unserer Studie hat also keinesfalls die differentialpsychologische Nullhypothese bewiesen, aber doch ein Argument gegen die Existenz eines starken Effekts ($f^2 \geq 0{,}35$) geliefert.

Weil G*Power keinen Plot mit der quadrierten (partiellen) Korrelation ρ^2 auf der X-Achse liefert, soll noch eine Gleichung zur Berechnung ρ^2 aus dem Effektstärkemaß f^2 angegeben werden. Aus der Definitionsgleichung für f^2

$$f^2 = \frac{\rho^2}{1-\rho^2}$$

folgt:

$$\rho^2 = \frac{f^2}{1+f^2}$$

10.1.2.3 Fehlende Werte

Fehlende Werte haben Einbußen bei der Teststärke und oft auch verzerrte Schätzwerte zur Folge, sodass einige Anstrengungen zur Vermeidung oder Reduktion des Problems angemessen sind. Wir haben bei der Berechnung des LOT-Werts geeignete Maßnahmen ergriffen, um die Anzahl fehlender Werte gering zu halten (vgl. Abschnitt 7.4).

Wer sich über die in SPSS und im Strukturgleichungsanalyseprogramm Amos enthaltenen Möglichkeiten zur Analyse und Behandlung fehlender Werte informieren möchte, kann ein ZIMK-Manuskript zu diesem Thema (Baltes-Götz 2013) auf dem Webserver der Universität Trier finden:

https://www.uni-trier.de/?id=23239

10.2 Prüfung der KFA-Hypothese per Vorzeichentest

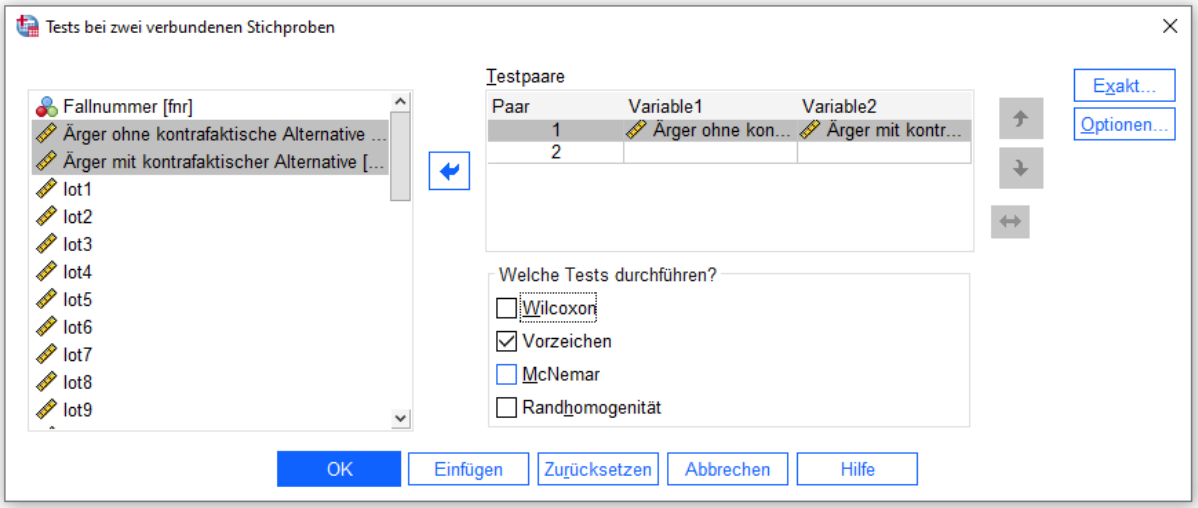
Nun wollen wir die allgemeinpsychologische Kernhypothese unserer Studie prüfen, dass der Ärger über ein ungünstiges Ereignis durch die mentale Verfügbarkeit kontrafaktischer (also positiver) Alternativen gesteigert wird. Aufgrund der Ausreißer- und Verteilungsanalyse in Abschnitt 9.3 haben wir uns dafür entschieden, statt des ursprünglich geplanten parametrischen t-Tests für abhängige Stichproben den nichtparametrischen **Vorzeichentest** zu verwenden.

Weil der Vorzeichentest weit weniger empfindlich auf Ausreißer reagiert als der parametrische t-Test, können wir den in Abschnitt 9.3 identifizierten kritischen Fall Nr. 4 in der Auswertung belassen. Damit vermeiden wir den Verdacht, die Daten zu unseren Gunsten bereinigt zu haben. Heben Sie also bitte die MD-Deklaration für den Wert -4 bei der Variablen AERGZ wieder auf.

Wir starten mit dem Menübefehl:

Analysieren > Nichtparametrische Tests >
Klassische Dialogfelder > Zwei verbundene Stichproben

In der folgenden Dialogbox wählen wir das **Testpaar** mit den zu vergleichenden Variablen und den gewünschten **Test**:



Wir erhalten das folgende Ergebnis:

Häufigkeiten

		N
Ärger mit kontrafaktischer Alternative - Ärger ohne kontrafaktische Alternative	Negative Differenzen [a]	2
	Positive Differenzen [b]	26
	Bindungen [c]	3
	Gesamt	31

a. Ärger mit kontrafaktischer Alternative < Ärger ohne kontrafaktische Alternative

b. Ärger mit kontrafaktischer Alternative > Ärger ohne kontrafaktische Alternative

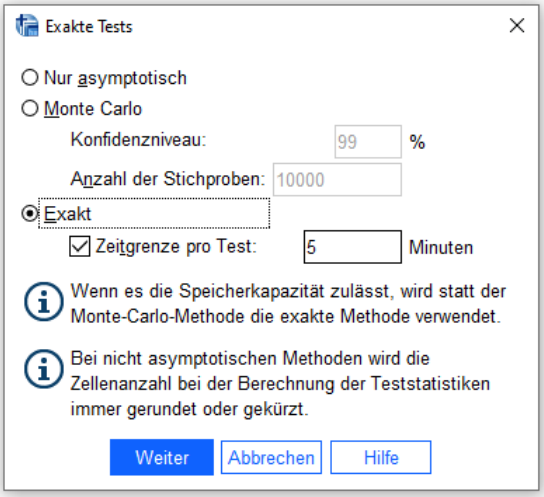
c. Ärger mit kontrafaktischer Alternative = Ärger ohne kontrafaktische Alternative

Teststatistiken[a]

	Ärger mit kontrafaktischer Alternative - Ärger ohne kontrafaktische Alternative
Z	-4,347
Asymp. Sig. (2-seitig)	,000

a. Vorzeichentest

In unserer kleinen Stichprobe kann anstelle der per Voreinstellung gelieferten *asymptotischen* Überschreitungswahrscheinlichkeit auch die *exakte* ohne großen Zeitaufwand berechnet werden. Das ist zwar nicht unbedingt erforderlich, weil die Verteilungs-Approximation schon ab $n \geq 20$ hinreichend genau ist, kann aber auch nicht schaden, wenn der Zeitaufwand im Rahmen bleibt. Nach einem Mausklick auf den Schalter **Exakt** in obiger Dialogbox kann der exakte Test folgendermaßen angefordert werden:



Die unserer gerichteten Fragestellung entsprechende einseitige Überschreitungswahrscheinlichkeit ist deutlich kleiner als die übliche Grenze von 0,05. Damit kann die KFA-Nullhypothese (*Kein Ärgerzuwachs durch eine kontrafaktische Alternative*) zurückgewiesen werden:

Statistik für Test[a]

	Ärger mit kontrafaktischer Alternative - Ärger ohne kontrafaktische Alternative
Z	-4,347
Asymptotische Signifikanz (2-seitig)	,000
Exakte Signifikanz (2-seitig)	,000
Exakte Signifikanz (1-seitig)	,000
Punkt-Wahrscheinlichkeit	,000

a. Vorzeichentest

Als gut mit dem Vorzeichentest harmonierendes Effektstärkemaß bietet sich der Anteil der Personen mit positivem AERGZ-Wert an: $P(\text{AERGZ} > 0)$. Wir erhalten die Schätzung:

$$\frac{26}{31} \approx 0{,}839$$

Für eine Variable mit den Werten

$$1, \text{ falls } \text{AERGZ} > 0$$
$$0, \text{ falls } \text{AERGZ} \leq 0$$

ergibt sich die eben vorgeschlagene Effektstärke als Anteil der Einser-Fälle, und über das in Abschnitt 5.8 beschriebene Verfahren erhalten wir sogar ein Vertrauensintervall für dieses Effektstärkemaß:

Konfidenzintervallübersicht

			95,0%-Konfidenzintervall	
Konfidenzintervalltyp	Parameter	Schätzer	Unterer Wert	Oberer Wert
Binomialerfolgsrate für eine Stichprobe (Jeffreys)	Wahrscheinlichkeit(ZG0=1).	,839	,682	,936

Man kann eine Variable mit den beschriebenen Werten folgendermaßen aus AERGZ berechnen:

```
COMPUTE esvt = aergz > 0.
EXECUTE.
```

Cohen (1988, S. 147) schlägt vor, zur Beschreibung der Effektstärke beim Vorzeichentest von der Wahrscheinlichkeit für positive Differenzen den Wert 0,5 zu subtrahieren:

$$g := P^{+} - 0{,}5$$

Er nennt zur Interpretation seines Effektstärkebegriffs die folgenden Orientierungswerte:

- kleiner Effekt: $g = 0{,}05$
- mittlerer Effekt: $g = 0{,}15$
- großer Effekt: $g = 0{,}25$

Für unsere KFA-Daten resultiert ein sehr großer Effekt:

$$0{,}839 - 0{,}5 \approx 0{,}339$$

G*Power 3.1 unterstützt eine Power-Analyse zum Vorzeichentest unter Verwendung von Cohens Effektstärkemaß, wobei die folgenden Einstellungen zu wählen sind:

- **Test family**: **Exact**
- **Statistical test**: **Proportion: Sign test (binomial test)**

Zur Beurteilung der empirischen Effektstärke im Sinne der in Abschnitt 2.3.2.4 verwendeten Definition ziehen wir trotz schiefer Verteilung den Mittelwert und die Standardabweichung von AERGZ heran, schließen aber den in Abschnitt 9.3 diagnostizierten extremen Wert aus. Die folgende Tabelle stammt von der in Abschnitt 9.2 beschriebenen Prozedur zur explorativen Datenanalyse:

Deskriptive Statistik

			Statistik	Standardfehler
Ärgerzuwachs durch die KFA	Mittelwert		2,20	,242
	95% Konfidenzintervall des Mittelwerts	Untergrenze	1,71	
		Obergrenze	2,69	
	5% getrimmtes Mittel		2,26	
	Median		2,00	
	Varianz		1,752	
	Standardabweichung		1,324	
	Minimum		-1	
	Maximum		4	
	Spannweite		5	
	Interquartilbereich		1	
	Schiefe		-,585	,427
	Kurtosis		-,059	,833

Setzt man die Schätzwerte für μ_z und σ_z in den Ausdruck zur d_z - Berechnung ein,

$$d_z := \frac{\mu_z}{\sigma_z}$$

ergibt sich ein sehr starker Effekt:

$$\frac{2{,}20}{1{,}324} = 1{,}66$$

Der geschätzte Effekt ist erheblich größer als der in Abschnitt 2.3.2.4 bei der Stichprobenumfangsplanung angenommene Effekt, weil die Standardabweichung von AERGZ deutlich unter dem angenommenen Wert von 2 (für die durch 10 dividierten Celsius-Werte) liegt.

Wie sich gleich in Abschnitt 10.3 zeigen wird, liefert die SPSS-Prozedur zum t-Test für abhängige Stichproben den d_z - Wert samt Vertrauensintervall.

Nach Klärung der zentralen Hypothesen ist unser Projekt eigentlich abgeschlossen, aber es gibt noch viele SPSS-Optionen kennenzulernen, und unsere Daten enthalten sicher auch noch einige interessante Details.

10.3 Übung

Für die Differenzvariable (GEWICHT - IDGEW) akzeptieren beide Normalverteilungstests die Nullhypothese (vgl. Abschnitt 9):

Tests auf Normalverteilung

	Kolmogorov-Smirnov [a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
RIDIFF	,092	31	,200 *	,984	31	,905

*. Dies ist eine untere Grenze der echten Signifikanz.

a. Signifikanzkorrektur nach Lilliefors

Folglich darf mit den Variablen GEWICHT und IDGEW ein t-Test für verbundene Stichproben zum folgenden Testproblem durchgeführt werden (vgl. Abschnitt 8.2):

H_0: Das Realgewicht der Trierer Studierenden liegt im Mittel nicht unter dem Idealgewicht nach der Formel „Größe - 100“.

versus

H_1: Die Trierer Studierenden sind in Relation zur Idealgewichtsformel „Größe - 100“ im Mittel zu leicht.

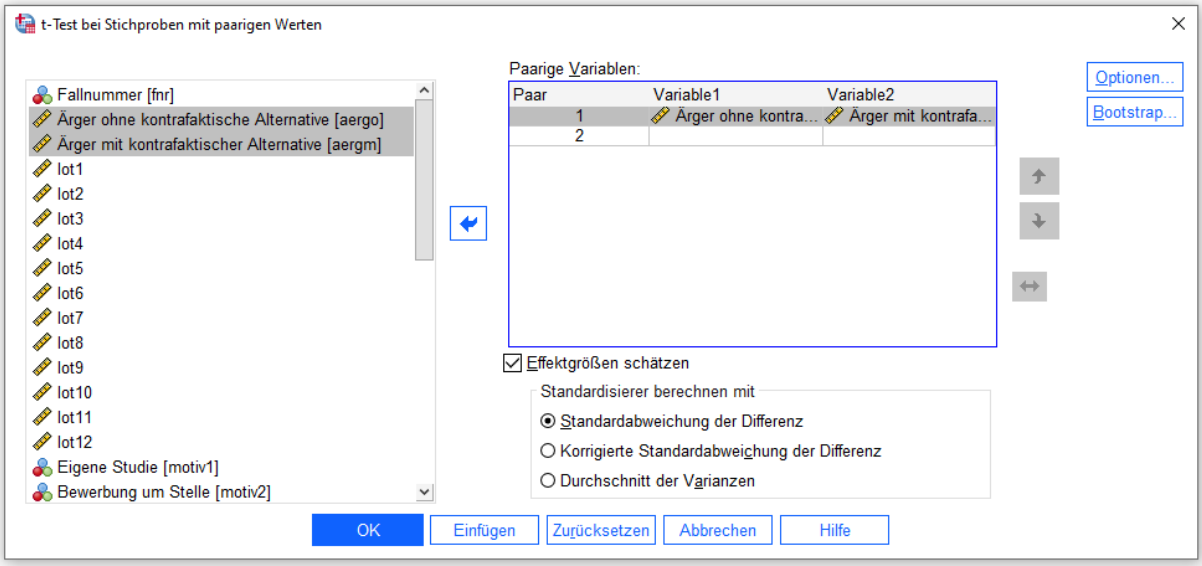
Mit den Symbolen μ_R für den Realgewichtsmittelwert und μ_I für den Idealgewichtsmittelwert kann man die beiden konkurrierenden Hypothesen kompakter notieren:

$$H_0 : \mu_R \geq \mu_I \quad \text{versus} \quad H_1 : \mu_R < \mu_I$$

Öffnen Sie mit dem Menübefehl

Analysieren > Mittelwerte vergleichen > t-Test bei verbundenen Stichproben

die zuständige Dialogbox, und bilden Sie (z. B. durch Markieren und Transportieren) ein **Paar** aus den beiden Gewichtsvariablen:



Die resultierende T_Z - Prüfstatistik übertrifft betragsmäßig den in Abschnitt 8.1.5 vorgestellten kritischen Wert 1,70 sehr deutlich, und dementsprechend fällt die Überschreitungswahrscheinlichkeit sehr klein aus, sodass die Nullhypothese zu verwerfen ist:

Test bei gepaarten Stichproben

		Gepaarte Differenzen					T	df	Signifikanz	
		Mittelwert	Std.-Abweichung	Standardfehler des Mittelwertes	95% Konfidenzintervall der Differenz Unterer Wert	Oberer Wert			Einseitiges p	Zweiseitiges p
Paaren 1	Körpergewicht (in kg) - Idealgewicht nach der Formel Größe - 100	-9,3226	6,1881	1,1114	-11,5924	-7,0528	-8,388	30	<,001	<,001

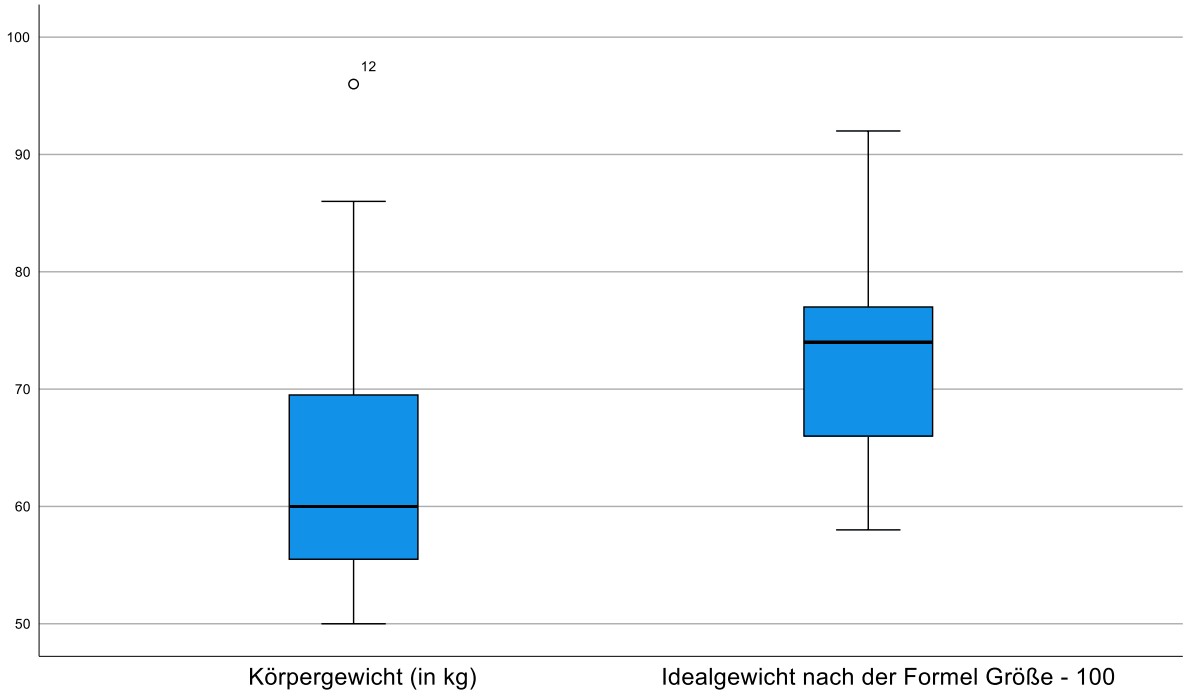
SPSS liefert seit der Version 27 für Cohens d_z die Punktschätzung und das Konfidenzintervall:

Effektgrößen bei Stichproben mit paarigen Werten

			Standardisierer [a]	Punktschätzung	95% Konfidenzintervall	
					Unterer Wert	Oberer Wert
Paaren 1	Körpergewicht (in kg) - Idealgewicht nach der Formel Größe - 100	Cohen's d	6,1881	-1,507	-2,018	-,983
		Hedges' Korrektur	6,2668	-1,488	-1,993	-,971

a. Der bei der Schätzung der Effektgrößen verwendete Nenner.
Cohen's d verwendet die Standardabweichung einer Stichprobe der Mittelwertdifferenz.
Hedges' Korrektur verwendet die Standardabweichung einer Stichprobe der Mittelwertdifferenz und einen Korrekturfaktor.

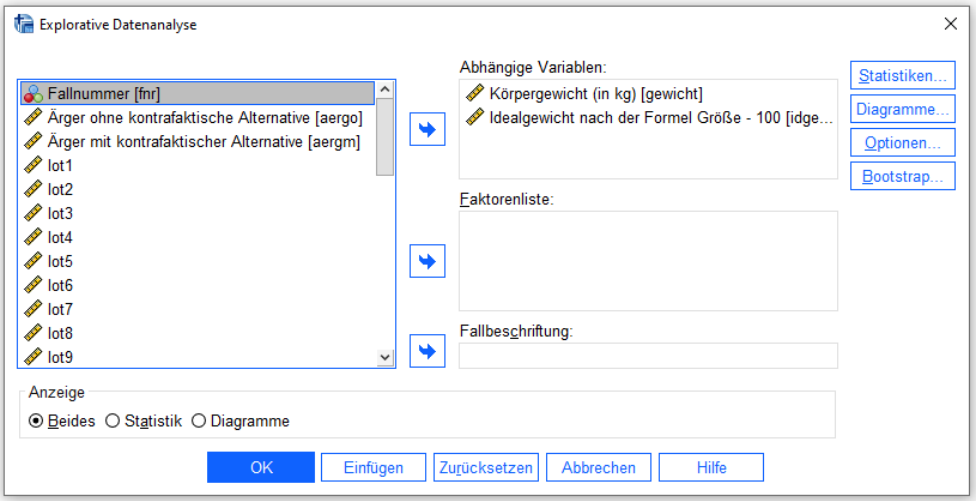
Zur grafischen Veranschaulichung der Ergebnisse eignet sich eine gemeinsame Darstellung der Boxplots zu den beiden Verteilungen:



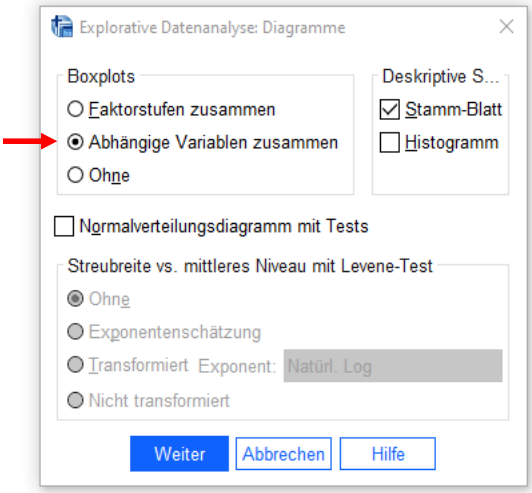
Diese Abbildung ist mit geringem Aufwand über die SPSS-Prozedur zur explorativen Datenanalyse zu erstellen. Öffnen Sie deren Dialogbox mit

Analysieren > Deskriptive Statistiken > Explorative Datenanalyse

und verwenden Sie GEWICHT sowie IDGEW als **abhängige Variablen**:



Veranlassen Sie im Subdialog **Diagramme**, dass die **abhängigen Variablen zusammen** dargestellt werden:



10.4 Arbeiten mit dem Ausgabefenster (Teil II)

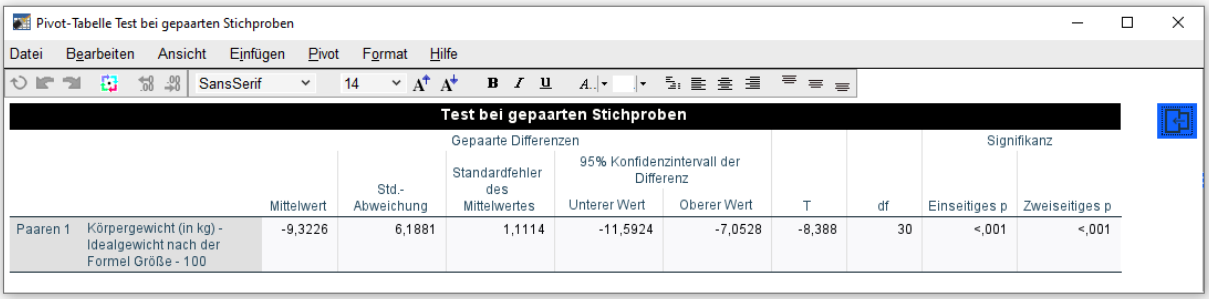
Unter dem *Pivotieren* einer Tabelle versteht SPSS u. a. die folgenden Operationen:

- Dargestellte Dimensionen neu auf die Zeilen, Spalten und Schichten der Tabelle verteilen
- Schachtelungsordnung bei den Zeilen-, Spalten- oder Schichtdimensionen ändern
- Kategorien einer Dimension gruppieren oder eine Gruppierung aufheben
- Kategorien einer Dimension ausblenden, verschieben oder vertauschen

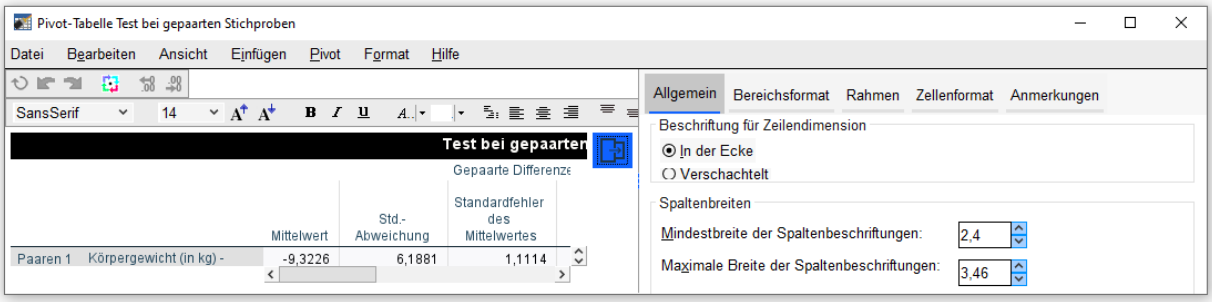
Im SPSS-Ausgabefenster sind die meisten Tabellen pivotierbar, und das zuständige Werkzeug ist der Pivot-Editor. Er unterstützt neben den Pivot-Operationen noch weitere Maßnahmen zur Gestaltung von Tabellen (z. B. Schriftarten wählen, Anzahl der Nachkommastellen bei Datenzellen ändern). Der Pivot-Editor ist leider nicht fehlerfrei, mit etwas Geduld aber doch produktiv einsetzbar.

10.4.1 Pivot-Editor starten

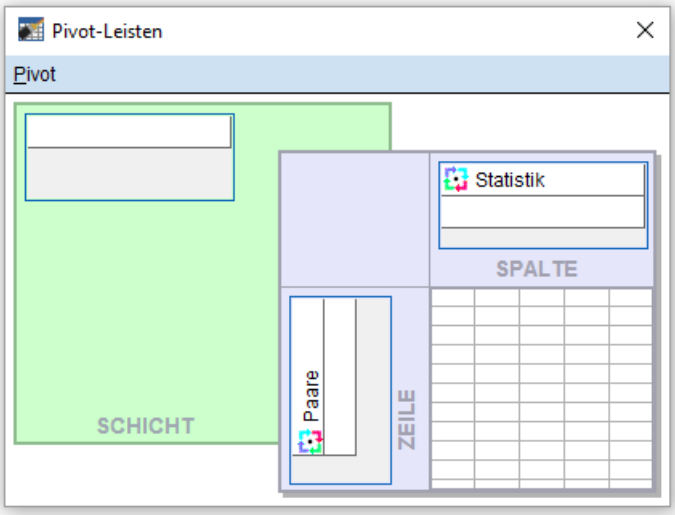
Um mit dem Editieren einer Tabelle zu beginnen, können Sie einen Doppelklick darauf setzen oder die Option **Bearbeiten** aus ihrem Kontextmenü wählen. Das folgende Fenster des Pivot-Editors zeigt die in einer Übung (siehe Abschnitt 10.3) von Ihnen erstellten Tabelle mit dem t-Test zum Vergleich von Real- und Idealgewicht:



Den vorläufig nicht benötigten Formatierungsbereich auf der rechten Seite schließen wir per Mausklick auf den Symbolschalter .



Für allgemeine Pivot-Operationen wird das folgende Fenster benötigt:



Es enthält je eine Ablagezone für die Schichten, Zeilen und Spalten der Tabelle und je einen Eintrag mit Pivotsymbol für die dargestellten Tabellendimensionen. Sollten Sie das Fenster vermissen, können Sie es mit dem folgenden Menübefehl aktivieren:

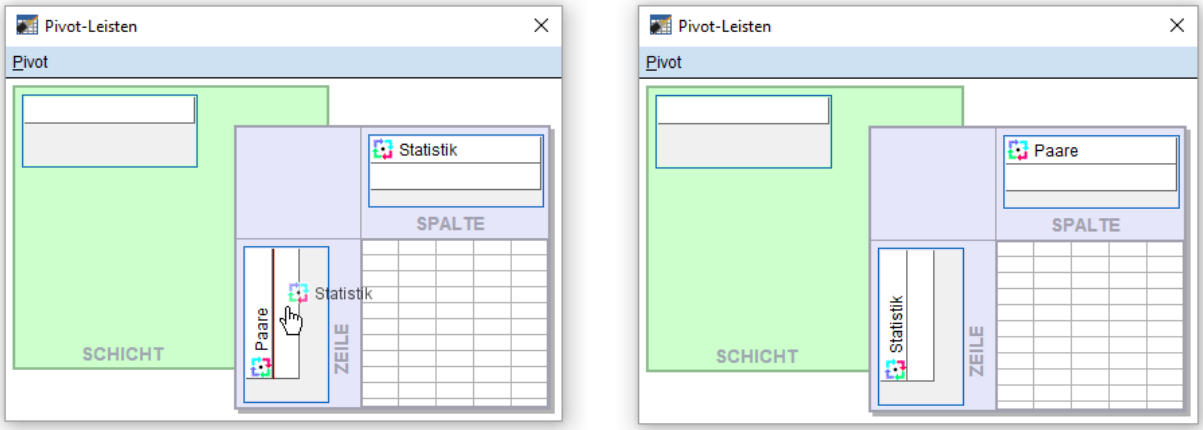
Pivot > Pivot-Leisten

Die Tabelle mit dem t-Test zum Vergleich von Real- und Idealgewicht enthält leider nur *eine* Schicht, sodass wir den Umgang mit Mehrschichttabellen nicht üben können. In den Zeilen der Tabelle wird die Dimension **Paare** dargestellt. Da wir nur ein einziges Variablenpaar untersucht haben, hat diese Dimension nur *eine* Kategorie. Die Spaltendimension **Statistik** sorgt mit ihren zahlreichen Kategorien für eine überbreite Tabelle, die schlecht auf eine DIN-A4 - Seite im Hochformat passt.

10.4.2 Dimensionen verschieben

Durch das Verschieben ihres Pivotsymbols kann man für eine Dimension neu festlegen, ob ihre Kategorien durch Spalten, Zeilen oder Schichten dargestellt werden sollen. Wenn in unserem Beispiel die beiden Pivotsymbole bzw. Dimensionen ihre Plätze tauschen,

Verschiebung der Pivotsymbole **Ergebnis**



benötigt die Tabelle in horizontaler Richtung deutlich weniger Platz:

Test bei gepaarten Stichproben

			Paaren 1 Körpergewicht (in kg) - Idealgewicht nach der Formel Größe - 100
Gepaarte Differenzen	Mittelwert		-9,3226
	Std.-Abweichung		6,1881
	Standardfehler des Mittelwertes		1,1114
	95% Konfidenzintervall der Differenz	Unterer Wert	-11,5924
		Oberer Wert	-7,0528
T			-8,388
df			30
Signifikanz	Einseitiges p		<,001
	Zweiseitiges p		,000

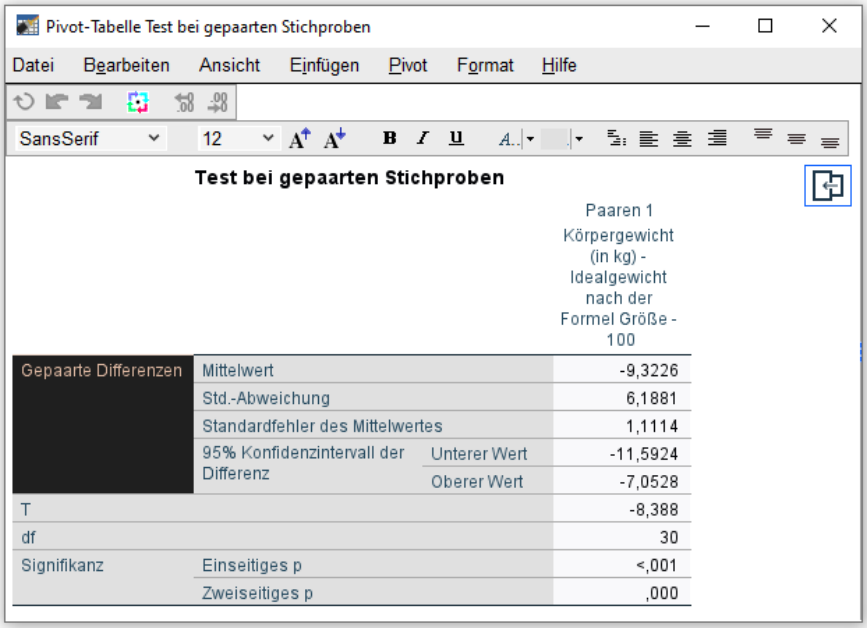
Wenn es lediglich um das Vertauschen von Zeilen und Spalten einer Tabelle geht, kann man an Stelle des Pivot-Werkzeugs auch den Menübefehl

Pivot > Zeilen und Spalten transponieren

verwenden.

10.4.3 Gruppierungen

Kategorien einer Dimension können zu einer Gruppe zusammengefasst und durch eine etikettierende Zelle hervorgehoben sein. In unserem Beispiel zeigt sich bei der Statistikdimension eine Gruppe mit dem Etikett **Gepaarte Differenzen** und bei der **Paare**-Dimension eine Gruppe mit dem Etikett **Paaren 1**:



Test bei gepaarten Stichproben

			Paaren 1 Körpergewicht (in kg) - Idealgewicht nach der Formel Größe - 100
Gepaarte Differenzen	Mittelwert		-9,3226
	Std.-Abweichung		6,1881
	Standardfehler des Mittelwertes		1,1114
	95% Konfidenzintervall der Differenz	Unterer Wert	-11,5924
		Oberer Wert	-7,0528
T			-8,388
df			30
Signifikanz	Einseitiges p		<,001
	Zweiseitiges p		,000

Eine unerwünschte Gruppierung lässt sich folgendermaßen aufheben:

- Rechtsklick auf das Kategorienetikett
- Aus dem Kontextmenü wählen: **Gruppierung aufheben**

So sieht die Beispieltabelle ohne die oben genannten, ziemlich überflüssigen Gruppierungen aus:

Test bei gepaarten Stichproben

		Körpergewicht (in kg) - Idealgewicht nach der Formel Größe - 100
Mittelwert		-9,3226
Std.-Abweichung		6,1881
Standardfehler des Mittelwertes		1,1114
95% Konfidenzintervall der Differenz	Unterer Wert	-11,5924
	Oberer Wert	-7,0528
T		-8,388
df		30
Signifikanz	Einseitiges p	<,001
	Zweiseitiges p	,000

Wenn Sie mehrere Kategorien einer Dimension zu einer Gruppe zusammenfassen wollen, können Sie folgendermaßen vorgehen:

- Alle Kategorien markieren
- Aus dem Kontextmenü zum markierten Bereich die Option **Gruppe** wählen

In der folgenden Version unserer Tabelle wurde eine Gruppe mit den drei Kategorien zum t-Test gebildet:

Test bei gepaarten Stichproben

			Körpergewicht (in kg) - Idealgewicht nach der Formel Größe - 100
Mittelwert			-9,3226
Std.-Abweichung			6,1881
Standardfehler des Mittelwertes			1,1114
95% Konfidenzintervall der Differenz	Unterer Wert		-11,5924
	Oberer Wert		-7,0528
Gruppenbeschriftung	T		-8,388
	df		30
	Signifikanz	Einseitiges p	<,001
		Zweiseitiges p	,000

Später werden wir die Beschriftung der Gruppenzelle vertikal innerhalb der Zelle zentrieren.

Auch selbst definierte Gruppierungen lassen sich über das Kontextmenü-Item **Gruppierung aufheben** wieder entfernen.

10.4.4 Kategorien aus- bzw. einblenden

Wenn eine SPSS-Tabelle zu ausführlich erscheint, können Kategorien einer Dimension ausgeblendet werden. In unserem Beispiel wollen wir auf den zweiseitigen Test verzichten:

Test bei gepaarten Stichproben

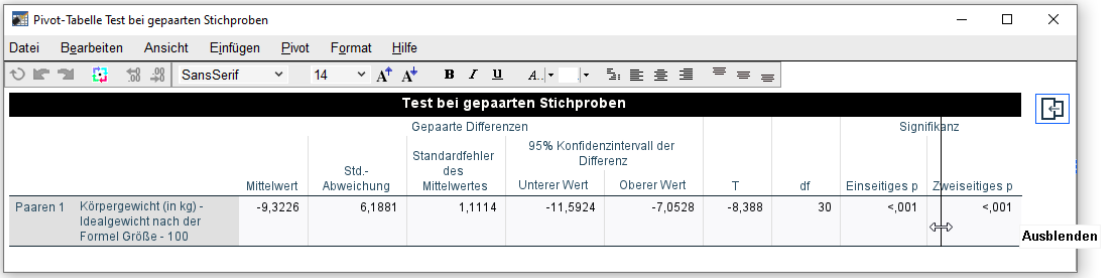
			Körpergewicht (in kg) - Idealgewicht nach der Formel Größe - 100
Mittelwert			-9,3226
Std.-Abweichung			6,1881
Standardfehler des Mittelwertes			1,1114
95% Konfidenzintervall der Differenz	Unterer Wert		-11,5924
	Oberer Wert		-7,0528
Signifikanztest	T		-8,388
	df		30
	Signifikanz	Einseitiges p	<,001

So lässt sich eine Kategorie ausblenden:

- Rechter Mausklick auf das Kategorienetikett
- Aus dem Kontextmenü wählen: **Auswählen > Datenzellen und Beschriftung**
- Erneuter rechter Mausklick auf das Kategorienetikett
- Aus dem Kontextmenü wählen: **Kategorie ausblenden**

In *Spalten* untergebrachte Kategorien kann man auch auf intuitive Weise eliminieren:

- linker Mausklick auf den rechten Spaltenrand, Maustaste gedrückt halten
- Spaltenbreite durch Verschieben der rechten Begrenzung reduzieren, bis die Quick-Info **Ausblenden** erscheint:



- Maustaste loslassen

Um vorher abgeschaltete Kategorien wieder einzublenden kann man (neben **Bearbeiten > Rückgängig**) den folgenden Menübefehl verwenden:

Ansicht > Alle Kategorien einblenden

10.4.5 Zellen modifizieren

10.4.5.1 Text editieren

Bei aktivem Pivot-Editor kann man nach einem Doppelklick auf eine Zelle den enthaltenen Text beliebig ändern. In unserem Beispiel sind folgende Änderungen sinnvoll:

- Der Titel sollte etwas informativer, und die Beschriftung der rechten Spalte sollte etwas sparsamer werden.
- Die Beschriftung zum p-Level sollte vereinfacht werden.

Das Ergebnis:

t-Test zum Vergleich von Real- und Idealgewicht

		Realgewicht - Idealgewicht
Mittelwert		-9,3226
Std.-Abweichung		6,1881
Standardfehler des Mittelwertes		1,1114
95% Konfidenzintervall der Differenz	Unterer Wert	-11,5924
	Oberer Wert	-7,0528
Signifikanztest	T	-8,388
	df	30
	p-Level (einseitig)	,000

10.4.5.2 Zellen zur weiteren Bearbeitung markieren

Mit dem Menübefehl **Bearbeiten > Auswählen** lassen sich Tabellenbestandteile (z. B. Tabellenkorpus, Datenzellen) zur weiteren Bearbeitung markieren. Außerdem stehen die windowsüblichen Markierungsmethoden per Maus und Tastatur zur Verfügung.

10.4.5.3 Zelleneigenschaften

Über die per Kontextmenü erreichbare Dialogbox mit den **Zelleneigenschaften** können zahlreiche Attribute der markierten Zellen beeinflusst werden:

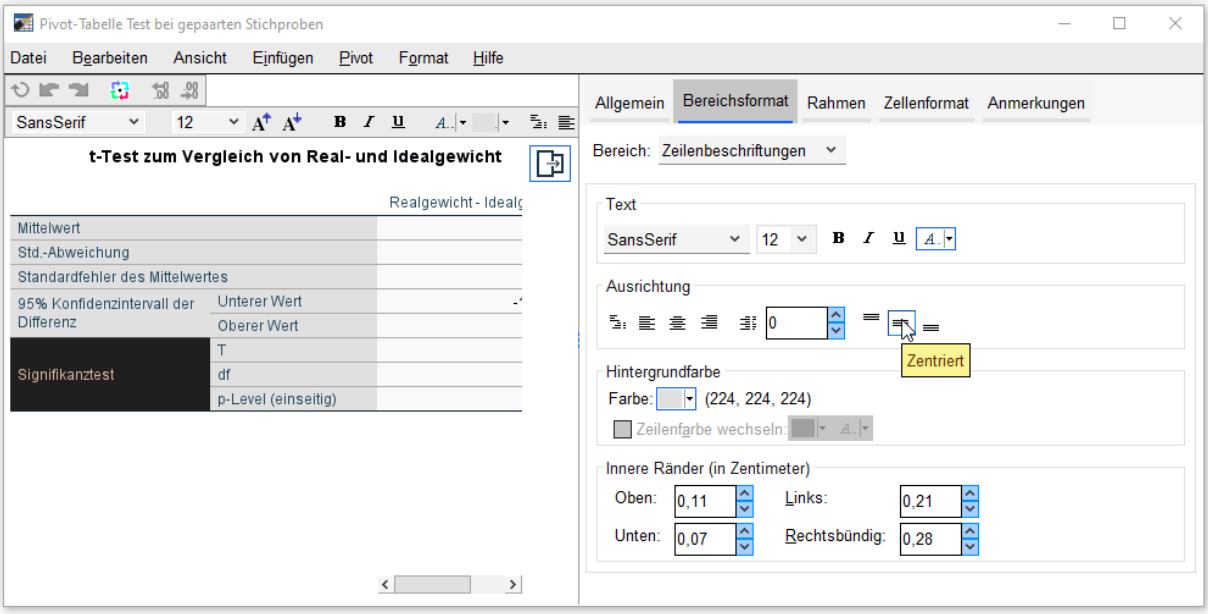
- Schriftart, Text- und Hintergrundfarbe
- Zahlenformate, Anzahl der Dezimalstellen
- Ausrichtung der Zellinhalte
- Randabstände der Zellinhalte

Um die Gruppenbeschriftung *Signifikanztest* in der Beispieltabelle vertikal zu zentrieren,

t-Test zum Vergleich von Real- und Idealgewicht

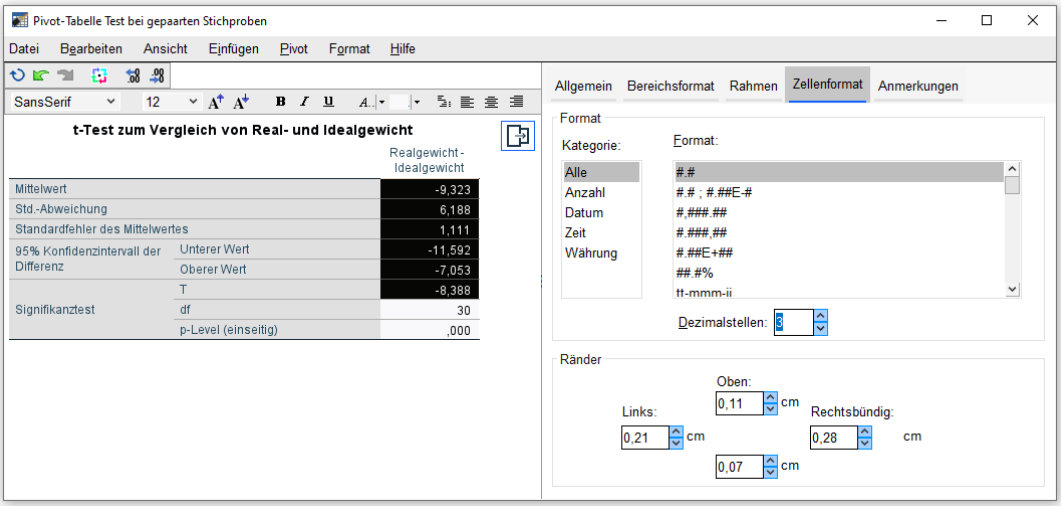
		Realgewicht - Idealgewicht
Mittelwert		-9,3226
Std.-Abweichung		6,1881
Standardfehler des Mittelwertes		1,1114
95% Konfidenzintervall der Differenz	Unterer Wert	-11,5924
	Oberer Wert	-7,0528
Signifikanztest	T	-8,388
	df	30
	p-Level (einseitig)	,000

wurde der Formatierungsbereich auf der rechten Seite des Pivot-Editors über einen Mausklick auf den blauen Schalter geöffnet. Dann wurde die zu verändernde Zelle markiert und auf der Registerkarte **Bereichsformat** die gewünschte **Ausrichtung** angefordert:



Außerdem wurde die Anzahl der Dezimalstellen folgendermaßen auf drei Stellen vereinheitlicht:

- Alle betroffenen Zellen markieren.
- Auf der **Zellenformat**-Registerkarte die gewünschte Anzahl der **Dezimalstellen** eintragen:



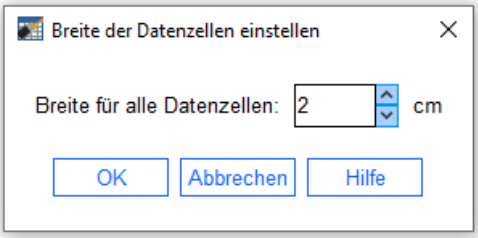
10.4.5.4 Spaltenbreite

Wenn sich der Mauszeiger über dem rechten Rand einer Spalte befindet, ändert er seine Form zu einem doppelseitigen Pfeil ⇔. Jetzt können Sie durch Klicken und Ziehen bei gedrückter linker Maustaste die Spaltengrenze verschieben und somit die Spaltenbreite ändern. Diese Technik haben wir oben schon dazu benutzt, Zellen oder Spalten auf die Breite null zu bringen und damit komplett auszublenden.

Über den Menübefehl

Format > Breite der Datenzellen

lässt sich die Breite sämtlicher Datenzellen einer Tabelle numerisch spezifizieren, z. B.:



Nach missratenen Gestaltungsbemühungen bringt eventuell der Menübefehl

Format > Automatisch anpassen

wieder ein akzeptables Ergebnis zu Stande.

