

Themengebiet: Statistik  
Betriebssystem: Windows  
Zielgruppe: Anwender

# Statistische Datenanalyse mit SPSS für Windows



## Grundlegende Konzepte und Techniken

**Version 2.7, 26. Februar 2009**

Universität Osnabrück

Rechenzentrum

Dipl.-Math. Frank Elsner

Mail: [Frank.Elsner@uni-osnabrueck.de](mailto:Frank.Elsner@uni-osnabrueck.de)

Internet: <http://www.home.uni-osnabrueck.de/elsner/Skripte/spss.pdf>

# Inhaltsverzeichnis

1	Einleitung.....	6
1.1	Voraussetzungen.....	6
1.2	Überblick über die behandelten Themen.....	6
1.3	Weiterführende Literatur.....	7
1.4	Download.....	7
1.5	Feedback.....	7
2	Besonderheiten für die Universitaet Osnabrück.....	8
2.1	Lizensierte SPSS Module.....	8
2.2	Miete von SPSS für Windows.....	8
2.3	Dokumentation.....	8
3	Überblick über SPSS für Windows.....	9
3.1	Funktionsumfang von SPSS für Windows.....	9
3.2	Typische Arbeitsschritte.....	9
4	Einführendes Beispiel.....	10
4.1	Fragebogen.....	10
4.2	Kennenlernen der Benutzeroberfläche.....	10
4.3	Definieren von Variablen.....	12
4.4	Begriffe.....	15
4.5	Anzeigen der Werte-Etiketten (Werte-Label).....	16
4.6	Speichern von Variablen- und Datenansicht.....	16
4.7	Übungen.....	17
4.8	Einlesen aus externen Quellen.....	18
5	Berechnen neuer Variablen.....	19
5.1	Beobachtungen und Variablen.....	19
5.2	Berechnen einer Altersgruppe.....	19
5.3	Erläuterungen.....	22
5.4	Übungen.....	22
6	Auswählen von Beobachtungen.....	24
6.1	Filtern von Beobachtungen.....	24
6.2	Übungen.....	25
7	Arbeiten mit Kalenderdaten.....	27
7.1	Definieren von Variablen mit Datentyp DATUM.....	27
7.2	Erläuterungen.....	28
7.3	Funktionen für den Datentyp Datum.....	28
7.4	Übungen.....	30

7.5 Berechnen von Durchschnittswerten .....	30
7.6 Übungen.....	31
8 Überblick über die deskriptive Statistik.....	33
8.1 Aufgaben der deskriptiven Statistik.....	33
8.2 Tabellen und Diagramme.....	33
8.3 Stichprobe und Grundgesamtheit.....	34
8.4 Auswahlmechanismen.....	35
8.5 Kennzeichen einer Stichprobe.....	35
8.6 Messung von Variablen.....	36
8.7 Meßniveau.....	37
8.8 Kenngrößen von Stichproben.....	37
8.9 Übungen.....	40
9 Erstellen von einfachen Tabellen .....	41
9.1 Berechnen von Häufigkeiten.....	41
9.2 Erstellen einer Kreuztabelle.....	43
9.3 Übungen.....	44
10 Berechnen von Kennzahlen.....	45
10.1 Berechnen einfacher Kennzahlen.....	45
10.2 Übungen.....	47
11 Erstellen von Diagrammen.....	49
11.1 Visualisieren von Datenmaterial.....	49
11.2 Erstellen eines einfachen Balkendiagramms.....	50
11.3 Erstellen eines gruppierten Balkendiagramms.....	53
11.4 Übungen.....	54
11.5 Erstellen eines Flächendiagramms.....	55
11.6 Erstellen eines gestapelten Flächendiagramms.....	57
11.7 Erstellen eines Histogramms.....	59
11.8 Übungen.....	60
11.9 Vergleichen von empirischen Verteilungen.....	61
11.10 Bearbeiten von Diagrammen.....	63
11.11 Übungen.....	64
12 Zufallsexperimente, Zufallsvariablen und Wahrscheinlichkeit.....	65
12.1 Zufallsexperiment und Wahrscheinlichkeit.....	65
12.2 Zufallsvariablen und ihre Verteilung.....	66
12.3 Vorher und Nachher.....	66
12.4 Aufgaben.....	67
13 Überblick über die mathematische Statistik.....	69