10.4.6 Tabellenvorlagen

Für die mittlerweile ziemlich brauchbare Beispieltabelle

t-Test zum Vergleich von Real- und Idealgewicht

		Realgewicht - Idealgewicht
Mittelwert		-9,323
Std.-Abweichung		6,188
Standardfehler des Mittelwertes		1,111
95% Konfidenzintervall der Differenz	Unterer Wert	-11,592
	Oberer Wert	-7,053
Signifikanztest	T	-8,388
	df	30
	p-Level (einseitig)	,000

kann nach dem Menübefehl

Format > Tabellenvorlagen

das Design einer Tabellenvorlage übernommen werden. So sieht unser Beispiel nach Anwendung der Vorlage **APA_SansSerif_10pt** aus:

t-Test zum Vergleich von Real- und Idealgewicht

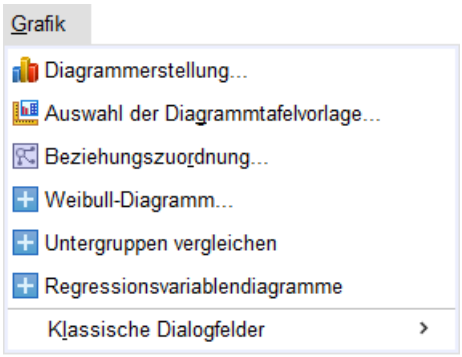
		Realgewicht - Idealgewicht
Mittelwert		-9,323
Std.-Abweichung		6,188
Standardfehler des Mittelwertes		1,111
95% Konfidenzintervall der Differenz	Unterer Wert	-11,592
	Oberer Wert	-7,053
Signifikanztest	T	-8,388
	df	30
	p-Level (einseitig)	,000

Im **Optionen**-Dialog von SPSS (erreichbar über **Bearbeiten > Optionen**) kann man auf der Registerkarte **Pivot-Tabellen** eine Tabellenvorlage durch Markieren zum Standard machen.

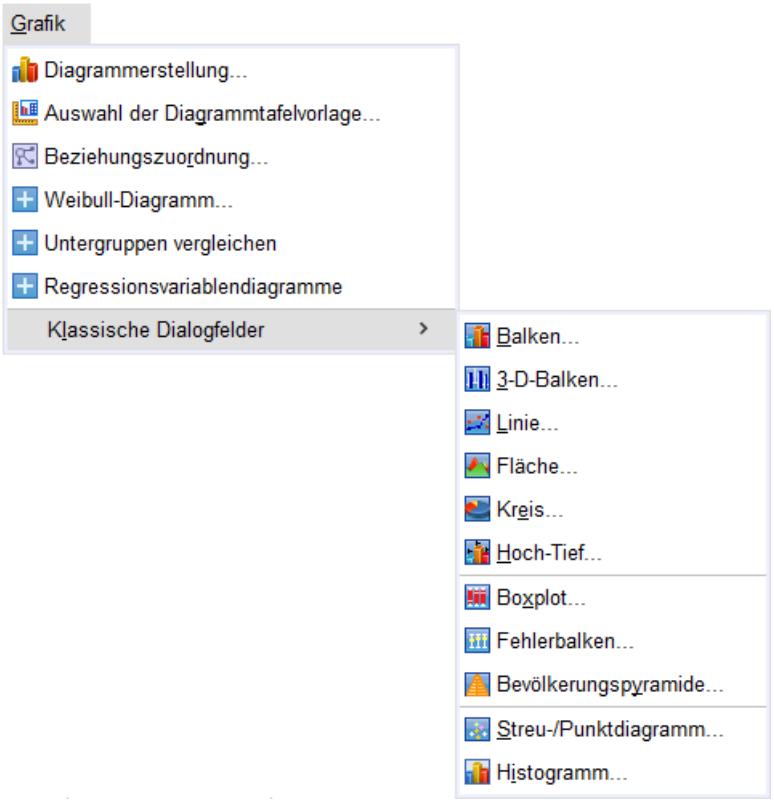
11 Diagrammerstellung am Beispiel des Streudiagramms

Wir haben schon einige grafische Darstellungsmöglichkeiten kennengelernt, die von SPSS im Rahmen von Statistikprozeduren angeboten werden (z. B. Balkendiagramm, Kreisdiagramm, Histogramm, Boxplot). Im aktuellen Kapitel arbeiten wir erstmals mit dem **Grafik**-Menü und vor allem mit dem Diagrammeditor zur individuellen Nachbearbeitung von Diagrammen.

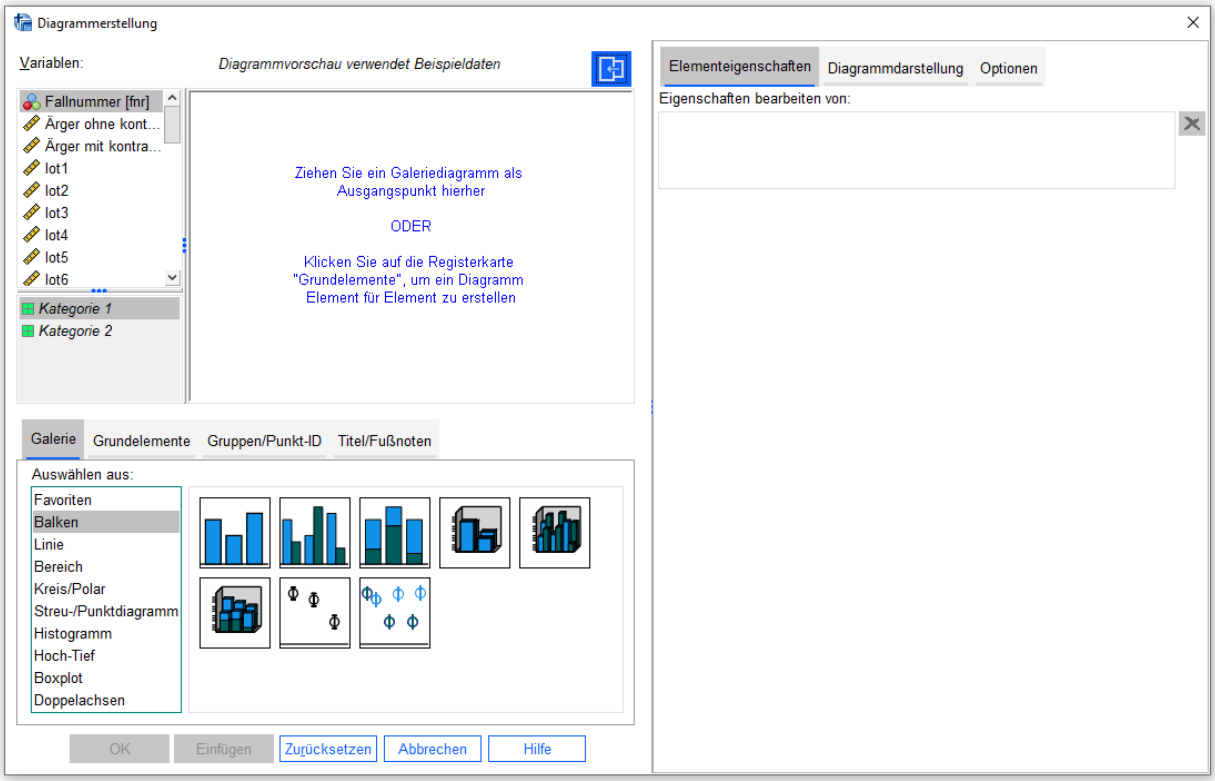
SPSS-Einsteiger werden vermutlich durch das **Grafik**-Menü leicht irritiert, weil hier mehrere Zugänge angeboten werden:



Über die **klassischen Dialogfelder**



(verknüpft mit dem SPSS-Kommando GRAPH) oder mit dem Dialog **Diagrammerstellung**



(verknüpft mit dem SPSS-Kommando GGRAPH und der *Graphics Programming Language* (GPL)) entstehen Diagramme, die anschließend mit dem **Diagrammeditor** modifiziert werden können.

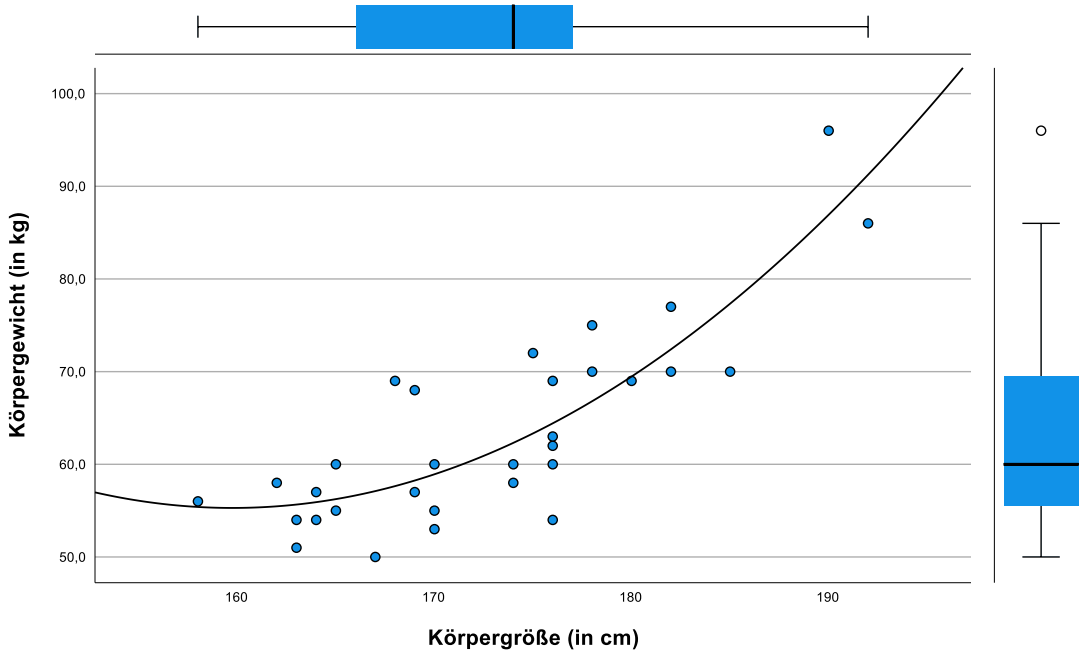
Über die **Auswahl der Diagrammtafelvorlage** ist ein weiterer Assistent zur Diagrammerstellung verfügbar. Obwohl im Hintergrund ebenfalls das SPSS-Kommando GGRAPH beteiligt ist, gehören die Diagramme zu einer anderen Ausgabekategorie und werden mit dem **Diagrammtafel-Editor** bearbeitet. Offenbar reduziert IBM/SPSS die Unterstützung für die Diagrammtafelvorlagen. Jedenfalls wird das zur Erstellung eigener Vorlagen konzipierte Programm *Visualization Designer* nicht mehr angeboten.[1]

Ein Blick in das Benutzerhandbuch zu SPSS 28 (siehe IBM Corp. 2021b, Kapitel 15) zeigt, dass der Hersteller von den möglichen Einstiegen in die Diagrammproduktion die **Diagrammerstellung** empfiehlt, die daher im Manuskript bevorzugt eingesetzt wird. Die Technik der Diagrammtafelvorlagen wird im Manuskript *nicht* behandelt.

Außerdem bietet das **Grafik**-Menü noch Diagramme für spezielle Aufgaben, z. B. die **Regressionsvariablendiagramme**. Im folgenden Beispiel wird die Regression von GEWICHT auf GROESSE zusammen mit den Randverteilungen der beiden Variablen beschrieben:

[1] Auf der Webseite https://developer.ibm.com/answers/questions/408492/spss-visualization-designer/ hat sich eine SPSS-Mitarbeiterin am 24.10.2017 folgendermaßen zum Visualization Designer geäußert:

> The software Visualization Designer is no longer developed nor sold by IBM. It was a standalone product that was used for creating special chart looks. I do therefore not think you need this software as you can create chartlooks with SPSS Statistics.



Von den zahlreichen in SPSS realisierbaren Diagrammtypen können im Manuskript aus Zeitgründen nur wenige Beispiele behandelt werden. Im aktuellen Kapitel 11 lernen Sie das Streudiagramm inklusive der Optionen zur Nachbearbeitung im Diagrammeditor kennen. Im weiteren Verlauf werden noch einige spezielle Balkendiagramme vorgestellt.

11.1 Streudiagramm anfordern

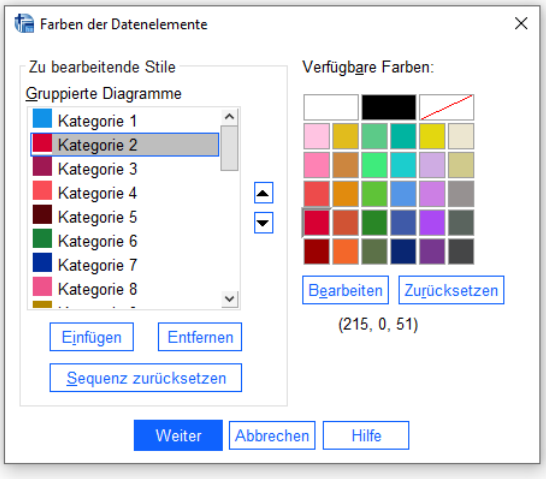
Um die empirische Regression von Gewicht auf Größe und Geschlecht betrachten zu können, fordern wir ein Streudiagramm mit diesen Variablen an. Dies tun wir (mit grundsätzlich identischem Ergebnis) sowohl mit der Dialogbox **Diagrammerstellung** als auch mit einem **klassischen Dialogfeld**.

11.1.1 Voreinstellungen für neue Diagramme modifizieren

Um geeignete Farben für gruppierte Diagramme einzustellen, sollte nach

Bearbeiten > Optionen > Diagramme > Farben

eine Anpassung vorgenommen werden. Im folgenden Dialog wird dafür gesorgt, dass die beiden ersten Kategorien farblich gut zu unterscheiden sind:

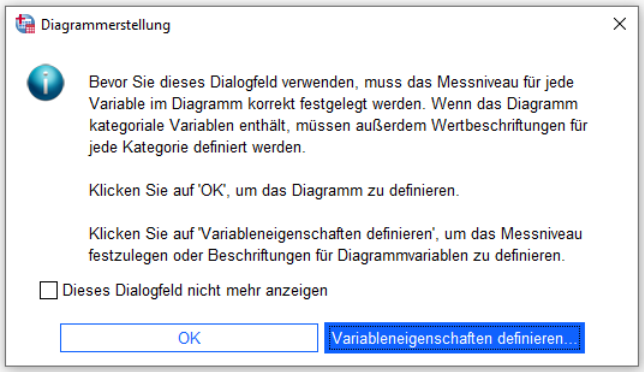


11.1.2 Diagrammerstellung

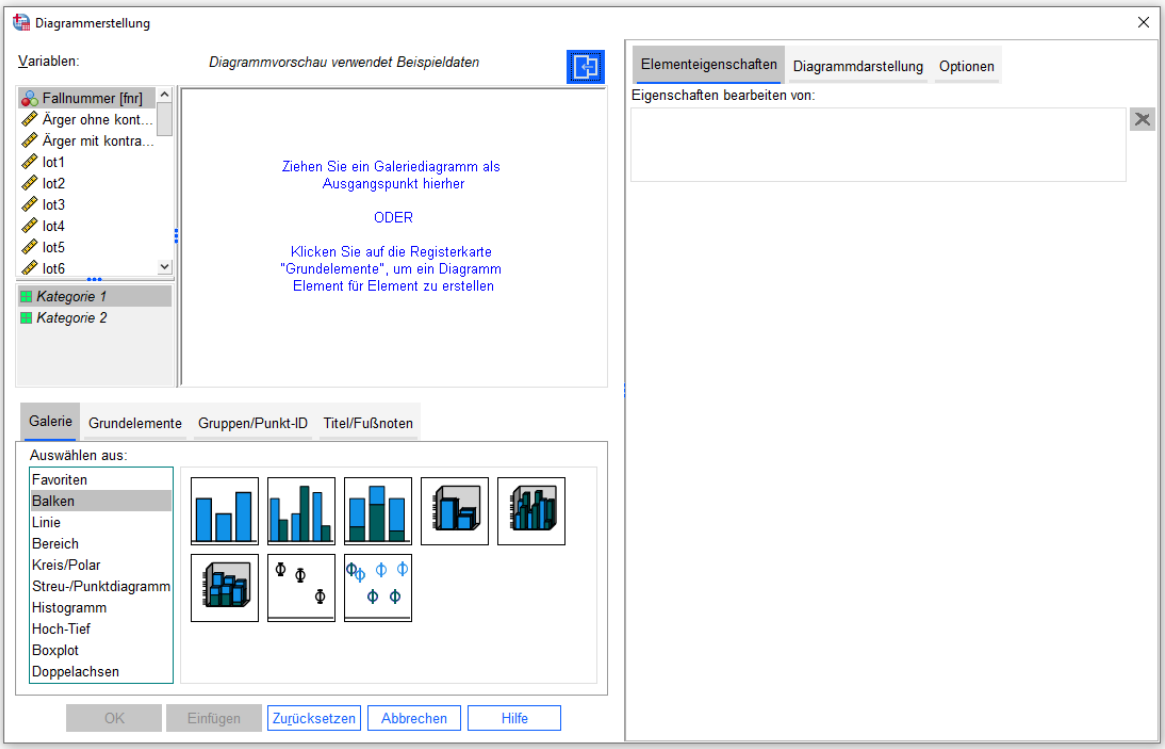
Nach dem Menübefehl

Grafik > Diagrammerstellung

informiert SPSS zunächst darüber, dass bei allen Variablen korrekt deklarierte Messniveaus und bei kategorialen (ordinalen oder nominalen) Variablen außerdem Wertbeschriftungen benötigt werden (zur Deklaration von Variablenattributen siehe Abschnitt 4.2.2):



Das Fenster **Diagrammerstellung**



unterstützt zwei Vorgehensweisen zur Definition eines neuen Diagramms:

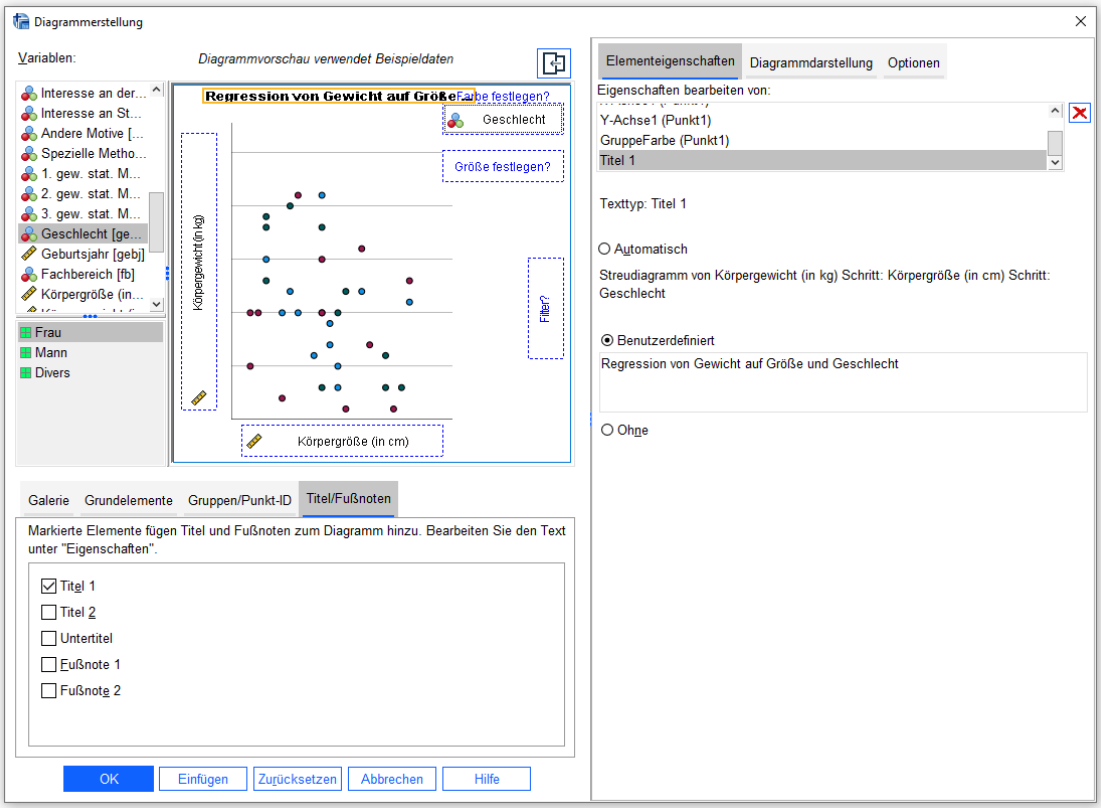
- Diagrammtyp aus der **Galerie** als Ausgangspunkt wählen und individuell gestalten
- Diagramm aus **Grundelementen** (z. B. Achsensystem, Linie) aufbauen

Wir erstellen das geplante Streudiagramm mit Hilfe der **Galerie**:

- Klicken Sie auf die Registerkarte **Galerie**, und wählen Sie den Typ **Streu-/Punkt-diagramm**.
- Befördern Sie das Symbol  zum **Streudiagramm** per Doppelklick oder Drag & Drop auf die **Diagrammvorschau** (Entwurfszone) über den Diagrammtypen.
- In der Diagrammvorschau erscheint ein Achsensystem mit Ablageflächen für
  - eine **X-Achsen**-Variable
  - eine **Y-Achsen**-Variable
  - eine Gruppierungsvariable (Beschriftung: **Farbe festlegen?**)
- Bringen Sie nun die drei Variablen GROESSE, GEWICHT und GESCHL in Position:
  - Ziehen Sie aus der **Variablen**-Liste in der linken oberen Fensterecke die Variable GROESSE auf die X-Achsen-Ablagefläche.
  - Ziehen Sie die Variable GEWICHT auf die Y-Achsen-Ablagefläche.
  - Ziehen Sie die Variable GESCHL auf die Gruppierungs-Ablagefläche mit der Beschriftung **Farbe festlegen**. Weibliche und männliche Datenpunkte werden unterschiedlich dargestellt und man kann ggf. geschlechtsbedingte Unterschiede in der Regression von Gewicht auf Größe erkennen.

  Zur Illustration erscheinen in der Diagrammvorschau künstliche Datenpunkte.
- Legen Sie einen Titel für das Diagramm fest:
  Auf der Registerkarte **Titel/Fußnoten** sollte bereits das Kontrollkästchen **Titel 1** markiert sein. Infolgedessen können Sie auf folgende Weise einen Titel für das Diagramm festlegen: Markieren Sie auf der Registerkarte **Elementeigenschaften** im rechten Teil des Fensters das Element **Titel 1**, wählen Sie die Option **Benutzerdefiniert**, und tragen Sie einen Titel ein.

Nun sollte die Dialogbox **Diagrammerstellung** ungefähr das folgende Bild zeigen:



Nach einem Klick auf den Schalter **OK** wird das Diagramm erstellt. Das Ergebnis ist in Abschnitt 11.2 zu sehen.

Von der Diagrammerstellung erhält man über den Schalter **Einfügen** als Syntax-Äquivalent zu der vorgenommenen Konfiguration ein GGRAPH-Kommando mitsamt dem darin verwendeten Inline-Code in der GPL (*Graphics Programming Language*), z. B.:

```
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=groesse gewicht geschl
   MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE
  /FITLINE TOTAL=NO SUBGROUP=NO.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: groesse=col(source(s), name("groesse"))
  DATA: gewicht=col(source(s), name("gewicht"))
  DATA: geschl=col(source(s), name("geschl"), unit.category())
  GUIDE: axis(dim(1), label("Körpergröße (in cm)"))
  GUIDE: axis(dim(2), label("Körpergewicht (in kg)"))
  GUIDE: legend(aesthetic(aesthetic.color.interior), label("Geschlecht"))
  GUIDE: text.title(label("Regression von Gewicht auf Größe und Geschlecht"))
  SCALE: cat(aesthetic(aesthetic.color.interior), include("1", "2", "3"))
  ELEMENT: point(position(groesse*gewicht), color.interior(geschl))
END GPL.
```

Lässt man das GGRAPH-Kommando zusammen mit dem GPL-Code in einem Syntaxfenster ausführen, wird die per Diagrammerstellung gestaltete Grafik aufgebaut. Durch das direkte Ändern von GGRAPH/GPL - Syntax gewinnt im Vergleich zum Dialog **Diagrammerstellung** zusätzliche Gestaltungsmöglichkeiten.

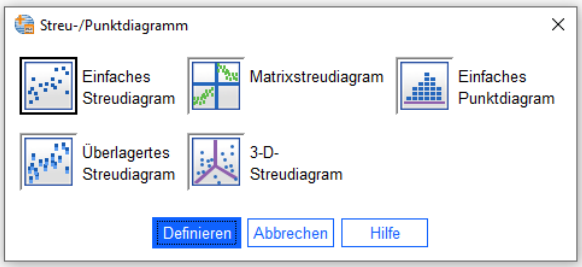
Über die GPL informiert ein im Internet frei verfügbares PDF-Dokument (IBM Corp. 2021c) mit ca. 400 Seiten. Im Kurs kann die GPL aus Zeitgründen leider nicht behandelt werden.

11.1.3 **Dialogbox** Einfaches Streudiagramm

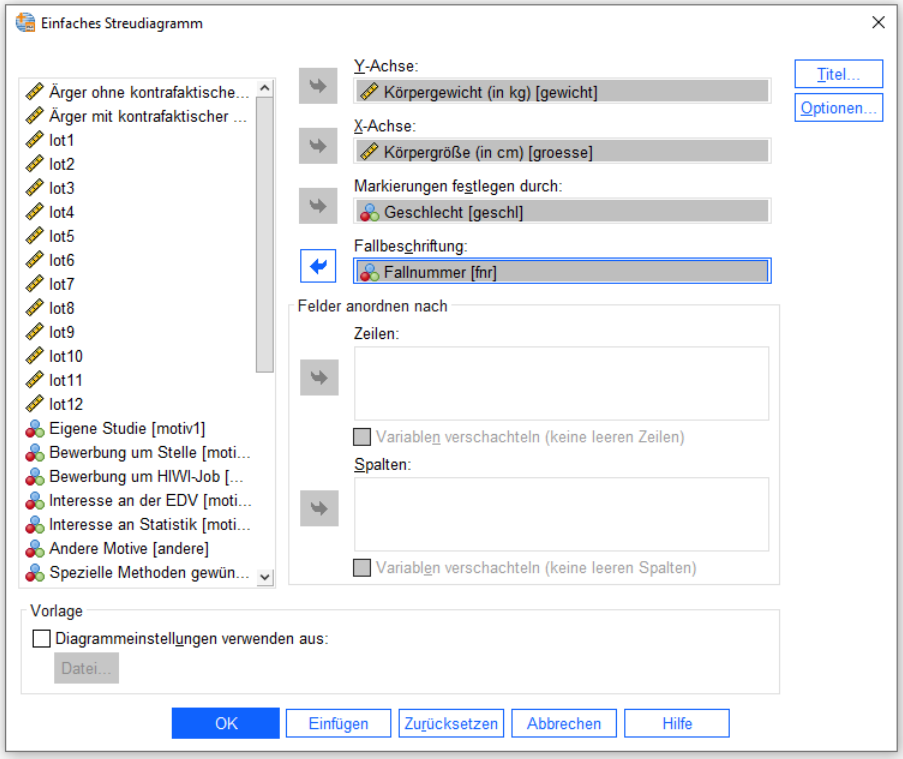
Wer sich mit der **Diagrammerstellung** nicht anfreunden kann, hat in der SPSS-Version 27 auch noch die **klassischen Dialogfelder** zur Verfügung, z. B. zum Erstellen eines Streudiagramms:

Grafik > Klassische Dialogfelder > Streu-/Punktdiagramm

In der nach diesem Menübefehl erscheinenden Palette akzeptieren wir für das Streudiagramm mit Gewicht, Größe und Geschlecht die voreingestellte einfache Variante



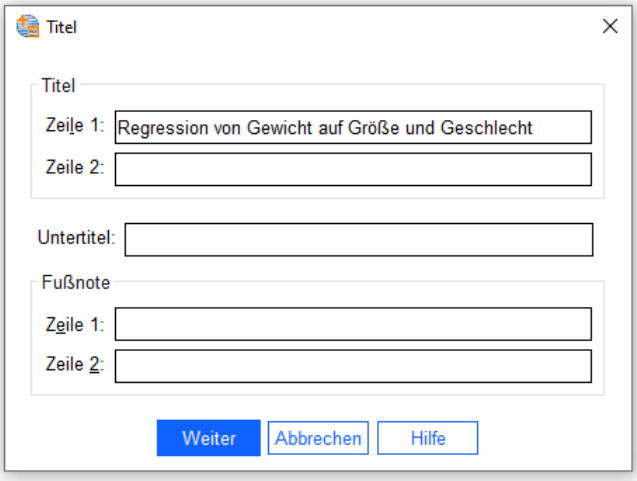
und wechseln per Mausklick auf den Schalter **Definieren** zur Dialogbox **Einfaches Streudiagramm**, wo die beteiligten Variablen per Drag & Drop oder Transportschalter ihre Rollen erhalten:



Durch die Verwendung von GESCHL zur **Markierung** werden weibliche und männliche Datenpunkte verschieden dargestellt, sodass geschlechtsbedingte Unterschiede bei der Regression von Gewicht auf Größe ggf. sichtbar werden.

Die Variable FNR soll später im **Datenbeschriftungsmodus** verwendet werden (siehe Abschnitt 11.2).

Nach einem Mausklick auf den Schalter **Titel** tragen wir eine Titelzeile ein:



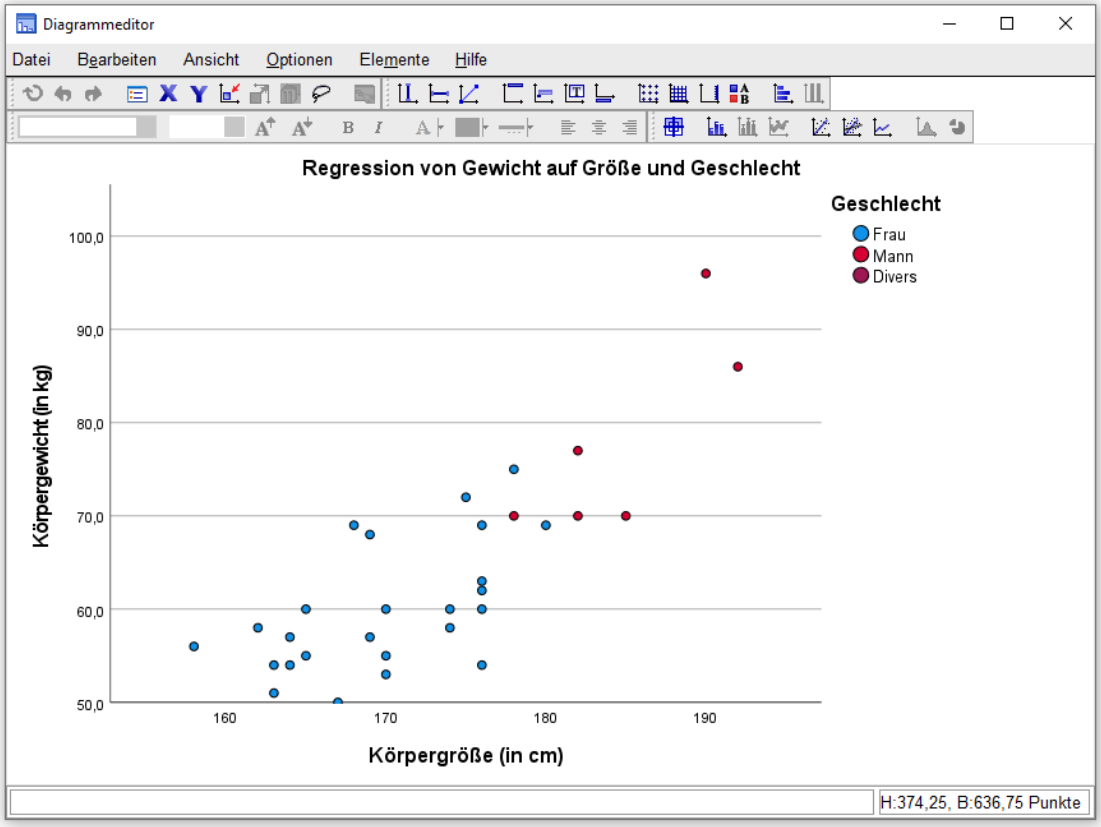
Quittieren Sie die Subdialogbox mit **Weiter** und die Hauptdialogbox mit **OK**, um das Diagramm zu erstellen.

Das Erstellen eines einfachen Streudiagramms fällt mit dem **klassischen Dialogfeld** etwas leichter als mit der **Diagrammerstellung** (siehe Abschnitt 11.1.1). Allerdings lassen sich per Diagrammerstellung mehr Darstellungswünsche realisieren. Zum einen bietet der komplexere

Dialog mehr Optionen, und zum anderen erhält man nach dem Quittieren des Dialogs einen Block mit editierbarer GPL-Syntax zum Diagramm (siehe Abschnitt 11.1.1).

11.2 Streudiagramm per Diagrammeditor modifizieren

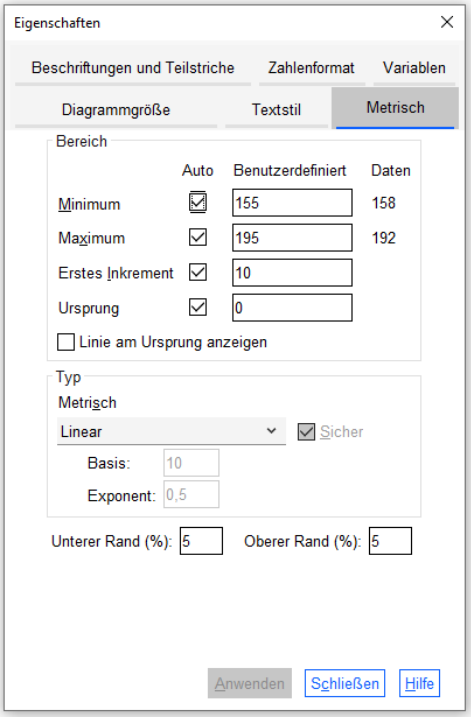
Die per Diagrammerstellung oder über das klassische Dialogfeld erstellten Streudiagramme sind im Wesentlichen identisch. Wir werden anschließend die per Diagrammerstellung entstandene Variante modifizieren. Wenn Sie im Ausgabefenster einen Doppelklick auf ein Diagramm setzen, wird es im **Diagrammeditor** geöffnet, z. B.:



Anschließend werden am Beispiel des Streudiagramms einige allgemeine Bedienungsoptionen des Diagrammeditors vorgestellt. Deren Effekte lassen sich über die Schalter (mehrstufig) rückgängig machen bzw. wiederherstellen.

11.2.1 Eigenschaftenfenster

Zum aktuell im Diagrammeditor markierten Objekt bzw. zur markierten Objektgruppe (erkennbar an einer optischen Hervorhebung) bietet das Eigenschaftenfenster



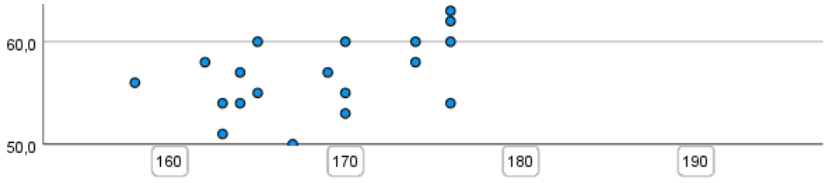
auf jeweils dynamisch zusammengestellten Registerkarten den Zugriff auf die modifizierbaren Attribute. Bei Bedarf kann das Eigenschaftenfenster per Doppelklick auf ein zu gestaltendes Objekt, über den Symbolleistenschalter, mit der Tastenkombination **Strg+T** oder mit den Menübefehl

Bearbeiten > Eigenschaften

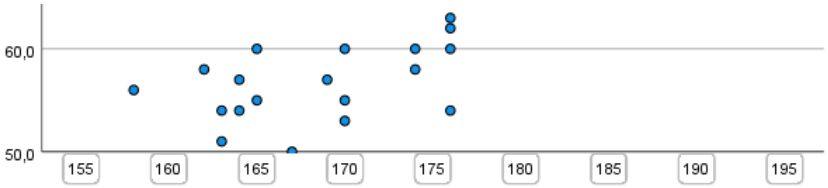
aktiviert werden.

Wer im Beispiel X-Achsenteilstrichwerte im Abstand von 5 cm wünscht, kann so vorgehen:

- X-Achsenteilstrichwerte per Mausklick auf einen Wert markieren



- im Eigenschaftsfenster die Registerkarte **Metrisch** wählen (siehe oben)
- als **erstes Inkrement** den Wert 5 eintragen
- Nach dem **Anwenden** erscheint das Ergebnis:

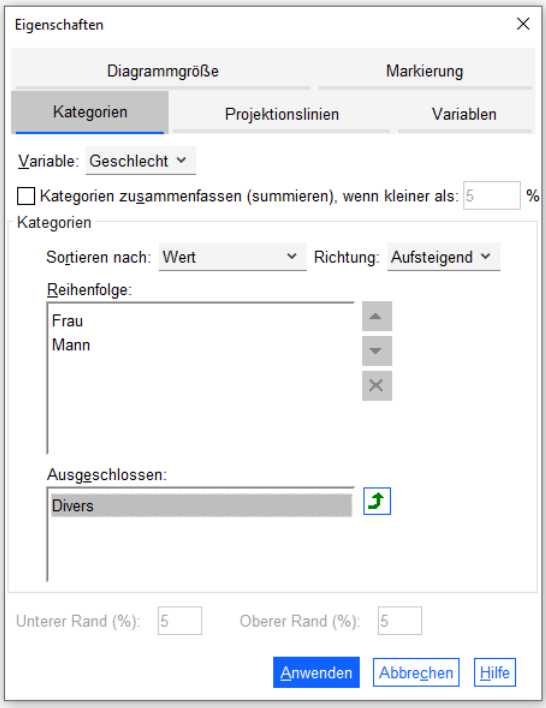


Im Eigenschaftenfenster zur **Y-Achse** sollte auf der Registerkarte **Metrisch** der untere Rand auf 5 gesetzt werden, damit alle Datenpunkte Abstand von der X-Achse halten.

11.2.2 Kategorien verwalten

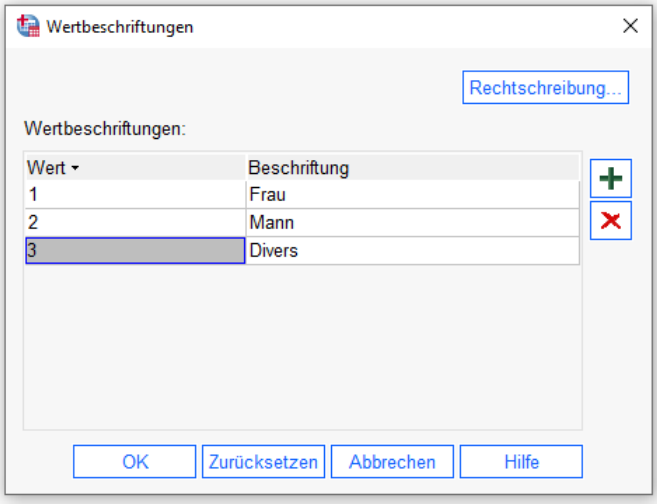
In der Legende taucht die unbesetzte GESCHL-Kategorie *Divers* auf, weil eine entsprechende Wertbeschriftung vorhanden ist. So lässt sich die unbesetzte Kategorie aus der Legende entfernen:

- Per Mausklick auf einen Datenpunkt werden alle Datenpunkte markiert.
- Im **Eigenschaften**-Fenster kann nun auf der Registerkarte **Kategorien** die unbesetzte Kategorie **ausgeschlossen** werden:



- Nach dem **Anwenden** ist die überflüssige Kategorie aus der Legende verschwunden.

Weil die in der Stichprobe nicht aufgetretene Geschlechts-Kategorie *Divers* bei etlichen Diagrammen und Tabellen für vermeidbaren Aufwand sorgt, entscheiden wir uns spontan dafür, die zugehörige Wertbeschriftung per Dateneditor zu löschen (vgl. Abschnitt 4.2.2.6).

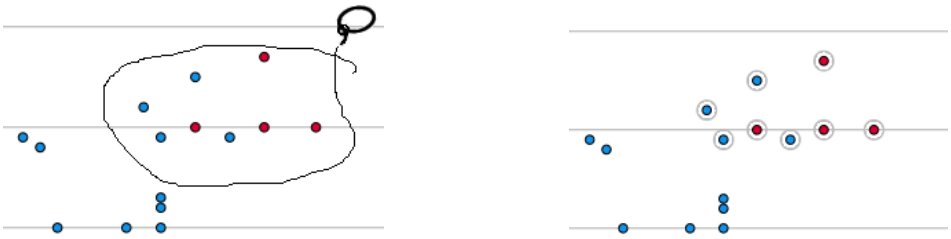


11.2.3 Markieren von gruppierten Objekten

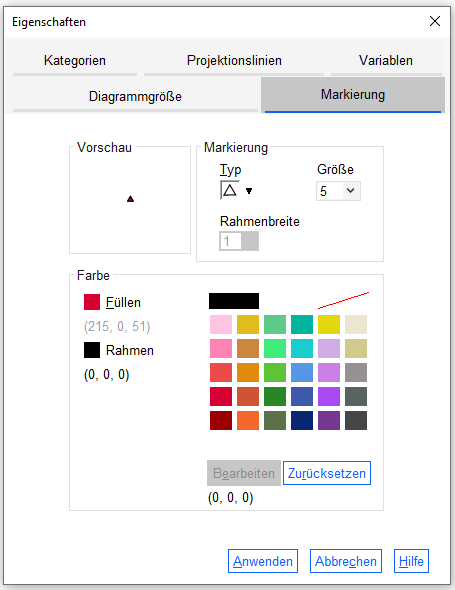
Sind die Objekte eines Typs gruppiert (z. B. die Datenpunkte in unserem Streudiagramm mit einer Gruppenaufteilung nach Geschlecht), dann wendet SPSS beim Markieren folgende Logik an:

- Ist gerade *kein* Objekt markiert, bewirkt ein Mausklick auf ein beliebiges Objekt aus einer beliebigen Gruppe das Markieren aller Objekte des Typs (aus sämtlichen Gruppen).
- Ein weiterer Mausklick auf ein Objekt des Typs schränkt die Markierung auf die betroffene Gruppe ein.
- Um die Komplettmarkierung zu einer anderen Gruppe wandern zu lassen, setzt man einen Mausklick auf ein Objekt dieser Gruppe.
- Eine alternative Möglichkeit zum Markieren aller Elemente einer Gruppe ist der Mausklick auf das zugehörige Symbol in der Legende, z. B.:

- Ist aktuell eine einzelne Gruppe markiert, kann ein einzelnes *Mitglied* dieser Gruppe per Mausklick markiert werden. Alternativ gelingt die Markierung eines einzelnen Datenpunkts über das Item **Auswählen > Diese Markierung** aus seinem Kontextmenü.
- Sobald ein einzelnes Objekt markiert ist, wandert bei weiteren Mausklicks die Einzelmarkierung über Gruppengrenzen hinweg zum getroffenen Objekt.
- Bei gedrückter **Strg**-Taste ist ein gruppenunabhängig kumulierendes Markieren möglich.
- Mit dem Lasso-Werkzeug kann man bei gedrückter linker Maustaste eine Linie um die zu markierenden Objekte (aus beliebigen Gruppen) ziehen, z. B.:



Für die markierten Datenpunkte lässt sich das zugehörige Symbol hinsichtlich Form, Größe, Randfarbe und/oder Füllfarbe ändern. So kann man z. B. für eine bessere Unterscheidbarkeit von Gruppen sorgen:



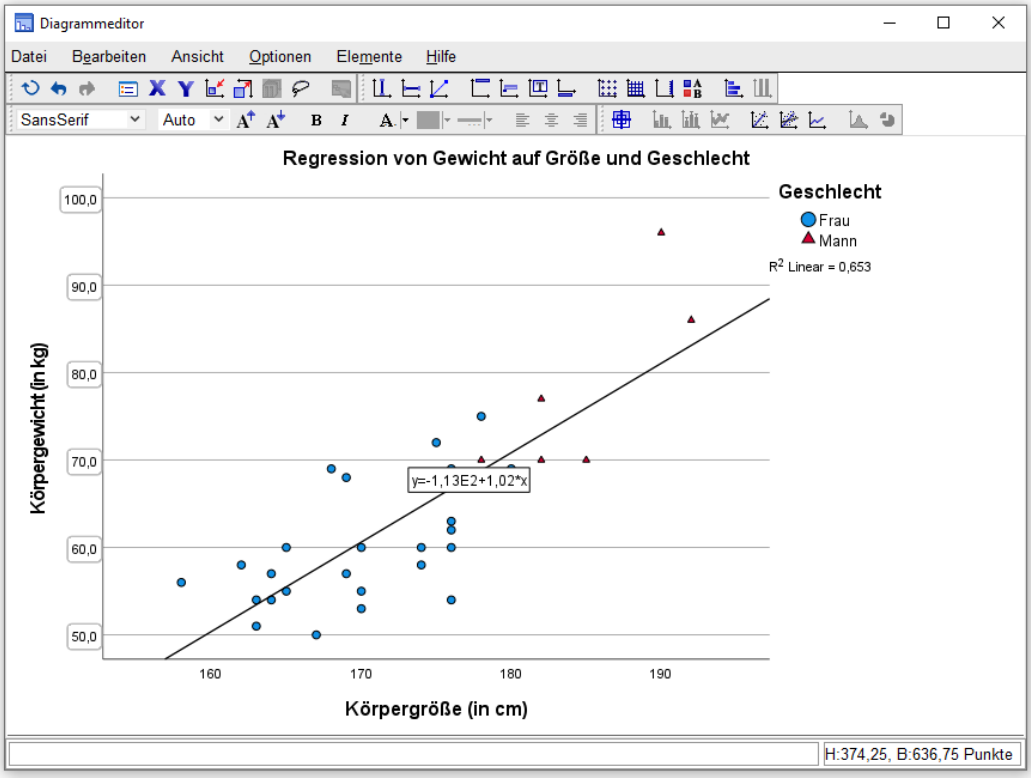
11.2.4 Menüs und Symbolleisten

Viele Gestaltungsmöglichkeiten sind über die Items **Optionen** und **Elemente** im Diagrammeditor-Hauptmenü sowie über äquivalente Symbolleistenschalter verfügbar (z. B. Anpassungs- oder Interpunktionslinien, Datenbeschriftungen, Legende, Anmerkungen). Außerdem ist zu allen Objekten ein Kontextmenü vorhanden.