13.1 Ziehen von Rückschlüssen aus einer Stichprobe.....	69
13.2 Durchführen von Schätzungen und Hypothesentests.....	70
13.3 Einschränken der gesuchten theoretischen Verteilung auf eine Klasse (parametrische Tests).....	71
13.4 Formulieren von Fragestellungen.....	72
13.5 Treffen von Entscheidungen anhand einer Entscheidungsregel.....	72
13.6 Entscheidungsregel.....	74
13.7 Übungen.....	75
14 Berechnen eines Vertrauensbereiches (Konfidenz-Intervalls).....	77
14.1 Interpretieren von Vertrauensbereichen.....	77
14.2 Berechnen eines Vertrauensbereichs.....	78
14.3 Ableiten der Formel für den Vertrauensbereich.....	80
14.4 Übungen.....	82
15 Testen der Unabhängigkeit .....	83
15.1 Berechnen der Chi-Quadrat-Testgröße.....	83
15.2 Ableiten der Formel für die Chi-Quadrat Testgröße.....	87
15.3 Übungen.....	88
16 Berechnen von Korrelationskoeffizienten.....	89
16.1 Festlegen eines Maßes für den linearen Zusammenhang.....	89
16.2 Ermitteln des Korrelationskoeffizientens.....	90
16.3 Übungen.....	92
17 Approximieren von x-y-Punkten durch Geraden (lineare Regression).....	94
17.1 Untersuchen eines möglichen linearen Zusammenhangs.....	94
17.2 Durchführen einer linearen Regression.....	95
17.3 Visualisieren der linearen Regression.....	98
17.4 Bewerten der Güte eines Regressionsmodells.....	99
17.5 Übungen.....	101
18 Vergleichen von 2 Gruppenmittelwerten (t-Test).....	102
18.1 Interpretieren von Unterschieden zwischen Gruppen.....	102
18.2 Testen auf gleiche Erwartungswerte.....	102
18.3 Übungen.....	104
19 Vergleichen mehrerer Gruppenmittelwerte (Varianz-Analyse).....	105
19.1 Aufstellen eines ein-faktoriellen Modells.....	105
19.2 Vergleichen von mehren unabhängigen Stichproben.....	106
19.3 Aufstellen eines mehr-faktoriellen Modells.....	108
19.4 Zurückführen der Varianz-Analyse auf ein lineares Modell.....	108
19.5 Übungen.....	108

20 Reduzieren der Variablenanzahl (Faktor-Analyse).....	110
20.1 Ermitteln von gemeinsamen Faktoren.....	110
20.2 Durchführen einer Faktoren-Analyse.....	111
20.3 Interpretieren der Faktoren in einem Streudiagramm.....	115
20.4 Reduzieren der Variablenanzahl.....	116
20.5 Übungen.....	117
21 Exploratives Analysieren von Daten .....	119
21.1 Exploratives Analysieren von Daten.....	119
21.2 Testen auf Normalverteilung.....	119
21.3 Testen auf Varianzhomogenität.....	122
21.4 Übungen.....	124
22 Zusammenfassen von Beobachtungen in Clustern (Cluster-Analyse).....	125
22.1 Zusammenfassen von Beobachtungen.....	125
22.2 Ermitteln von hierarchisch geordneten Clustern.....	127
22.3 Übungen.....	131
23 Anhang.....	132
23.1 Bundestagswahlen	
.....	132

# 1 Einleitung

In diesem Kapitel wird ein Überblick über Voraussetzungen, Zielsetzung und Aufbau dieses Skriptes gegeben.

## 1.1 Voraussetzungen

Dieses Skript wendet sich an Benutzer, die mit dem Programm **SPSS für Windows** (im folgenden kurz mit **SPSS** bezeichnet) menügeführt Daten einlesen und im Anschluß statistische Datenanalysen durchführen wollen.