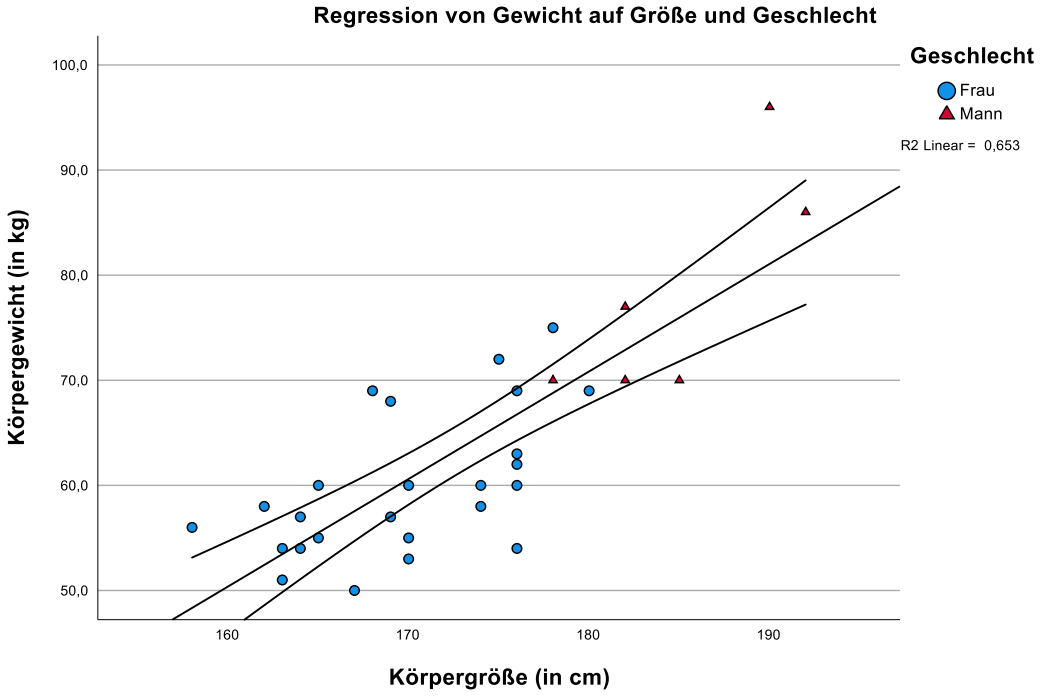
Im Beispiel bietet es sich an, über den Symbolleistenschalter  oder den Menübefehl

Elemente > Anpassungslinie bei Gesamtsumme

die empirische Regressionsgerade einzeichnen zu lassen:

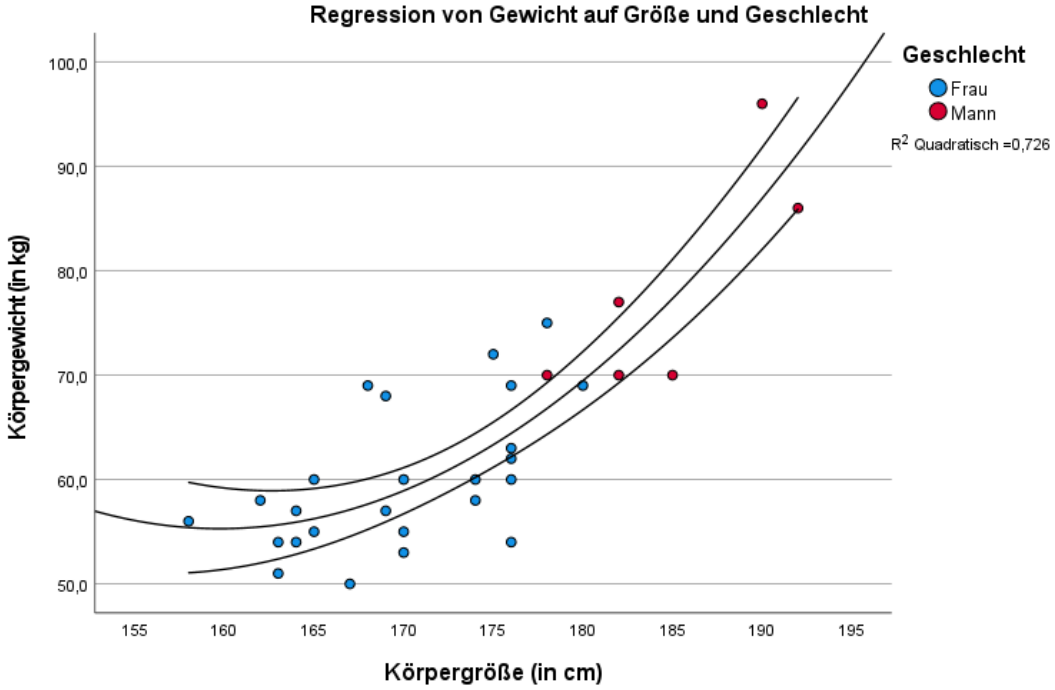


Bei markierter Regressionsgeraden kann man im Eigenschaftsfenster auf der Registerkarte **Anpassungslinie** die von SPSS übereifrig eingetragene **Beschriftung** mit der Regressionsgleichung entfernen und außerdem **Konfidenzintervalle** für die **Mittelwerte** anzeigen lassen:



So wird erkennbar, dass die Unsicherheit bzgl. des für eine bestimmte Größe erwarteten mittleren Gewichts mit der Entfernung vom Mittelwert des Regressors steigt.

Auf derselben Registerkarte lässt sich auch die **Anpassungsmethode** ändern, z. B. von **Linear** auf **Quadratisch**:[1]



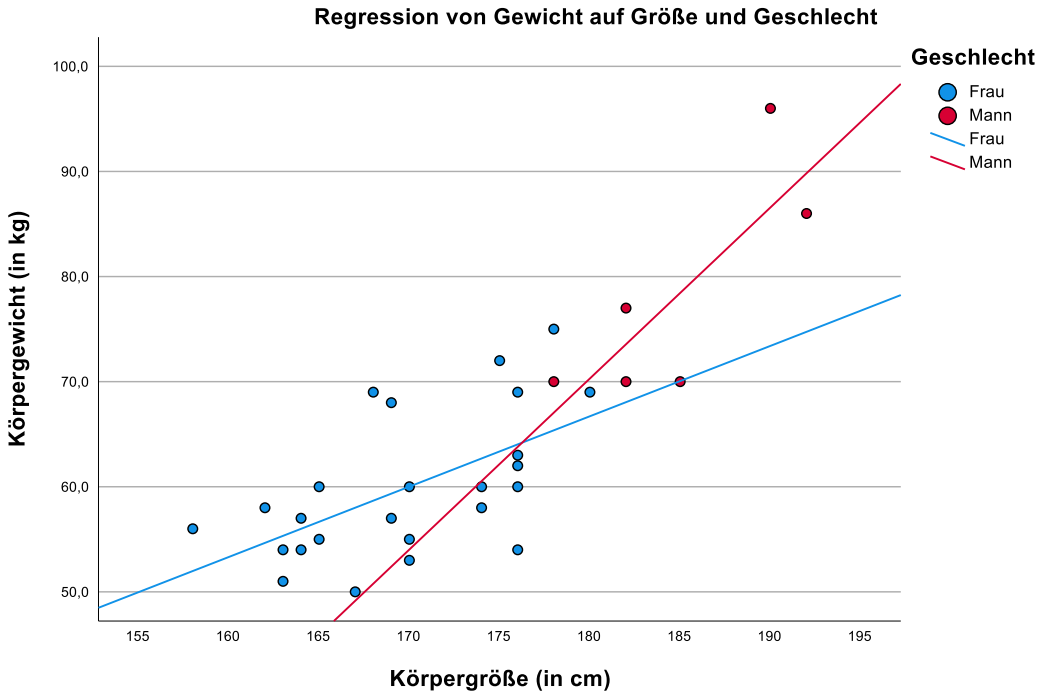
[1] Dieses Diagramm musste im Bitmap-Format in das Textverarbeitungsprogramm übernommen werden, weil nach der Übernahme im EMF-Format reproduzierbar das Exportieren des Manuskripts im PDF-Format scheiterte. Daher ist die Qualität reduziert (vgl. Abschnitt 5.4.4).

Im Beispiel lässt der erhebliche R^2 - Anstieg vermuten, dass die Regression bei kleinen Personen (oder bei Frauen?) flacher verläuft als bei großen Personen (oder bei Männern?).

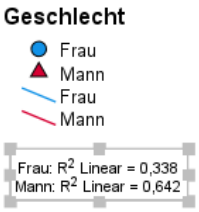
Über das Symbol oder den Menübefehl

Elemente > Anpassungslinie bei Untergruppen

erhält man gruppenspezifische (geschlechtsbedingte) Regressionsgeraden:



Die folgende Anmerkung mit gruppenspezifischen R^2 - Werten



wird nach dem Zwischenablagentransfer in ein Textverarbeitungsprogramm im EMF-Format falsch dargestellt:

Frau: R2 Linear = 0,338
Mann: R2 Linear = 0,642

Daher wurde die Anmerkung folgendermaßen gelöscht:

- Kontextmenü öffnen
- **Löschen**

Man erkennt in der Grafik einen Geschlechtsunterschied hinsichtlich der Regressionssteigung, der eventuell durch Unterschiede im Körperbau zu erklären ist:

Bei zwei Männern mit 10 cm Größenunterschied besteht möglicherweise ein stärkerer Gewichtsunterschied als bei zwei Frauen mit derselben Größendifferenz. Nach dieser Vermutung *moderiert* Geschlecht den Effekt von Größe auf Gewicht.

Allerdings kennen wir schon eine Alternative zu dieser Deutung:

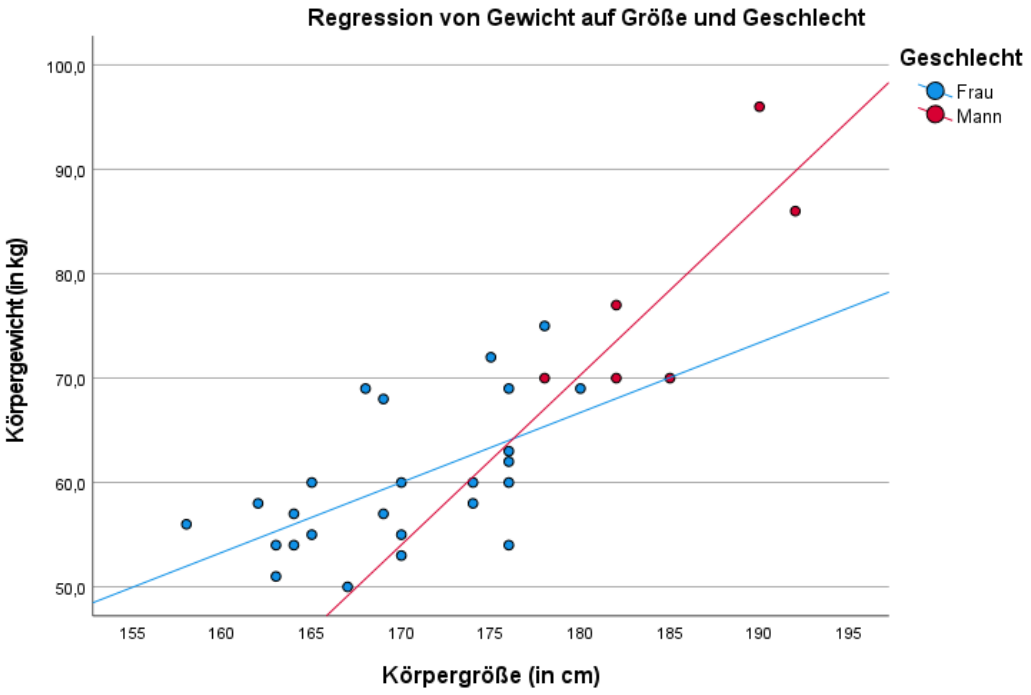
- Eventuell folgt die Regression von Gewicht auf Größe unabhängig vom Geschlecht einem quadratischen Modell (siehe oben),
- und das Geschlecht hat einen Effekt auf die Größe (also einen indirekten Effekt auf das Gewicht).

Über die Analyse von Moderatoreffekten und von indirekten Effekten über die SPSS-Regressions-Prozedur und das SPSS-Makro PROCESS von Andrew Hayes informiert ein ZIMK-Manuskript (Baltes-Götz 2020a), das auf dem Webserver der Universität Trier zu finden ist:

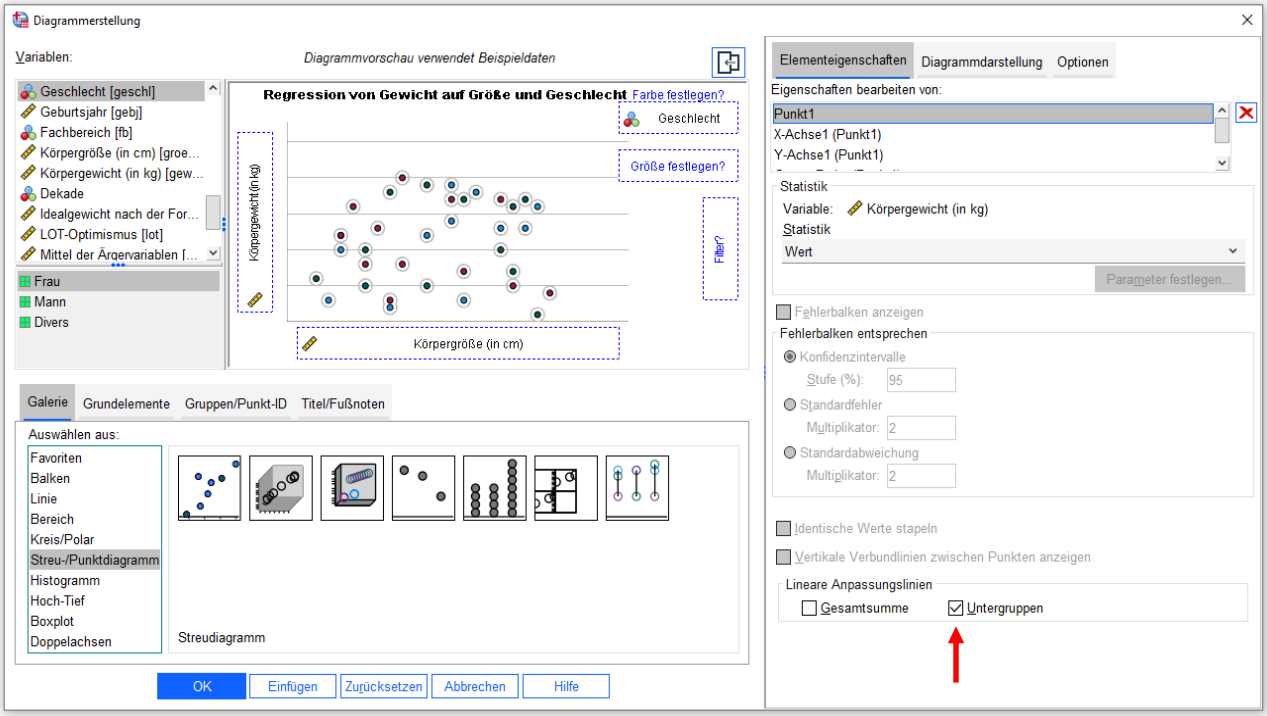
https://www.uni-trier.de/?id=22528

Um separate Legenden für die Markierungen und die Linientypen

zu vermeiden und das folgende Ergebnis zu erzielen,



muss man die gruppen-spezifischen Regressionsgeraden schon im Dialog der **Diagrammerstellung** anfordern:



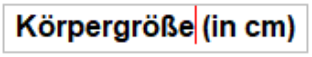
11.2.5 Beschriftungen

Bei manchen Beschriftungen in einem SPSS-Diagramm lässt sich nach dem Markieren die Größe des umgebenden Rechtecks über acht Anfasser ändern:

Solche Beschriftungen lassen sich auch verschieben, wobei die Transportfunktionalität des Mauszeigers am Rand aktiv wird, signalisiert durch die Zeigergestalt .

Unbewegliche Beschriftungen besitzen einen Markierungsrahmen *ohne* Anfasser, z. B.:

Viele Beschriftungen lassen sich ändern. Um in den Eingabemodus zu gelangen, muss man auf das bereits markierte Textfeld einen weiteren Mausklick setzen und erhält dann bei veränderlichen Texten eine Schreibmarke, z. B.:



Zum Beenden der Texteingabe drückt man die **Enter**-Taste oder setzt einen Mausklick außerhalb des Markierungsrahmens.

Über die Schaltfläche (de)aktiviert man das Werkzeug zur Datenbeschriftung, das zu angeklickten Datenpunkten den Wert der vereinbarten Fallbeschriftungsvariablen oder aber die laufende Datenblattzeilennummer in die Grafik einfügt bzw. wieder entfernt, z. B.:

Nach einem rechten Mausklick auf einen Datenpunkt mit dem Fallbeschriftungswerkzeug kann man per Kontextmenü veranlassen, dass die zugehörige Zeile im Datenfenster markiert wird.

11.3 Diagramme verwenden

Wie Tabellen lassen sich auch Diagramme aus dem Ausgabefenster über die Windows-Zwischenablage in andere Anwendungen übertragen (siehe Abschnitt 5.4.4 zu Beeinflussung der Diagrammformate).

Als Ausgabefensterbestandteile lassen sich Diagramme sichern, drucken oder exportieren.

Zur Verwendung als **Vorlage** kann man ein Diagramm aus dem Diagrammeditor mit dem Menübefehl

Datei > Diagrammvorlage speichern

in eine Datei mit der Namenserweiterung **sgt** sichern. Es bestehen etliche Möglichkeiten, eine Vorlage auf andere Diagramme anzuwenden:

- per Diagrammerstellung auf der Registerkarte **Diagrammerstellung** über
  Vorlage > Diagrammvorlagendatei verwenden
- beim Erstellen eines Diagramms über ein **klassisches Dialogfeld** (siehe Dialogbox **Einfaches Streudiagramm** in Abschnitt 11.1.3)
- im Diagrammeditor über
  Datei > Diagrammvorlage zuweisen
- im SPSS-Optionendialog als Standardvorlage für alle Diagramme über
  Bearbeiten > Optionen > Diagramme > Diagrammvorlage

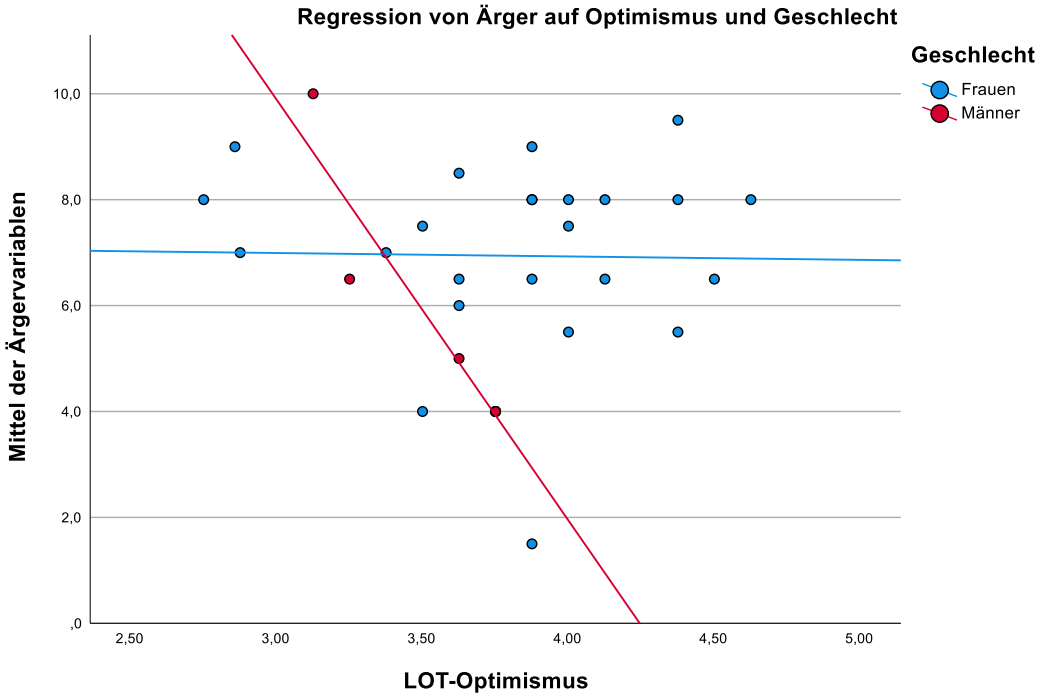
Natürlich ist die Verwendung einer Vorlage auch beim Erstellen eines neuen Diagramms per Syntax möglich, z. B.:

```
GRAPH
  /SCATTERPLOT(BIVAR)=groesse WITH gewicht
  /MISSING=LISTWISE
  /TEMPLATE='C:\Program Files\IBM\SPSS Statistics\28\Looks\APA_Styles.sgt'.
```

Der Installationsordner von SPSS 28 (z. B. **C:\Program Files\IBM\SPSS Statistics\28**) enthält im Unterordner **Looks** einige Diagrammvorlagen, z. B. die Vorlage **APA_Styles.sgt** für Diagramme nach den Richtlinien der **American Psychological Association** (APA).

11.4 Übung

Nach dem „Scheitern“ der differentialpsychologischen Hypothese (siehe Abschnitt 10.1) wird man versuchen, aus den Daten Hinweise für eine mögliche Verbesserung der Hypothese zu gewinnen. Erzeugen Sie ein Streudiagramm mit den Variablen AERGAM und LOT, und verwenden Sie wie in obigem Beispiel GESCHL als Markierungsvariable. Die Manuskriptstichprobe liefert mit eingezeichneten Regressionsgeraden für die Untergruppen das folgende Ergebnis:



Während bei den Frauen offenbar *kein* Zusammenhang zwischen LOT und AERGAM besteht, zeigt sich bei den Männern ein Effekt im Sinne der differentialpsychologischen Hypothese. Allerdings sollten wir die Beobachtung sehr zurückhaltend interpretieren, weil unsere Stichprobe lediglich sechs Männer enthält. Immerhin resultiert bei einer regressionsanalytischen Auswertung für den Moderatoreffekt (siehe Baltes-Götz 2020a) eine relativ kleine Überschreitungswahrscheinlichkeit (0,01):

Koeffizienten [a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten			95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Standardfehler	Beta	T	Sig.	Untergrenze	Obergrenze
1	(Konstante)	-19,356	11,285		-1,715	,098	-42,512	3,799
	LOT-Optimismus	7,818	3,121	1,863	2,505	,019	1,414	14,222
	Geschlecht	26,543	10,211	5,426	2,600	,015	5,592	47,494
	Geschlecht * LOT	-7,883	2,860	-5,633	-2,756	,010	-13,751	-2,015

a. Abhängige Variable: Mittel der Ärgervariablen

Hier haben wir es aber **nicht** mit dem signifikanten Ergebnis eines statistischen Tests zu tun, sondern mit einem deskriptiven Maß zu einer interessanten Vermutung, die sich bei der explorativen Datenanalyse ergeben hat. Eine Testentscheidung über die Moderatorhypothese ist nur in einer unabhängigen Stichprobe möglich (siehe Abschnitt 17.3).

12 T-Test für unabhängige Stichproben

In diesem Abschnitt interessieren wir uns für Geschlechtsunterschiede beim Body Mass Index und führen mit unseren Variablen GESCHL und BMI einen t-Test für unabhängige Stichproben zum folgenden Hypothesenpaar durch:

H_0: Bei Frauen ist der BMI-Mittelwert mindestens genauso groß wie bei Männern.

versus

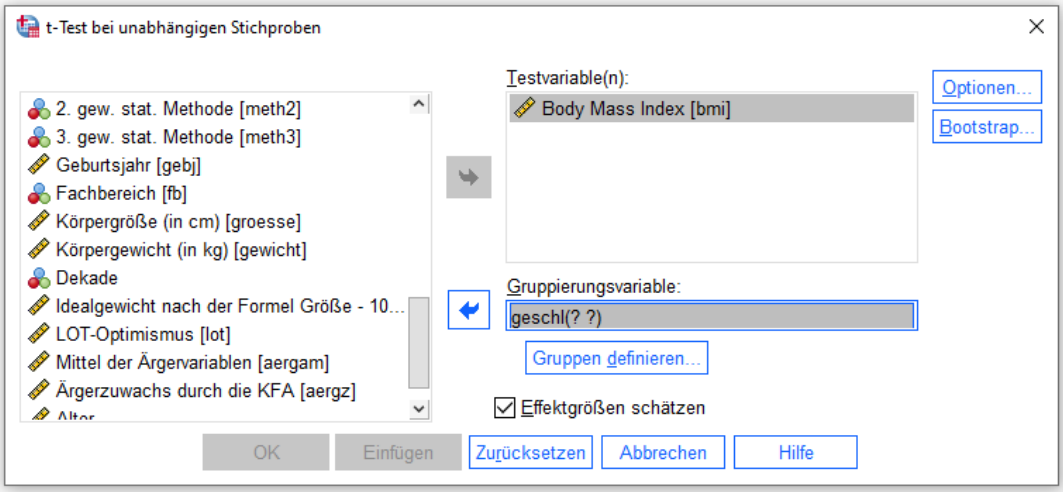
H_1: Bei Frauen ist der BMI-Mittelwert niedriger als bei Männern.

12.1 T-Test anfordern

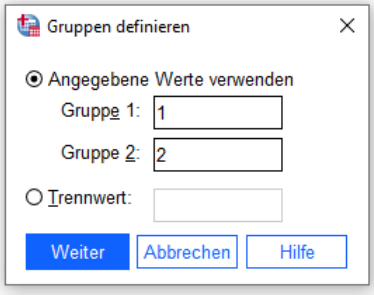
Fordern Sie mit dem folgenden Menübefehl die zugständige Dialogbox an:

Analysieren > Mittelwerte vergleichen > t-Test bei unabhängigen Stichproben

Transportieren Sie den BMI in die Liste der **Testvariable(n)**, und verwenden Sie Geschlecht als **Gruppenvariable**:



Über den Schalter **Gruppen definieren** erreicht man die folgende Dialogbox, um die beiden zu vergleichenden Gruppen über ihre Werte bei der Gruppenvariablen festzulegen:



12.2 Interpretation

Wir erhalten die folgenden deskriptiven Statistiken:

Gruppenstatistiken

	Geschlecht	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Body Mass Index	Frau	25	20,7488	1,89347	,37869
	Mann	6	22,8078	2,17495	,88792

Bei den Männern fällt der BMI-Mittelwert im H_1 – Sinn um ca. 2 Punkte höher aus.

Zunächst ist die Frage zu klären, welche der beiden in der folgenden Tabelle angebotenen t-Test – Varianten (*mit* bzw. *ohne* Voraussetzung der Varianzhomogenität) zu verwenden ist:

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		t-Test für die Mittelwertgleichheit						
						Signifikanz				95% Konfidenzintervall der Differenz
		F	Sig.	T	df	Einseitiges p	Zweiseitiges p	Mittlere Differenz	Differenz für Standardfehler	Unterer Wert	Oberer Wert
Body Mass Index	Varianzen sind gleich	,000	,989	-3,674	18	<,001	,002	-3,24410	,88296	-5,09913	-1,38906
	Varianzen sind nicht gleich			-3,499	4,383	,011	,021	-3,24410	,92728	-5,73232	-,75588

Als Entscheidungshilfe berechnet SPSS den **Levene-Test der Varianzhomogenität**, der in unserem Fall durch eine empirische Überschreitungswahrscheinlichkeit von 0,94 (> 0,05) seine Nullhypothese gleicher Varianzen klar akzeptiert. Der somit verwendbare klassische t-Test *mit* vorausgesetzter Varianzhomogenität ermittelt eine Überschreitungswahrscheinlichkeit unterhalb der kritischen Grenze von 0,05. Damit ist die Nullhypothese zu verwerfen, sofern die Voraussetzungen des Tests hinreichend erfüllt sind.

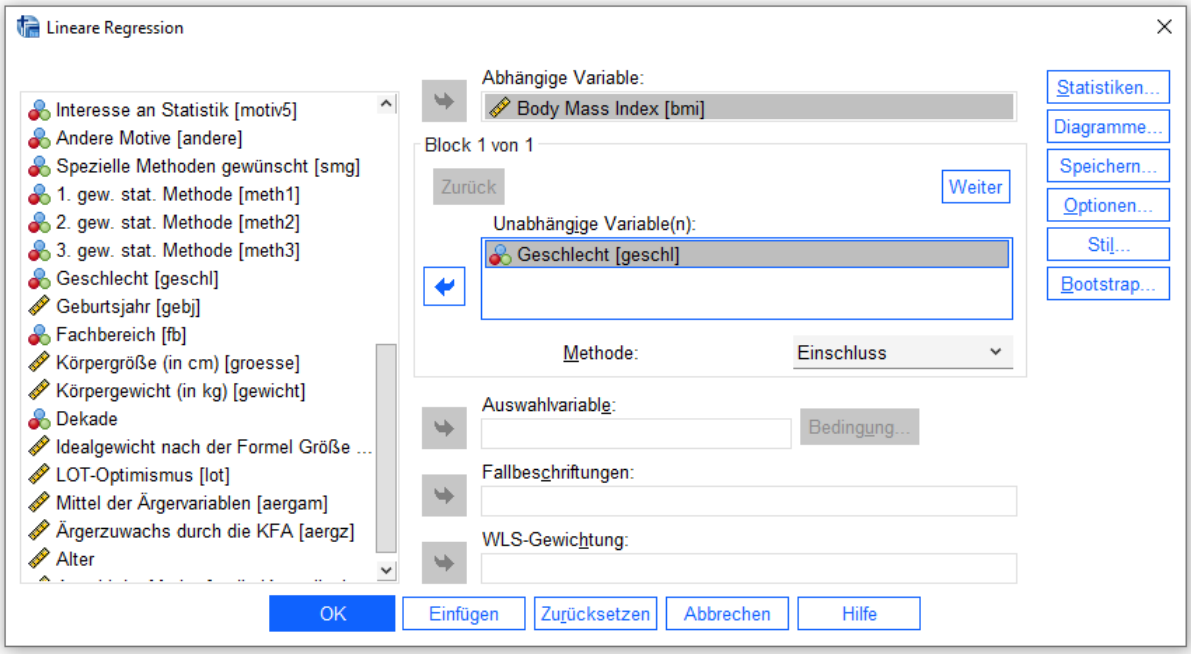
12.3 Prüfung der Voraussetzungen

Die in Abschnitt 8.2 diskutierten Voraussetzungen der linearen Regression gelten auch bei Verwendung von kategorialen Regressoren. Beim t-Test für unabhängige Stichproben kommt ein *dichotomer* Regressor zum Einsatz, sodass die Linearität auf jeden Fall erfüllt ist. Nachdem die Varianzhomogenität der Residuen geklärt ist, und deren Unabhängigkeit aufgrund des querschnittlichen Designs angenommen werden darf, bleibt von den Voraussetzungen der Analyse noch die Normalität der Residuen zu untersuchen.

Um die Verteilung der Residuen mit geringem technischem Aufwand per Histogramm beurteilen zu können, führen wir den t-Test für unabhängige Stichproben mit der Prozedur für die lineare Regression erneut durch. Diese Prozedur beherrscht als Spezialfall auch den klassischen t-Test (mit angenommener Varianzhomogenität) und bietet generell die Ausgabe eines Histogramms für die standardisierten Residuen an. Nach dem Menübefehl

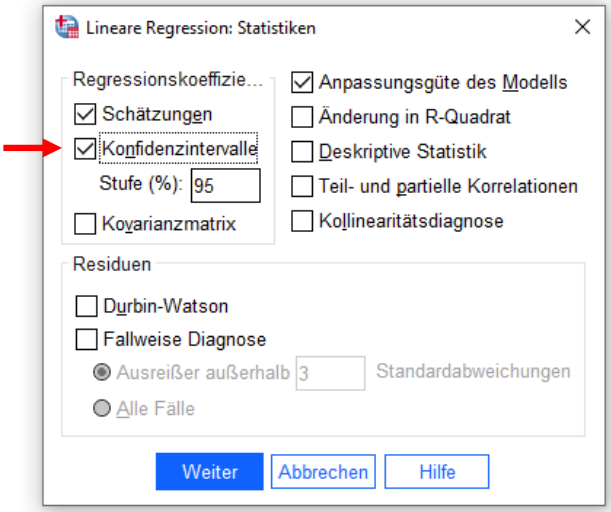
Analysieren > Regression > Linear

wählen wir die **abhängige Variable** BMI und die **unabhängige Variable** GESCHL:

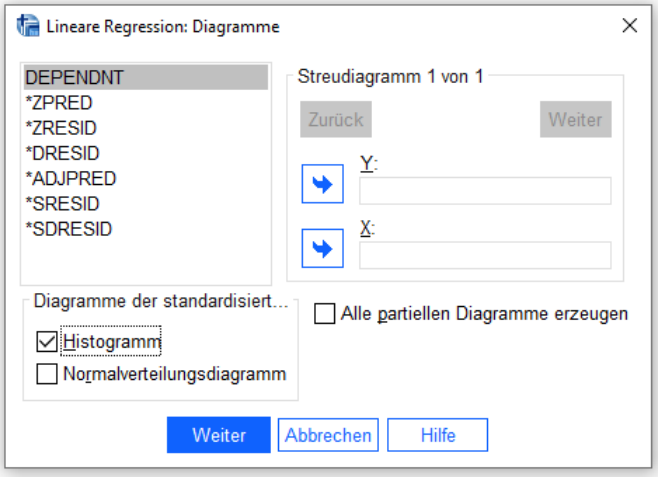


Hätten wir eine Zeichenfolgenvariable für das nominalskalierte Merkmal Geschlecht verwendet, wäre übrigens die Verwendung als unabhängige Variable in der Regressionsprozedur gescheitert (vgl. Abschnitt 2.4.3.1).

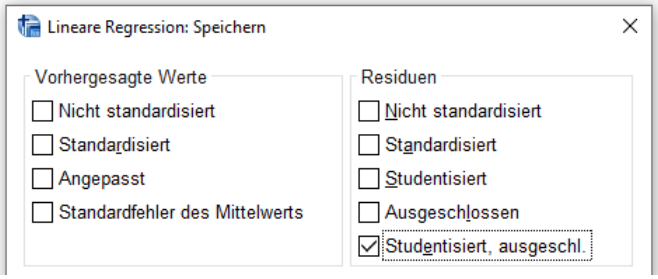
In der **Statistiken** - Subdialogbox verlangen wir die Berechnung von **Konfidenzintervallen** zu den Regressionskoeffizienten:



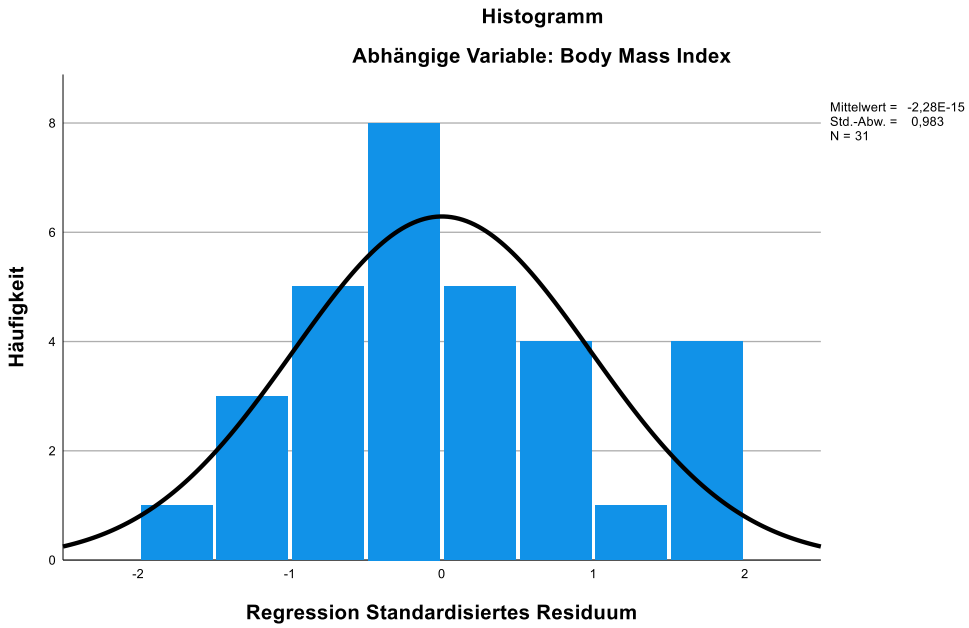
In der Subdialogbox **Diagramme** fordern wir ein **Histogramm** für die standardisierten Residuen an:



Im Subdialog **Speichern** sorgen wir dafür, dass SPSS eine neue Variable mit den ausgelassen-studentisierten Residuen in die Arbeitsdatei schreibt:



Das resultierende Histogramm gibt wenig Anlass zur Sorge bzgl. der Normalverteilungsannahme:



Die über eine explorative Datenanalyse (siehe Abschnitt 9.2) für die abgespeicherten ausgelassen-studentisierten Residuen durchgeführten Signifikanztests zur Normalitäts-Nullhypothese bestätigen den visuellen Eindruck:[1]

Tests auf Normalverteilung

	Kolmogorov-Smirnov [a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
Studentized Deleted Residual	,132	31	,184	,965	31	,383

a. Signifikanzkorrektur nach Lilliefors

In der Koeffiziententabelle der linearen Regression findet sich erwartungsgemäß das t-Test - Ergebnis wieder, dessen Interpretierbarkeit mittlerweile bestätigt ist:

Koeffizienten [a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten			95,0% Konfidenzintervalle für B	
		Regressions-koeffizientB	Std.-Fehler	Beta	T	Sig.	Untergrenze	Obergrenze
1	(Konstante)	18,690	1,112		16,813	,000	16,416	20,963
	Geschlecht	2,059	,884	,397	2,329	,027	,251	3,867

a. Abhängige Variable: Body Mass Index

Der Regressionskoeffizient von 2,06 zum Geschlecht ist als Gruppenunterschied zu interpretieren, und sein Konfidenzintervall [0,251; 3,867] wurde (vom irrelevanten Vorzeichen abgesehen) von der t-Test - Prozedur identisch berechnet.

Durch die Unterschiedlichkeit der beiden Teilstichprobenumfänge wird übrigens *keine* Voraussetzung des linearen Modells verletzt. Man sollte bei der Untersuchungs*planung* nach Möglichkeit für gleich große Teilstichproben sorgen, damit der t-Test seine optimale Power bei fixem Gesamtstichprobenumfang erreicht. Bei einer vorhandenen Stichprobe ist es für solche Optimierungsbemühungen zu spät.

12.4 Empirische Effektstärke

Das Ergebnis eines Signifikanztests hängt stark vom Stichprobenumfang ab und sollte daher möglichst durch eine Beurteilung der Effektstärke ergänzt werden, damit ein Ergebnis auf praktische Bedeutsamkeit beurteilt werden kann. Beim t-Test für unabhängige Stichproben ist die Effektstärke d auf Populationsebene unter der Annahme identischer Standardabweichungen folgendermaßen definiert:

$$d := \frac{\mu_1 - \mu_2}{\sigma}$$

Die Differenz der beiden Erwartungswerte wird zum Zweck der Normierung durch die gemeinsame Standardabweichung dividiert.

[1] Das Studentisieren ist auch bei einem dichotom-kategorialen Regressor sinnvoll. Es berücksichtigt bei der Standardisierung der Residuen die Hebelwirkungen der Fälle auf die Schätzergebnisse, die bei Fällen aus der schwächer besetzten Kategorie größer ist.

Wir beobachten eine starke Verwandtschaft zum Effektstärkenbegriff d_z beim t-Test für abhängige Stichproben (vgl. Abschnitt 2.3.2.4). Nach dem Vorschlag von Cohen (1988, S. 40) gelten außerdem zur Beurteilung der Effektstärke für beide t-Test - Varianten identische Richtwerte:

- kleiner Effekt: $d = 0{,}2$
- mittlerer Effekt: $d = 0{,}5$
- großer Effekt: $d = 0{,}8$

Zur Schätzung der empirischen Effektstärke eignet sich bei homogenen Varianzen die folgende Formel mit der gemittelten („gepoolten“) Standardabweichung im Nenner:

$$\hat{d} := \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}}}$$

Für unser Beispiel erhalten wir einen starken Effekt:

$$\hat{d} = \frac{22{,}8078 - 20{,}7488}{\sqrt{\frac{(6-1)4{,}7304 + (25-1)3{,}5852}{6 + 25 - 2}}} = \frac{2{,}059}{1{,}9449} = 1{,}05866104$$

Seit der Version 27 berechnet SPSS beim t-Test für unabhängige Stichproben die Effektstärke und liefert auch gleich ein Konfidenzintervall mit:

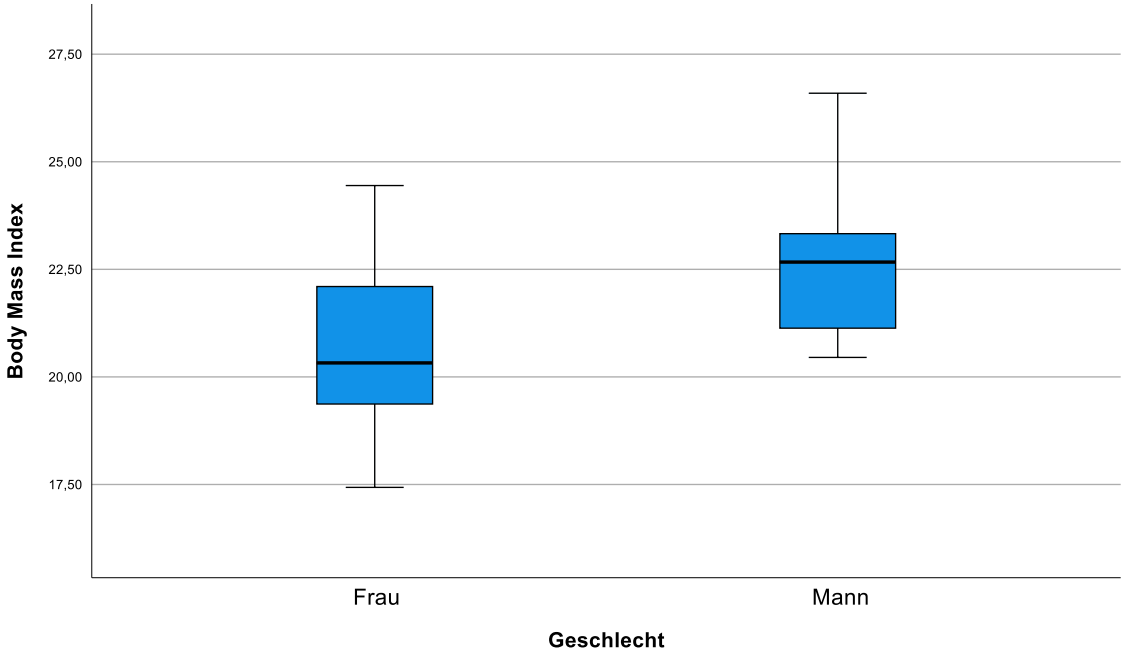
Effektgrößen bei unabhängigen Stichproben

				95% Konfidenzintervall	
		Standardisierer [a]	Punktschätzung	Unterer Wert	Oberer Wert
Body Mass Index	Cohen's d	1,94491	-1,059	-1,982	-,119
	Hedges' Korrektur	1,99708	-1,031	-1,930	-,116
	Glass' Delta	2,17495	-,947	-1,973	,143

a. Der bei der Schätzung der Effektgrößen verwendete Nenner.
Cohen's d verwendet die zusammengefasste Standardabweichung.
Hedges' Korrektur verwendet die zusammengefasste Standardabweichung und einen Korrekturfaktor.
Glass' Delta verwendet die Standardabweichung einer Stichprobe von der Kontrollgruppe.