Grundlegende Kenntnisse über Microsoft **Windows** werden vorausgesetzt wie auch grundlegende **wahrscheinlichkeitstheoretische und statistische Kenntnisse** wie sie in der Oberstufe des Gymnasiums oder in einführenden Veranstaltungen an der Universität vermittelt werden.

Dieses Skript ist als Begleitmaterial zu einem Kurs des Rechenzentrums mit Theorie- und Praxis-Anteilen konzipiert und ist deshalb nur mit Einschränkungen zum Selbststudium geeignet.

Der Autor hat sich übrigens entschlossen, vom üblichen (passiven) deutschen Lehrbuchstil - „Man öffnet ...“, „Die Seite wird gespeichert.“ usw. - abzuweichen und den Leser immer „persönlich“ anzusprechen - „Starten Sie das Programm ...“ .

## 1.2 Überblick über die behandelten Themen

In diesem Skript werden folgende Fragestellungen behandelt:

1. Wie können Sie Variablen definieren und die zugehörigen Datenwerte direkt im SPSS Dateneditor erfassen?
2. Wie können Sie Daten tabellarisch und grafisch darstellen, um sich einen Überblick über das Datenmaterial zu verschaffen und damit Anregungen für weiterführende Analysen zu erhalten?
3. Wie können Sie beschreibende Statistiken der Stichprobe wie z.B. Mittelwert, empirische Varianz, empirische Standardabweichung oder Spannweite berechnen?
4. Wie können Sie Stichproben mit statistischen Verfahren untersuchen und die von SPSS gelieferten Ergebnisse interpretieren (Stichworte: Test auf Unabhängigkeit, Konfidenz-Intervall, t-Test, Varianz-Analyse, ...)?

In jedem Kapitel wird zunächst anhand eines Fallbeispiels eine kurze motivierende Einführung in gegeben. Im Anschluß wird die prinzipielle Vorgehensweise anhand des Fallbeispiels erläutert. Gelegentlich wird der statistische Hinter-

grund in einem anschließenden Abschnitt ausführlicher dargestellt, um die Interpretation der Ergebnisse fundierter begründen zu können.

Jedes Kapitel enthält zahlreiche Übungen zur Vertiefung des Stoffes. Alle im Handbuch genannten Übungsdateien stehen maschinenlesbar zur Verfügung (siehe Download). Anspruchsvollere Aufgaben in Übungen sind durch den Hinweis „[zusätzlich]“ gekennzeichnet und können bei Zeitmangel ausgelassen werden.

## 1.3 Weiterführende Literatur

In diesem Skript wird der Funktionsumfang von SPSS nur in ausgewählten Teichbereichen vorgestellt, zum anderen ist die Beschreibung des Menüsystems sehr kurz gehalten. Abhängig von Ihren konkreten Aufgaben finden Sie weitere Informationen in der kostenlos im Lieferumfang enthaltenen, vollständigen SPSS Dokumentation (Adobe Acrobat Portable Document Format [PDF]), im integrierten SPSS Hilfesystem oder in weiterführender Literatur zur Wahrscheinlichkeitsrechnung und Statistik und zum Programm SPSS.

## 1.4 Download

Das Skript und Material zu diesem Skript finden Sie unter:

- Skript <http://www.home.uni-osnabrueck.de/elsner/Skripte/spss.pdf>
- Material <http://www.home.uni-osnabrueck.de/elsner/Skripte/Material/SPSS>  
(nur nach vorheriger Authentifizierung)
- Handbücher: <https://softdist.rz.uni-osnabrueck.de/SPSS/>  
(nur nach vorheriger Authentifizierung)

## 1.5 Feedback

Falls Sie Anregungen oder Kommentare zu diesem Skript haben oder einfach nur Lob oder Kritik loswerden wollen, schicken Sie doch bitte einfach eine Mail an:

[Frank.Elsner@uni-osnabrueck.de](mailto:Frank.Elsner@uni-osnabrueck.de)

## 2 Besonderheiten für die Universitaet Osnabrück

Die folgenden Hinweise sind nur für **Studenten und Mitarbeiter der Universität Osnabrück** relevant.

### 2.1 Lizenzierte SPSS Module

Als Grundlage dieses Skriptes dient **SPSS für Windows**, Version 16, deutsch.

Die an der Universität Osnabrück verfügbare Lizenz von **SPSS für Windows** enthält sämtliche verfügbaren Module. Die meisten der behandelten Funktionen oder Menüpunkte sind bereits in den Vorgängerversionen ab Version 12.0 vorhanden.

Beachten Sie, daß andere Universitäten oder Firmen ggf. weniger Module lizenziert haben, so daß dementsprechend weniger Menüpunkte oder Funktionen zur Verfügung stehen.

Weitere Informationen zu SPSS finden Sie beim Hersteller:

SPSS GmbH: <http://www.spss.com>

### 2.2 Miete von SPSS für Windows

Studenten und Mitarbeiter der Universität Osnabrück können **SPSS** für jeweils ein Jahr zur Miete im Sekretariat des Rechenzentrums erwerben.

Nähere Informationen finden Sie auf der Heimatseite des Rechenzentrums unter A-Z, dort: Software zum Erwerb:

<http://www.rz.uni-osnabrueck.de>

### 2.3 Dokumentation

Die Handbücher zu **SPSS** stehen auf dem Installations-Datenträger von SPSS im Adobe Acrobat PDF Format zur Verfügung. Das Rechenzentrum stellt die Handbücher zusätzlich innerhalb des Hochschulnetzes zum Download zur Verfügung:

Download Handbücher:

<https://softdist.rz.uni-osnabrueck.de/SPSS/>

Die gedruckte Fassung kann in einigen Buchhandlungen erworben werden, wie auch umfangreiche Sekundärliteratur zu SPSS.

## 3 Überblick über SPSS für Windows

In diesem Kapitel erhalten Sie einen kurzen Überblick über den Funktionsumfang und die Bedienung von **SPSS für Windows** sowie über typische Arbeitsschritte anhand eines einfachen Beispiels.

### 3.1 Funktionsumfang von SPSS für Windows

**SPSS für Windows** ist ein modular aufgebautes **Statistik-Analyse-System**. Es besteht aus einem Basis-System, das bereits das komplette Daten- und Dateimanagement, sämtliche Grafiktypen und eine breite Palette an statistischen Funktionen umfasst, und kann durch zusätzliche Module, die die statistische Leistungsfähigkeit des Basis-Systems erweitern, ergänzt werden.

### 3.2 Typische Arbeitsschritte

Die Vorbereitungen für eine statistische Datenanalyse (wie z. B. Auswahl der befragten Personen bzw. Meßaufbau, Design eines Fragebogens, Kodierung der Antworten) sind nicht Gegenstand dieses Skriptes. Einen kurzen Überblick liefert die Broschüre **SPSS Survey Tipps**.

Sie führen bei einer statistischen Datenanalyse in der Regel die folgenden Schritte durch:

1. **Definieren von Variablen** und ggf. **Zuordnen von beschreibenden Namen** (Etiketten, Umschreibungen) für Variablen (*variable labels*) und Datenwerte (*value labels*), um die spätere Text- und Grafik-Ausgabe aussagekräftiger zu gestalten
2. **Erfassen** der kodierten Daten, **Kontrollieren** auf Eingabefehler und **Speichern** in eine SPSS Arbeitsdatei „MeineDaten.sav“ auf Festplatte
3. (*Optional*) **Transformieren** der Daten in eine zweckmäßigere Form bzw. **Erzeugen** von neuen Variablen bzw. **Zusammenfassen** von Daten
4. (*Optional*)- **Auswählen** einer Teilmenge von Fällen und/oder Variablen für die folgende Analysen
5. Tabellarisches **Darstellen** und **Berechnen** von statistischen Kennzahlen zur Vorbereitung von statistischen Analysen
6. Grafisches Darstellen (**Visualisieren**) zur Vorbereitung von statistischen Analysen
7. **Analysieren** der Daten mit Verfahren der mathematischen Statistik

# 4 Einführendes Beispiel

In diesem Kapitel wird als Basis ein einfacher Fragebogen für eine Befragung von Passanten verwendet – der Fragebogen thematisiert in einfacher Form die sogenannte "Sonntagsfrage": Welche Partei würden Sie wählen, wenn am nächsten Sonntag Bundestagswahlen stattfinden würden?