12.5 Grafische Veranschaulichung

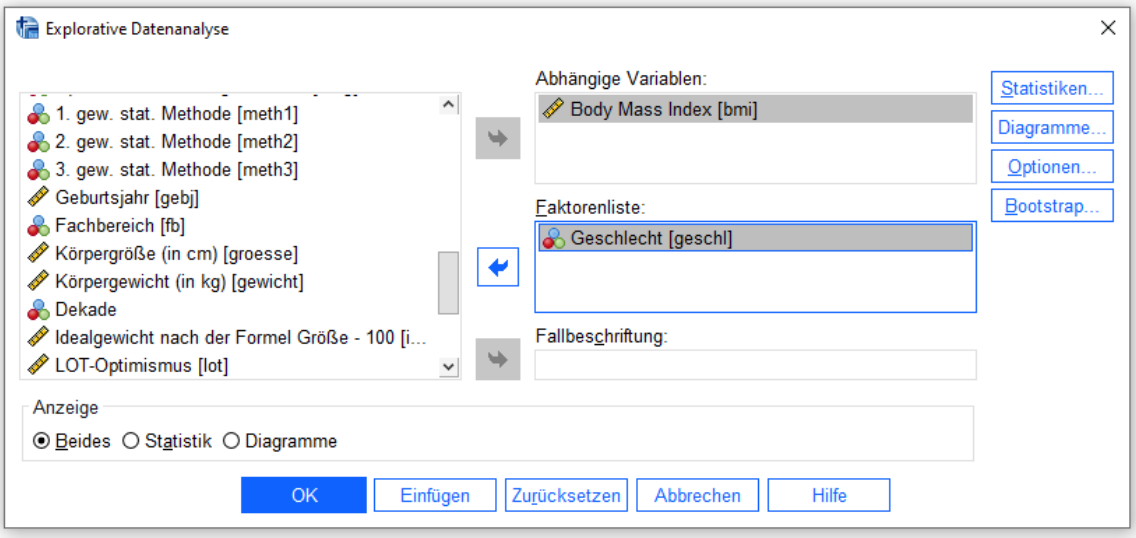
Zur grafischen Veranschaulichung des Geschlechtsunterschieds beim BMI eignet sich z. B. das folgende Paar von gruppenspezifischen Box-Plots:



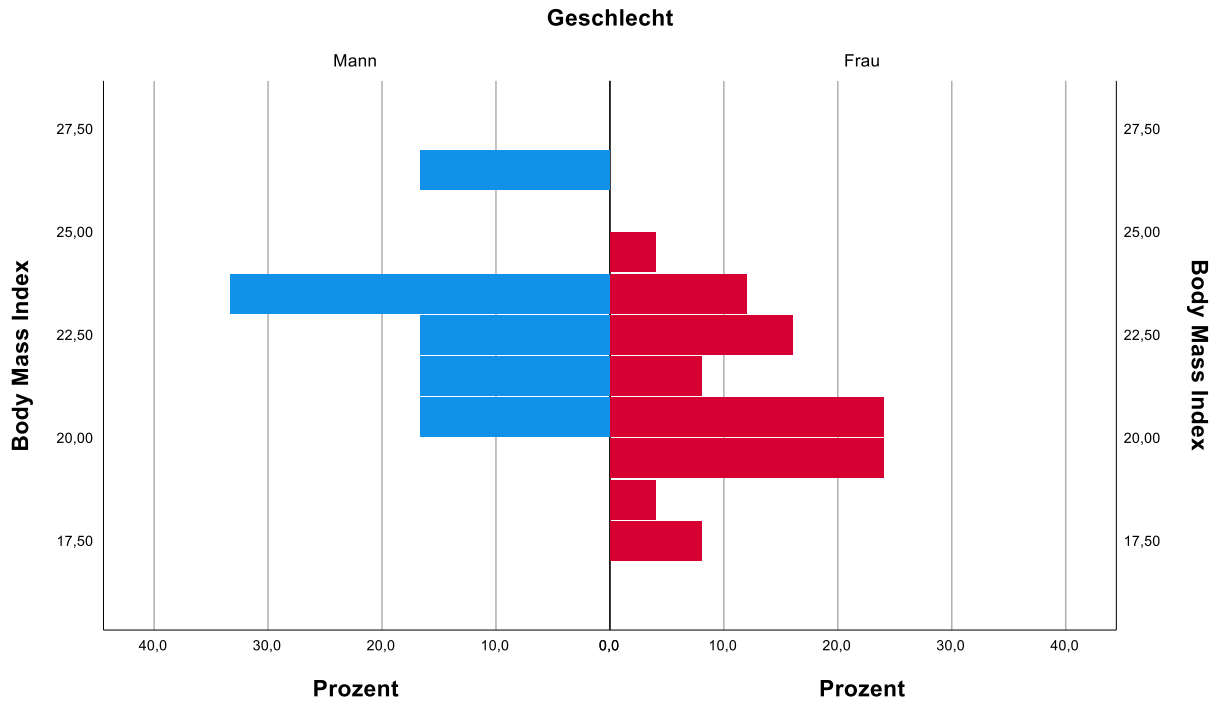
Diese Abbildung ist mit geringem Aufwand über die SPSS-Prozedur zur explorativen Datenanalyse zu erstellen. Öffnen Sie deren Dialogbox mit

Analysieren > Deskriptive Statistiken > Explorative Datenanalyse

und beziehen Sie neben der abhängigen Variablen BMI den Faktor GESCHL ein:



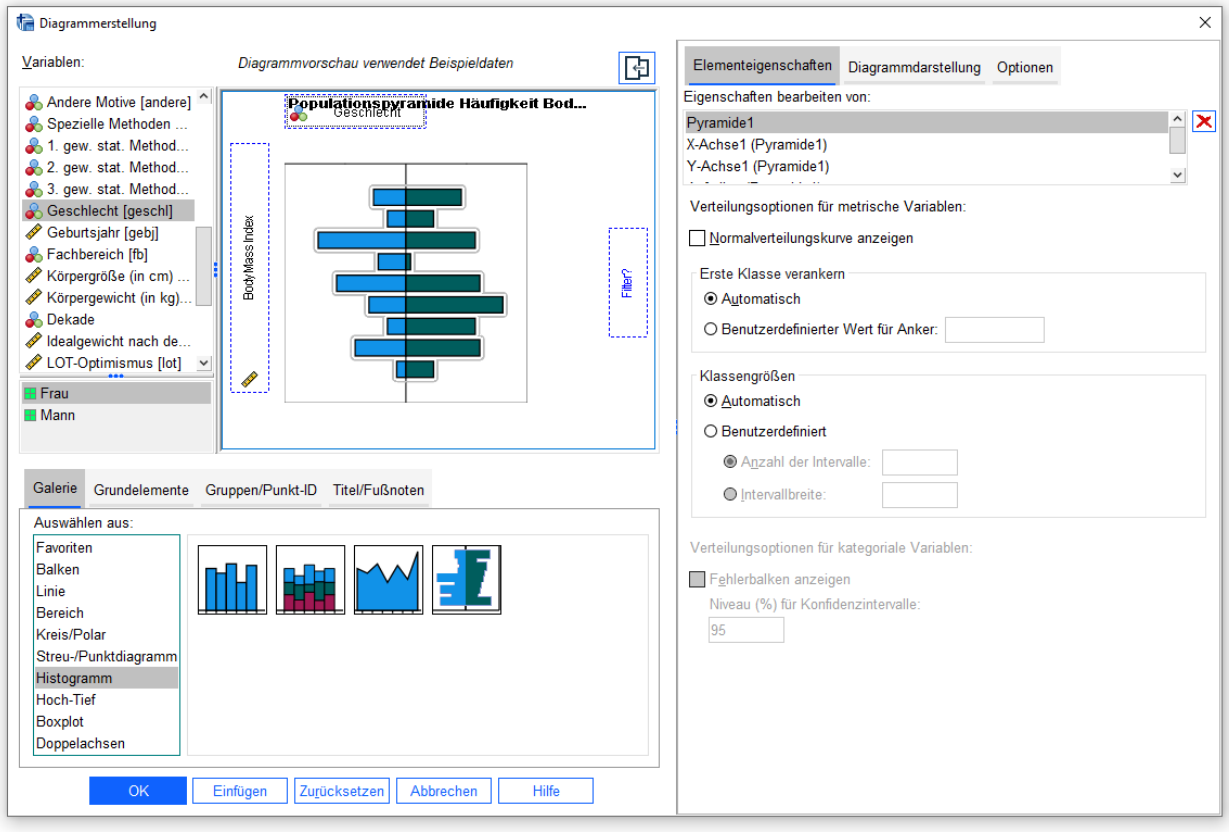
Bei stark asymmetrischen Verteilungen kann die Boxplot-Darstellung wegen zahlreicher Ausreißer unanschaulich werden. Nicht nur in dieser Situation kommt als Alternative das folgende Zweigruppen-Histogramm in Frage, das eine informative und optisch attraktive Gegenüberstellung der beiden Verteilungen liefert:



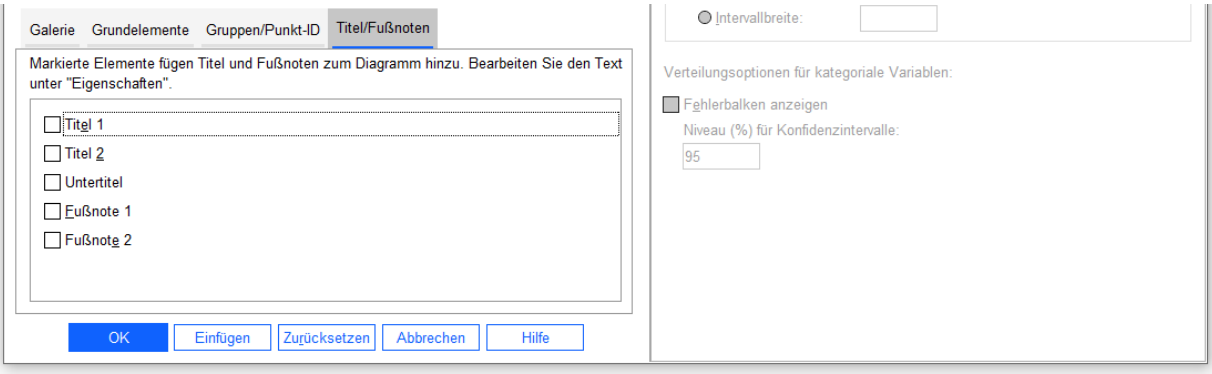
Öffnen Sie zum Erstellen dieser Abbildung über den Menübefehl

Grafik > Diagrammerstellung

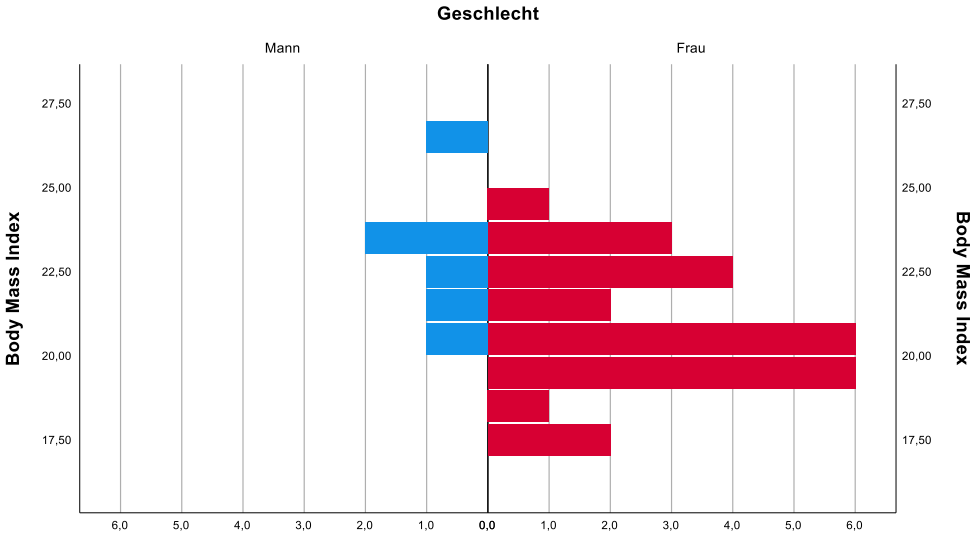
die Diagrammerstellung, und wählen Sie aus der **Galerie** die **Histogramm**-Variante mit dem Namen **Populationspyramide**. Verwenden Sie Geschlecht als **Teilungsvariable** und den Body Mass Index als **Verteilungsvariable**:



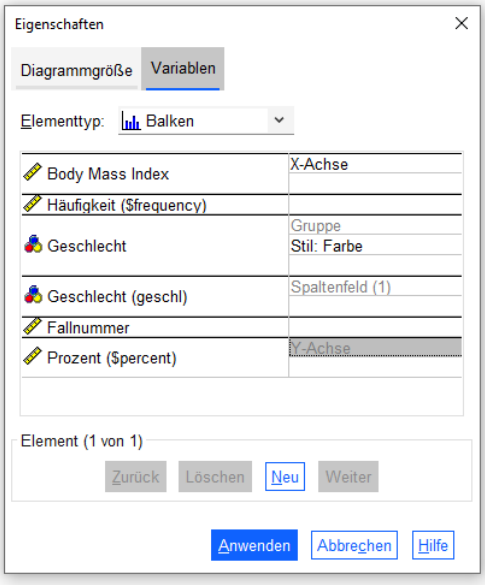
Schalten Sie auf der Registerkarte **Titel/Fußnoten** den **Titel 1** ab:



Am produzierten Diagramm stört die Verwendung von *absoluten* Häufigkeiten statt der zur Verteilungsbeschreibung günstigeren *relativen* Häufigkeiten:

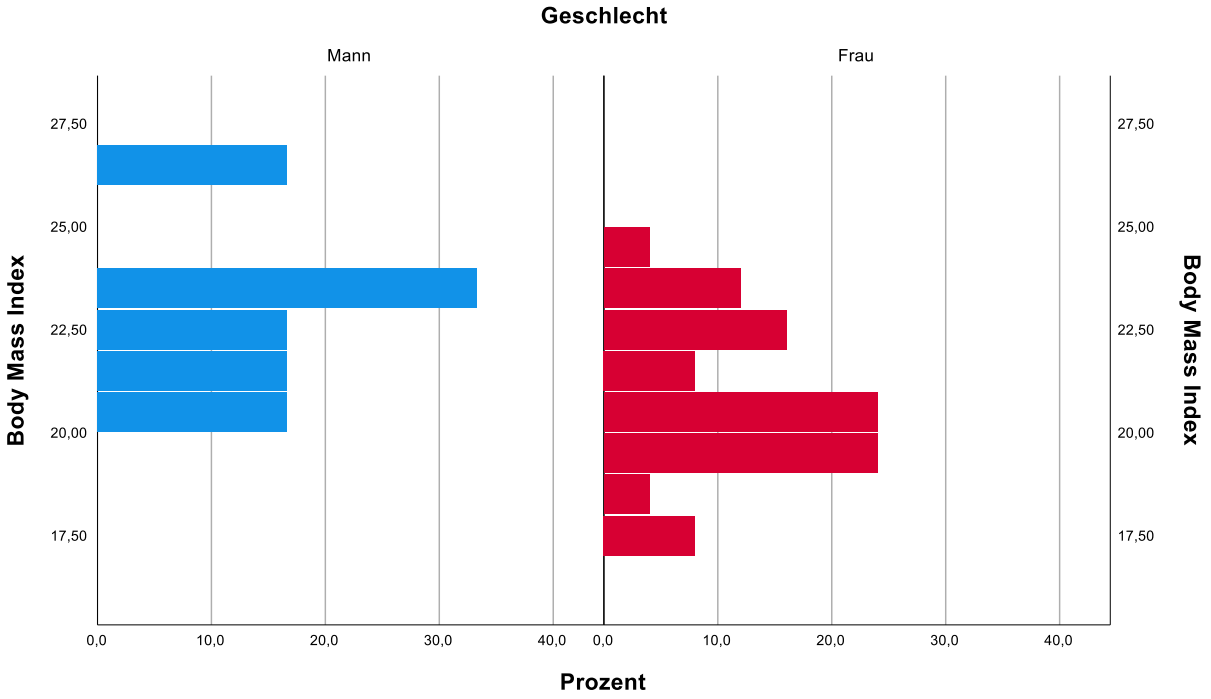


Um diesen Mangel zu beheben, öffnet man den Diagrammeditor per Doppelklick auf die Grafik und wählt auf der Registerkarte **Variablen** des **Eigenschaften**-Fensters für die **Prozent**-Werte die Rolle **Y-Achse**:

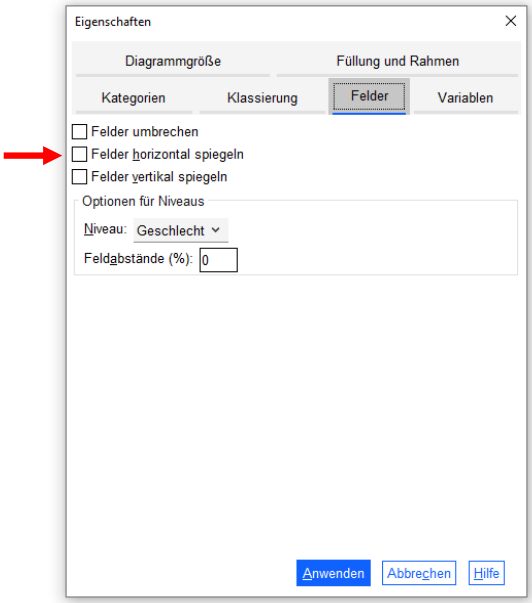


Nach einem Klick auf den Schalter **Anwenden** entsteht das oben abgebildete Zweigruppen-Histogramm.

Wer die folgende Anordnung der beiden Histogramme



bevorzugt, muss im **Eigenschaften**-Fenster auf der Registerkarte **Felder** die Markierung des Kontrollkästchens **Felder horizontal spiegeln** entfernen:



13 Fälle auswählen

Es kommt durchaus vor, dass man sich bei einer Analyse auf eine Teilstichprobe beschränken möchte. Bei unserer KFA-Studie ist es von Interesse, die Personen mit einem *negativen* KFA-Effekt (AERGZ < 0) näher kennenzulernen. Wir können dazu nach geeigneter Fallauswahl einen Bericht mit interessanten Variablenausprägungen anfordern.

13.1 Auswahl über eine Bedingung

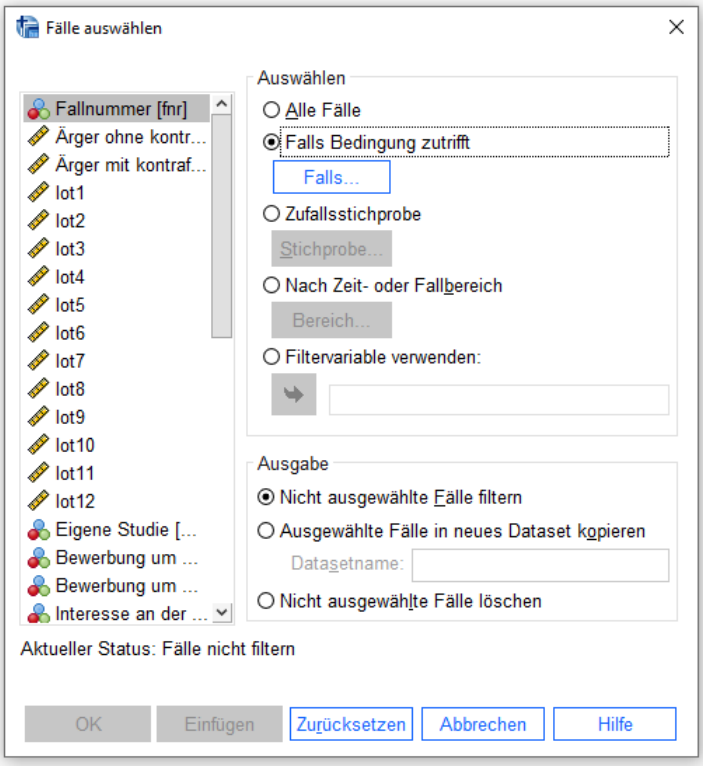
Man kann Fälle in Abhängigkeit von einer Bedingung ...

- temporär deaktivieren,
- aus der Arbeitsdatei entfernen
- oder in ein neues Datenblatt kopieren.

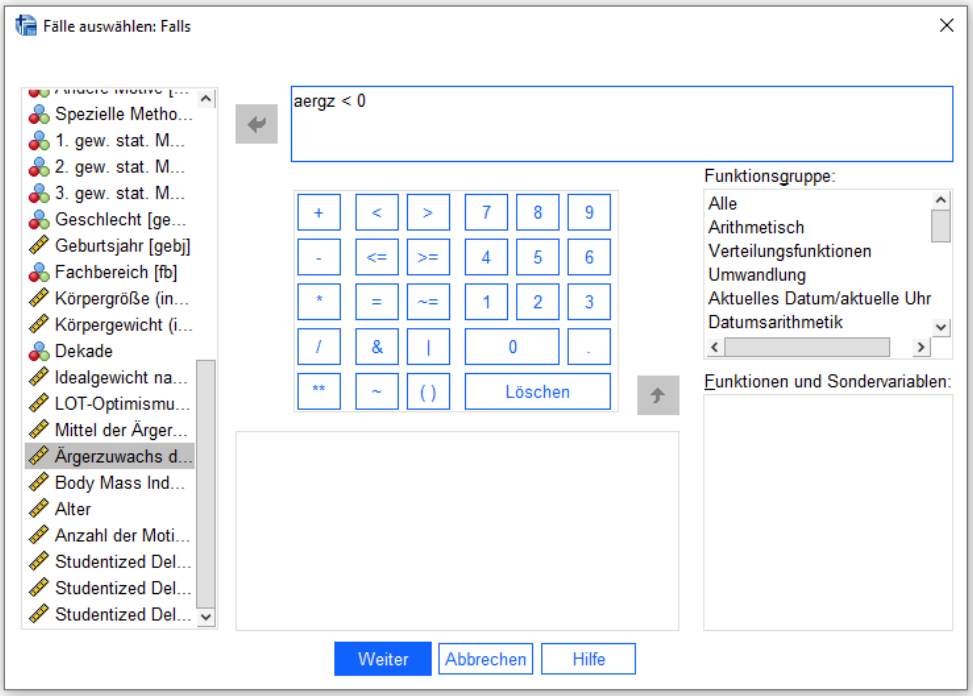
Die zuständige Dialogbox erreichen Sie über den Symbolleistenschalter  im Dateneditor oder mit den Menübefehl:

Daten > Fälle auswählen

Um ein Auswahlkriterium zu formulieren, müssen Sie im Optionenfeld **Auswählen** die Alternative **Falls Bedingung zutrifft** wählen und anschließend die zuständige Subdialogbox mit dem **Falls**-Schalter aktivieren:



Im **Falls**-Dialog ist ein logischer Ausdruck (vgl. Abschnitt 7.5.2) als Auswahlkriterium zu definieren, z. B.:

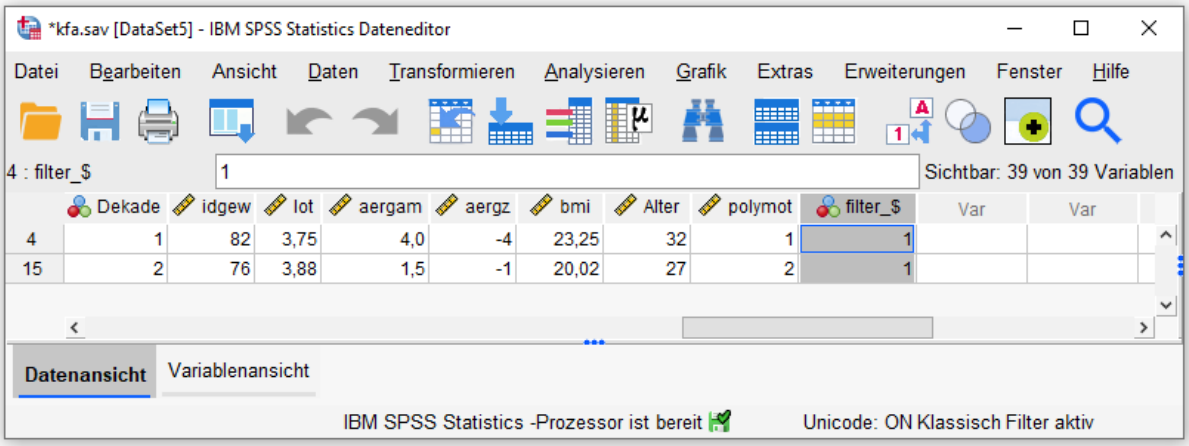


Wenn Sie nach erfolgreicher Definition des Auswahlkriteriums **Weiter** machen, können Sie im Optionenfeld **Ausgabe** des Hauptdialogs (siehe oben) entscheiden, was mit den Positiv- bzw. Negativfällen geschehen soll:

- **Nicht ausgewählte Fälle filtern**
  SPSS erzeugt aufgrund des logischen Ausdrucks eine Hilfsvariable namens FILTER_$ mit folgenden Werten:

  1 falls bei einem Fall der logische Ausdruck wahr ist,
  0 falls bei einem Fall der logische Ausdruck falsch ist,
  SYSMIS sonst (also bei unbestimmtem Ausdruck).

  Diese Variable wird als Filter aktiviert, d. h. bis zur Deaktivierung des Filters werden bei allen Analysen nur noch Fälle mit dem Wert 1 bei FILTER_$ einbezogen. Die in den einstweiligen Ruhezustand versetzten Fälle (mit den Werten 0 oder SYSMIS bei FILTER_$) werden im Datenfenster nicht mehr angezeigt:



Dass ein **Filter aktiv** ist, sodass nicht alle Fälle des Datenblatts zu sehen sind, wird in der Statuszeile dokumentiert.

Filter wirken sich nur bei statistischen und grafischen Analysen aus. Bei Datentransformationen werden hingegen auch die ausgefilterten Fälle einbezogen. Für eine bedingte Datentransformation ist die in Abschnitt 7.5 beschriebene Vorgehensweise zu verwenden.

Wenn ein Filter aktiv ist, wird dies in der Statuszeile angezeigt (siehe obiges Bildschirmfoto). Um den Filter im weiteren Verlauf der aktuellen SPSS-Sitzung zu *deaktivieren*, müssen Sie die Dialogbox **Fälle auswählen** erneut aufrufen und im **Auswählen**-Optionenfeld den Ausgangszustand **Alle Fälle** wiederherstellen.

Durch das Einrichten oder Modifizieren eines Filters wird die Variable FILTER_$ erstellt oder verändert. Folglich fragt SPSS am Ende der Sitzung nach, ob das veränderte Datenblatt gespeichert werden soll. Wenn Sie zustimmen, landet die Variable FILTER_$ in der Datendatei. Beim nächsten Öffnen dieser Datei ist allerdings *kein* Filter aktiv. Um die durch FILTER_$ definierte Fallauswahl zu reaktivieren, muss diese Variable in der Dialogbox **Fälle auswählen** als **Filtervariable verwendet** werden. Weil Filtervariablen mit beliebigem Namen akzeptiert werden, kann man in einer SPSS-Datendatei mehrere Filtervariablen bereithalten. Außerdem kann man die einem Filter zugrunde liegende und über den Schalter **Einfügen** im Dialog **Fälle auswählen** verfügbare Syntax abspeichern und später wiederverwenden.

- **Ausgewählte Fälle in neues Dataset kopieren**
  Die Fälle mit dem Wert 1 bei FILTER_$ werden in ein neues Datenblatt kopiert.
- **Nicht ausgewählte Fälle löschen**
  Die Fälle mit dem Wert 0 oder SYSMIS bei FILTER_$ werden aus der Arbeitsdatei entfernt. Aus einer eventuell zugeordneten *externen* Datendatei (z. B. auf der Festplatte) verschwinden die Fälle dabei *nicht*. Überlegen Sie gründlich, ob Sie das teilentleerte Datenblatt irgendwann (z. B. beim Verlassen von SPSS) in die zugeordnete Datei sichern wollen.

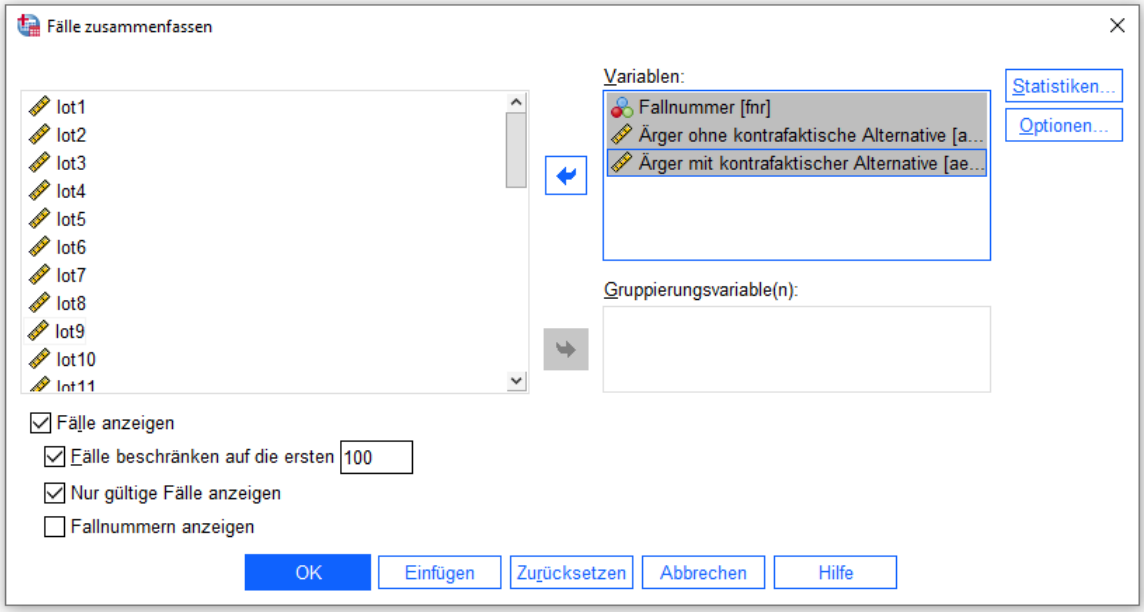
Man kann im Dialog **Fälle auswählen** eine Auswahl von Positivfällen auch über eine **Zufallsstichprobe** gewinnen oder einen **Fallbereich** festlegen, z. B. zur Beschränkung auf die ersten *n* Fälle.

13.2 Bericht anfordern

Gelegentlich benötigt man für eine bestimmte Teilmenge von Fällen eine übersichtliche Liste mit den Ausprägungen bestimmter Variablen. Um z. B. für Personen mit negativem AERGZ-Wert eine Liste mit den Variablen FNR, AERGO und AERGM zu erhalten, vereinbart man zunächst eine Fallauswahl mit der Filterbedingung „AERGZ < 0“ (siehe Abschnitt 13.1) und fordert dann über

Analysieren > Berichte > Fallzusammenfassungen

die gewünschte Auflistung an:



Wir erhalten die folgende Liste mit 2 Fällen:

Zusammenfassung von Fällen[a]

		Fallnummer	Ärger ohne kontrafaktische Alternative	Ärger mit kontrafaktischer Alternative
1		4	6	2
2		15	2	1
Insgesamt	N	2	2	2

a. Begrenzt auf die ersten 100 Fälle.

14 Analyse von Kreuztabellen

In diesem Kapitel untersuchen wir Geschlechtsunterschiede bei der Wahl des Studienfachs mit Hilfe der Kreuztabellenanalyse.

14.1 Untersuchungsplanung

In unserer Datendatei **kfa.sav** enthält die Fachbereichsvariable (FB) Information über die Studienfächer der Untersuchungsteilnehmer auf einem angemessenen Aggregationsniveau. Ihre Werte stehen für die folgenden Fachbereiche der Universität Trier:

Wert	Fachbereich mit den Fächern
1	I: Pädagogik, Philosophie, Psychologie
2	II: Sprachorientierte Fächer
3	III: Historische und politische Wissenschaften
4	IV: BWL, Ethnologie, Informatik, Mathematik, Soziologie, VWL, Wirtsch.-Informatik
5	V: Jura
6	VI: Geowissenschaften

Mit Hilfe dieser Variablen sollen Geschlechtsunterschiede bei der Wahl des Studienfachs untersucht werden, wobei die empirisch zu prüfende Nullhypothese so formuliert werden kann:

Die Merkmale Geschlecht und Fachbereich sind unabhängig voneinander.

Die Unabhängigkeitsbehauptung der Nullhypothese bedeutet, dass sich aus dem Wissen über das Geschlecht eines Untersuchungsteilnehmers keinerlei Information über seine Fachbereichszugehörigkeit ableiten lässt, dass also die bedingten Fachbereichsverteilungen bei Frauen und Männern identisch sind. Zur Illustration des *Unabhängigkeitsbegriffs* wird hier auf eine *Verteilungshomogenität* verwiesen. Später folgen noch einige Erläuterungen zu den beiden Begriffen und zu ihrer Beziehung.

Unsere Nullhypothesenformulierung ist „zweiseitig“, wozu es auch gar keine Alternative gibt, weil die Fachbereichsvariable mehr als zwei Stufen hat. Hier ist die (anschließend wegen unzureichender Zellbesetzungen zu bemängelnde) (2 × 6) - Kreuztabelle mit den Daten aus der Manuskriptstichprobe zu sehen:

	Fachbereich					
Geschlecht	I	II	III	IV	V	VI
Frau	17	0	2	3	0	3
Mann	2	0	0	3	0	1

Bei (2 × 2) - Kreuztabellen sind aber auch gerichtete Hypothesen möglich (siehe Abschnitt 14.4.5.2). Beschränkt man sich z. B. auf eine Betrachtung der Fachbereiche I und IV, lässt sich die Alternativhypothese formulieren, dass der Frauenanteil im FB I größer sei als im FB IV.

Weil es bei der aktuellen Fragestellung um den Zusammenhang zwischen den beiden *nominalskalierten* Merkmalen Fachbereich und Geschlecht geht, wählen wir als Auswertungsmethode die Kreuztabellenanalyse. Diese Methode ist recht beliebt, wobei Anwendungsfälle gelegentlich durch das wenig empfehlenswerte künstliche Kategorisieren von metrischen Variablen herbeigeführt werden. Hoffentlich trägt die ausführliche Behandlung der Methode im aktuellen Kapitel nicht dazu bei, die Kreuztabellenanalyse als Universalwerkzeug der Statistik erscheinen zu lassen. Sie ist adä-

quat zur Untersuchung von Geschlechtsunterschieden bei der Wahl des Studienfachs, weil es hier um die Beziehung von zwei nominalskalierten Merkmalen geht. Man kann die Fragestellung so auffassen, dass ...

- ein nominalskaliertes Kriterium (Studienfachpräferenz) aufgeklärt werden soll,
- wobei nur ein *einziger* Prädiktor interessiert (Geschlecht),
- der ebenfalls nominales Messniveau besitzt.

Bei vielen Analysen mit einem nominalskalierten Kriterium werden aber Methoden benötigt, die …

- eine beliebige Anzahl von Prädiktoren erlauben,
- neben nominalskalierten auch metrische Prädiktoren unterstützen, also die in metrischen Prädiktoren enthaltenen Informationen komplett verwerten,
- eine flexible Modellierung ermöglichen (z. B. mit Wechselwirkungen zwischen Prädiktoren).

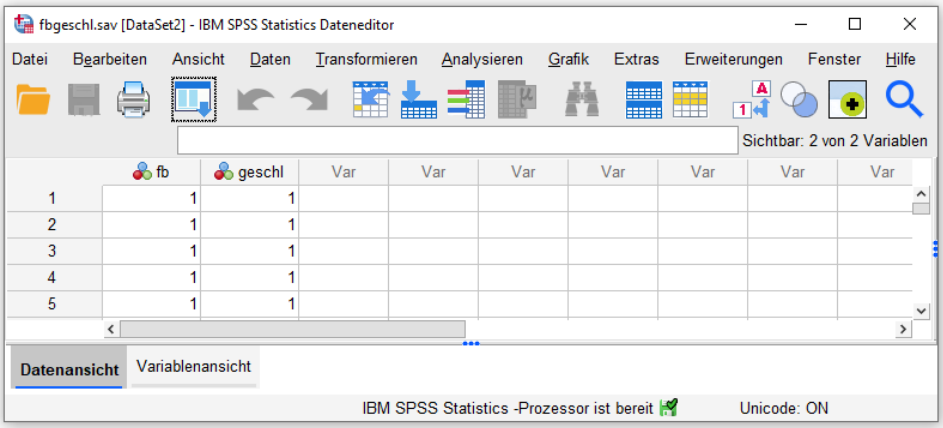
Mit der **logistischen Regressionsanalyse** steht für kategoriale oder ordinale Kriterien ein Verfahren bereit, das Modelle mit beliebig vielen kategorialen oder metrischen Regressoren erlaubt und auch Wechselwirkungen unterstützt. Nähere Informationen finden Sie z. B. in einem ZIMK-Manuskript (Baltes-Götz 2012), das auf dem Webserver der Universität Trier zu finden ist:

https://www.uni-trier.de/?id=22513

Leider erweist sich die Manuskriptstichprobe bei näherer Betrachtung als ungeeignet zur Untersuchung von Geschlechtsunterschieden bei der Wahl des Studienfachs:

- Sie ist sehr klein, was zu einer geringen Teststärke führt.
- Die Stichprobe ist wenig repräsentativ, weil sie nur an SPSS interessierte Personen enthält. Folglich sind die Fachbereiche II, III und V (fast) nicht vertreten (siehe obige Tabelle).

Analoge Verhältnisse bestehen in der Regel auch in den Kursstichproben des statistischen Praktikums mit SPSS. Daher wurde eine Zufallsstichprobe der Größe $n = 283$ aus der Datenbank mit allen Studierenden der Universität Trier im Wintersemester 1993/94 gezogen.[1] Bei jedem Fall wurden die Merkmale Geschlecht (Variable GESCHL) und Fachbereich (Variable FB) festgestellt. Die SPSS-Datendatei **fbgeschl.sav** mit den beiden Variablen befindet sich an der im Vorwort für Kursdateien vereinbarten Stelle.



[1] Aufmerksame Leser(innen) werden zu Recht fragen, warum nicht *alle* Trierer Studierenden im damaligen Wintersemester einbezogen wurden. Eine größere Stichprobe bringt stabilere Ergebnisse und hätte in dieser speziellen Situation kaum mehr „gekostet“. Allerdings habe ich aus didaktischen Gründen eine Stichprobe mit „typischem“ Umfang vorgezogen.

Wir können die Stichprobengröße nicht ändern, wollen aber die daraus resultierende Power des geplanten Hypothesentests abschätzen. Dazu verwenden wir erneut das Programm **G*Power 3.1**, das schon bei der Stichprobenumfangsplanung in Abschnitt 2.3 zum Einsatz kam. Auf den Pool-PCs der Universität Trier unter dem Betriebssystem Windows ist G*Power 3.1 im Startmenü so zu finden

Statistik > GPower

G*Power arbeitet bei der Kreuztabellenanalyse mit dem folgenden Effektstärkeindex *W* (Cohen 1988, S. 215ff)

$$W := \sqrt{\sum_{i=1}^{z}\sum_{j=1}^{s}\frac{(p_{ij}^{(1)} - p_{ij}^{(0)})^2}{p_{ij}^{(0)}}}$$

Darin bedeuten:

z, s	=	Anzahl der Zeilen bzw. Spalten
$p_{ij}^{(1)}$	=	Wahrscheinlichkeit der Zelle *ij* unter der Alternativhypothese
$p_{ij}^{(0)}$	=	Wahrscheinlichkeit der Zelle *ij* unter der Nullhypothese

Unter der Alternativhypothese sind die $z \times s$ Zellwahrscheinlichkeiten unrestringiert und müssen sich lediglich zu 1 aufaddieren. Unter der Nullhypothese ist die Wahrscheinlichkeit der Zelle *ij* identisch mit dem Produkt aus der Wahrscheinlichkeit der Zeile *i* und der Wahrscheinlichkeit der Spalte *j* (siehe Abschnitt 14.3). Während es sich bei $p_{ij}^{(1)}$ um die tatsächliche Wahrscheinlichkeit der Zelle (*i*, *j*) handelt, ist $p_{ij}^{(0)}$ eine unter der Unabhängigkeitsannahme aus den Wahrscheinlichkeiten der Zeile *i* und der Spalte *j* berechnete Wahrscheinlichkeit.

Im Effektstärkeindex *W* werden Diskrepanzen zwischen den Zellwahrscheinlichkeiten unter der Alternativ- bzw. Nullhypothese über alle ($z \times s$) Zellen der Kreuztabelle aufsummiert. Für jede Zelle *ij* wird die quadrierte Differenz von $p_{ij}^{(1)}$ und $p_{ij}^{(0)}$ durch die Nullhypothesenwahrscheinlichkeit $p_{ij}^{(0)}$ dividiert. Eine quadrierte Differenz geht also umso stärker in den Index ein, je kleiner die Nullhypothesenwahrscheinlichkeit der Zelle ist. Insgesamt wird quantifiziert, wie weit die unrestringierten Zellwahrscheinlichkeiten von den Wahrscheinlichkeiten unter der Restriktion der Nullhypothese entfernt sind.

In Abschnitt 14.4.1 wird sich enge Beziehungen herausstellen zwischen dem Effektstärkeindex *W* und ...

- Pearsons χ_P^2 - Prüfgröße zur Unabhängigkeitshypothese
- Cramers V (einem Maß der Assoziationsstärke für zwei nominalskalierte Variablen)

Weil uns keine Information über die Effektstärke in der Population vorliegt, nehmen wir einen *mittleren* Wert von $W = 0{,}3$ nach der von Cohen (1988, S. 227) vorgeschlagenen Konvention

- kleiner Effekt: $W = 0{,}1$
- mittlerer Effekt: $W = 0{,}3$
- großer Effekt: $W = 0{,}5$

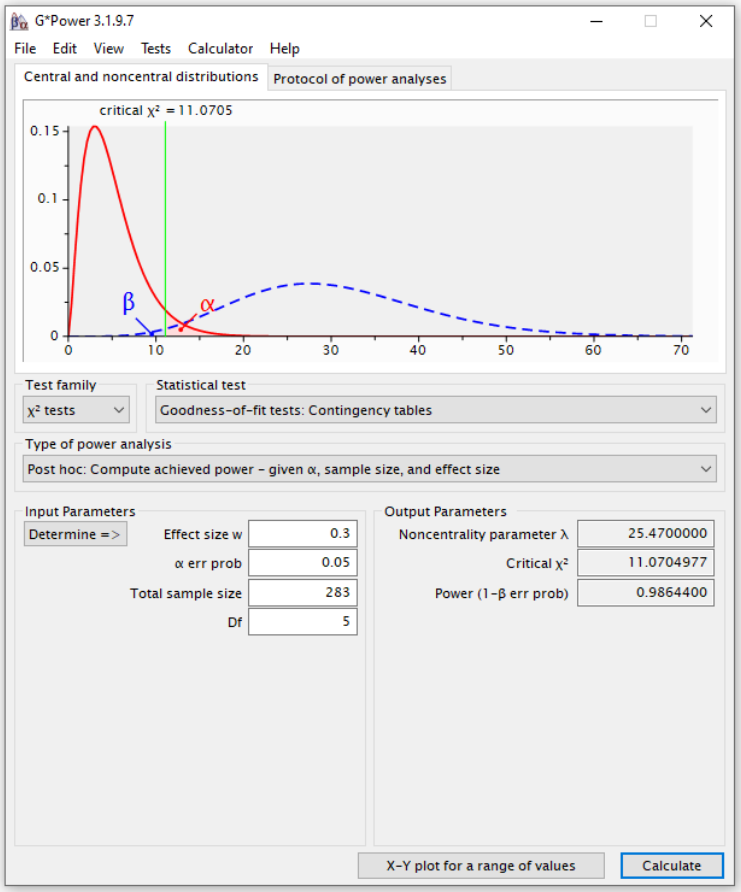
als inhaltlich relevant an.

Wir wählen in G*Power 3.1 folgende Einstellungen:

- **Test family** **χ^2-Tests**
- **Statistical test** **Goodness-of-fit tests Contingency tables**
- **Type of power analysis** **Post hoc**
- **Effect size w** 0.3
- **α err prob** 0.05
- **Total sample size** 283
- **Df** 5

  Warum bei einer Tabelle mit zwei Zeilen und sechs Spalten gerade fünf Freiheitsgrade (engl.: *degrees of freedom*) resultieren, erfahren Sie in Abschnitt 14.4.1.

Wir erhalten eine erfreulich hohe Power von 0,99:



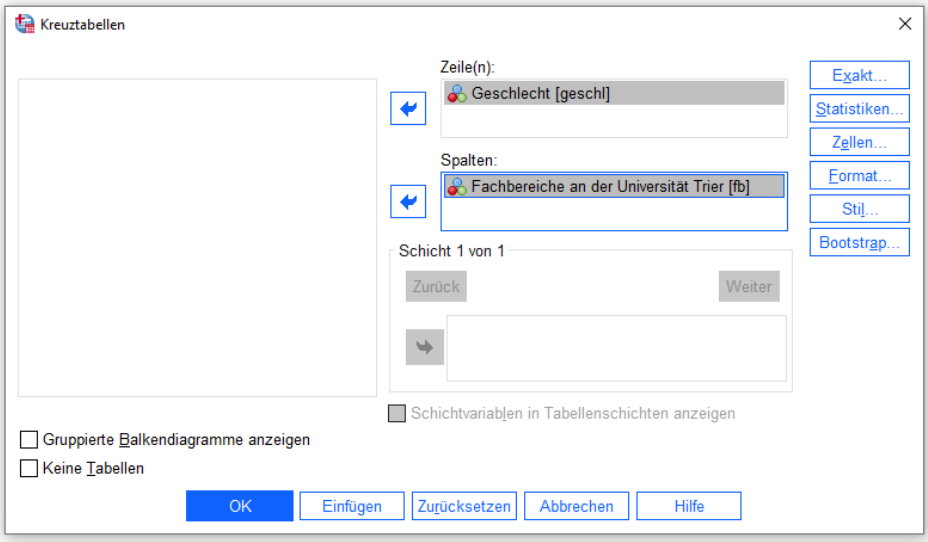
Wenn in der Population ein Effekt von der angenommenen mittleren Stärke existiert, werden wir mit hoher Wahrscheinlichkeit ein signifikantes Testergebnis zum Nachteil der Unabhängigkeits-Nullhypothese erhalten.

14.2 Beschreibung der bivariaten Häufigkeitsverteilung

Die SPSS-Dialogbox zur Analyse zweidimensionaler Kontingenztabellen erscheint nach dem Menübefehl:

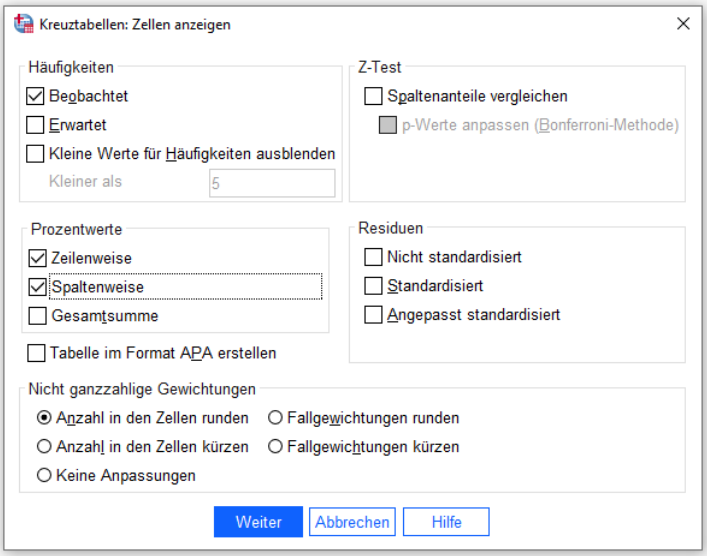
Analysieren > Deskriptive Statistiken > Kreuztabellen

Wir wählen GESCHL als Zeilen- und FB als Spaltenvariable:



Im Beispiel liegt es nahe, GESCHL als unabhängige und FB als abhängige Variable zu betrachten. Zur Frage, wie in einer solchen Lage die Zeilen- und die Spaltenposition der Tabelle zu vergeben sind, hat sich keine eindeutige Konvention etabliert (vgl. Gehring & Weins 2004, S. 90). Wir verwenden die platzsparende Anordnung.

In der **Zellen**-Subdialogbox kann man u. a. zeilen- und spaltenbezogene **Prozentwerte** für die Zellen der Kontingenztabelle anfordern:



Aufgrund dieser Spezifikationen erhalten wir für unsere Stichprobe die folgende Kreuztabelle:[1]

[1] Die Tabelle wurde mit dem Pivot-Editor (vgl. Abschnitt 10.4) durch Aufhebung der Gruppierung Geschlecht und gekürzte Kategorienbeschriftungen etwas schlanker gemacht.

Geschlecht * Fachbereiche an der Universität Trier Kreuztabelle

		Fachbereiche an der Universität Trier						
		I	II	III	IV	V	VI	Gesamt
Frauen	Anzahl	29	26	18	22	26	23	144
	% von Geschlecht	20,1%	18,1%	12,5%	15,3%	18,1%	16,0%	100,0%
	% von Fachbereich	63,0%	66,7%	50,0%	31,0%	54,2%	53,5%	50,9%
Männer	Anzahl	17	13	18	49	22	20	139
	% von Geschlecht	12,2%	9,4%	12,9%	35,3%	15,8%	14,4%	100,0%
	% von Fachbereich	37,0%	33,3%	50,0%	69,0%	45,8%	46,5%	49,1%
Gesamt	Anzahl	46	39	36	71	48	43	283
	% von Geschlecht	16,3%	13,8%	12,7%	25,1%	17,0%	15,2%	100,0%
	% von Fachbereich	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Durch die Einträge in den (2 × 6) Zellen wird die gemeinsame Verteilung der beiden Variablen GESCHL und FB beschrieben:

- Oben ... steht die absolute Häufigkeit der Zelle
  Z. B. befanden sich in der Stichprobe 17 Männer aus dem Fachbereich I.
- In der Mitte ... steht der prozentuale Anteil der Zelle an allen Fällen in der zugehörigen Zeile.
  Z. B. gehörten von den 139 *männlichen* Untersuchungsteilnehmern 12,2% zum Fachbereich I. Diese auf die Zeile bezogenen relativen Häufigkeiten beschreiben die bedingte Verteilung der Spaltenvariablen (FB) für einen festen Wert der Zeilenvariablen (GESCHL). Wir erhalten z. B. für die Männer die folgende bedingte FB-Verteilung:

I	II	III	IV	V	VI
12,2%	9,4%	12,9%	35,3%	15,8%	14,4%

- Unten ... steht der prozentuale Anteil der Zelle an allen Fällen in der zugehörigen Spalte
  Z. B. waren von den 46 Personen aus dem Fachbereich I 37% Männer. Diese auf die Spalte bezogenen relativen Häufigkeiten beschreiben die bedingte Verteilung der Zeilenvariablen (GESCHL) für einen festen Wert der Spaltenvariablen (FB). Wir erhalten z. B. für den Fachbereich I die folgende bedingte Geschlechtsverteilung:

Frauen	63%
Männer	37%

In der **Zellen**-Subdialogbox können noch weitere Informationen zu den Zellen angefordert werden (z. B. die erwarteten Häufigkeiten unter der Nullhypothese).

Beim Vergleich der fachbereichsbedingten Geschlechtsverteilungen zeigen sich erhebliche Unterschiede:

- In den Fachbereichen I und II dominieren die Frauen mit einem Anteil von 63% bzw. 66,7%.
- Im Fachbereich IV sind die Frauen mit einem Anteil von 31% in der Minderheit.
- In den übrigen Fachbereichen III, V und VI zeigt sich ein ziemlich ausgeglichenes Geschlechtsverhältnis.

Die Betrachtung der fachbereichsbedingten Geschlechtsverteilungen, also der Geschlechtsverteilungen innerhalb der einzelnen Fachbereiche, steht nicht im Widerspruch zur Annahme, dass sich

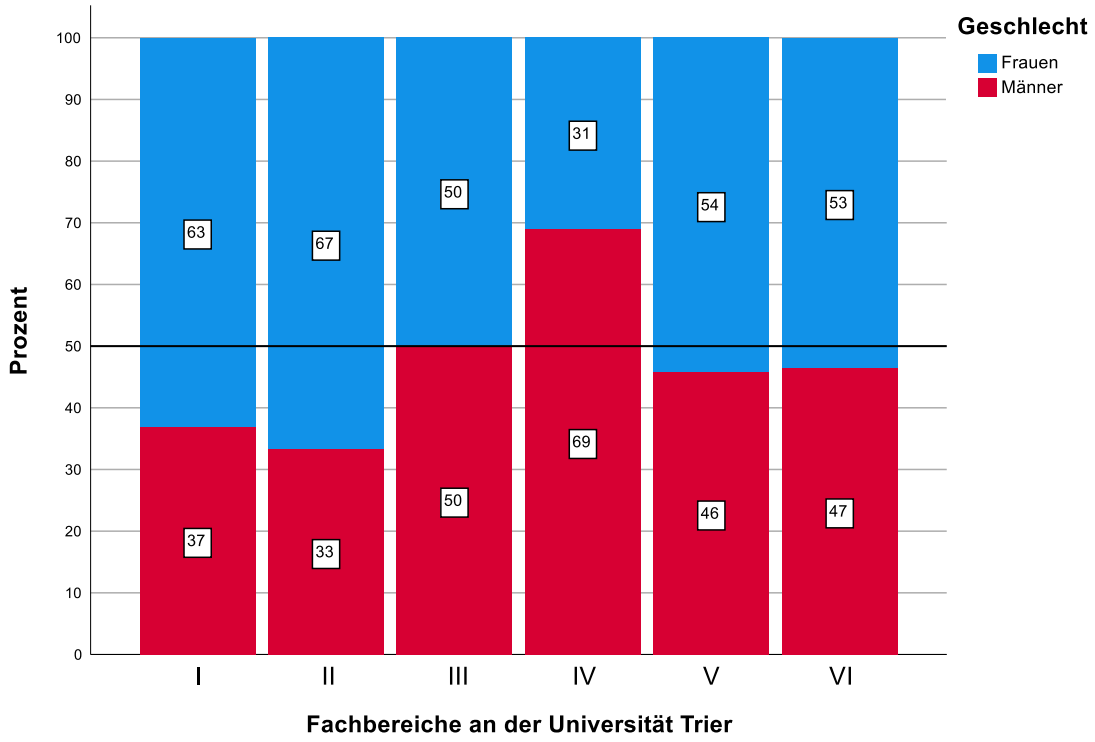
das Geschlecht einer Person auf ihre Wahl eines Fachbereichs, also auf ihre FB-Ausprägung auswirkt. Hier wird der wahrscheinlichkeitstheoretische Begriff der *bedingten Verteilung* verwendet. Die Unabhängigkeitshypothese der Kreuztabellenanalyse ist symmetrisch und äquivalent ...

- zur Homogenität der geschlechts-bedingten FB-Verteilungen und
- zur Homogenität der fachbereichs-bedingten Geschlechtsverteilungen.

Im Beispiel wird die Abweichung der empirischen bivariaten Verteilung von der Unabhängigkeitsannahme besonders offensichtlich durch die Betrachtung fachbereichsbedingten Geschlechtsverteilungen, weil diese bedingten Verteilungen durch *eine* Wahrscheinlichkeit beschrieben werden können.

Weil die Kreuztabellenanalyse als Verfahren zur Untersuchung der Assoziation von zwei kategorialen Merkmalen komplett symmetrisch arbeitet, spielt die Vermutung zur kausalen Einflussrichtung für die Durchführung des Verfahrens keine Rolle.

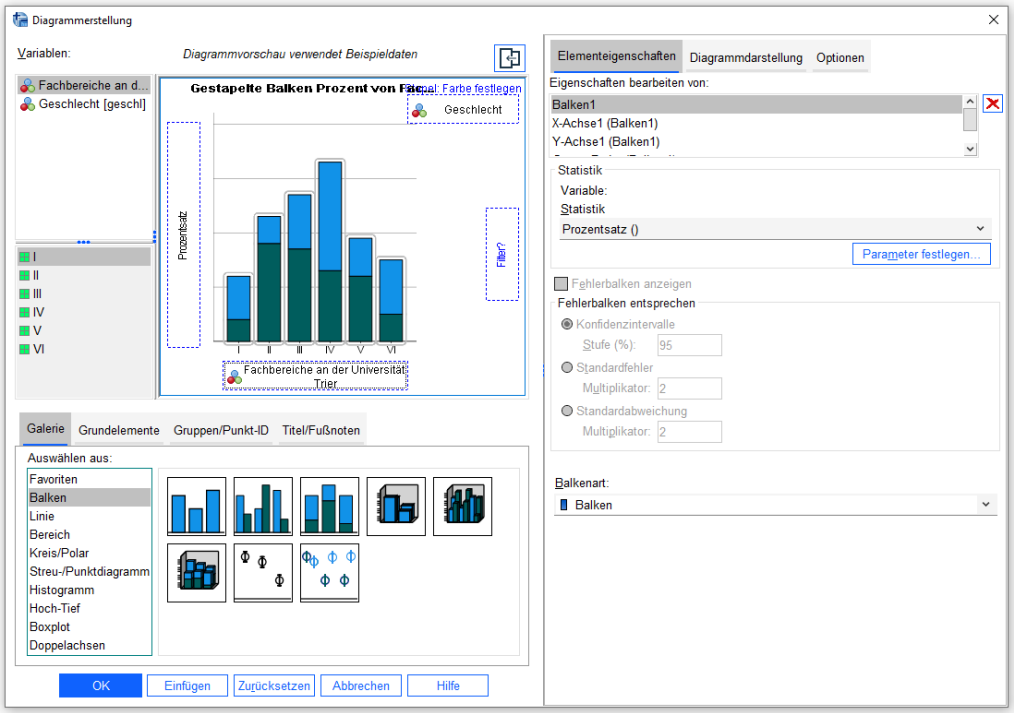
Das folgende gestapelte Balkendiagramm veranschaulicht die fachbereichsbedingten Geschlechtsverteilungen:



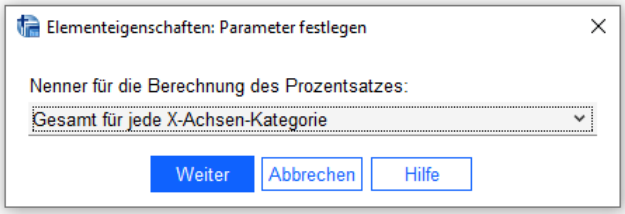
Zur Produktion dieses Diagramms starten wir mit

Grafik > Diagrammerstellung

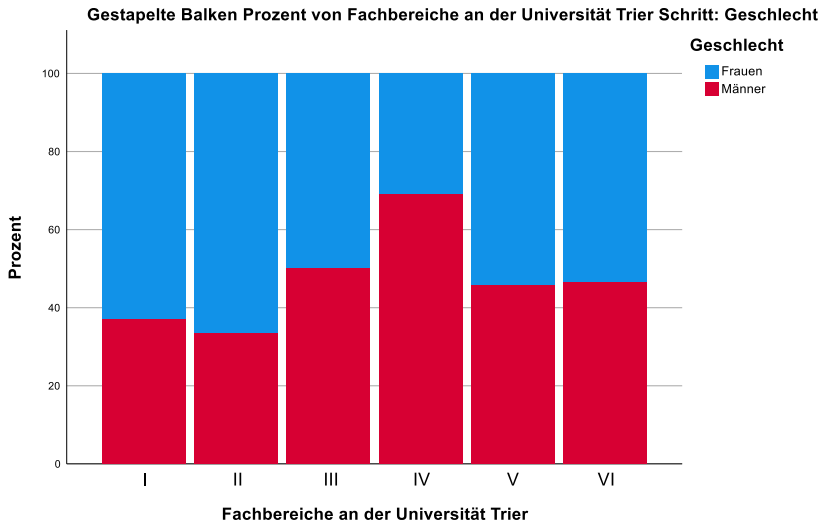
und wählen in der **Galerie** von den **Balken** die gestapelte Variante . Dann befördern wir GESCHL auf die **Stapel**-Ablage und FB auf die **X-Achse**:



Auf der Registerkarte mit den **Elementeigenschaften** entscheiden wir uns für den **Prozentsatz()** als darzustellende **Statistik**. Nach einem Klick auf den Schalter **Parameter festlegen** wählen wir als **Nenner für die Berechnung des Prozentsatzes** die Häufigkeit der zugehörigen **X-Achsen-Kategorie**:

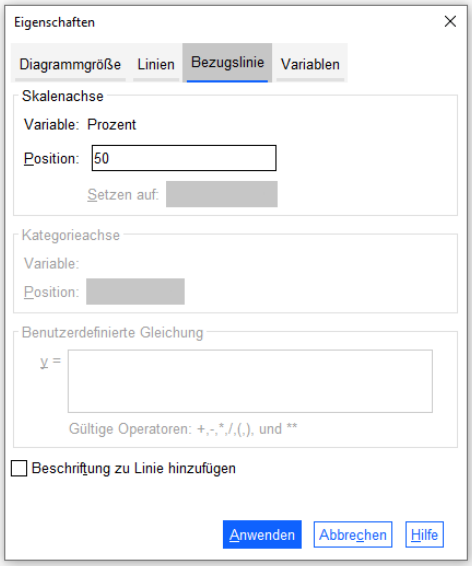


Wir lassen das Diagramm mit **OK** erstellen. Den folgenden Rohling

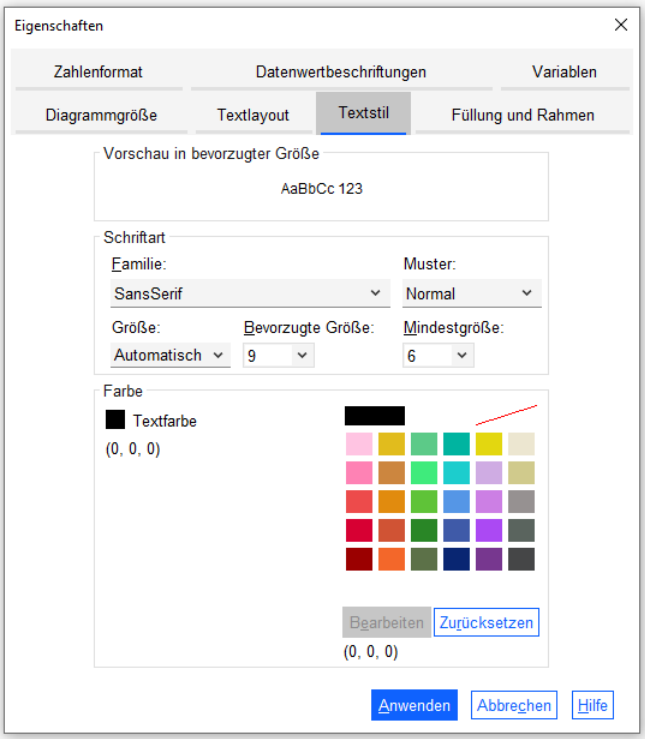


öffnen wir per Doppelklick im Diagrammeditor, um noch einige Details zu optimieren:

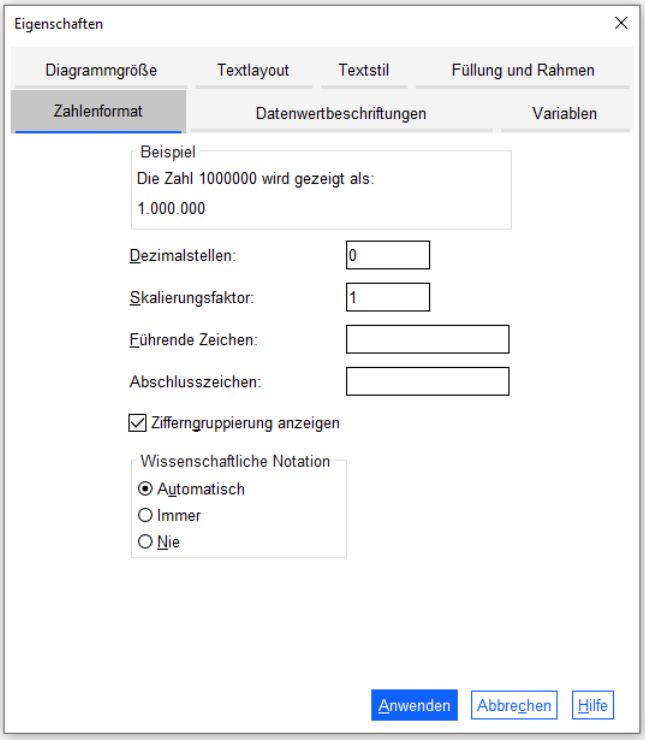
- Der Titel wird gelöscht.
- Um eine Y-Achsen - Beschriftung im 10er - Abstand zu erreichen, wird nach der Markierung einer Y-Achsen - Teilstrichbeschriftung auf der **Eigenschaften**-Fenster - Registerkarte **Metrisch** das **erste Inkrement** auf 10 gesetzt. Auf derselben Registerkarte wird der **obere Rand** auf 5 gesetzt. Diese Einstellungen werden über den Schalter **Anwenden** realisiert.
- Über den Menübefehl **Optionen > Bezugslinie für y-Achse** wird die 50% - Marke hervorgehoben:



- Bei markierten Balken sorgen wir über **Elemente > Datenbeschriftungen einblenden** oder den Symbolschalter für eine Anzeige der Prozentwerte. Über die **Eigenschaften**-Fenster - Registerkarte **Textstil** erhalten diese Beschriftungen die **bevorzugte Größe** 9:



Auf der Registerkarte **Zahlenformat** sorgen wir für 0 **Dezimalstellen** und entfernen das **Abschlusszeichen**:



14.3 Die Unabhängigkeits- bzw. Homogenitätshypothese

Hypothesen zum Zusammenhang zwischen zwei nominalskalierten Merkmalen lassen sich auf letztlich äquivalente Weise durch Verwendung verschiedener wahrscheinlichkeitstheoretischer Begriffe formulieren. Dies soll an unserem Beispiel demonstriert werden, damit Sie die Äquivalenz verstehen und auszunutzen lernen. Es ist ja generell sinnvoll, einen Sachverhalt aus verschiedenen Blickrichtungen zu betrachten.

1. Formulierung: Unabhängigkeitshypothese

H_0: Die Merkmale Geschlecht und Fachbereich sind unabhängig.
Die Wahrscheinlichkeit für jedes Verbundereignis (z. B. Mann im Fachbereich V) ist gleich dem Produkt aus den Wahrscheinlichkeiten der beiden Randereignisse (im Beispiel: Mann, Fachbereich V).

H_1: Die Merkmale Geschlecht und Fachbereich sind abhängig.
Die Wahrscheinlichkeit für mindestens ein Verbundereignis ist ungleich dem Produkt aus den Wahrscheinlichkeiten der Randereignisse.

2. Formulierung: Homogenitätshypothese

H_0: Die geschlechts-bedingten FB-Verteilungen sind gleich.

H_1: Die geschlechts-bedingten FB-Verteilungen sind nicht gleich.

Man kann leicht zeigen (vgl. Hartung 1989, S. 412): Perfekte Homogenität liegt genau dann vor, wenn die Merkmale Geschlecht und Fachbereich unabhängig sind. Außerdem lässt sich die Homogenitätshypothese äquivalent auch über die fachbereichs-bedingten Geschlechtsverteilungen formulieren:

H_0: Die Frauenanteile sind in allen Fachbereichen gleich.

H_1: Die Frauenanteile sind nicht in allen Fachbereichen gleich.

14.4 Testverfahren

14.4.1 Asymptotische χ^2 - Tests

Die bekannteste Prüfgröße zur Testung der Unabhängigkeits- bzw. Homogenitätshypothese ist die folgende χ^2_P **- Teststatistik nach Pearson**:

$$\chi^2_P := \sum_{i=1}^{z}\sum_{j=1}^{s}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}, \quad \text{mit } m_{ij} := \frac{n_{i.}\cdot n_{.j}}{n}$$

Darin bedeuten:

z, s	=	Anzahl der Zeilen bzw. Spalten
n_{ij}	=	beobachtete Häufigkeit in Zelle ij
m_{ij}	=	geschätzte erwartete Häufigkeit in Zelle ij unter der H_0
$n_{i.}$	=	beobachtete Häufigkeit in Zeile i
$n_{.j}$	=	beobachtete Häufigkeit in Spalte j
n	=	Umfang der Gesamtstichprobe

Die Formel zur Schätzung der erwarteten Häufigkeiten m_{ij} unter der Nullhypothese

$$m_{ij} := \frac{n_{i.}\cdot n_{.j}}{n}$$

ist leicht nachvollziehbar. Zunächst soll die Wahrscheinlichkeit $p_{ij}^{(0)}$ der Zelle ij unter der H_0 bestimmt werden. Da es sich hier um ein Verbundereignis aus zwei *unabhängigen* (H_0!) Einzelereignissen handelt (Zeile i und Spalte j), ergibt sich $p_{ij}^{(0)}$ als Produkt der Wahrscheinlichkeiten $p_{i.}$ und $p_{.j}$ für die beiden verknüpften Einzelereignisse:

$$p_{ij}^{(0)} = p_{i.}\cdot p_{.j}$$

Die Wahrscheinlichkeiten $p_{i.}$ und $p_{.j}$ sind allerdings nicht bekannt, sondern müssen durch die entsprechenden relativen Häufigkeiten in der Stichprobe geschätzt werden.[1] Die Wahrscheinlichkeit $p_{i.}$ zur Zeile i wird geschätzt durch die relative Häufigkeit der Zeile i in der Stichprobe:

$$\hat{p}_{i.} := \frac{n_{i.}}{n}$$

Analog ergibt sich die geschätzte Wahrscheinlichkeit $p_{.j}$ der Spalte j:

$$\hat{p}_{.j} := \frac{n_{.j}}{n}$$

[1] Diese Formulierung geht davon aus, dass man *eine* Stichprobe gezogen und bei jedem Fall die *beiden* Merkmale Geschlecht und Fachbereich beobachtet hat. Ein anderes Stichprobenmodell läge vor, wenn man in jedem Fachbereich eine Stichprobe der festen Größe 50 gezogen und bei jedem Fall das *eine* Merkmal Geschlecht beobachtet hätte. Dann wären die Randwahrscheinlichkeiten der FB-Kategorien bekannt. Allerdings bleiben auch unter dem alternativen Stichprobenmodell alle vorgestellten Rechnungen und Entscheidungsregeln korrekt.

Damit gilt für die geschätzte Wahrscheinlichkeit der Zelle *ij* unter der H_0:

$$\hat{p}_{ij}^{(0)} = \hat{p}_{i.} \cdot \hat{p}_{.j} = \frac{n_{i.}}{n}\frac{n_{.j}}{n} = \frac{n_{i.} \cdot n_{.j}}{n^2}$$

Die Wahrscheinlichkeit $p_{ij}^{(0)}$ der Zelle *ij* unter der Nullhypothese lässt sich interpretieren als Erwartungswert der Indikator-Zufallsvariablen X_{ij} zur Zelle *ij* beim Ziehen *eines* Falles bei gültiger H_0:

- Tritt die Zelle *ij* auf, nimmt X_{ij} den Wert 1 an,
- anderenfalls nimmt X_{ij} den Wert 0 an.

Werden *n* Fälle unabhängig gezogen, realisieren sich *n* unabhängige Zufallsvariablen $X_{ij}^{(k)}$, *k* = 1, .., *n*, mit dem identischem Erwartungswert $p_{ij}^{(0)}$. Deren Summe

$$\sum_{k=1}^{n} X_{ij}^{(k)}$$

ergibt die Häufigkeit der Zelle *ij*, sodass der Erwartungswert der Summenvariablen

$$\mathrm{E}(\sum_{k=1}^{n} X_{ij}^{(k)}) = \sum_{k=1}^{n} \mathrm{E}(X_{ij}^{(k)}) = \sum_{k=1}^{n} p_{ij}^{(0)} = n \cdot p_{ij}^{(0)}$$

die erwartete Häufigkeit der Zelle *ij* liefert.

In Pearsons Statistik werden die Abweichungen der beobachteten Häufigkeiten von den geschätzten erwarteten Häufigkeiten unter der H_0 quadriert. Durch das Quadrieren werden größere Diskrepanzen besonders stark gewichtet. Jede quadrierte Abweichung wird außerdem *normiert*, indem sie durch die erwartete Häufigkeit unter der H_0 dividiert wird. Steht etwa dem erwarteten Wert 5 die beobachtete Häufigkeit 15 gegenüber, so resultiert die quadrierte und normierte Diskrepanz 20:

$$\frac{(15-5)^2}{5} = 20$$

Dieselbe Abweichung einer beobachteten Häufigkeit 2010 vom erwarteten Wert 2000 erbringt jedoch sinnvollerweise nur eine quadrierte und normierte Diskrepanz von 0,05:

$$\frac{(2010-2000)^2}{2000} = 0{,}05$$

Der χ^2_{P}-Wert ist offenbar, wie es in Abschnitt 8.1.3 von einer Teststatistik gefordert wird, indikativ für Abweichungen der Stichprobendaten von der Nullhypothese. Mit $\frac{n_{ij}}{n}$ als geschätzter Wahrscheinlichkeit $\hat{p}_{ij}^{(1)}$ der Zelle *ij* unter der Alternativhypothese (beliebige Multinomialverteilung der Häufigkeiten in den *z*·*s* Zellen) und

$$\frac{m_{ij}}{n} = \frac{n_{i.} \cdot n_{.j}}{n^2} = \frac{n_{i.}}{n}\frac{n_{.j}}{n} = \hat{p}_{i.} \cdot \hat{p}_{.j}$$

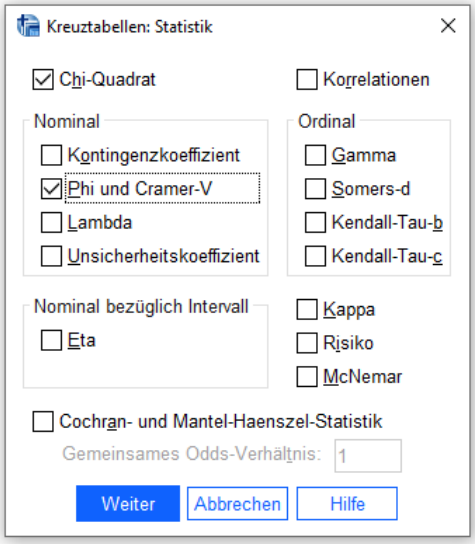
als geschätzter Wahrscheinlichkeit $\hat{p}_{ij}^{(0)}$ der Zelle *ij* unter der Nullhypothese (siehe oben) zeigt sich ein enger Bezug zwischen Pearsons χ^2_{P}-Prüfgröße und dem Effektstärkeindex *W* (vgl. Abschnitt 14.1):

$$\chi^2_P = \sum_{i=1}^{z}\sum_{j=1}^{s}\frac{(n_{ij}-m_{ij})^2}{m_{ij}} = n\sum_{i=1}^{z}\sum_{j=1}^{s}\frac{(\frac{n_{ij}}{n}-\frac{m_{ij}}{n})^2}{\frac{m_{ij}}{n}} = n\sum_{i=1}^{z}\sum_{j=1}^{s}\frac{(\hat{p}_{ij}^{(1)}-\hat{p}_{ij}^{(0)})^2}{\hat{p}_{ij}^{(0)}} = n\hat{W}^2$$

Die aus einer Stichprobe berechnete Prüfgröße χ^2_P ist also im Kern ein Schätzer für die quadrierte Effektstärke in der Population und damit bestens geeignet, eine Abweichung der wahren Populationsverteilung von der Nullhypothese aufzudecken.

Außerdem erfüllt die χ^2_P-Teststatistik nach Pearson auch die Verteilungsbedingung aus Abschnitt 8.1.3, wenn auch nur approximativ. Unter der Nullhypothese ist die χ^2_P-Statistik asymptotisch, d. h. für $n \to \infty$, χ^2-verteilt mit $df = (z - 1) \cdot (s - 1)$ Freiheitsgraden.[1] Für unsere Kreuztabelle zur Studienfachpräferenz - Hypothese erhalten wir also: $df = 1 \cdot 5 = 5$. Folglich kann für Pearsons χ^2_P-Statistik eine empirische Überschreitungswahrscheinlichkeit berechnet und nach den Regeln aus Abschnitt 8.1 ein Signifikanztest durchgeführt werden.

In SPSS wird die χ^2_P-Statistik samt Signifikanztest mit dem Kontrollkästchen **Chi-Quadrat** in der **Kreuztabellen**-Subdialogbox **Statistik** angefordert:



Zur Beurteilung der empirischen Effektstärke wählen wir zusätzlich **Phi und Cramer-V** (siehe Abschnitt 14.4.2).

Wir erhalten die folgenden Testergebnisse:

[1] In diesem Satz treten zwei Symbole mit ähnlicher Gestalt aber verschiedener Bedeutung auf: χ^2_P steht für eine Prüfgröße, mit χ^2 ist hingegen eine theoretische Verteilung gemeint.

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)
Chi-Quadrat nach Pearson	18,191 [a]	5	,003
Likelihood-Quotient	18,570	5	,002
Zusammenhang linear-mit-linear	3,197	1	,074
Anzahl der gültigen Fälle	283		

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 17,68.

Es ergibt sich ein χ^2_P-Wert von 18,191, der bei df = 5 unter der H_0 eine Überschreitungswahrscheinlichkeit (**Asymptotische Signifikanz**) von ca. 0,003 besitzt. Ein χ^2_P - Wert ≥ 18,19 bei df = 5 ist also unter der H_0 wenig wahrscheinlich. Insbesondere ist die empirisch ermittelte Überschreitungswahrscheinlichkeit deutlich kleiner als die üblicherweise akzeptierte Irrtumswahrscheinlichkeit erster Art von $\alpha = 0{,}05$. Folglich entscheidet sich der χ^2_P - Test klar gegen die H_0. In Abschnitt 8.1 wurde dieses Argumentationsmuster der Inferenzstatistik erläutert.

Neben der χ^2_P-Statistik nach Pearson berechnet SPSS noch die alternative Prüfgröße χ^2_{LQ}, die auf dem **Likelihood-Quotienten - Prinzip** basiert. Letztere ist unter der H_0 ebenfalls asymptotisch, d. h. für $n \to \infty$, χ^2 - verteilt mit $df = (z\text{-}1)\cdot(s\text{-}1)$ Freiheitsgraden, und trotz unterschiedlicher Herleitungen sind die beiden Statistiken asymptotisch äquivalent, d. h. mit wachsender Stichprobengröße werden sie immer ähnlicher. Während bei größeren Stichproben wegen der asymptotischen Äquivalenz die Entscheidung zwischen den beiden Prüfgrößen beliebig ist, sprechen einige Befunde dafür, bei kleineren Stichproben die χ^2_P-Statistik nach Pearson wegen der besseren Verteilungsapproximation zu bevorzugen (siehe z. B. Hartung 1989, S. 439). Damit ist es also sinnvoll, die χ^2_P-Statistik nach Pearson grundsätzlich gegenüber der Likelihood-Quotienten - Prüfgröße zu bevorzugen. SPSS liefert stets beide Prüfgrößen. In unserem Fall sind die Unterschiede geringfügig und für die Testentscheidung irrelevant.

Die Pearson- und die Likelihood-Quotienten - Statistik zur Beurteilung der Unabhängigkeits- bzw. Homogenitätshypothese sind nur **asymptotisch**, d. h. für $n \to \infty$, χ^2-verteilt. Für die Zulässigkeit der zugehörigen Hypothesentests setzt man üblicherweise voraus, dass alle **erwarteten** Häufigkeiten m_{ij} mindestens gleich 5 sind. SPSS protokolliert die minimale erwartete Häufigkeit in einer Fußnote zur Tabelle mit den **Chi-Quadrat-Tests**. In unserem Fall beträgt sie 17,68, sodass keine Einwände gegen Tests auf Basis der χ^2_P - oder χ^2_{LQ}-Statistik bestehen.

Manche Autoren formulieren etwas schwächere Voraussetzungen für die erwarteten Häufigkeiten. Siegel (1976, S. 107) verlangt z. B. für χ^2_P-Tests mit $df > 1$, dass die beiden folgenden Bedingungen erfüllt sind:

- Weniger als 20% der Zellen haben eine erwartete Häufigkeit kleiner als 5.
- Keine Zelle hat eine erwartete Häufigkeit kleiner als 1.

Neben den beiden Statistiken zur Prüfung der Unabhängigkeits- bzw. Homogenitätshypothese liefert SPSS unter der Bezeichnung **Zusammenhang linear-mit-linear** auch noch den χ^2_{MH}-Wert nach **Mantel-Haenszel** samt Überschreitungswahrscheinlichkeit. Hier geht es um eine Entscheidung zwischen den Hypothesen:

H_0: Es existiert keine lineare Beziehung zwischen den beiden Variablen.

versus

H_1: Es existiert eine lineare Beziehung zwischen den beiden Variablen.

Die Prüfgröße basiert auf der Produkt-Moment - Korrelation zwischen den beiden Variablen:

$$\chi^2_{MH} := r^2 (n-1)$$

Üblicherweise gilt der Mantel-Haenszel - Test als anwendbar, sofern beide Variablen mindestens ordinale Skalenqualität besitzen (siehe z. B. Norušis 2011b). Da wir zwei nominalskalierte Variablen betrachten, ist diese Statistik bei unserer Tabelle sinnlos.

14.4.2 Schätzung der Effektstärke

Zur Beurteilung der empirischen Effektstärke haben wir in der Kreuztabellen-Subdialogbox **Statistik** (siehe Abschnitt 14.4.1) **Phi und Cramer-V** angefordert.

Der bei (2 × 2) - Tabellen als Zusammenhangsmaß taugliche **Phi-Koeffizient** steht in folgendem Zusammenhang zu Pearsons χ^2_P-Statistik (siehe z. B. Bortz & Schuster 2010, S. 174):

$$\phi = \sqrt{\frac{\chi^2_P}{n}}$$

Nach einer Rechnung aus Abschnitt 14.4.1 ist ϕ damit ein Schätzer für die Effektstärke W (vgl. Abschnitt 14.1):

$$\phi = \hat{W}$$

W und ϕ haben den maximalen Wert

$$\sqrt{q-1}$$

mit

$$q := \text{Min}(z, s)$$

Bei $q > 2$ können W und ϕ also größer als 1 werden (vgl. Cohen 1988, Abschnitt 7.2). In der Definition von Cramers V wird folgendermaßen für den Maximalwert 1 gesorgt:

$$V := \sqrt{\frac{\chi^2_P}{n(q-1)}} = \sqrt{\frac{\chi^2_P}{n}} \frac{1}{\sqrt{q-1}} = \hat{W} \frac{1}{\sqrt{q-1}}$$

Bei der FB-GESCHL - Analyse (mit Min(z, s) = 2) ist Cramers V identisch mit dem Phi-Koeffizienten, und wir erhalten den Wert 0,254:

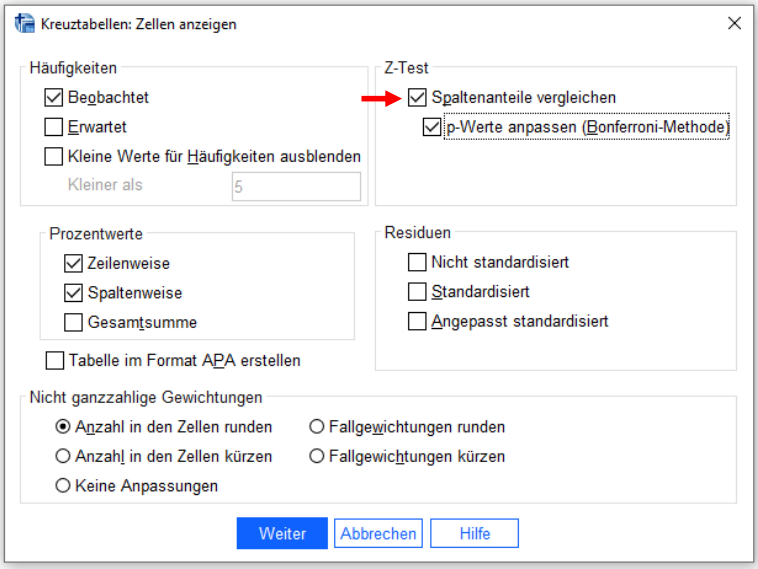
Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	,254	,003
	Cramer-V	,254	,003
Anzahl der gültigen Fälle		283	

Er ist nicht weit entfernt vom Wert 0,3, den wir in Abschnitt 14.1 für den Effektstärkeindex W angenommen haben.

14.4.3 Einzelvergleiche der Spaltenanteile

Nachdem die Nullhypothese identischer Frauenanteile in den Fachbereichen der Universität Trier verworfen werden konnte, sind paarweise Einzelvergleiche von Interesse. SPSS bietet solche Einzelvergleiche der Spaltenanteile in der Subdialogbox **Zellen** an:



Durch eine Bonferroni-Adjustierung wird dafür gesorgt, dass trotz der Durchführung mehrerer Tests (z. B. 15 Paarvergleiche bei 6 Fachbereichen) die Wahrscheinlichkeit, bei Gültigkeit sämtlicher Nullhypothesen einen oder mehrere α-Fehler zu begehen, unter 5% bleibt. Allerdings ist das Bonferroni-Verfahren sehr konservativ, z. B. weil auf die Bonferroni-Holm - Verfeinerung verzichtet wird.[1]

Im Beispiel

Geschlecht * Fachbereiche an der Universität Trier Kreuztabelle

		Fachbereiche an der Universität Trier						
		I	II	III	IV	V	VI	Gesamt
Frauen	Anzahl	29 a	26 a	18 a, b	22 b	26 a, b	23 a, b	144
	% von Geschlecht	20,1%	18,1%	12,5%	15,3%	18,1%	16,0%	100,0%
	% von Fachbereich	63,0%	66,7%	50,0%	31,0%	54,2%	53,5%	50,9%
Männer	Anzahl	17 a	13 a	18 a, b	49 b	22 a, b	20 a, b	139
	% von Geschlecht	12,2%	9,4%	12,9%	35,3%	15,8%	14,4%	100,0%
	% von Fachbereich	37,0%	33,3%	50,0%	69,0%	45,8%	46,5%	49,1%
Gesamt	Anzahl	46	39	36	71	48	43	283
	% von Geschlecht	16,3%	13,8%	12,7%	25,1%	17,0%	15,2%	100,0%
	% von Fachbereich	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Jeder tiefgestellte Buchstabe gibt eine Teilmenge von Fachbereiche an der Universität Trier Kategorien an, deren Spaltenanteile sich auf dem ,05-Niveau nicht signifikant voneinander unterscheiden.

sind signifikante Unterschiede für folgende Fachbereichspaare nachzuweisen:

[1] Siehe z. B.: https://en.wikipedia.org/wiki/Holm–Bonferroni_method

- I vs. IV
- II vs. IV

Für diese Fachbereichspaare haben die Häufigkeiten zu den weiblichen oder männlichen Zellen *keinen* gemeinsamen Index.

14.4.4 Exakte Tests

Für die (2 × 2) - Kreuztabelle gibt es seit Jahrzehnten mit dem **exakten Test von Fisher** eine gute Alternative zu den approximativen χ^2 - Tests. Wie sein Name sagt, kommt Fishers Test ohne Approximation aus und ist daher bei *jeder* Stichprobe anwendbar. Erfreulicherweise bietet SPSS exakte Tests auch für beliebige ($z \times s$) - Kreuztabellen, sofern das Zusatzmodul **Exact Tests** im Lizenzumfang enthalten ist.

Eine ausführliche Beschreibung (Baltes-Götz 1998) der statistischen Verfahren, die durch das SPSS-Zusatzmodul **Exact Tests** implementiert werden, ist auf dem Webserver der Universität Trier zu finden:

https://www.uni-trier.de/?id=22571

Allerdings sind die asymptotischen Verfahren keinesfalls obsolet, weil der exakte Test für ($z \times s$) - Kreuztabellen wegen seines hohen Rechenaufwands bei großen Stichproben nicht durchführbar ist. Insgesamt steht für die meisten Situationen ein angemessenes Verfahren zur Verfügung:

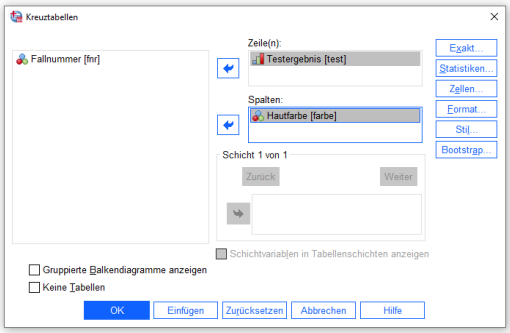
- Wenn die Anwendbarkeitskriterien für die asymptotischen Verfahren erfüllt sind (siehe Abschnitt 14.4.1), dann sollten Sie den Pearson-Test verwenden.
- Anderenfalls sollten Sie einen exakten Test versuchen.

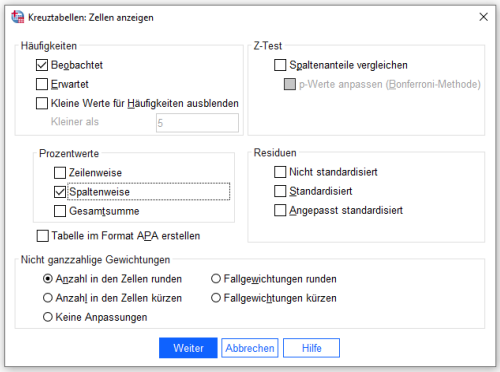
Wenn bei einer Kreuztabelle einerseits die Minimalanforderungen an die erwarteten Häufigkeiten nicht erfüllt sind, und andererseits der exakte Test aufgrund des insgesamt zu großen Stichprobenumfangs scheitert, dann können Sie die verantwortlichen schwach besetzten Zeilen bzw. Spalten entweder löschen oder miteinander bzw. mit anderen Zeilen/Spalten zusammenlegen.

In einem Anwendungsbeispiel sollen Daten aus dem SPSS-Handbuch zum Modul **Exact Tests** (Mehta & Patel 2010, S. 2) analysiert werden. Es handelt sich um Prüfungsergebnisse weißer, schwarzer, asiatischer und hispanoider Feuerwehrbewerber in einer amerikanischen Kleinstadt. Diese Kreuztabelle

		Hautfarbe				
		Weiß	Schwarz	Asiatisch	Mittel- und Südamerika	Gesamt
Anzahl	Bestanden	5	2	2	0	9
	Unklar	0	1	0	1	2
	Durchgefallen	0	2	3	4	9
	Gesamt	5	5	5	5	20
Prozent	Bestanden	100,0%	40,0%	40,0%	0,0%	45,0%
	Unklar	0,0%	20,0%	0,0%	20,0%	10,0%
	Durchgefallen	0,0%	40,0%	60,0%	80,0%	45,0%
	Gesamt	100,0%	100,0%	100,0%	100,0%	100,0%

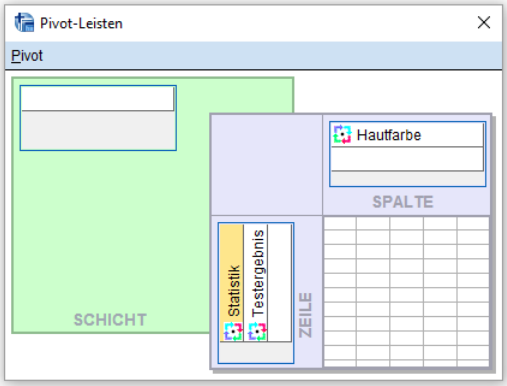
wurde mit den folgenden Dialogboxen angefordert:





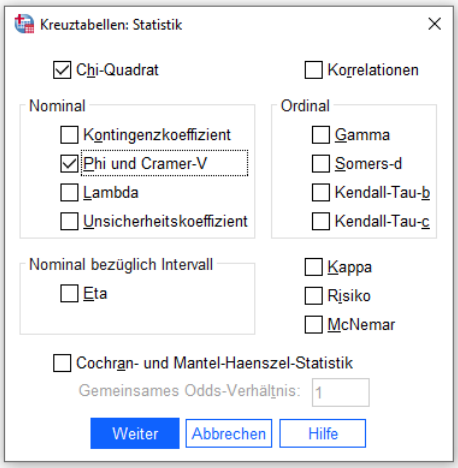
Der Tabellenrohling wurde per Doppelklick im Pivot-Editor (vgl. Abschnitt 10.4) geöffnet und folgendermaßen modifiziert:

- Für die beiden Zeilendimensionen (Testergebnis, Statistik) wurde per Pivot-Werkzeug die Schachtelungsordnung geändert:



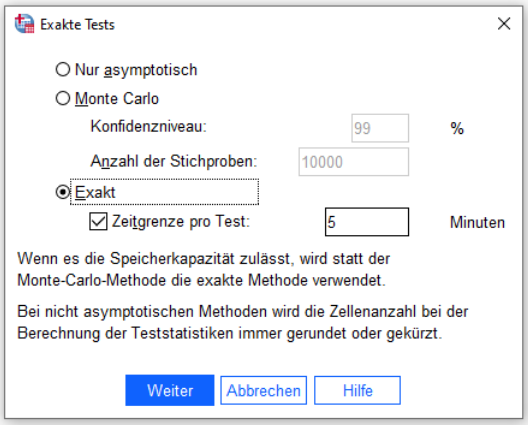
- Die Gruppierungszelle zur **Testergebnis**-Dimension wurde entfernt (Kontextmenü-Item **Gruppierung aufheben**).
- Die untere Statistik-Kategorie hat die neue Beschriftung *Prozent* erhalten.
- Die Titelzeile wurde gelöscht

Um χ^2 - Tests zur Nullhypothese (Unabhängigkeit der Prüfungsergebnisse von der Hautfarbe) anzufordern, klickt man in der Dialogbox zur Kreuztabellenanalyse auf den **Statistiken**-Schalter und markiert dann im folgenden Subdialog das Kontrollkästchen **Chi-Quadrat**:



Im selben Dialog wird über das Kontrollkästchen **Phi und Cramer-V** eine Schätzung der Effektstärke angeboten.

Nach einem Mausklick auf den **Exakt**-Schalter in der Dialogbox zur Kreuztabellenanalyse kann man in der folgenden Subdialogbox die **exakte** Testmethode wählen:



Daraufhin erhält man neben den approximativen Ergebnissen auch exakte Überschreitungswahrscheinlichkeiten für die Pearson- und die Likelihood-Quotienten - Teststatistik. Außerdem führt SPSS noch eine Verallgemeinerung des exakten Tests von Fisher durch, der in seiner klassischen Variante auf (2×2) - Tabellen beschränkt ist:

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)	Exakte Sig. (zweiseitig)	Exakte Sig. (einseitig)	Punkt-Wahrschein-lichkeit
Pearson-Chi-Quadrat	11,556 [a]	6	,073	,040		
Likelihood-Quotient	15,673	6	,016	,040		
Exakter Test nach Fisher-Freeman-Halton	11,239			,040		
Zusammenhang linear-mit-linear	8,276 [b]	1	,004	,004	,002	,001
Anzahl der gültigen Fälle	20					

a. 12 Zellen (100,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist ,50.

b. Die standardisierte Statistik ist 2,877.

Die approximativen χ^2 - Unabhängigkeitstests (Pearson und Likelihood-Quotient) sind nicht anwendbar, weil in allen 12 Zellen die erwartete Häufigkeit kleiner als 5 ist. Wer dieses Problem ignoriert, andererseits aber weiß, dass der Pearson-Test gegenüber dem Likelihood-Quotienten - Test speziell bei kleinen Stichproben wegen der besseren Verteilungsapproximation zu bevorzugen ist, gelangt zu einer falschen Testentscheidung. Der asymptotische Pearson-χ^2 - Test empfiehlt durch eine Überschreitungswahrscheinlichkeit von 0,073, die Nullhypothese beizubehalten. Die exakte Überschreitungswahrscheinlichkeit zur Pearson-Prüfgröße beträgt hingegen 0,04, was zur Ablehnung der Nullhypothese führt.

Die exakten Überschreitungswahrscheinlichkeiten zu den drei in Frage kommenden Signifikanztests müssen nicht in jedem Fall übereinstimmen. Nachträglich die kleinste Überschreitungswahrscheinlichkeit zu wählen, ist *nicht* zulässig. Wer den approximativen Pearson-Test gemäß obiger Empfehlung routinemäßig bei der allgemeinen $(z \times s)$ - Kreuztabelle verwendet, sofern er zulässig

ist, sollte dessen Prüfgröße auch bei der exakten Berechnung der Überschreitungswahrscheinlichkeit zugrunde legen.

Dass es trotz der winzigen Stichprobe zu einem signifikanten Ergebnis gereicht hat, liegt nicht nur am exakten Testverfahren, sondern auch an der erheblichen Effektstärke. Mit dem Phi-Koeffizienten erhalten wir eine geschätzte Effektstärke von 0,76:

Symmetrische Maße

		Wert	Näherungsweise Signifikanz	Exakte Signifikanz
Nominal- bzgl. Nominalmaß	Phi	,760	,073	,040
	Cramer-V	,537	,073	,040
Anzahl der gültigen Fälle		20		

Diese ist nach Cohen (1988, S. 227) als *sehr groß* zu beurteilen (vgl. Abschnitt 14.1).

14.4.5 Besonderheiten bei (2 × 2) - Tabellen

14.4.5.1 Ein klarer Fall für den exakten Test von Fisher

Im Spezialfall der (2 × 2) - Tabelle ist Fishers Test nicht nur *exakt*, sondern er besitzt sogar unter den sogenannten *unverfälschten* Tests die besten Güteeigenschaften. Daher sollten Sie in dieser Situation grundsätzlich Fishers Test verwenden. Die oben beschriebenen Rechenzeitprobleme bei exakten Tests für allgemeine ($z \times s$) - Kreuztabellen treten bei Fischers Test für die (2 × 2) - Tabelle *nicht* auf.

Für eine Teststärkeanalyse mit dem Programm G*Power 3.1 (vgl. Abschnitt 2.3.2) wählt man bei Fishers exaktem Test für die (2 × 2) - Tabelle:

- **Test family**: **Exact**
- **Statistical test**: **Proportions: … (Fisher's exact test)**

14.4.5.2 Gerichtete Hypothesen

Bei einer (2 × 2) - Tabelle lässt sich auch eine *gerichtete* (einseitige) Hypothese über den Zusammenhang zwischen den beiden Merkmalen formulieren. Wenn wir uns z. B. beim Vergleich der Frauenanteile unter den Studierenden der Universität Trier auf die Fachbereiche III und IV beschränken, können wir das folgende Testproblem untersuchen:

H_0: Der Frauenanteil ist im FB IV mindestens genauso groß wie im FB III.

H_1: Der Frauenanteil ist im FB IV kleiner als im FB III.

Aus den (z. B. per Filterbedingung, vgl. Abschnitt 12) eingeschränkten Beispieldaten (Datei **fbgeschl.sav**) erhalten wir folgende Ergebnisse:

Geschlecht * Fachbereiche an der Universität Trier Kreuztabelle

		Fachbereiche an der Universität Trier		
		III	IV	Gesamt
Frauen	Anzahl	18	22	40
	% von Geschlecht	45,0%	55,0%	100,0%
	% von Fachbereich	50,0%	31,0%	37,4%
Männer	Anzahl	18	49	67
	% von Geschlecht	26,9%	73,1%	100,0%
	% von Fachbereich	50,0%	69,0%	62,6%
Gesamt	Anzahl	36	71	107
	% von Geschlecht	33,6%	66,4%	100,0%
	% von Fachbereich	100,0%	100,0%	100,0%

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)	Exakte Sig. (zweiseitig)	Exakte Sig. (einseitig)
Pearson-Chi-Quadrat	3,689 [a]	1	,055		
Kontinuitätskorrektur [b]	2,922	1	,087		
Likelihood-Quotient	3,643	1	,056		
Exakter Test nach Fisher				,061	,044
Zusammenhang linear-mit-linear	3,655	1	,056		
Anzahl der gültigen Fälle	107				

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 13,46.

b. Wird nur für eine 2x2-Tabelle berechnet

Wie wir bereits wissen, beträgt der Frauenanteil im FB III 50% und im FB IV 31%; die deskriptiven Statistiken fallen also klar im Sinne der Alternativhypothese aus. Der nach Abschnitt 14.4.5.1 zu verwendende exakte Test von Fisher liefert für die *zweiseitige* Fragestellung eine Überschreitungswahrscheinlichkeit von 0,061, sodass die Nullhypothese beibehalten werden müsste. Bei *einseitiger* Testung erhalten wir jedoch eine Überschreitungswahrscheinlichkeit von 0,044, sodass die Nullhypothese verworfen werden kann.

Beachten Sie abschließend noch, dass sich bei Fishers Test die einseitige Überschreitungswahrscheinlichkeit *nicht* durch Halbieren der zweiseitigen ergibt. Die in Abschnitt 8.1 für den Spezialfall des t-Tests angegebene Regel zur Berechnung der einseitigen Überschreitungswahrscheinlichkeit aus der zweiseitigen darf also nicht generalisiert werden. Allerdings können Sie sich darauf verlassen, dass SPSS neben der zweiseitigen Überschreitungswahrscheinlichkeit auch die einseitige ausgeben wird, wenn die beiden nicht in einer einfachen Beziehung zueinander stehen.

14.4.5.3 Kontinuitätskorrektur nach Yates

Bei (2×2) - Tabellen berechnet SPSS traditionell auch eine χ^2_Y-Größe mit Kontinuitätskorrektur nach Yates. Sie soll bei kleineren Stichproben der Pearson-χ^2_P - Statistik überlegen sein. Gemäß Abschnitt 14.4.5.1 ist sie allerdings irrelevant, weil in der (2×2) - Situation Fishers exakter Tests in jedem Fall zu bevorzugen ist.

15 Fälle gewichten

Per Voreinstellung bezieht SPSS bei statistischen Auswertungen *alle* Fälle mit dem Gewicht 1 ein. In Abschnitt 12 haben Sie schon eine Möglichkeit kennengelernt, Fälle aufgrund von Filterkriterien temporär oder permanent aus der Arbeitsdatei ausschließen. Nun erfahren Sie, wie man die Fälle individuell gewichtet, sodass sie bei statistischen Analysen unterschiedlichen Einfluss auf die Ergebnisse haben.

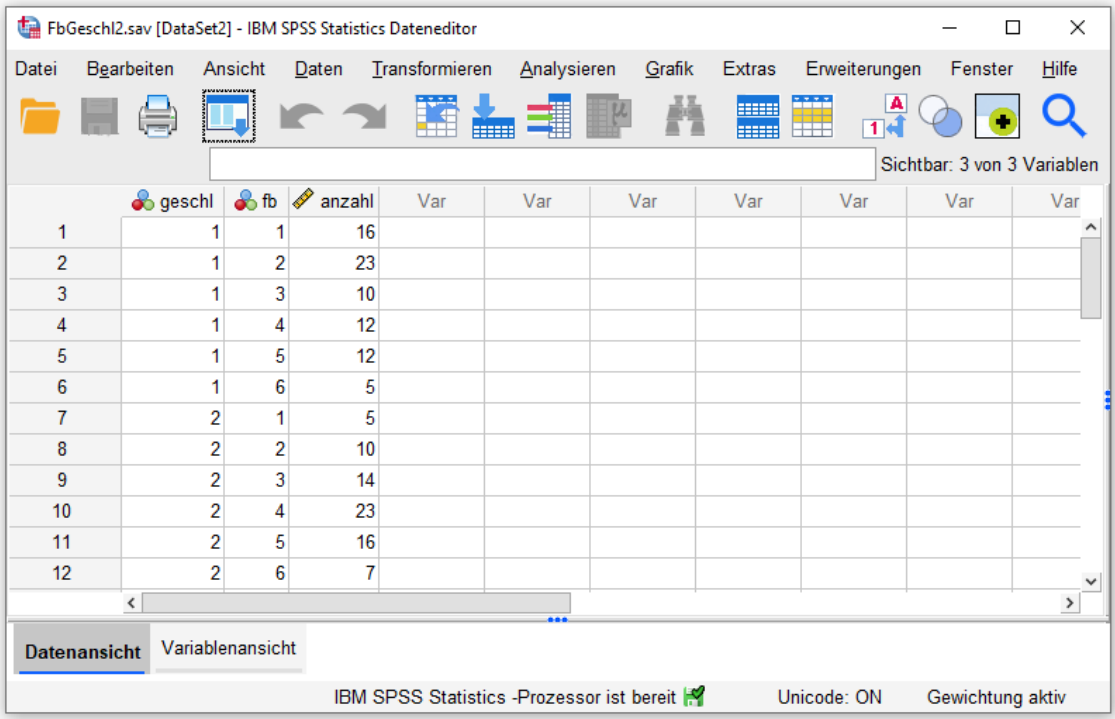
15.1 Beispiel

Die Möglichkeit, von 1 verschiedene Fallgewichte zu verwenden, d. h. z. B. einem Fall des Gewicht 16 zuzuschreiben und so zu tun, als seien 16 Fälle mit genau gleichen Variablenausprägungen vorhanden, erscheint zunächst sinnlos. Aber erinnern wir uns an die (Geschlecht × Fachbereich) - Kreuztabelle aus Abschnitt 14. Zur Verwendung in einer späteren Übungsaufgabe betrachten wir hier eine strukturell identische Tabelle, die auf einer anderen Zufallsstichprobe der Größe $n = 153$ beruht:

	Fachbereich					
Geschlecht	I	II	III	IV	V	VI
Frau	16	23	10	12	12	5
Mann	5	10	14	23	16	7

Um mit den in Abschnitt 14 behandelten χ^2 - Tests anhand dieser Stichprobendaten prüfen zu können, ob in den Fachbereichen die Geschlechtsverteilungen verschieden sind, brauchen Sie nach unserem bisherigen Kenntnisstand eine Arbeitsdatei, in der z. B. 16 Fälle mit dem Geschlecht 1 und dem Fachbereich 1 enthalten sind, 23 Fälle mit Geschlecht 1 und Fachbereich 2 usw. Wir haben jedoch lediglich die obige Tabelle zur Verfügung. Statt nun mühselig 153 Fälle im Dateneditor einzutippen, können wir von der Möglichkeit der Fallgewichtung folgendermaßen Gebrauch machen:

- Wir sorgen für ein leeres Datenblatt, z. B. über den Menübefehl
  Datei > Neu > Daten
  Dort definieren wir die Variablen GESCHL (Geschlecht), FB (Fachbereich) und ANZAHL.
- Jede Zelle der (Geschlecht × Fachbereich) - Kreuztabelle wird im SPSS-Datenblatt als *ein* Fall behandelt. Der erste Fall erhält z. B. für die drei Variablen GESCHL, FB und ANZAHL die Werte 1, 1 und 16:



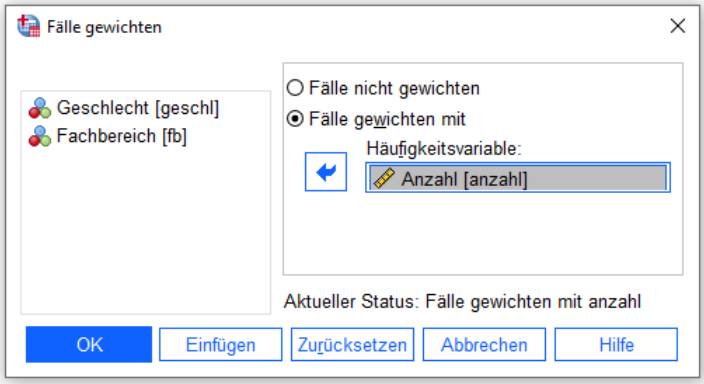
- Die Fälle werden mit der Variablen ANZAHL gewichtet. Damit tun wir z. B. so, als seien 16 Fälle mit dem Geschlecht 1 und dem Fachbereich 1 vorhanden gewesen. Aber das stimmt ja wirklich. Offenbar ist die Fallgewichtung doch nicht sinnlos.

Um eine Gewichtsvariable zu vereinbaren, rufen wir mit dem Menübefehl

Daten > Fälle gewichten

eine Dialogbox auf, die folgende Optionen anbietet:

- **Fälle nicht gewichten**
  Damit wird eine bestehende Gewichtung wieder aufgehoben.
- **Fälle gewichten mit**
  Die gewünschte Variable wird mit dem Transportschalter oder per Drag & Drop in die Position der **Häufigkeitsvariablen** gebracht, z. B.:



In der Dialogbox wird außerdem angezeigt, ob momentan eine Gewichtungsvariable vereinbart ist. Dieselbe Information erscheint auch in der Statuszeile des Datenfensters (siehe oben).

Beim Einsatz von Gewichtungsvariablen ist zu beachten:

- Zur Gewichtung kann natürlich nur eine *numerische* Variable verwendet werden; diese darf allerdings auch gebrochene Werte enthalten. Negative und fehlende Werte werden auf 0 gesetzt, d. h. die betroffenen Fälle werden nicht berücksichtigt, solange die Gewichtungsvariable aktiv ist.
- Ist beim Speichern der Arbeitsdatei eine Gewichtung aktiv, so wird diese abgespeichert und ist bei späterer Verwendung der Datendatei in Kraft.
- Bei der in diesem Abschnitt beschriebenen Anwendung der Gewichtungsoption wird dafür gesorgt, dass alle tatsächlich in der Studie vorhandenen Beobachtungen mit dem Gewicht 1 in die Kreuztabellenanalyse eingehen. Wenn die vorhandenen Beobachtungen individuelle Gewichte ($\neq 1$) erhalten, werden natürlich Signifikanztests erheblich beeinflusst. Auf jeden Fall muss dann die Summe der Gewichte gerade den Stichprobenumfang ergeben.

15.2 Übung

Prüfen Sie anhand der Daten aus der Tabelle am Anfang von Abschnitt 15.1 die Nullhypothese, dass die Merkmale Geschlecht und Fachbereich unabhängig sind.

16 Auswertung von Mehrfachwahlfragen

In Abschnitt 2.4.2.3 wurde betont, dass mit einer Mehrfachwahlfrage nicht etwa *ein* Merkmal mit mehreren Ausprägungen erfasst wird, das wohl durch manche Köpfe bzw. Alpträume spukt, sondern *mehrere* inhaltlich verwandte, dichotome Merkmale.

Grundsätzlich besteht kein Bedarf an speziellen Auswertungsverfahren für die mit Mehrfachwahlfragen erhobenen Variablen. Es ist allerdings gelegentlich sinnvoll, eine Häufigkeits- oder Kreuztabellenanalyse für *alle* Mitglieder einer Familie dichotomer Variablen (ob aus einer Mehrfachwahlfrage entstanden oder wie auch immer) in gleicher Form auszuführen. Für diese Situation bietet SPSS Rationalisierungsmöglichkeiten, die in diesem Kapitel vorgestellt werden. Außerdem kann SPSS für die mit einem sparsamen Set aus kategorialen Variablen erfassten dichotomen Merkmale Häufigkeits- und Kreuztabellenanalysen durchführen, ohne dass zuvor die dichotomen Variablen zu den Merkmalen (durch Datentransformationsanweisungen) konstruiert werden müssen (vgl. Abschnitt 16.4).

16.1 Mehrfachantwortsets definieren

Im Teil 3a unseres Fragebogens haben die Teilnehmer für fünf konkrete Motive, das statistische Praktikum mit SPSS zu besuchen, und eine Restkategorie alles zutreffende angekreuzt. Es liegt nahe, eine kompakte Tabelle zu erstellen, aus der für die einzelnen Motive abzulesen ist, wie häufig sie gewählt worden sind. Man könnte die Zustimmungsfrequenzen bei den Motiv-Variablen von der Häufigkeitsanalyse (**Analysieren > Deskriptive Statistiken > Häufigkeiten**) bestimmen lassen und die sechs resultierenden Häufigkeitstabellen (eine pro Variable) manuell zusammenfassen. Eine zeitökonomischere Alternative besteht darin, die Kombitabelle von SPSS erstellen zu lassen. Für die Manuskriptstichprobe resultiert die folgende Tabelle:

		Anzahl	Prozent
Motive zur Kursteilnahme	Eigene Studie	23	76,7%
	Bewerbung um Stelle	1	3,3%
	Bewerbung um HIWI-Job	1	3,3%
	Interesse an der EDV	5	16,7%
	Interesse an Statistik	10	33,3%
	Andere Motive	1	3,3%

Es zeigt sich etwa, dass 23 Personen (= 76,7% von den 30 Fällen mit gültigen Werten bei den Motiv-Variablen) dem ersten Motiv zugestimmt haben. Ein Fall, auf den wir später noch eingehen müssen, fand keines von den fünf konkreten Motive für sich passend und markierte die Restkategorie (*Andere Motive*).

Zum Erstellen der obigen Tabelle wird eine **Variablengruppe** benötigt, die zuvor definiert werden muss. Wählen Sie dazu den Menübefehl:[1]

Daten > Mehrfachantwortensets definieren

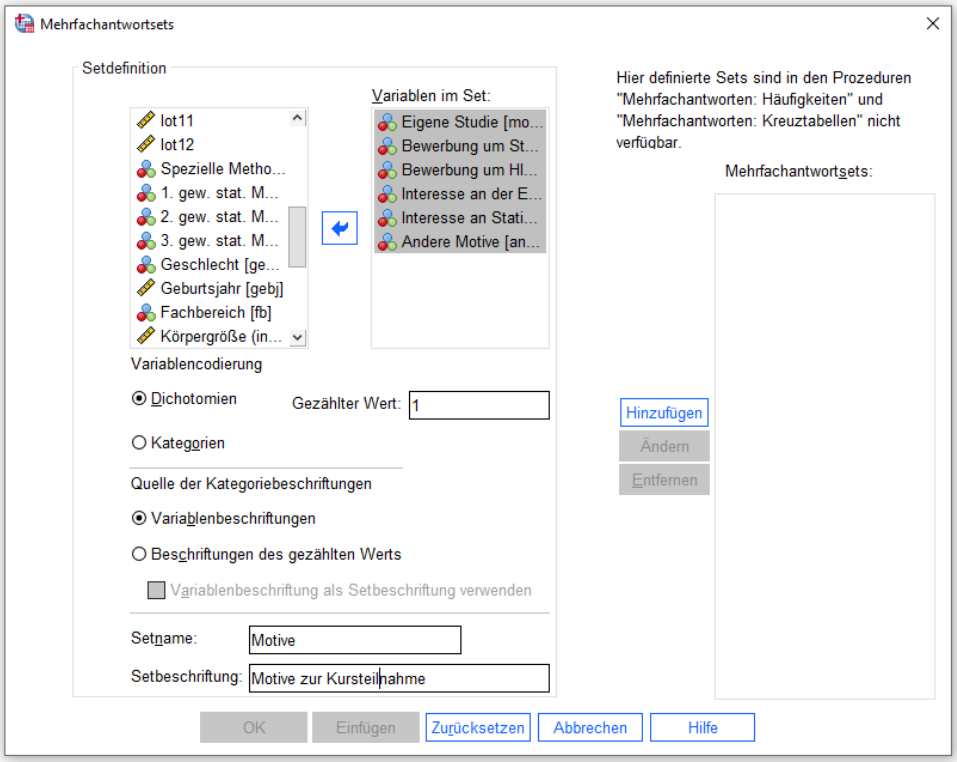
In der nun erscheinenden Dialogbox sind folgende Aktionen nötig:

[1] Alternativ können Sie auch den folgenden Menübefehl verwenden:

Analysieren > Tabellen > Mehrfachantwortsets

- Befördern Sie die Variablen MOTIV1 bis MOTIV5 sowie ANDERE in die Liste der **Variablen im Set**.
- Wählen Sie im Rahmen **Variablencodierung** die Option **Dichotomien** mit dem **gezählten Wert** 1.
- Vereinbaren Sie den **Setnamen** *Motive* und die **Setbeschriftung** *Motive zur Kursteilnahme*. Die letztlich entstehende Variablengruppe enthält daraufhin den Namen $MOTIVE.

Danach müsste Ihre Dialogbox ungefähr so aussehen:



Nehmen Sie mit **Hinzufügen** das neue Set in die Liste der **Mehrfachantwortensets** auf, und quittieren Sie die Dialogbox mit **OK**.

Auf die beschriebene Weise definierte Mehrfachantwortensets werden in der Arbeitsdatei gespeichert und können in die zugeordnete Datendatei gesichert werden, sodass sie beim späteren Öffnen der Datei wieder zur Verfügung stehen.

Bei der Set-Definition kommt das SPSS-Kommando MRSETS zum Einsatz, das mit Hilfe der Dialogbox **Mehrfachantwortensets definieren** über den Schalter **Einfügen** erzeugt werden kann, z. B.:

```
MRSETS
  /MDGROUP NAME=$Motive LABEL='Motive zur Kursteilnahme'
   CATEGORYLABELS=VARLABELS
   VARIABLES=motiv1 motiv2 motiv3 motiv4 motiv5 andere VALUE=1
  /DISPLAY NAME=[$Motive].
```

Bei wichtigen Sets sollte das definierende MRSETS-Kommando in das Transformationsprogramm zum Erstellen der Fertigdatendatei aufgenommen werden (vgl. Abschnitte 7.1.1 und 7.7).

Über den Menübefehl

Analysieren > Mehrfachantworten > Variablensets definieren

ist noch eine ältere Möglichkeit zur Set-Definition verfügbar, die wesentliche Nachteile gegenüber der oben beschriebenen Lösung hat:

- Die Set-Definitionen gehen beim Schließen des zugehörigen Datenblatts verloren, können also *nicht* in einer Datendatei gespeichert werden.
- Die Set-Definitionen können nur eingeschränkt verwendet werden:
  - über den Menübefehl **Analysieren > Mehrfachantworten > Häufigkeiten**
  - über den Menübefehl **Analysieren > Mehrfachantworten > Kreuztabellen**
  - über das Kommando MULT RESPONSE

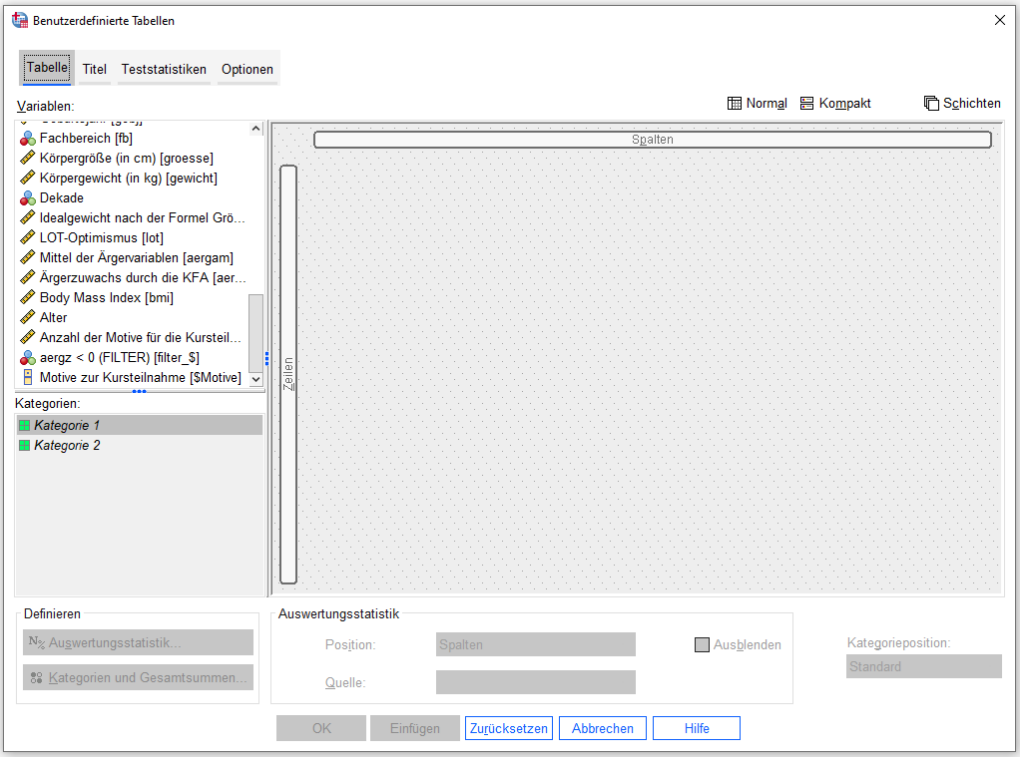
Die veraltete Option zur Set-Definition besitzt aus Traditions- bzw. Kompatibilitätsgründen eine prominente Position im Menüsystem von SPSS und wird leider in vielen Büchern ausschließlich beschrieben (siehe z. B. Action et al. 2009, S. 166ff; Wagner 2017, S. 32f).

16.2 Häufigkeitstabellen für Mehrfachantwortsets

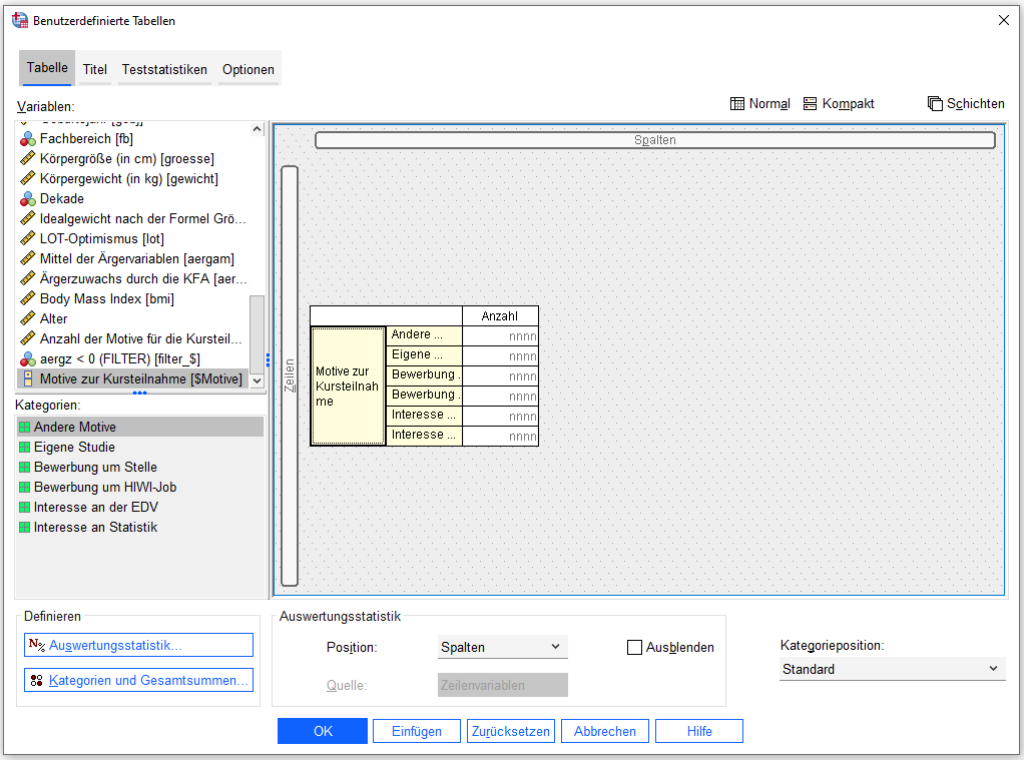
Unter Verwendung der per MRSETS-Kommando erzeugten Variablengruppe $MOTIVE lässt sich die in Abschnitt 16.1 präsentierte Tabelle mit den Häufigkeitsverteilungen der Set-Variablen über den Menübefehl

Analysieren > Tabellen > Benutzerdefinierte Tabellen

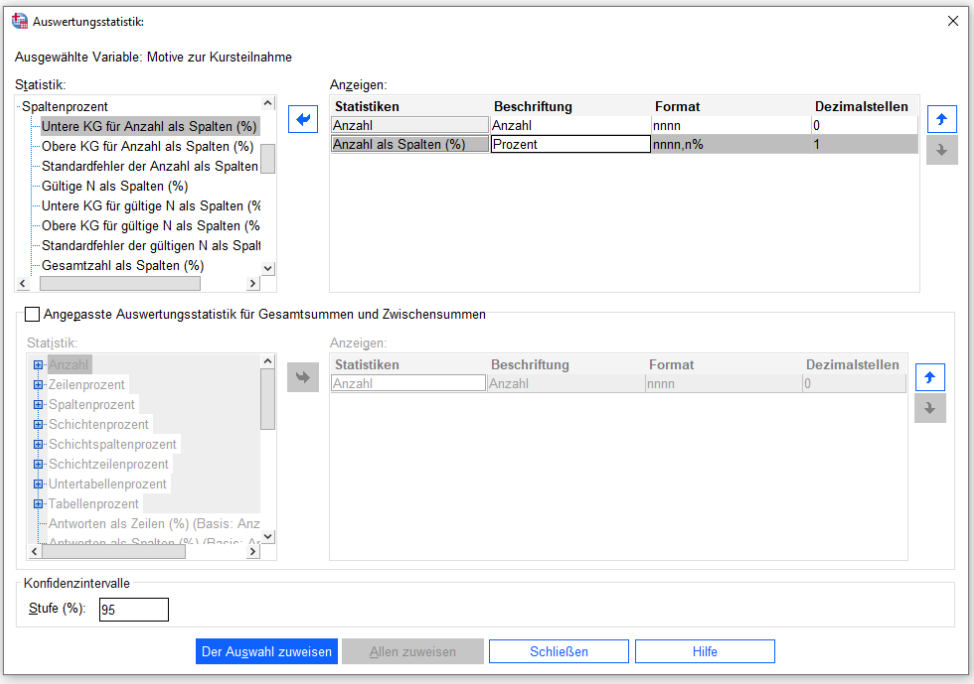
und den folgenden Dialog anfordern:



Wir befördern die unterhalb der Variablenliste zu findende Variablengruppe $MOTIVE per Drag & Drop auf die **Zeilen**-Ablagezone:



Um in der resultierenden Tabelle die voreingestellte **Anzahl**-Spalte durch eine Prozentspalte zu ergänzen, klicken wir auf den Schalter **Auswertungsstatistik** und befördern aus der **Statistik**-Liste das Element **Anzahl als Spalten%** in den **Anzeigen**-Bereich:



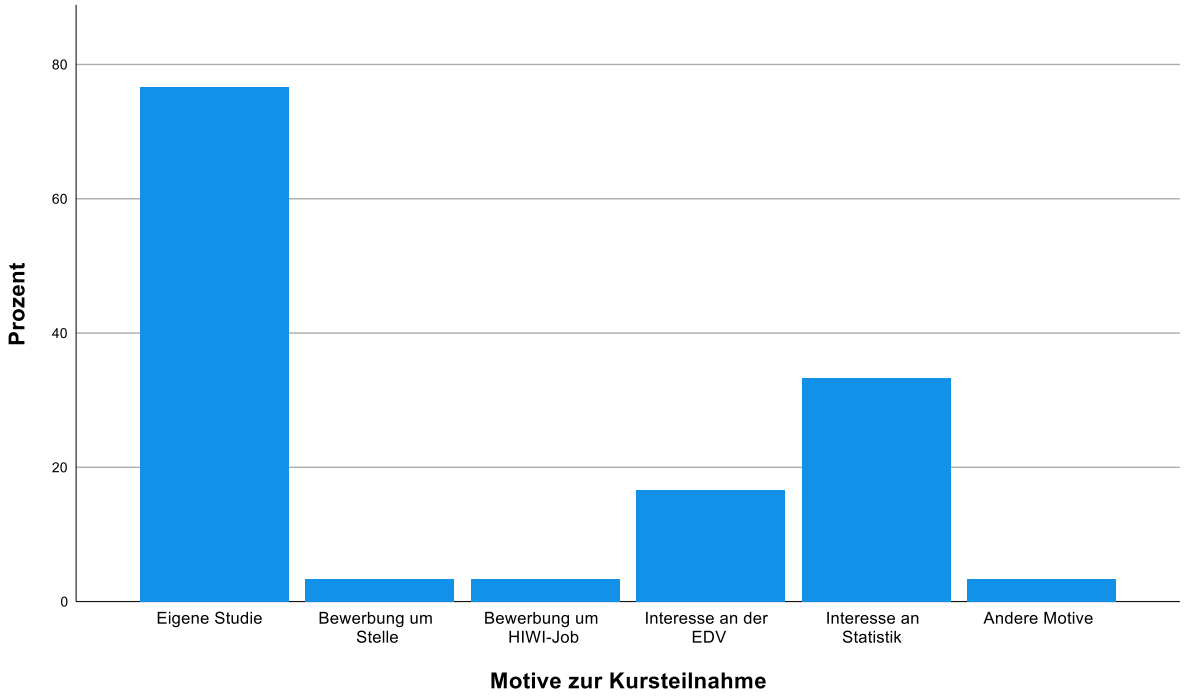
Wie ändern die vorgeschlagene **Beschriftung**, quittieren mit dem Schalter **Der Auswahl zuweisen** und **schließen** den Dialog **Auswertungsstatistik**. Wird anschließend der Dialog **Benutzerdefinierte Tabellen** mit **OK** quittiert, erscheint die fertige Tabelle im Ausgabefenster.

Entfernt man die Variable ANDERE zur Restkategorie der sonstigen Motive aus dem Set $MOTIVE, dann resultieren folgende Ergebnisse mit abweichenden Prozentwerten:

		Anzahl	Prozent
Motive zur Kursteilnahme	Eigene Studie	23	79,3%
	Bewerbung um Stelle	1	3,4%
	Bewerbung um HIWI-Job	1	3,4%
	Interesse an der EDV	5	17,2%
	Interesse an Statistik	10	34,5%

Des Rätsels Lösung ist eine SPSS-Eigenart bei der Analyse von Mehrfachantwortensets aus dichotomen Variablen: Als gültig werden nur solche Fälle betrachtet, die bei mindestens einer Set-Variablen den zu zählenden Wert besitzen (bei uns also die 1). Daher wird neben dem Fall 13 mit SYSMIS bei den Variablen MOTIV1 bis MOTIV5 auch der dritte Fall ausgeschlossen, der *alle konkreten Motive verneint*, aber die Restkategorie markiert hat. Folglich wird in obiger Ausgabe z. B. zum Motiv 1 meldet, dass 79,3% der Fälle (23 von 29) zugestimmt hätten. Offenbar erwartet SPSS von einem Mehrfachantwortenset, dass jede redliche Auskunftsperson mindestens einer Option zustimmen muss. Wir haben bei der Fragebogenkonstruktion mit Blick auf die Unterscheidbarkeit von verneinenden und fehlenden Antworten darauf geachtet, die Menge der Antwortmöglichkeiten im Teil 3a zu komplettieren (siehe Abschnitt 2.4.3.2.4). Für diese Befragungstechnik stellt sich nun eine weitere Begründung heraus.

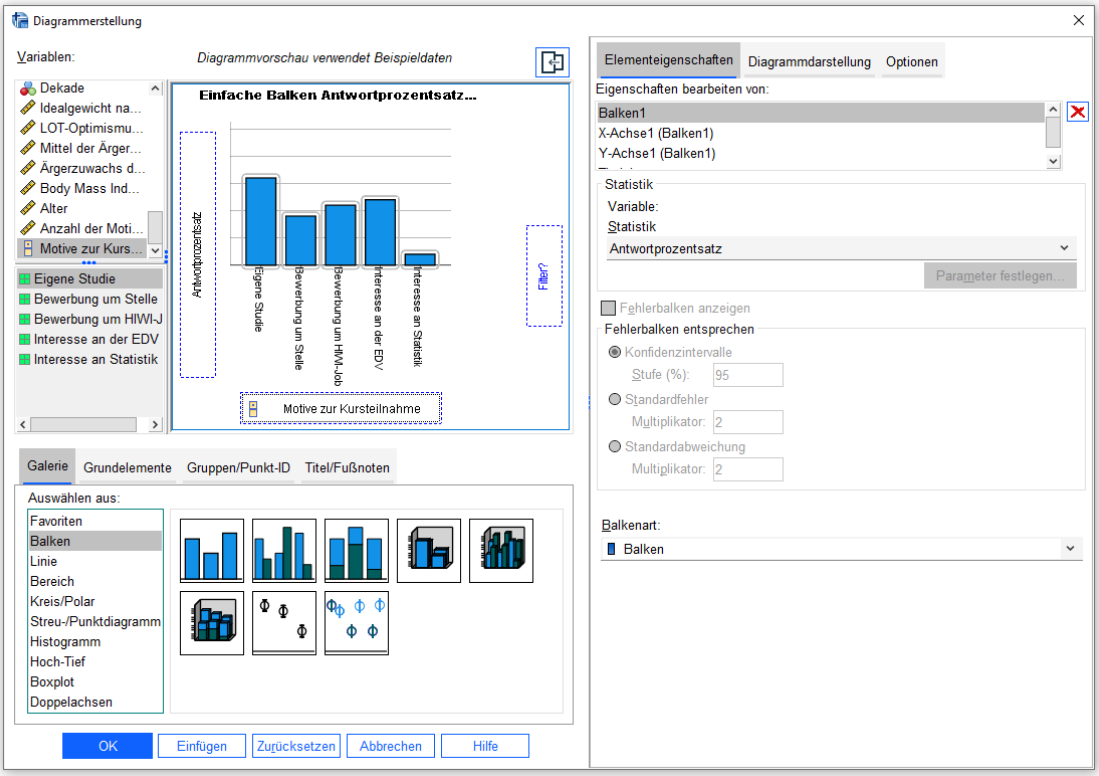
Um ein Balkendiagramm mit den Zustimmungsanteilen für alle Motivvariablen



zu produzieren, wählt man nach dem Menübefehl

Grafik > Diagrammerstellung

aus der **Galerie** die **einfachen Balken** und befördert das Mehrfachantwortenset $MOTIVE auf die **X-Achsen**-Ablagezone:



Auf der Registerkarte mit den **Elementeigenschaften** wählt man den **Antwortprozentsatz** als darzustellende **Statistik** (vorletzte Option in der Drop-Down - Liste).

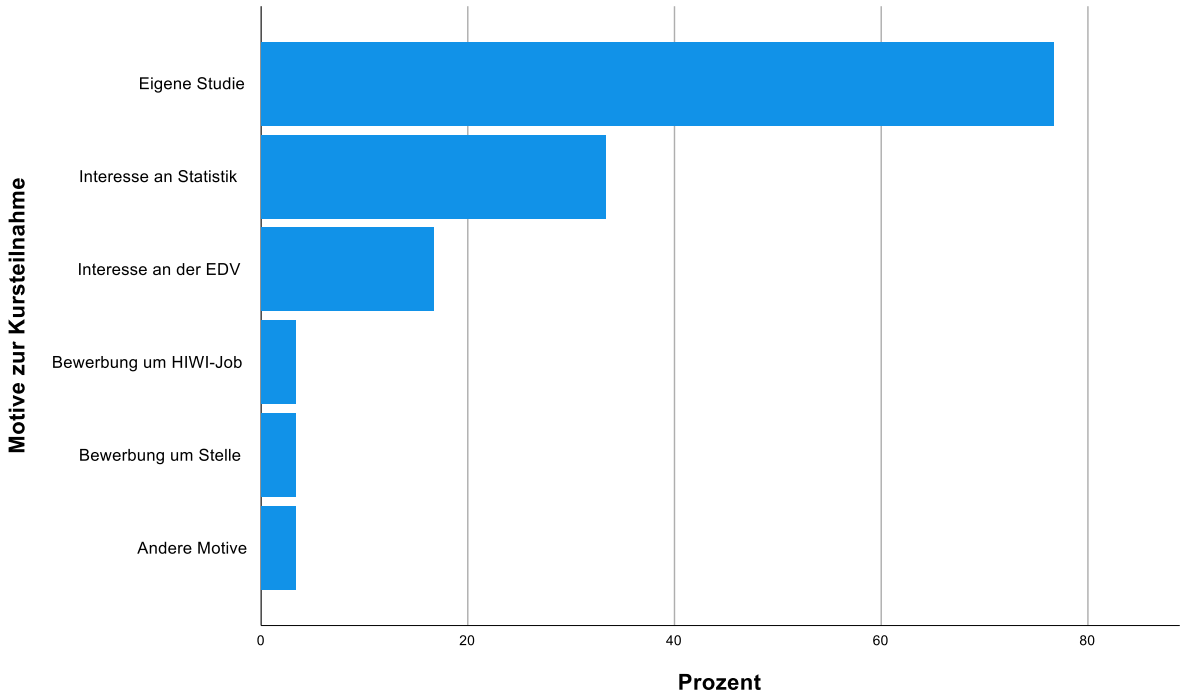
Nach dem Quittieren der **Diagrammerstellung** mit **OK** wird im Diagrammeditor die Detailgestaltung vorgenommen, z. B.:

- Änderung der Kategorienreihenfolge auf der bei markierten Balken verfügbaren Registerkarte **Kategorien**
- Beschriftung der Y-Achse
- Verkleinerung der **bevorzugten Größe** für die Balkenbeschriftungen (von 12 auf 10)
- Löschen des Titels

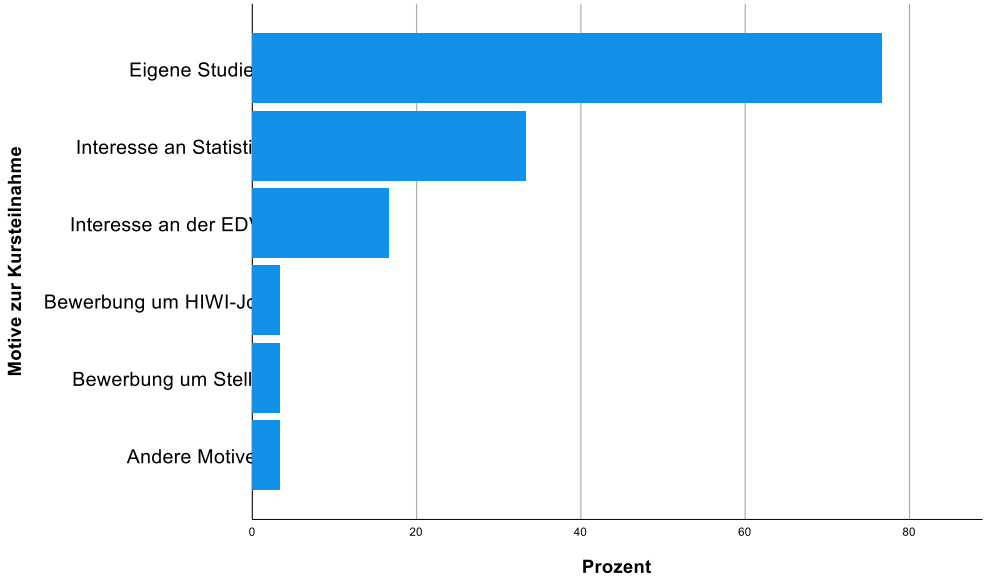
Für die folgende Darstellungsalternative, die bei einer größeren Anzahl von Optionen und/oder längeren Beschriftungen von Vorteil ist, wurde zunächst bei markierten Balken auf der Registerkarte **Kategorien** für die Motive ein aufsteigendes **Sortieren nach Statistik** angefordert. Anschließend wurde über den Symbolschalter bzw. mit dem Menübefehl

Optionen > Diagramm transponieren

das Koordinatensystem transponiert:



Durch die Reduktion der **bevorzugten Größe** für die Balkenbeschriftungen (von 12 auf 10) wird ein fehlerhaftes Ergebnis beim Metafile-Transfer von SPSS zu Microsoft Word verhindert:



16.3 Kreuztabellen für Mehrfachantwortsets

Wenn wir uns für Geschlechtsunterschiede bei der Zustimmung zu den Motiven interessieren (z. B.: Wer interessiert sich mehr für Statistik?), sind *sechs* (2×2) - Tabellen zu analysieren. Über den aus Kapitel 14 bekannten Menübefehl **Analysieren > Deskriptive Statistiken > Kreuztabellen** erhalten wir z. B. für das Statistikmotiv das folgende Ergebnis:[1]

Interesse an Statistik * Geschlecht Kreuztabelle

			Geschlecht		
			Frau	Mann	Gesamt
Interesse an Statistik	Ja	N	9	1	10
		%	37,5%	16,7%	33,3%
	Nein	N	15	5	20
		%	62,5%	83,3%	66,7%
Gesamt		N	24	6	30
		%	100,0%	100,0%	100,0%

Weil die Motivvariablen nur zwei Ausprägungen besitzen, sind die Ergebnisse zur Nein-Kategorie überflüssig. Es genügt zu wissen, dass 37,5% von den 24 Frauen und 16,7% von den sechs Männern ein Interesse an Statistik angegeben haben. Unter Verzicht auf die redundanten Zeilen lässt sich eine kompakte Darstellung der Geschlechtsunterschiede bei *allen* Kursmotiven erstellen:

		Geschlecht					
		Frau		Mann		Gesamt	
		Anzahl	Prozent	Anzahl	Prozent	Anzahl	Prozent
Motive zur Kursteilnahme	Eigene Studie	19	79,2%	4	66,7%	23	76,7%
	Bewerbung um Stelle	1	4,2%	0	0,0%	1	3,3%
	Bewerbung um HIWI-Job	0	0,0%	1	16,7%	1	3,3%
	Interesse an der EDV	3	12,5%	2	33,3%	5	16,7%
	Interesse an Statistik	9	37,5%	1	16,7%	10	33,3%
	Andere Motive	1	4,2%	0	0,0%	1	3,3%
	Gesamt	24	100,0%	6	100,0%	30	100,0%

Beachten Sie bitte: Dies ist nicht *eine* (6×2) - Kontingenztabelle, sondern dies ist eine Kombitabelle mit der essentiellen deskriptiven Informationen aus *sechs* (2×2) - Kontingenztabellen.

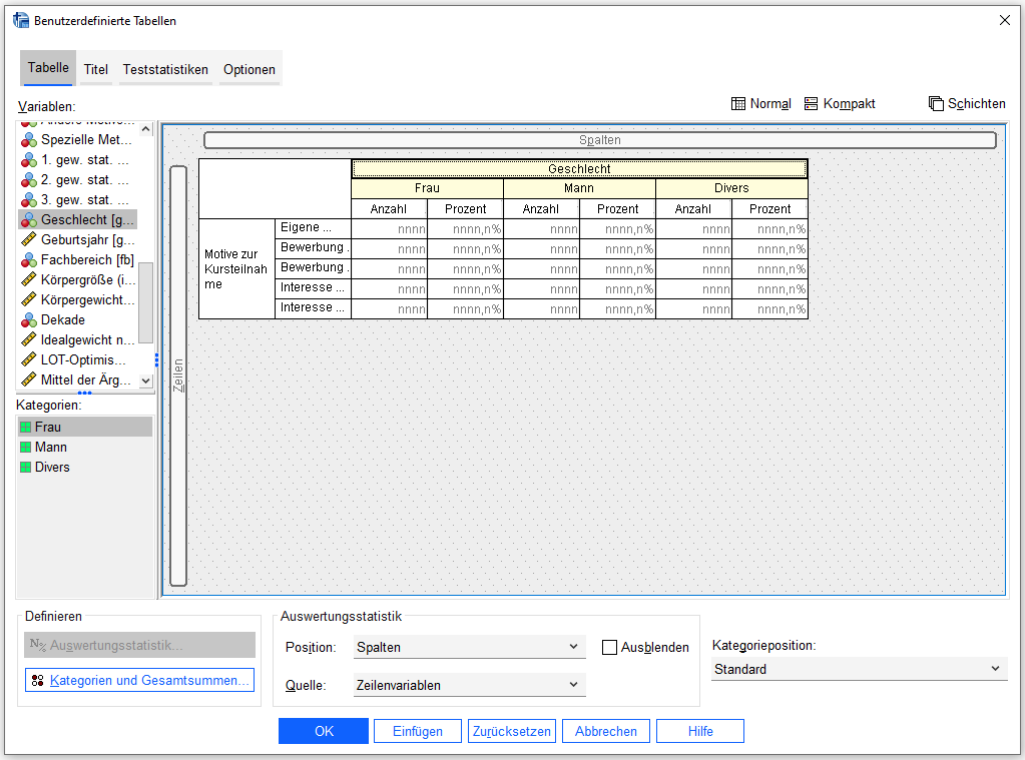
Bei der Erstellung dieser Tabelle würde die unbesetzte GESCHL-Kategorie *Divers* stören. Um ein analoges Problem zu lösen, haben wir schon in Abschnitt 11.2.2 die Beschriftung zu dem in unserer Stichprobe nicht aufgetretenen GESCHL-Wert 3 gelöscht.

Wir öffnen über

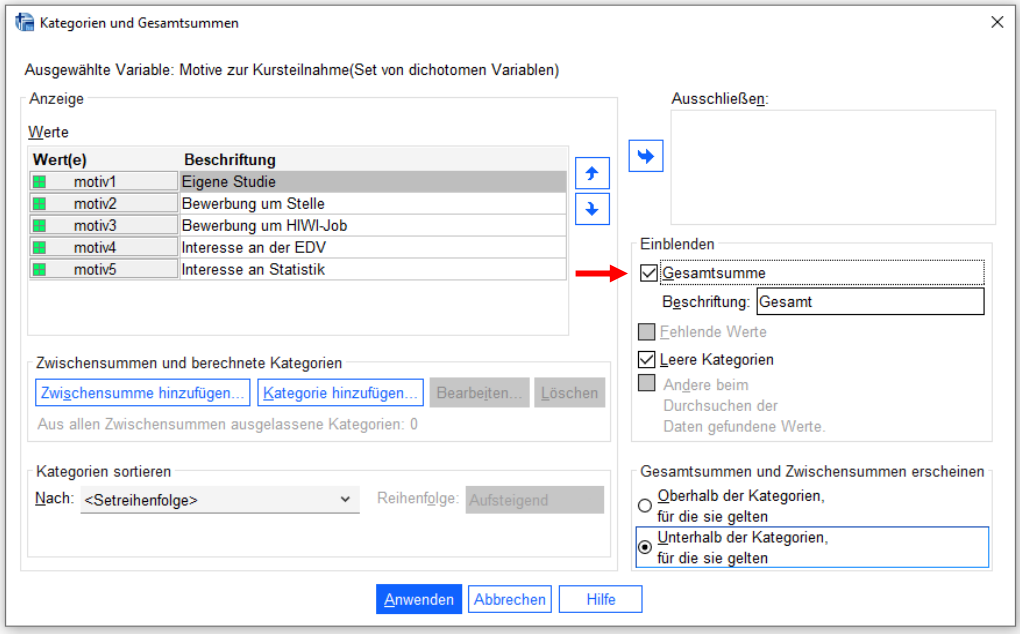
Analysieren > Tabellen > Benutzerdefinierte Tabellen

erneut die Dialogbox für benutzerdefinierte Tabellen. Ausgehend von dem in Abschnitt 16.2 erreichten Bearbeitungszustand bewegen wir die Variable GESCHL auf die **Spalten**-Ablagezone:

[1] Die Tabelle wurde mit dem Pivot-Editor nachbearbeitet (vgl. Abschnitt 10.4).



Bei markierter Zeilen- bzw. Spaltendimension (Motive zur Kursteilnahme bzw. Geschlecht) öffnen wir jeweils über den Schalter **Kategorien und Gesamtsummen** den folgenden Dialog



und markieren das Kontrollkästchen **Gesamtsumme**. So resultiert schließlich die oben gezeigte Tabelle.

Auch bei den Kreuztabellen ist die in Abschnitt 16.2 beschriebene MD-Konzeption der SPSS-Mehrfachwahl-Auswertung zu beachten. Wäre nicht die Variable ANDERE Mitglied im Set $MOTIVE, dann würde SPSS in der Kombitabelle nur noch diejenigen Fälle berücksichtigen, die mindestens ein konkret abgefragtes Motiv bejaht haben.

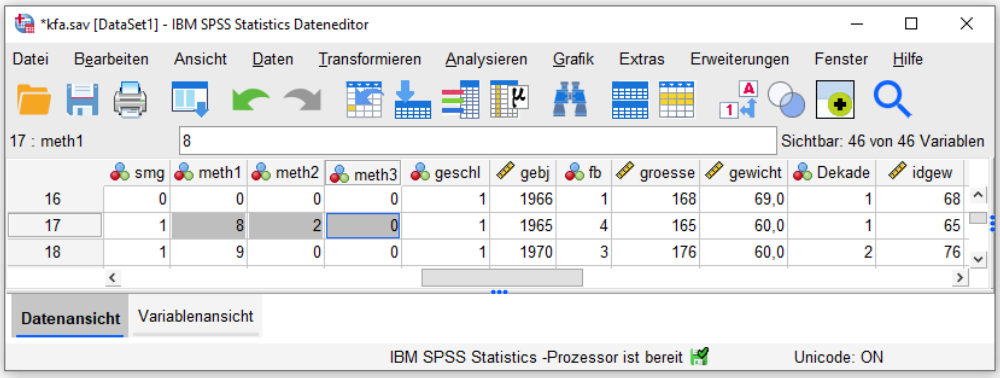
16.4 Ein sparsames Set kategorialer Variablen expandieren

In Abschnitt 2.4.2.3 wurde das sparsame Set aus kategorialen Variablen für Mehrfachwahlfragen mit sehr vielen Antwortmöglichkeiten zur Vereinfachung der Erfassung empfohlen. Zwar ist diese Datenstruktur kein Nachteil bei den Analyseprozeduren, die in den Abschnitten 16.2 und 16.3 beschrieben wurden, doch ist bei anderen Auswertungen ein vollständiges Set aus dichotomen Variablen erforderlich. In dieser Situation kann man das sparsame Set mit Hilfe der SPSS-Kommandosprache „expandieren“. Die folgenden Kommandos erzeugen zu unseren Variablen METH1 bis METH3 die acht dichotomen Variablen STAT1 bis STAT8, die für jeweils eine bestimmte statistische Methode festhalten, ob sie genannt worden ist (Wert 1) oder nicht (Wert 0):

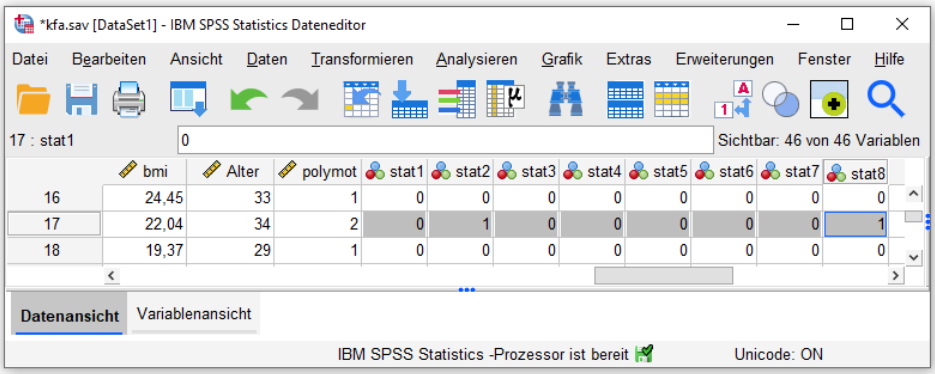
```
do repeat  stat = stat1 to stat8 /n = 1 to 8.
 do if (meth1 = n) or (meth2 = n) or (meth3 = n).
  compute stat = 1.
 else.
  compute stat = 0.
 end if.
end repeat.
formats  stat1 to stat8 (f8.0).
variable width  stat1 to stat8 (5).
variable level  stat1 to stat8 (nominal).
variable role
  /target  stat1 to stat8.
execute.
```

Bei den Daten der Manuskriptstichprobe steht z. B. die Variable STAT2 für die Regressionsanalyse, weil nach dem vervollständigten Codierplan (vgl. Abschnitt 4.2.9) bei einer von den Variablen METH1 bis METH3 eine Zwei zu notieren war, wenn ein Fall im Fragebogenteil 3b die Regressionsanalyse genannt hatte.

Beim Fall Nr. 17 wurden die genannten Methodenwünsche 8 (= logistische Regression) und 2 (= Regressionsanalyse) folgendermaßen mit dem sparsamen Set kategorialer Variablen METH1 bis METH3 erfasst:



Daraus ergeben sich folgende Werte für die Variablen STAT1 bis STAT8:



In obiger Syntax werden zwei ausgesprochen nützliche Kontrollstrukturen der SPSS-Kommandosprache verwendet:

Schleife für strukturgleiche Transformationen

Die (DO REPEAT - END REPEAT) - Schleife wird achtmal ausgeführt, wobei im *i*-ten Umlauf die beiden Stellvertreter STAT und N gerade mit dem *i*-ten Element der jeweils zugehörigen Liste identisch sind.

Fallunterscheidung

In der folgenden Tabelle wird das SPSS-Verhalten beim Ausführen der (DO IF - ELSE - END IF) - Struktur in Abhängigkeit vom Wahrheitswert des logischen Ausdruck beschrieben:

Wert des logischen Ausdrucks	Aktion
wahr, z. B. im ersten Schleifenumlauf bei METH1 = 1, METH2 = 2, METH3 = 0	Das erste COMPUTE-Kommando wird ausgeführt, und die Variable STAT1 erhält den Wert 1.
falsch, z. B. im ersten Schleifenumlauf bei METH1 = 3, METH2 = 5, METH3 = 8	Das zweite COMPUTE-Kommando wird ausgeführt, und die Variable STAT1 erhält den Wert 0.
unbestimmt, z. B. im ersten Schleifenumlauf bei METH1=METH2=METH3=SYSMIS	Die Variable STAT1 behält den Initialisierungswert SYSMIS.

17 Datendateien im Textformat einlesen

Gelegentlich sind Daten auszuwerten, die in Textdateien vorliegen. Dabei können zwei Dateiformate auftreten:

- separierte Daten
  Für jeden Fall befinden sich alle Variablenausprägungen hintereinander in einer Zeile, wobei zwischen zwei Werten ein Trennzeichen steht. Die Reihenfolge der Variablen ist für alle Fälle gleich.
- positionierte Daten
  Für jeden Fall liegt eine identische Anzahl von Datenzeilen vor. Jede Variable hat eine individuelle, feste Breite (z. B. 2 Stellen) und eine feste Position (z. B. Zeile 2, Spalten 8-9).

Zum Importieren von Textdatendateien stellt SPSS einen Assistenten zur Verfügung, der mit

Datei > Daten importieren > Textdaten

gestartet wird. Er kommt aber auch dann zum Einsatz, wenn Sie nach

Datei > Öffnen > Daten

eine Textdatendatei wählen.

An der im Vorwort vereinbarten Stelle finden Sie die Dateien **kfar-kv-sep.txt** und **kfar-kv-pos.txt** mit separierten bzw. positionierten KFA-Daten von 77 Fällen. Es bietet sich an, diese Daten einzulesen, um die in Abschnitt 11.4 durch grafische Datenexploration gewonnene Moderatorversion der differentialpsychologischen Hypothese anhand einer unabhängigen Stichprobe zu überprüfen.

17.1 Import von separierten Textdaten

Separierte Textdaten lassen sich sehr bequem importieren, zumal sie üblicherweise durch eine Zeile mit den Variablennamen eingeleitet werden. Hier ist der Anfang der Datei **kfar-kv-sep.txt** zu sehen:

```
FNR     GESCHL  GEBJ    FB      GROESSE GEWICHT AERGO   AERGM   LOT1    LOT2    ...
1       1       77      1       158     48      6       6       4       3       ...
2       1       77      1       159     55      4       8       3       4       ...
3       1       74      4       160     48      3       8       4       3       ...
4       1       75      1       165     78      2       2       5       5       ...
.       .       .       .       .       .       .       .       .       .       ...
.       .       .       .       .       .       .       .       .       .       ...
```

Gehen Sie folgendermaßen vor, um die Daten zu importieren:

Textimport-Assistenten starten und Datei auswählen

Nach dem Start des Textimport-Assistenten über den Menübefehl

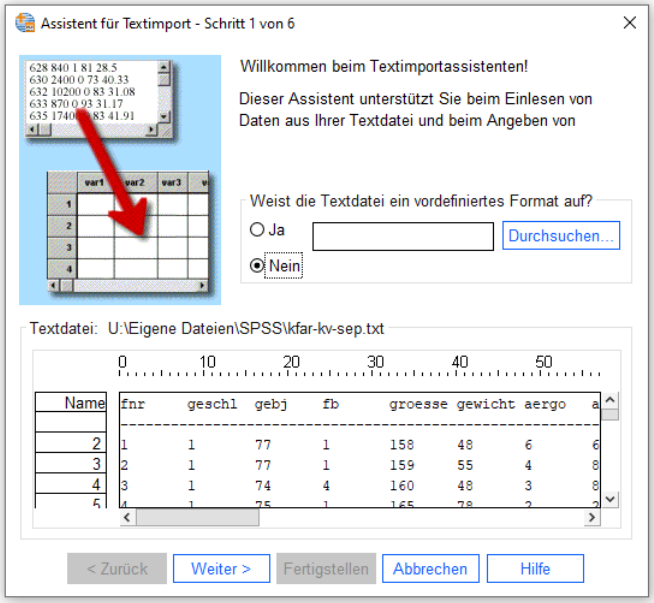
Datei > Daten importieren > Textdaten

ist zunächst die Eingabedatei zu wählen:

Der Textimportassistent unterstützt die Codierungen UTF-8 sowie ANSI (lokale Codierung) und erkennt diese Codierungen automatisch. Wir klicken auf **Öffnen**.

Schritt 1

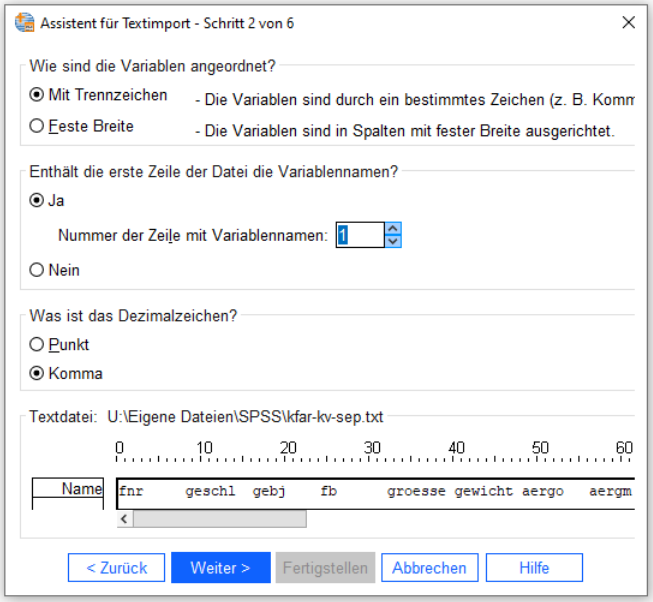
Im ersten Schritt zeigt der Assistent den Anfang der Datei und akzeptiert ggf. ein **vordefiniertes Format** aus früheren Assistenteneinsätzen, das die Dateistruktur beschreibt.



Da wir auf eine solche Vorarbeit nicht zurückgreifen können, machen wir **weiter**.

Schritt 2

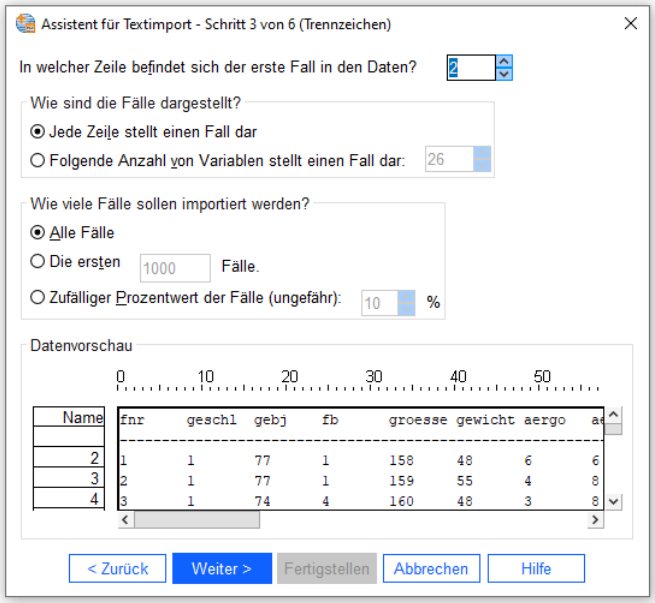
Im zweiten Schritt informieren wir den Assistenten darüber, dass **Trennzeichen** für Ordnung in der Datei sorgen, und dass die erste Zeile die **Variablennamen** enthält:



Falls in der zu importierenden Datei ein **Dezimaltrennzeichen** verwendet wird, muss es in diesem Schritt korrekt eingestellt werden. Die Datei **kfar-kv-sep.txt** enthält keine Dezimaltrennzeichen (auch nicht bei der Variablen GEWICHT).

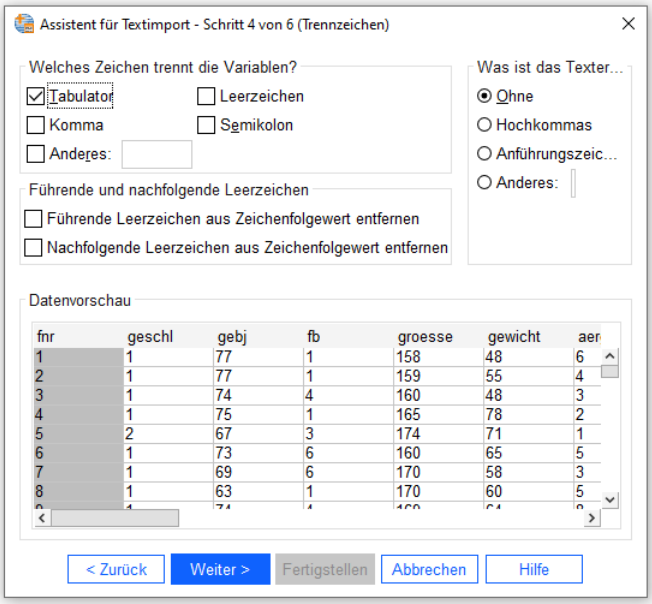
Schritt 3

Der erste Fall befindet sich in der zweiten Zeile der Datei (hinter der einleitenden Zeile mit den Variablennamen), und jeder Fall belegt genau eine Zeile:



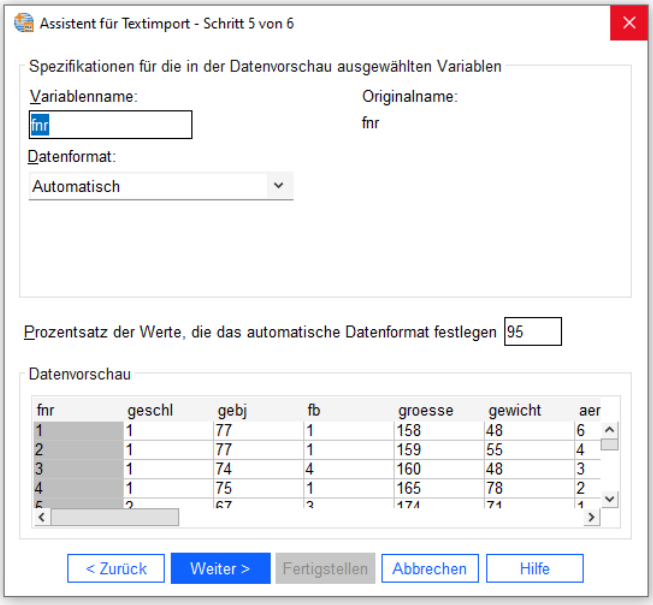
Schritt 4

SPSS erkennt korrekt, dass in der Datei **kfar-kv-sep.txt** nur der **Tabulator** als Trennzeichen zum Einsatz kommt:



Schritt 5

Im fünften Assistentenschritt kann für alle Variablen ein **Datenformat** festgelegt oder auf die automatische Erkennung vertraut werden, wobei der Prozentsatz der Fälle, **die das automatische Datenformat festlegen**, eingestellt werden kann (Voreinstellung: 95%):

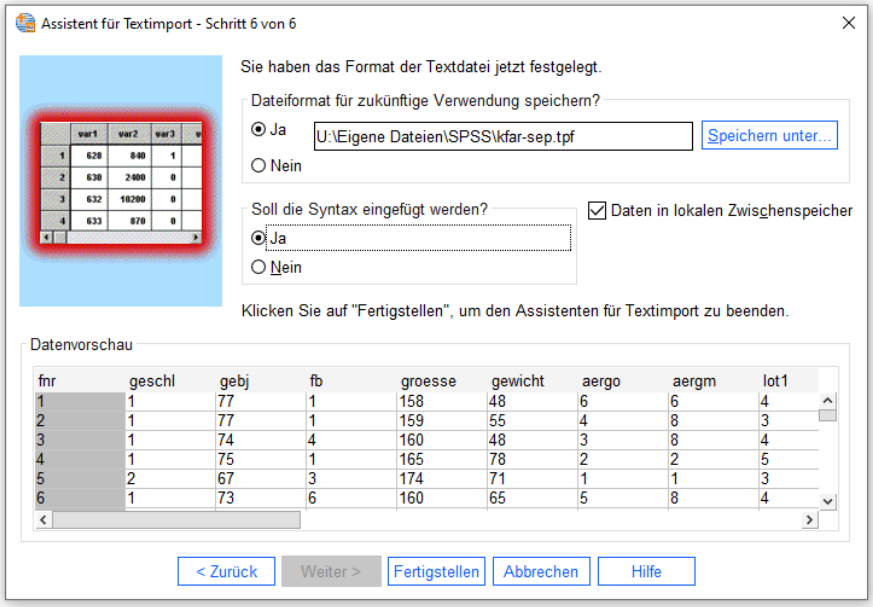


Schritt 6

Der Assistent bietet zwei Möglichkeiten zum Konservieren einer Dateispezifikation:

- **Dateiformat für zukünftige Verwendung speichern?**
  Es entsteht eine Textassistent-Formatdatei (Erweiterung **.tpf**), die bei einem späteren Assistenteneinsatz im ersten Schritt angegeben werden kann (siehe oben).
- **Soll die Syntax eingefügt werden?**
  Die für den Datenimport erforderlichen SPSS-Kommandos werden in ein Syntaxfenster geschrieben. Es bietet sich an, zusätzliche Kommandos zu ergänzen, z. B. zum Deklarieren von MD-Indikatoren, die in den Textdaten vorhanden sind. Später kann mit Hilfe des entstandenen SPSS-Programms der Import mit allen erforderlichen Zusatzmaßnahmen automatisiert ausgeführt werden.

Es spricht nichts dagegen, *beide* Konservierungsoptionen zu verwenden:



In der vom Textimport-Assistenten erzeugten Syntax spielt das GET DATA - Kommando die entscheidende Rolle.

In der Datei **kfar-kv-sep.txt** sind alle Fälle komplett. Generell dürfen Sie nach dem Einlesen von separierten Textdaten auf keinen Fall die Deklaration der eventuell vorhandenen **MD-Indikatoren** vergessen. Dies sollte per Syntax mit dem Kommando MISSING VALUES geschehen, z. B.:

```
MISSING VALUES geschl lot1 to lot12 (9).
```

Weil das Kommando MISSING VALUES (nicht nur beim Einlesen von Textdaten) von großer Bedeutung ist, soll seine allgemeine Syntax hier beschrieben werden:

MISSING VALUES {*varlist* | ALL} (*valuelist*) [[/]{*varlist* | ALL} ...]**.**

varlist Dieser Platzhalter steht für eine Liste vorhandener Variablen, deren MD-Indikatoren festgelegt werden sollen (vgl. Abschnitt 19.1.5.2).

valuelist Dieser Platzhalter steht für eine Liste von Werten, die als MD-Indikatoren vereinbart werden sollen. In der Werteliste sind folgende Schlüsselwörter erlaubt:

- LO, LOWEST
  Kleinster Wert
- HI, HIGHEST
  Größter Wert
- THRU
  Verbindungswort für Wertebereiche, z. B.:

```
7 THRU 9
7 THRU HI
```

Durch eine leere Werteliste werden für die betroffenen Variablen alle benutzerdefinierten MD-Indikatoren aufgehoben, z. B.:

```
MISSING VALUES partei ().
```

In der Syntaxbeschreibung werden folgende generelle Regeln verwendet:

- Ist von mehreren Ausdrücken genau einer zu wählen, werden diese Ausdrücke zwischen geschweiften Klammern und voneinander durch senkrechte Striche getrennt präsentiert.
- Eckige Klammern schließen optionale Ausdrücke ein

17.2 Import von positionierten Textdaten (feste Breite)

Die Datei **kfar-kv-pos.txt** enthält dieselben KFA-Daten, die in Abschnitt 17.2 aus einer separierten Datei gelesen wurden. Hier sind die Werte eines Falles auf zwei Zeilen verteilt, und jede Variable hat eine feste Position im Datensatz eines Falles (z. B. Variable AERGO in Zeile 2, Spalten 5-6).

```
11 177115848
12  6 6 431214542432 110000
21 177115955
22  4 8 343335442442 110010
31 174416048
32  3 8 433224443342 100010
41 175116578
42  2 2 553125544531 100100
 .  .  .  .  .  .  .  .  .
 .  .  .  .  .  .  .  .  .
```

Die für uns relevanten Variablen haben folgende Positionen:

Variable	Datenzeile	Spalten
GESCHL	1	5
AERGO	2	5-6
AERGM	2	7-8
LOT01-LOT12	2	10-21

Alle übrigen Variablen können wir ignorieren.

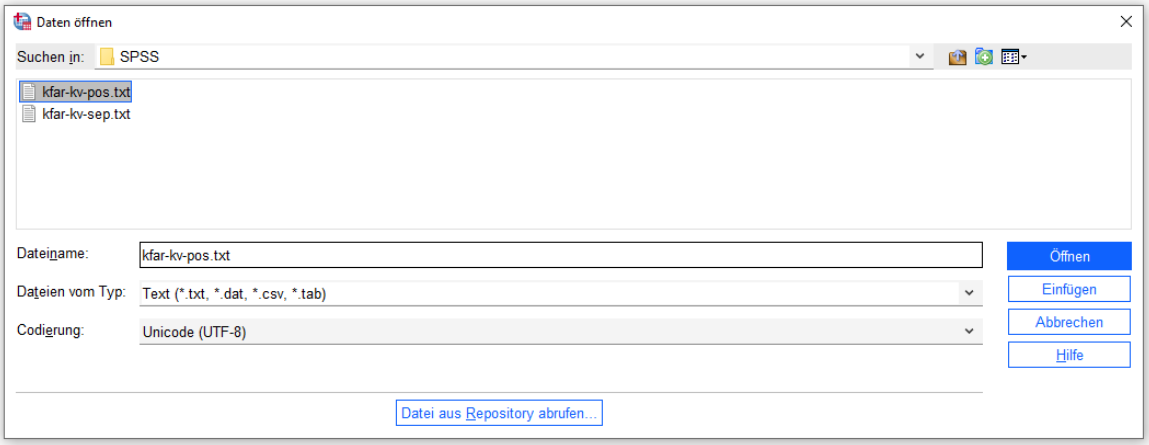
Gehen Sie folgendermaßen vor, um die relevanten Daten zu importieren:

Textimport-Assistenten starten und Datei auswählen

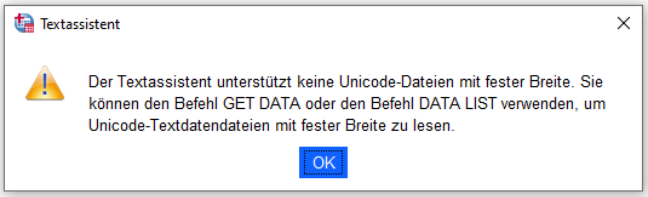
Nach dem Start des Textimport-Assistenten über den Menübefehl

Datei > Daten importieren > Textdaten

ist zunächst die Eingabedatei zu wählen:



Der Assistent unterstützt bei positionierten Textdateien keine Unicode-Codierung:

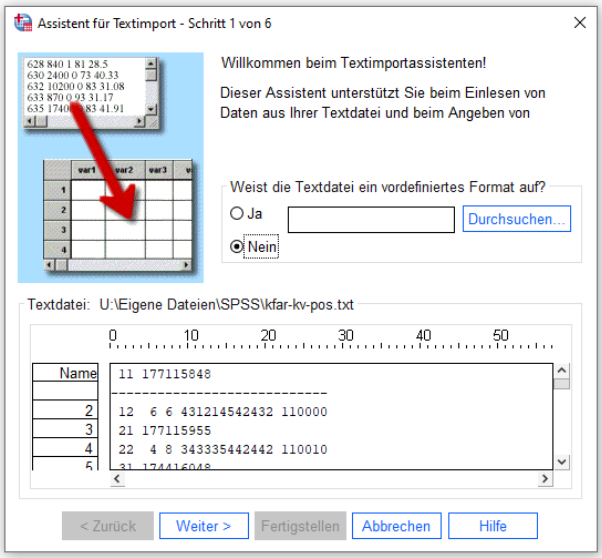


In der Regel enthalten Textdateien mit positionierten Daten nur Zahlen, sodass nichts gegen die Verwendung des kompatiblen ANSI-Codes (oft als **Lokale Kodierung** bezeichnet) spricht. Sind in einer positionierten Textdatei aber auch Zeichenketten und/oder Variablennamen mit Unicode-Codierung vorhanden, dann muss ein direkt (im Syntax-Fenster) verfasstes GET DATA - oder DATA LIST - Kommando zum Einlesen verwendet werden.

Wenn unter Windows10 trotz alle Bemühungen zur Änderung der Kodierung das Problem bestehen bleibt, kann es helfen, die Textdatei per Drag & Drop zu öffnen.

Schritt 1

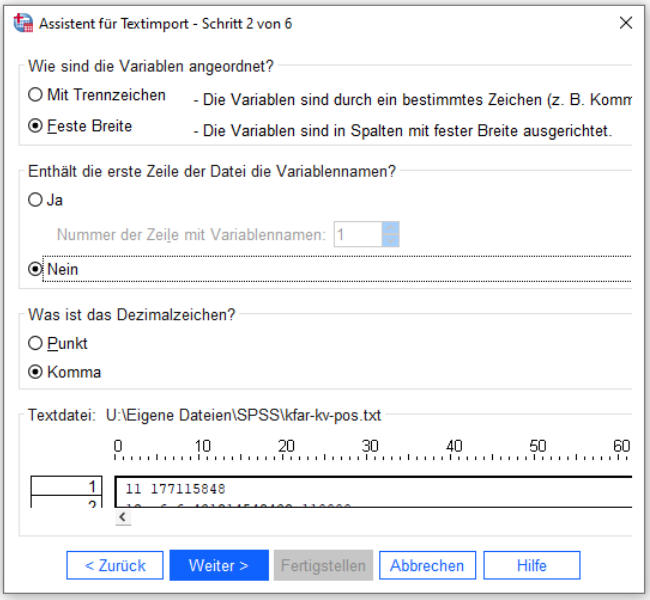
Im ersten Schritt zeigt der Assistent den Anfang unserer Datei und akzeptiert ggf. ein **vordefiniertes Format** aus früheren Assistenteneinsätzen, das die Dateistruktur beschreibt.



Da wir auf eine solche Vorarbeit *nicht* zurückgreifen können, machen wir **weiter**.

Schritt 2

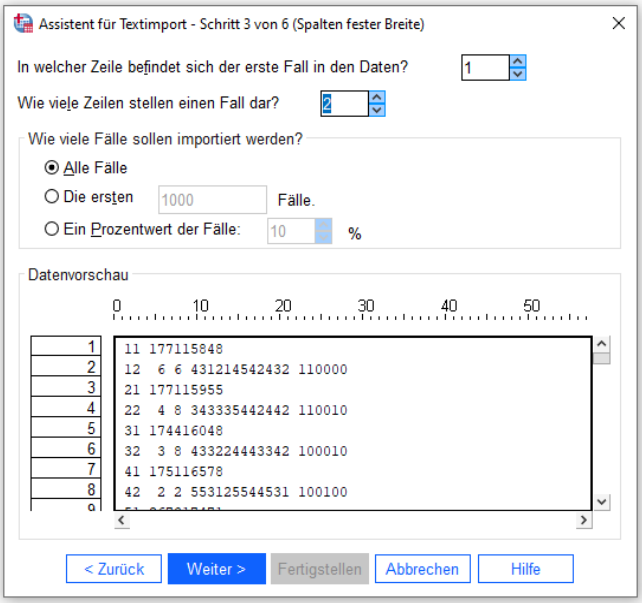
Im zweiten Schritt teilen wir mit, dass die Variablen in unserer Eingabedatei feste Positionen bzw. eine **feste Breite** besitzen:



Von der Möglichkeit, in der **ersten Zeile der Datei die Variablennamen** zu transportieren, wird in unserem Beispiel kein Gebrauch gemacht. Falls in der zu importierenden Datei ein Dezimaltrennzeichen verwendet wird, muss es in diesem Schritt korrekt eingestellt werden. Die Datei **kfar-kv-pos.txt** enthält keine Dezimaltrennzeichen (auch nicht bei der Variablen GEWICHT).

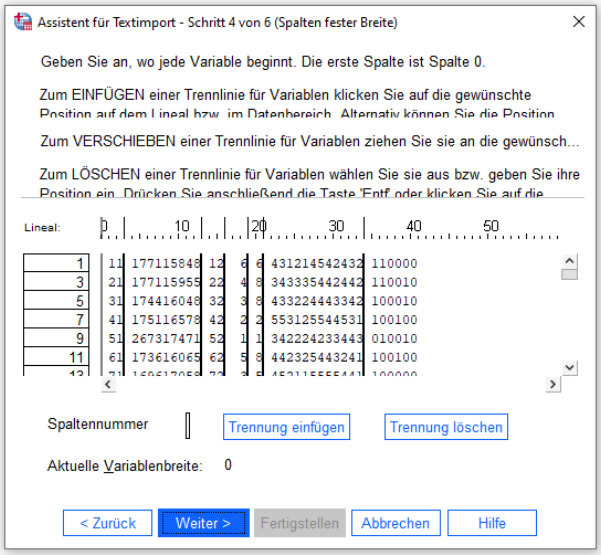
Schritt 3

Da unsere Datei keinen Vorspann enthält, **befindet sich der erste Fall** in Zeile 1. Allerdings befindet er sich dort nicht komplett, weil jeweils zwei **Zeilen einen Fall darstellen**:



Schritt 4

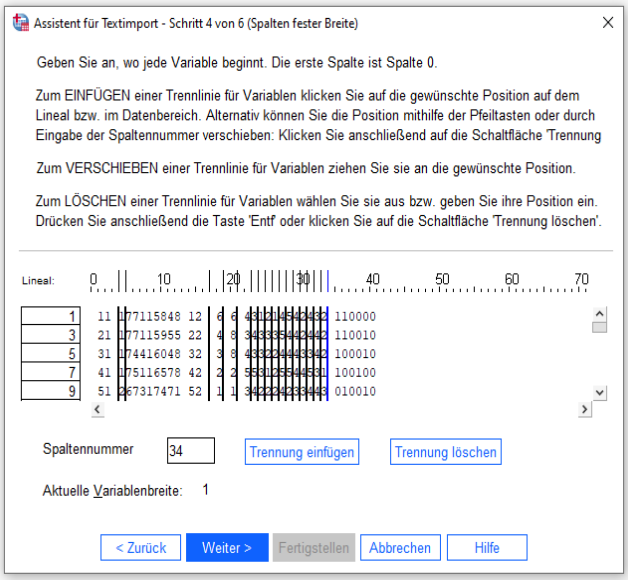
Nun müssen wir die Positionen der einzulesenden Variablen durch Setzen, Verschieben und Löschen von Trennlinien festlegen, wobei alle Zeilen eines Falles hintereinander angezeigt werden. Der Assistentenvorschlag orientiert sich an Leerzeichen und würde im Beispiel zu 7, teilweise unbrauchbaren Variablen führen:



Hinweise zur Benutzung der Trennlinien:

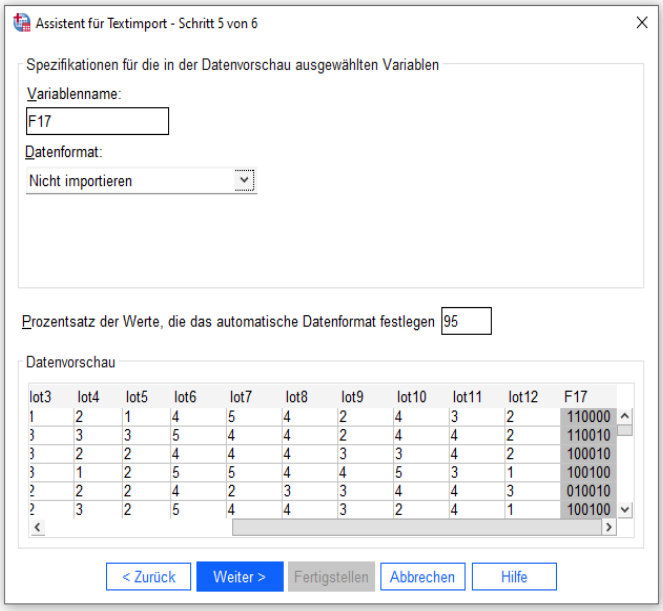
- Neue Trennlinie einfügen
  Klicken Sie innerhalb der Datenzone auf die gewünschte Spaltenposition.
- Trennlinie verschieben
  Klicken Sie innerhalb der Datenzone auf die Trennlinie und verschieben Sie diese bei fest gehaltener Maustaste.
- Trennlinie löschen
  Trennlinie per Mausklick markieren und anschließend **Trennung löschen**

So kann im Beispiel das Einlesen der Variablen GESCHL, AERGO, AERGM und LOT1 bis LOT12 vorbereitet werden:



Schritt 5

Im fünften Assistentenschritt kann man die von SPSS erkannten bzw. vorgeschlagenen Variablennamen ändern und ein **Datenformat** festlegen. Zum Umbenennen ist jeweils genau eine Spalte zu markieren. Das Datenformat lässt sich auch für eine markierte Variablen*liste* wählen. Wird auf die automatische Formaterkennung vertraut, kann der Prozentsatz der Fälle, **die das automatische Datenformat festlegen**, eingestellt werden (Voreinstellung: 95%). Mit dem speziellen Datenformat **Nicht importieren** können überflüssige Variablen ausgeschlossen werden, z. B.:



Schritt 6

Im letzten Assistentendialog werden die schon in Abschnitt 17.2 vorstellten Optionen zum Konservieren der Importspezifikation angeboten.

Das vom Textimport-Assistenten erzeugte GET DATA – Kommando verwendet eine Spaltenzählung ab **0**. So erhält z. B. die in der fünften Spalte der ersten Zeile eines Falles befindliche Variable GESCHL die Spaltenpositionsangabe 4-4.

```
PRESERVE.
 SET DECIMAL COMMA.

GET DATA  /TYPE=TXT
  /FILE="U:\Eigene Dateien\SPSS\kfar-kv-pos.txt"
  /ENCODING='Locale'
  /FIXCASE=2
  /ARRANGEMENT=FIXED
  /FIRSTCASE=1
  /VARIABLES=
  /1 geschl 4-4 AUTO
  F2 5-12 8X
  /2 aergo 4-5 AUTO
  aergm 6-7 AUTO
  lot1 8-9 AUTO
  lot2 10-10 AUTO
  lot3 11-11 AUTO
```

```
    lot4 12-12 AUTO
    lot5 13-13 AUTO
    lot6 14-14 AUTO
    lot7 15-15 AUTO
    lot8 16-16 AUTO
    lot9 17-17 AUTO
    lot10 18-18 AUTO
    lot11 19-19 AUTO
    lot12 20-20 AUTO
    F17 21-27 7X.
  RESTORE.

  CACHE.
  EXECUTE.
  DATASET NAME DataSet11 WINDOW=FRONT.
```

Auch nach dem Einlesen von positionierten Textdaten dürfen Sie auf keinen Fall die Deklaration der dort eventuell verwendeten **MD-Indikatoren** vergessen. Dies sollte per Syntax mit dem Kommando MISSING VALUES geschehen, das in Abschnitt 17.2 beschrieben wird, z. B.:

```
MISSING VALUES geschl lot1 to lot12 (9).
```

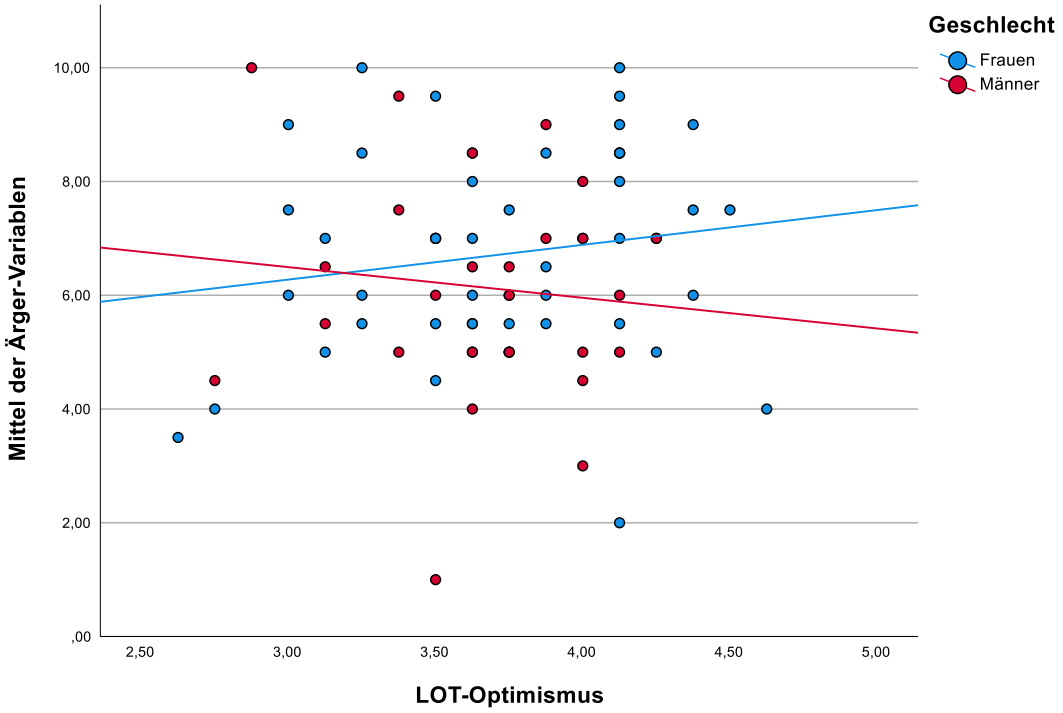
In der Datei **kfar-kv-pos.txt** sind allerdings alle Fälle komplett.

17.3 Überprüfung der revidierten differentialpsychologischen Hypothese

Um mit den in Abschnitt 17.2 bzw. Abschnitt 17.1 importierten Daten die revidierte differentialpsychologische Hypothese prüfen zu können, sind zunächst einige Datentransformationen erforderlich, wobei wir die erforderlichen Kommandos teilweise aus dem Transformationsprogramm **kfat.sps** übernehmen können:

```
* Labels für GESCHL.
VARIABLE LABELS geschl Geschlecht.

* LOT-Fragen Umcodieren.
RECODE
  lot3 lot4 lot5 lot12  (5=1)  (4=2)  (2=4)  (1=5).
EXECUTE.

* LOT berechnen.
COMPUTE lot = MEAN.6(lot1,lot3,lot4,lot5,lot8,lot9,lot11,lot12).
VARIABLE LABELS lot 'LOT-Optimismus'.
EXECUTE.

* AERGAM berechnen.
COMPUTE aergam = (aergo + aergm)/2.
VARIABLE LABELS aergam 'Mittel der Ärger-Variablen'.
EXECUTE.

* Produktvariable für die Moderatorhypothese.
COMPUTE geslot = geschl * lot.
VARIABLE LABELS geslot 'GESCHL * LOT'.
EXECUTE.
```

Auch in der neuen Stichprobe scheint das Geschlecht die Regression von AERGAM auf LOT im erwarteten Sinn zu moderieren:



Allerdings wird der Interaktionseffekt in der Moderatoranalyse (vgl. Abschnitt 11.4) *nicht* signifikant:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten			95,0% Konfidenzintervalle für B	
		Regressions-koeffizient B	Std.-Fehler	Beta	T	Sig.	Untergrenze	Obergrenze
1	(Konstante)	,773	5,562		,139	,890	-10,313	11,858
	LOT-Optimismus	1,761	1,493	,413	1,180	,242	-1,214	4,737
	Geschlecht	3,670	4,130	,949	,889	,377	-4,561	11,900
	GESCHL * LOT	-1,150	1,118	-1,120	-1,029	,307	-3,378	1,077

a. Abhängige Variable: Mittel der Ärger-Variablen

Zwar darf die die angegebene Überschreitungswahrscheinlichkeit zum Produktterm (p = 0,307) halbiert werden, weil sie zu einem zweiseitigen Test gehört, doch liegt auch das einseitige p-Level deutlich über der kritischen Grenze (0,05).

Weitere Versuche zur Rettung der differentialpsychologischen Hypothese könnten sich z. B. auf eventuelle Mängel bei der Operationalisierung der theoretischen Begriffe (Ärger und Optimismus) konzentrieren. Allerdings muss auch die theoretische Fundierung kritisch hinterfragt werden.

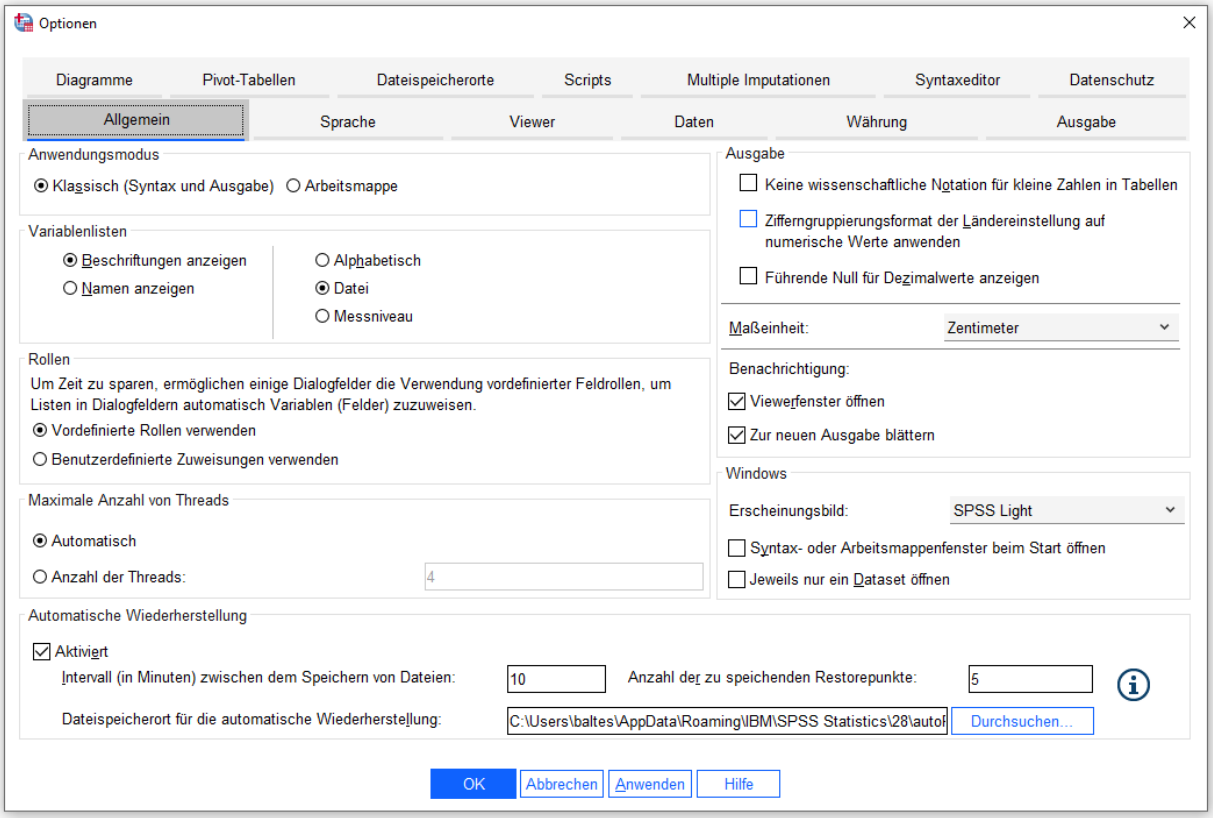
18 Einstellungen modifizieren

Das Standardverhalten von SPSS für Windows lässt sich auf vielfältige Weise den individuellen Bedürfnissen anpassen, was wir bei passender Gelegenheit auch schon getan haben.

Über den Menübefehl

Bearbeiten > Optionen

erhalten Sie die folgende Dialogbox mit Optionen zur SPSS-Konfiguration:



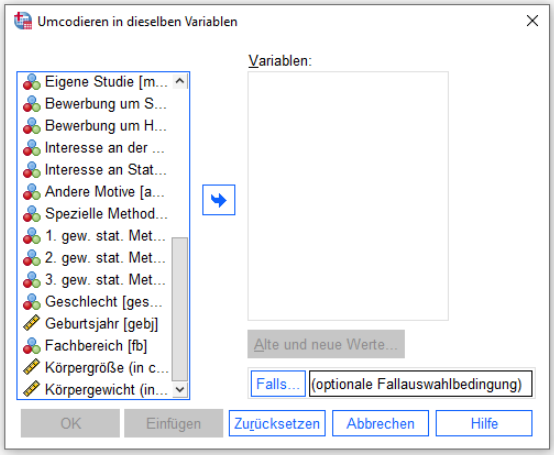
18.1 Allgemein

Auf dem Registerblatt **Allgemein** sind u. a. folgende Optionen von Relevanz:

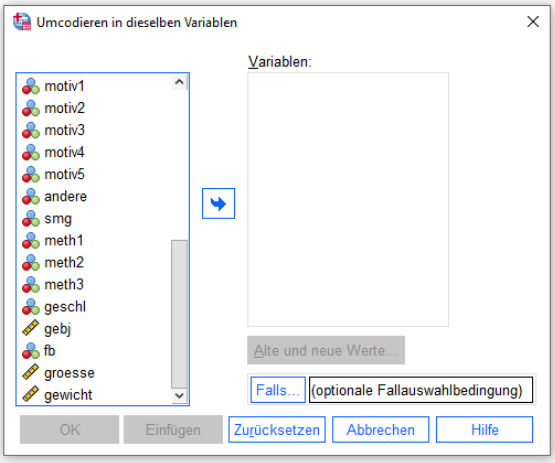
Variablenlisten

Bei den Listen auswählbarer Variablen in Dialogboxen verwendet SPSS folgende Voreinstellungen:

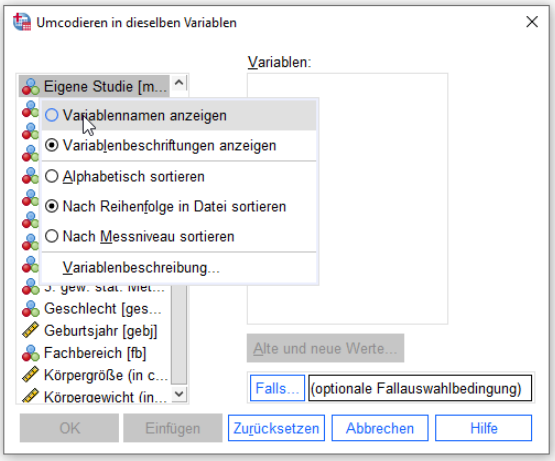
- Besitzt eine Variable eine Beschriftung, wird dieses vorrangig präsentiert, und der Variablenname erscheint hinter dem Label zwischen eckigen Klammern. Dabei werden manche Dialogboxen zur Analysespezifikation etwas unübersichtlich, z. B.:



Mit der Option **Namen anzeigen** im Bereich **Variablenlisten** kann man auf die kompaktere Darstellung *ohne* Labels umschalten, was bei Verwendung aussagekräftiger Variablennamen zu empfehlen ist, z. B.:



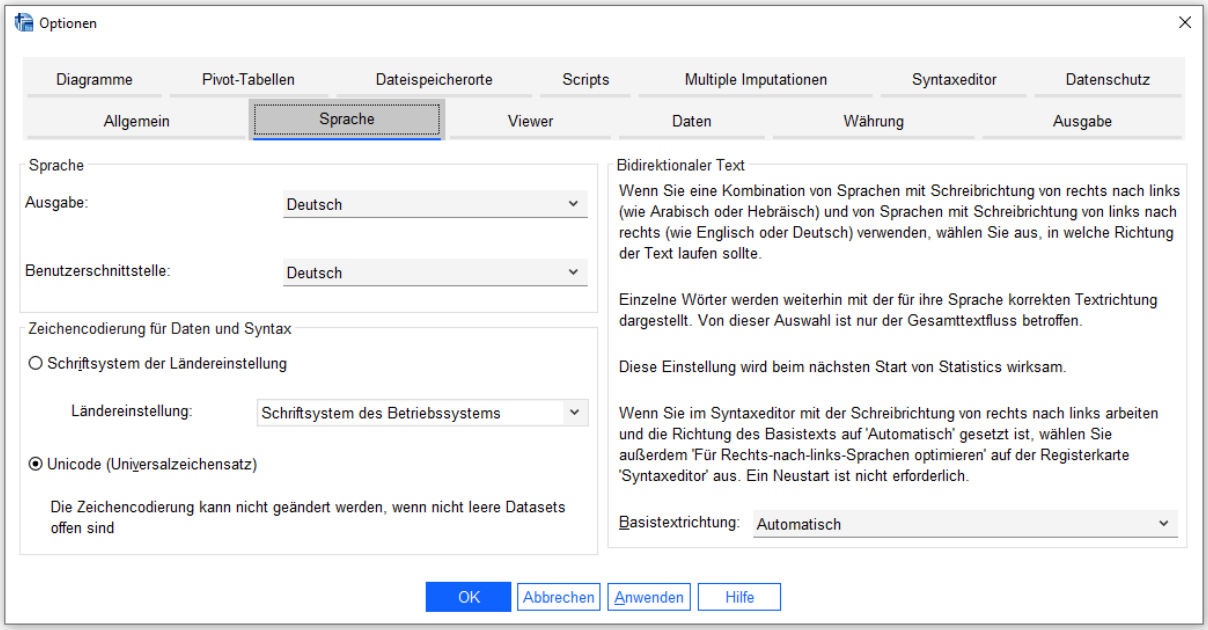
Man kann übrigens die Beschriftungsanzeige auch temporär (mit Gültigkeit für die aktuelle Dialogbox) abschalten, indem man aus dem Kontextmenü zur Variablenliste das Item **Variablennamen anzeigen** wählt, z. B.:



- Per Voreinstellung sind die Variablen in Dialogboxen genauso angeordnet wie in der **Datei**. Dies erlaubt in der Regel ein bequemes Arbeiten, weil gemeinsam zu analysierende und damit in Dialogboxen gemeinsam auszuwählende Variablen oft in der Arbeitsdatei hintereinander stehen. Bei der Arbeit mit einer unbekannten Datendatei findet man (namentlich bekannte) Variablen jedoch leichter bei **alphabetischer** Sortierung.

18.2 Sprache

Auf dem Registerblatt **Sprache**



sind u. a. folgende Optionen von Relevanz:

Sprache der Ausgabe bzw. Benutzerschnittstelle

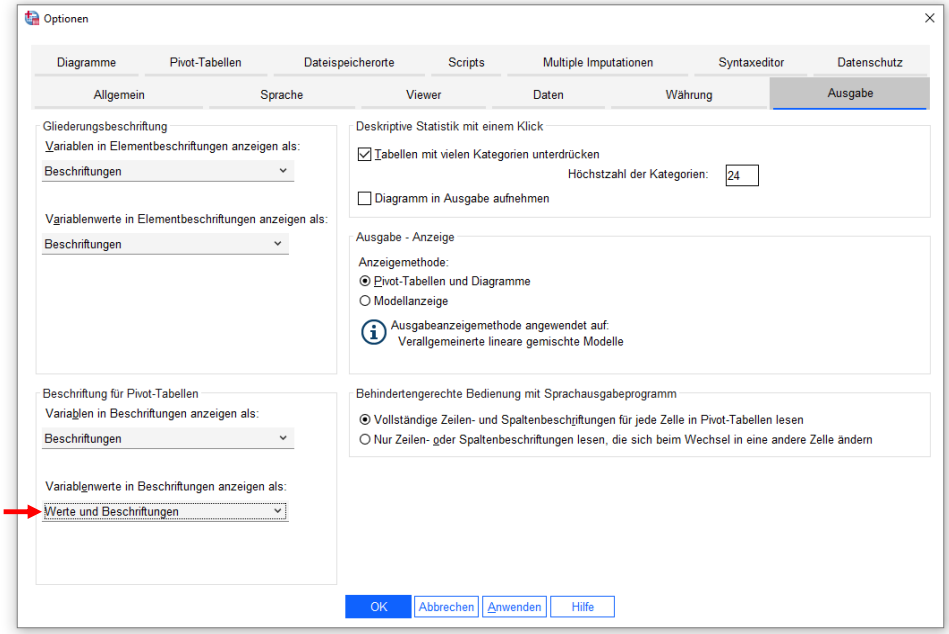
SPSS erlaubt für die **Ausgabe** (Beschriftung der Tabellen) und für die **Benutzerschnittstelle** (Menüs und Dialogboxen) die Wahl zwischen diversen Sprachen.

Zeichencodierung für Daten und Syntax

SPSS kann bei Texten in Daten- und Syntaxdateien den Unicode-Zeichensatz unterstützen, sodass z. B. kyrillische oder japanische Zeichen möglich sind. Eventuell führt die Unicode-Einstellung aber bei älteren Datendateien zu falsch angezeigten Zeichen (z. B. Umlauten). Auf der Registerkarte **Sprache** lässt sich die Codierung umschalten, sofern kein nichtleeres Datenfenster geöffnet ist.

18.3 Ausgabe

Auf dem Registerblatt **Ausgabe** kann man z. B. veranlassen, dass in Pivot-Tabellen neben den Wertbeschriftungen auch die Werte selbst angezeigt werden:



Außerdem kann man zur **Ausgabe-Anzeige** für verallgemeinerte lineare gemischte Modelle Tests die bis SPSS 25 als Voreinstellung verwendete, bei Anwendern aber wenig beliebten **Modellanzeige** anstelle von **Pivot-Tabellen und Diagrammen** wählen.

18.4 Dateispeicherorte

Auf dem Registerblatt **Dateispeicherorte** kann man u. a. einstellen:

Startordner für die Dialogfelder "Öffnen" und "Speichern"

Auf den Pool-PCs an der Universität Trier ist die Einstellung

U:\Eigene Dateien\SPSS

sinnvoll, nachdem der Ordner **SPSS** unterhalb von **U:\Eigene Dateien** angelegt worden ist.

Sitzungsjournal

Per Voreinstellung protokolliert SPSS alle Kommandos, die Sie während einer Sitzung per Dialogbox oder via Syntaxfenster abschicken, in einer sogenannten **Journaldatei**. Diese Datei kann sehr nützlich sein, weil sie die Kommando-Äquivalente zu praktisch allen Arbeiten früherer Sitzungen enthält. Per Voreinstellung wird beim Start einer SPSS-Sitzung eine vorhandene Journaldatei *nicht* überschrieben, sondern die neuen Kommandos werden am Ende angehängt. Falls die Datei zu groß wird, muss sie gelegentlich verkleinert oder gelöscht werden.

19 Anhang

19.1 Weitere Hinweise zur SPSS-Kommandosprache

In Kapitel 6 wurden nur oberflächliche Hinweise zur SPSS-Kommandosprache gegeben. Diese sollten genügen für Anwender, die nicht frei programmieren, sondern nur gelegentlich ein von SPSS automatisch erzeugtes Kommando modifizieren wollen. Der aktuelle Abschnitt ist für ambitionierte Anwender gedacht, die bereit sind, SPSS-Programme zu schreiben, ...

- um auch die ausschließlich per Syntax verfügbaren SPSS-Leistungen nutzen zu können,
- um rationeller mit SPSS zu arbeiten.

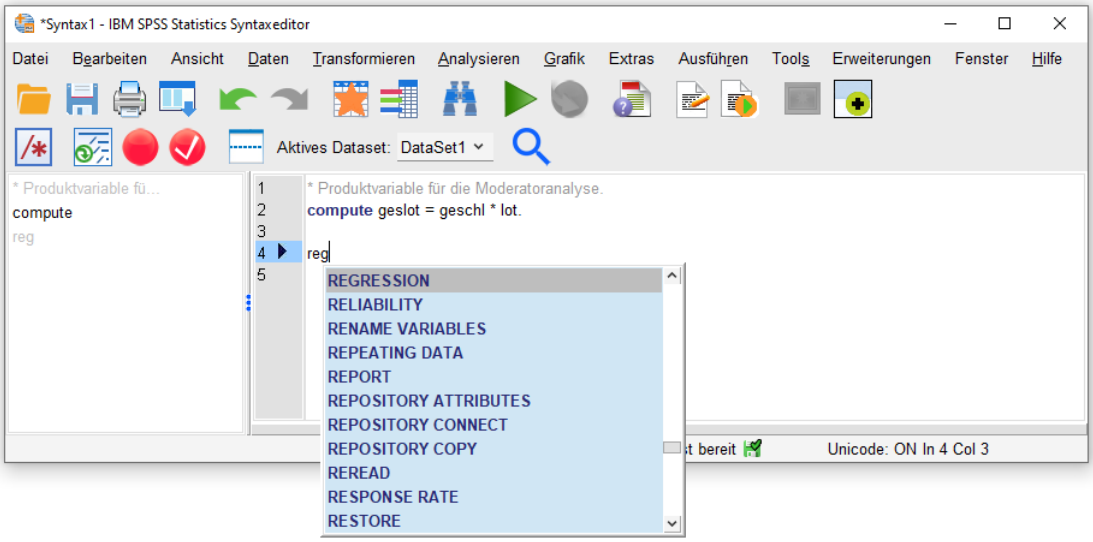
Das wichtigste Hilfsmittel für das Arbeiten mit der SPSS-Kommandosprache ist die *Command Syntax Reference*, die als PDF-Dokument mit ca. 2400 Seiten über das Hilfesystem verfügbar ist:

Hilfe > Befehlssyntaxreferenz

Hier findet man ausführliche Beschreibungen der SPSS-Kommandos mit zahlreichen Beispielen und wertvollen Literaturhinweisen zu den realisierten statistischen Methoden. Das Kapitel 2 (*Universals*) enthält eine Einführung in generelle Themen im Zusammenhang mit Kommandos, Dateien, Variablen und Transformationen. Nützliche Informationen bietet auch eine von der Firma IBM SPSS kostenlos zur Verfügung gestellte syntaxorientierte Einführung in die Datenverwaltung mit SPSS (IBM Corp. 2016).

19.1.1 Hilfsmittel für das Arbeiten mit der SPSS-Kommandosprache

Ein Syntaxfenster zum Erstellen oder Modifizieren von SPSS-Kommandos



erscheint bei Bedarf spontan nach einem Mausklick auf den in vielen Dialogboxen vorhandenen **Einfügen**-Schalter, um die automatisch erzeugte, zur Dialogbox äquivalente Syntax aufzunehmen. Über den Menübefehl

Datei > Neu > Syntax

lässt sich jederzeit ein Syntaxfenster explizit anfordern. Über

Datei > Öffnen > Syntax

oder

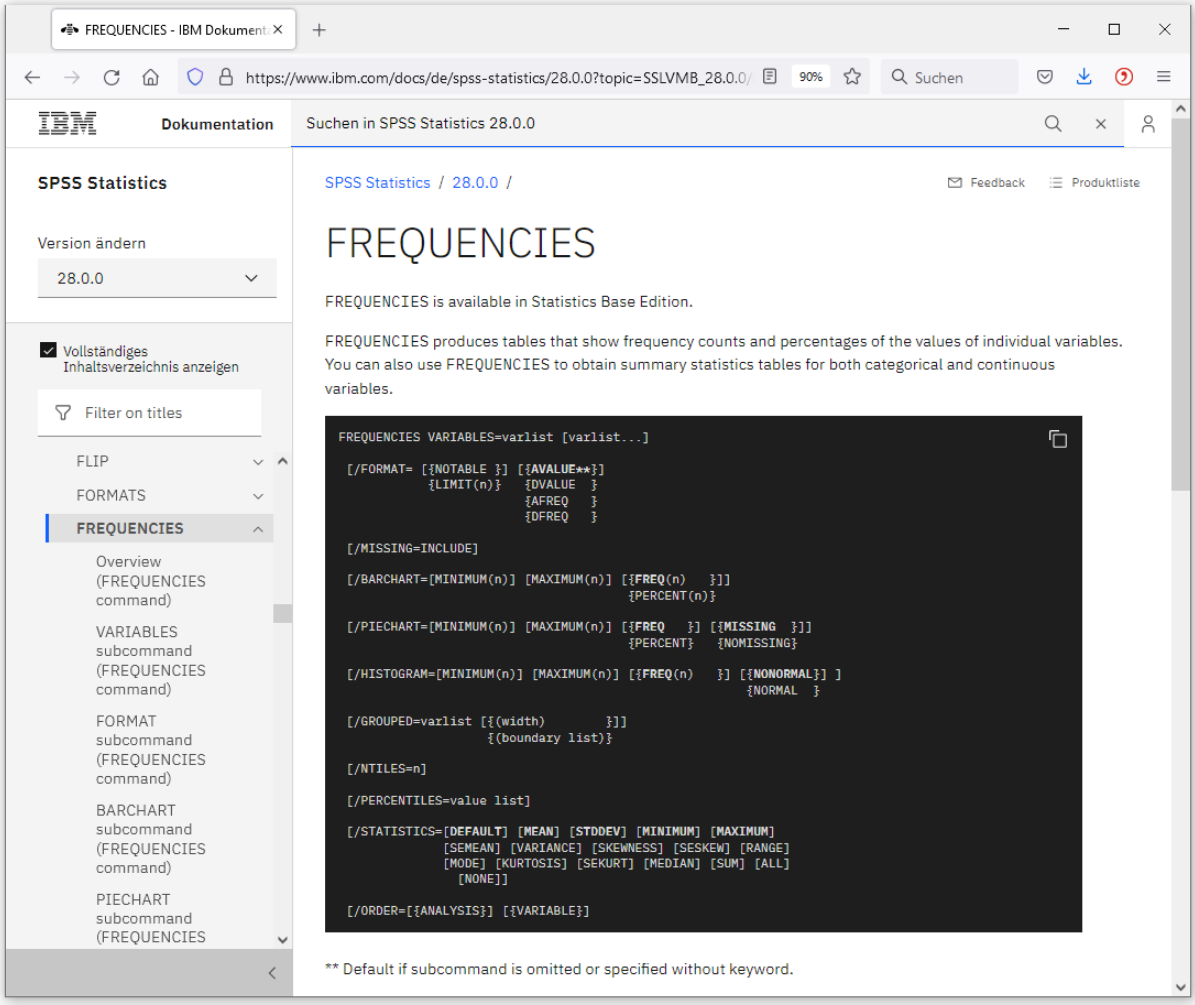
Datei > Zuletzt verwendete Dateien

erhält man ein Syntaxfenster mit dem Inhalt einer vorhandenen Syntaxdatei.

Das Syntaxfenster erleichtert die Bearbeitung von Kommandos u. a. durch …

- eine Navigationszone zur Orientierung in längeren Programmen
- Zeilennummern
- farbliche Unterscheidung verschiedener Syntaxbestandteile
- eine intelligente Syntaxvervollständigung (siehe vorheriges Bildschirmfoto)
- Haltepunkte

Außerdem bietet es ein bequemes Verfahren, Syntaxinformationen zu einem konkreten Kommando anzufordern: Setzen Sie die Schreibmarke auf das Kommando, und klicken Sie dann auf das Symbol  . Zum FREQUENCIES-Kommando, das der **Häufigkeiten**-Dialogbox zugrunde liegt, erscheint z. B. die folgende Erläuterung:



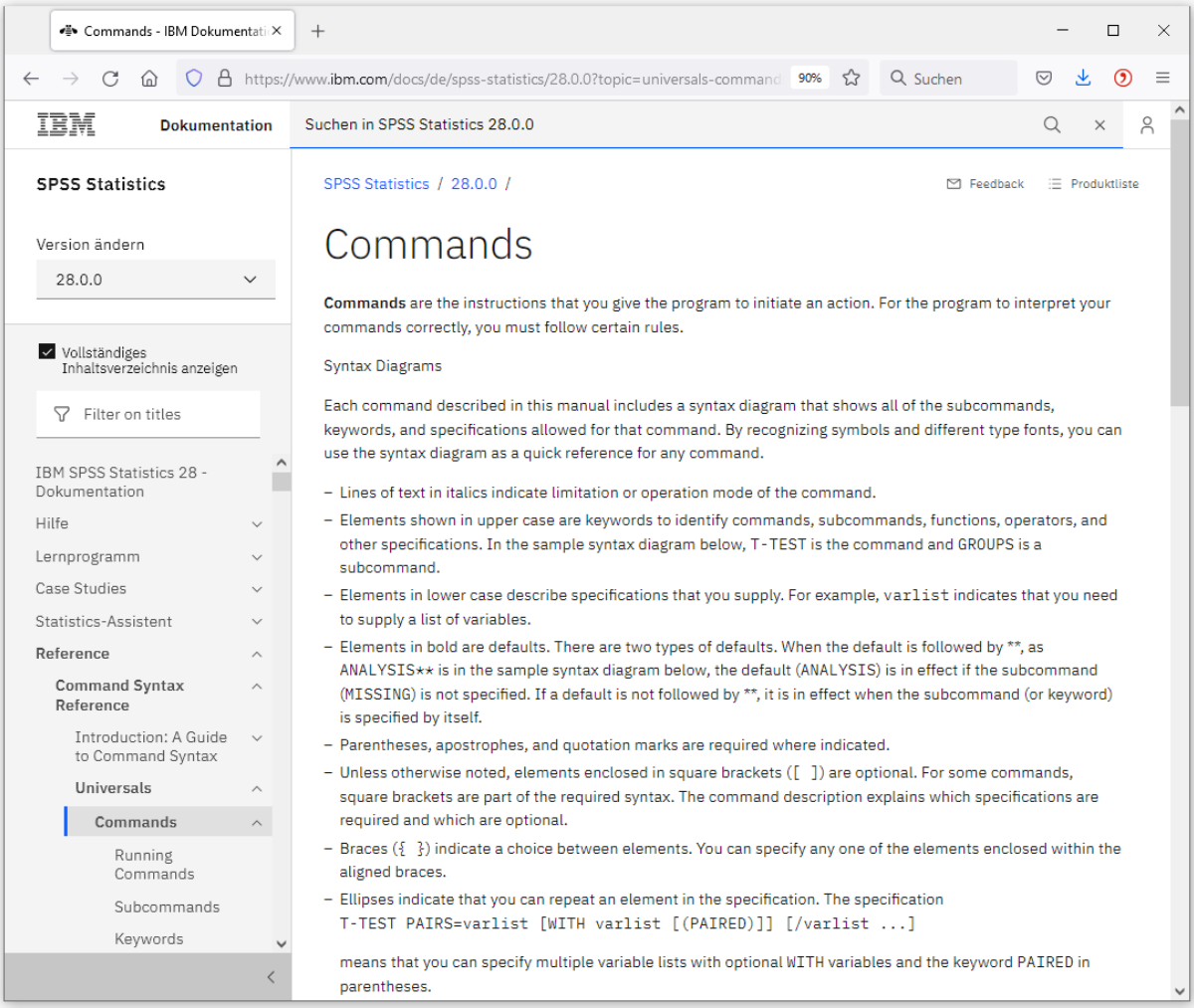
Sie startet mit dem Syntaxdiagramm zum beschriebenen Kommando. Wie ein solches Syntaxdiagramm zu lesen ist, beschreibt der folgende Abschnitt.

19.1.2 Interpretation von Syntaxdiagrammen

Mit einem Syntaxdiagramm wird die allgemeine Form eines Kommandos definiert und somit festgelegt, wie konkrete Beispiele gebildet werden dürfen. Solche Syntaxdiagramme werden auch im weiteren Verlauf dieses Abschnitts benutzt, um Bestandteile der SPSS-Kommandosprache zu erläutern. In den Syntaxdiagrammen treten einige Metazeichen auf (z. B. "[", "{"), die nicht zur Kommandosprache selbst gehören, sondern diese Sprache beschreiben. Die Bedeutung dieser Metazeichen müssen Sie kennen, um Syntaxdiagramme richtig interpretieren zu können. Im Hilfesystem finden Sie eine Erläuterung, indem Sie nach

Hilfe > Themen

den Suchbegriff *syntax diagrams* eintippen und dann in der Trefferliste auf **Commands** klicken:



19.1.3 Aufbau von SPSS-Programmen

Welche Kommandos SPSS für das Erstellen von Programmen bereithalten muss, ergibt sich aus unseren Zielvorstellungen. Wir möchten …

- empirische Daten in ein Datenblatt einlesen,
- gegebenenfalls Variablen verändern und neue Variablen erstellen,
- eventuell ein modifiziertes Datenblatt in eine Datendatei sichern,
- statistische Analysen durchführen oder grafische Darstellungen kreieren.

Darüber hinaus haben wir gelegentlich Sonderwünsche hinsichtlich der Arbeitsweise von SPSS.

Orientiert an den gerade skizzierten Teilaufgaben kann man die verfügbaren SPSS-Kommandos in folgende Gruppen einteilen:

- **Dateidefinitions-Kommandos**
  Sie dienen zum Einlesen von Daten in ein Datenblatt. Als Beispiel haben wir das GET-Kommando kennengelernt. Wenn ein Programm kein Dateidefinitions-Kommando enthält, wenn es also nicht selbst für das Einlesen seiner Daten sorgt, kann es nur ausgeführt werden, wenn zuvor ein Datenblatt erzeugt worden ist.
- **Transformations-Kommandos**
  Diese Kommandos dienen zur Veränderung oder Neuberechnung von Variablen bzw. zur Auswahl von Fällen für die weitere Verarbeitung.
- **Prozedur-Kommandos**
  Damit werden statistische Analysen, grafische Präsentationen oder Dateibearbeitungen (z. B. Sortieren der Fälle) angefordert. Ein Beispiel ist das FREQUENCIES-Kommando. Als Spezialfall gehört auch das Kommando SAVE, das den Inhalt der Arbeitdatei in eine Datendatei sichert, zu den Prozedur-Kommandos.
- **Dienst-Kommandos**
  Damit kann man u. a. die Arbeitsweise von SPSS beeinflussen (z. B. Startwert des Pseudozufallszahlengenerators setzen) und verschiedene Informationen anfordern.

In folgendem SPSS-Programm treten Kommandos aus allen Gruppen auf:

```
comment Größe und Gewicht.                          | Dienst-Kommando

get file = 'kfa.sav'.                               | Dateidef.-Kommando

frequencies var = groesse gewicht                   | Prozedur-
 /statistics = all /histogram = normal.             | Kommando

compute ideal = groesse - 100.                      | Transformations-
                                                    | Kommando

t-test pairs = gewicht ideal.                       | Prozedur-
                                                    | Kommando
```

Ein SPSS-Programm darf selbstverständlich beliebig viele Prozeduren verwenden oder auch mehrere Datenblätter definieren und jeweils durch Transformationen verändern.

19.1.4 Aufbau eines einzelnen SPSS-Kommandos

Die wichtigsten Regeln für SPSS-Kommandos:

- Ein Kommando besteht aus einem Namen und den zugehörigen Spezifikationen:

kommandoname spezifikationen

  - Der **Kommandoname** kann aus *einem* Wort bestehen oder aus mehreren Wörtern.
    Beispiele:
    - FREQUENCIES
    - GET DATA
  - Die **Spezifikationen** dürfen enthalten:
    - Schlüsselwörter (z. B. VARIABLES)
    - Variablennamen
    - Zahlen
    - Zeichenfolgen (z. B. Variablenlabel)
    - Operatoren (z. B. "+")
    - spezielle Begrenzungszeichen: ( ) = ' "

    Zwischen diesen Elementen ist mindestens ein Leerzeichen erforderlich. Ausnahme:
    Die speziellen Begrenzungszeichen, die arithmetischen Operatoren und manche Vergleichsoperatoren (z. B. ">") sind selbstbegrenzend, d. h. davor und danach sind keine Leerzeichen nötig (aber erlaubt).
    Statt eines Leerzeichens darf man meist verwenden:
    - beliebig viele Leerzeichen,
    - ein Komma,
    - einen Zeilenwechsel, wobei aber keine Leerzeile entstehen darf.

    Dies ermöglicht eine übersichtliche Programmgestaltung.
- Jedes Kommando muss in einer neuen Zeile beginnen und mit einem Punkt enden. Ein Kommando muss dabei keinesfalls in der ersten Spalte beginnen, sondern darf eingerückt werden. Von dieser Möglichkeit sollte man z. B. bei DO REPEAT - Schleifenkonstruktionen Gebrauch machen.
  Beispiel:
  ```
  do repeat  mc=mc001 to mc100.
    compute  mc=normal(1).
  end repeat.
  ```
  Hier werden 100 unabhängige, normalverteilte Zufallsvariablen erzeugt. Durch das Einrücken wird deutlich gemacht, dass die COMPUTE-Anweisung innerhalb der DO REPEAT - Schleife steht.
- Ein Kommando kann sich über beliebig viele Fortsetzungszeilen erstrecken. *Innerhalb* eines Kommandos sind aber keine Leerzeilen erlaubt.
- Eine Syntaxzeile sollte maximal 256 Zeichen enthalten, um in allen Kontexten ausführbar zu sein.
- Die Verwendung von Groß- oder Kleinbuchstaben ist beliebig.
- Schlüsselwörter dürfen meist bis auf die ersten drei Zeichen abgekürzt werden.
  Beispiel: "`fre`" für "`frequencies`"

- Bei den meisten Kommandos sind die Spezifikationen in Subkommandos unterteilt. Diese beginnen mit einem Subkommandonamen, meist gefolgt von einem Gleichheitszeichen, und sind durch Schrägstriche voneinander getrennt.
  Beispiel:
```
frequencies var=lot01 /format=notable
  /statistics=all.
```

Merken Sie sich aus dieser Liste für den Anfang vor allem:

JEDES KOMMANDO MUSS IN EINER NEUEN ZEILE BEGINNEN UND MIT EINEM PUNKT ENDEN.

19.1.5 Regeln für Variablenlisten

19.1.5.1 Abkürzende Spezifikation einer Serie von Variablen

In Transformations- oder Prozedurkommandos soll häufig eine Folge **bereits existierender** und **in der Arbeitsdatei hintereinander stehender** Variablen angesprochen werden. Dies ermöglicht das **aufrufende TO**, dessen Syntax im Folgenden erläutert wird:

vara TO *varb*

vara, *varb* — Namen bereits vorhandener Variablen, wobei *vara* in der Arbeitsdatei vor *varb* stehen muss.

Beispiele:

- `frequencies variables = alter to beruf.`
  Für alle Variablen, die in der Arbeitsdatei von ALTER bis BERUF positioniert sind, werden Häufigkeitstabellen erstellt.
- `frequencies variables = frage1 to frage3.`
  Wenn in der Arbeitsdatei zwischen FRAGE1 und FRAGE3 1500 beliebig benannte Variablen stehen, dann bewirkt dieses Kommando 1502 Häufigkeitstabellen!

19.1.5.2 Der Platzhalter varlist

In folgendem Syntaxdiagramm wird der in SPSS-Kommandos häufig auftretende Platzhalter *varlist* definiert:

{*varname* | *varname_1* TO *varname_2*} [{...]

varname, *varname_1*, *varname_2* — Variablennamen

Beispiel:
```
missing values  nieder01 to hoehe ozon mess1 to mess4 (9).
```
Hier wird mit dem MISSING VALUES - Kommando für alle aufgelisteten Variablen die 9 als MD-Indikator vereinbart.

Literaturverzeichnis

Action, C., Miller, R., Fullerton, D. & Maltby, J. (2009). *SPSS for Social Scientists* (2nd ed.). Basingstoke: Palgrave MacMillan.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2nd ed.). Hoboken, NJ: Wiley

Akremi, L. & Baur, N. (2011). Kreuztabellen und Kontingenzanalysen. In: Akremi, L., Baur, N. & Fromm, S. (Hrsg.). *Datenanalyse mit SPSS für Fortgeschrittene 1*. (3 Aufl.). Wiesbaden: VS Verlag.

Antonakis, J., Bendahan, S., Jacquart, P. & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21, 1086–1120.

APA (2010). *Publication Manual* (6th ed.). Washington, DC: American Psychological Association.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2008). *Multivariate Analysemethoden* (12. Aufl.). Berlin: Springer.

Backhaus, K., Erichson, B., & Weiber, R. (2015). *Fortgeschrittene multivariate Analysemethoden* (3. Aufl.). Berlin: Springer.

Baltes-Götz, B. (1998). *Exakte Tests mit SPSS*. Online-Dokument: http://www.uni-trier.de/index.php?id=22571

Baltes-Götz, B. (2012). *Logistische Regressionsanalyse mit SPSS*. Online-Dokument: http://www.uni-trier.de/index.php?id=22513

Baltes-Götz, B. (2013). *Behandlung fehlender Werte in SPSS und Amos*. Online-Dokument: http://www.uni-trier.de/index.php?id=23239

Baltes-Götz, B. (2015). *Analyse von Strukturgleichungsmodellen mit Amos 18*. Online-Dokument: http://www.uni-trier.de/index.php?id=22640

Baltes-Götz, B. (2016a). Regressionsmodelle für Paneldaten. In A. Geissler & M. Schneider (Hrsg.). *Zwischen artes liberales und artes digitales*, S. 157-181. Marburg: Tectum-Verlag.

Baltes-Götz, B. (2016b). *Generalisierte lineare Modelle und GEE-Modelle in SPSS*. Online-Dokument: http://www.uni-trier.de/index.php?id=51455

Baltes-Götz, B. (2019). *Lineare Regressionsanalyse mit SPSS*. Online-Dokument: http://www.uni-trier.de/index.php?id=22489

Baltes-Götz, B. (2020a). *Mediator- und Moderatoranalyse per multipler Regression mit SPSS*. und PROCESS. Online-Dokument: http://www.uni-trier.de/index.php?id=22528

Baltes-Götz, B. (2020b). *Analyse von hierarchischen linearen Modellen mit SPSS*. Online-Dokument: http://www.uni-trier.de/index.php?id=39127

Baltes-Götz, B. (2021). *Online-Umfragen mit Enterprise Feedback Suite Survey*. Online-Dokument: http://www.uni-trier.de/index.php?id=52985

Baur, N. & Fromm, S. (2011). Nützliche Software und Fundorte für Daten. In: Akremi, L., Baur, N. & Fromm, S. (Hrsg.). *Datenanalyse mit SPSS für Fortgeschrittene 1*. (3 Aufl.). Wiesbaden: VS Verlag.

Bortz, J. & Döring, N. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Berlin: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin: Springer.

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. In R. L. Launer & G. N. Wilkinson (Eds.). *Robustness in Statistics*, pp. 201-236. New York: Academic Press.

Brandstätter, E. (1999). Konfidenzintervalle als Alternative zu Signifikanztests. *Methods of Psychological Research Online*, Vol.4, No.2. Online-Dokument: http://www.dgps.de/fachgruppen/methoden/mpr-online/issue7/art3/article.html

Brosius, F. (2013). *SPSS 21*. Heidelberg: mitp.

Brosius, F. (2018). *SPSS: Umfassendes Handbuch zu Statistik und Datenanalyse*. Heidelberg: mitp.

Brüderl, J. (2010). Kausalanalyse mit Paneldaten. In C. Wolf & H. Best (Hrsg.). *Handbuch der sozialwissenschaftlichen Datenanalyse*, S. 963-994. Wiesbaden: VS Verlag.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson.

Bühner, M & Ziegler, M (2017). *Statistik für Psychologen und Sozialwissenschaftler* (2. Aufl.). München: Pearson.

Bühl, A. (2016). *SPSS 23. Einführung in die moderne Datenanalyse* (15. Aufl.). München: Pearson.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Academic Press.

Cohen, J., Cohen, P., West, S.G. & Aiken, L. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). Mahwah: Lawrence Erlbaum Associates.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: SIAM-Monograph #38.

Eid, M., Gollwitzer, M. & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5. Aufl.). Weinheim: Beltz.

Faik, J. (2018). *Statistik mit SPSS: Alles in einem Band für Dummies, Weinheim:* Wiley.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149-1160.

Field, A. (2017). *Discovering Statistics Using SPSS* (5th ed.). London: SAGE Publications.

Frees, B. & Koch, W. (2018). ARD/ZDF-Onlinestudie 2018: Zuwachs bei medialer Internetnutzung und Kommunikation. *Media Perspektiven*, 9/2018. Online-Dokument: http://www.ard-zdf-onlinestudie.de/files/2018/0918_Frees_Koch.pdf

Gehring, U.W. & Weins, C. (2004). *Grundkurs Statistik für Politologen* (4. Aufl.). Wiesbaden: VS Verlag.

Griffith, A. (2010). *SPSS For Dummies*, (2nd ed.). Indianapolis, Indiana: Wiley.

Hartung, J. (1989). *Statistik* (7. Auflage). München: Oldenbourg.

Hayes, A.F. (2018). *Mediation, Moderation, and Conditional Process Analysis*. (2nd ed.). New York: Guilford Press.

IBM Corp. (2016). *Programming and Data Management for IBM SPSS Statistics 24*. Online-Dokument: https://community.ibm.com/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=6484b416-0961-d151-0873-855f06879567&forceDialog=0

IBM Corp. (2021a). *IBM SPSS Data Preparation 28*. Online-Dokument: https://www.ibm.com/docs/SSLVMB_28.0.0/pdf/de/IBM_SPSS_Data_Preparation.pdf

IBM Corp. (2021b). *Benutzerhandbuch zum IBM SPSS Statistics 28 Core-System*. Online-Dokument: https://www.ibm.com/docs/SSLVMB_28.0.0/pdf/de/IBM_SPSS_Statistics_Core_System_User_Guide.pdf

IBM Corp. (2021c). *GPL Reference Guide for IBM SPSS Statistics*. Online-Dokument: https://www.ibm.com/docs/SSLVMB_28.0.0/pdf/GPL_Reference_Guide_for_IBM_SPSS_Statistics.pdf

Jacob, R., Heinz, A. & Décieux, J.P. (2013). *Umfrage* (3. Aufl.). München: Oldenbourg.

Kahneman, D. & Miller, D.T. (1986) Norm theory: comparing reality to its alternatives. *Psychological Review*, 93, 136-153.

Kerchoff, A.C. (1974). *Ambition and attainment*. Rose Monograph Series.

Lehmann, E.L. (1993). The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? *Journal of the American Statistical Association*, 88(424), 1242-1249.

Liang, K. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.

Liddell, T.M. & Kruschke, J.K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79, 328-348.

Lück, D (2011). Mängel im Datensatz beseitigen. In: Akremi, L., Baur, N. & Fromm, S. (Hrsg.). *Datenanalyse mit SPSS für Fortgeschrittene 1*. (3 Aufl.). Wiesbaden: VS Verlag.

MacCallum, R.C., Zhang, S., Preacher, K.J. & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7(1), 19-40.

Mehta, C.R. & Patel, N.R (2010). *IBM SPSS Exact Tests*. Online-Dokument: https://www.researchgate.net/publication/265357333_SPSS_exact_tests

Norušis, M.J. (2012a). *IBM SPSS Statistics 19 Guide to Data Analysis*. Upper Saddle River, NJ: Prentice Hall.

Norušis, M.J. (2012b). *IBM SPSS Statistics 19 Statistical Procedures Companion*. Upper Saddle River, NJ: Prentice Hall.

Norušis, M.J. (2012c). *IBM SPSS Statistics 19 Advanced Statistical Procedures*. Upper Saddle River, NJ: Prentice Hall.

Pedhazur, E.J. & Pedhazur Schmelkin L. (1991). *Measurement, design, and analysis. An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Rasch, B., Friese, M., Hofmann, W. & Naumann, E. (2014). *Quantitative Methoden* (Band 1 und 2, 4. Aufl.). Berlin: Springer.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Rumsey, D. (2008). *Übungsbuch Statistik für DUMMIES*. Weinheim: Wiley.

Scheier, M.F. & Carver, C.S. (1985). Optimism, Coping, Health: Assessment and implications of generalized outcome expectancies. *Health Psychology*, 4, 219-247.

Schnell, R., Hill, P. B. & Esser, E. (2018). *Methoden der empirischen Sozialforschung* (11. Aufl.). München: Oldenbourg.

Siegel, S. (1976). *Nichtparametrische statistische Methoden*. Frankfurt: Fachbuchhandlung für Psychologie

Snijders, T. A. B. & Bosker, R. L. (2012). *Multilevel Analysis*. Los Angeles, CA: SAGE Publications.

Statistisches Bundesamt. (2016). *Statistik und Wissenschaft. Demografische Standards*. Online-Dokument: https://www.destatis.de/DE/Methoden/Demografische-Regionale-Standards/textbaustein-demografische-standards.html

Sullivan, G.M. & Artino Jr. A.R. (2013). Analyzing and Interpreting Data From Likert-Type Scales, *Journal of Graduate Medical Education*, 5(4), 541-542.

Tabachnik, B.G. & Fidell, L.S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.

Urban, D. & Fiebig, J. (2015). *Quantitative Meta-Analyse zur Überprüfung sozialwissenschaftlicher Hypothesen*. Weinheim: Beltz Juventa.

Vlaeminck, S., et al. (2015). *Auffinden, zitieren, dokumentieren. Forschungsdaten in den Sozial- und Wirtschaftswissenschaften*. Version 2.0. GESIS Leibniz Institute for the Social Sciences. Online-Dokument: http://doi.org/10.4232/10.fisuzida2015.2

Wagner, W.E. (2017). *Using IBM SPSS Statistics for Research Methods and Social Science Statistics*. Thousand Oaks, CA: SAGE Publications.

Wallis, W.A. & Roberts, H.V. (1956). *Statistics, a new approach*. Glencoe, Ill.: The Free Press.

Warner, R.M. (2013). *Applied statistics: from bivariate through multivariate techniques* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Wooldridge, J. M. (2013). *Introductory Econometrics: A modern approach* (5th ed.). Cengage Learning.

Stichwortregister

E

F

G

H

I

J

K

L

M

S

T

U

V

W

Z