**Zielsetzung**

Wie kann ich einen Fragebogen in eine SPSS Datendatei umsetzen?

## 4.1 Fragebogen

Der Fragebogen ist folgendermaßen aufgebaut:

Name der Variablen	Fragetext bzw. Hinweise	Kodierung (Wert) und Etiketten	Antwort (kodiert)
id	Fragebogen-Kennung; Eindeutige Kennung des Fragebogens, Interviewer-ID + laufende Nummer des Fragebogens	001-0001, 001-0002 usw.	□□□-□□□□
sex	Geschlecht; <i>(wird vom Interviewer ausgefüllt)</i>	1 (weiblich) 2 (männlich) 0 (keine Angabe)	□
alter	Alter; Wie alt sind Sie?	<i>nnn</i> -1 (keine Angabe)	□□□ / □ (k.A.)
partei	Partei; Welche Partei würden Sie wählen, wenn am nächsten Sonntag eine Bundestagswahl stattfinden würde?	1 (CDU/CSU) 2 (FDP) 3 (SPD) 4 (Grüne/Bündnis 90) 5 (PDS) 6 (Republikaner) 7 (Sonstige) -1 (keine Angabe)	□□ / □ (k.A.)

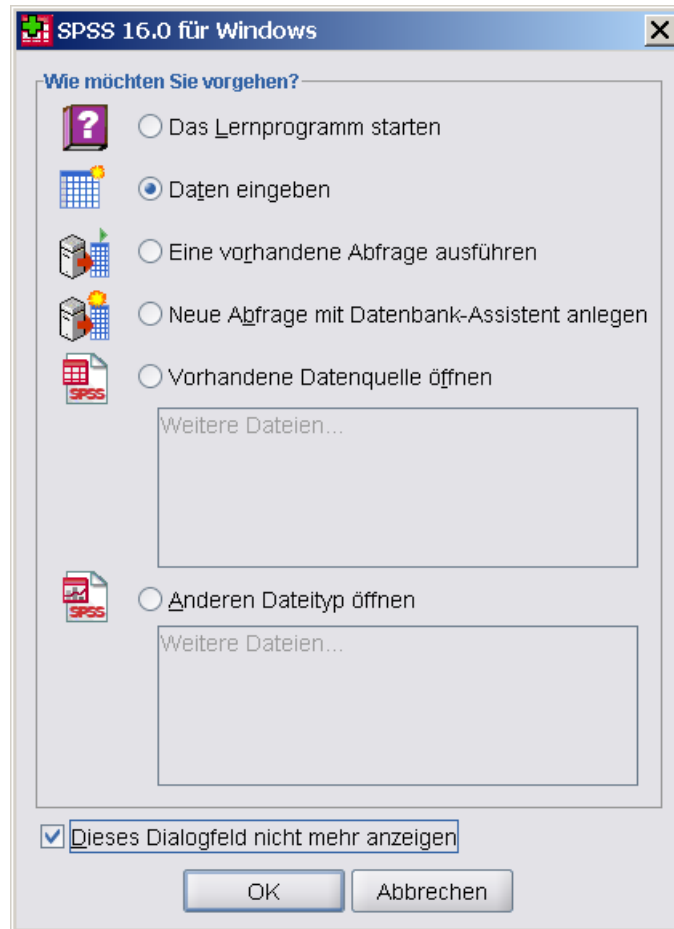
Tabelle 1: Fragebogen zur Sonntagsumfrage (fragebogen-01.rtf)

## 4.2 Kennenlernen der Benutzeroberfläche

Starten Sie eine SPSS Sitzung über den folgenden Menüpunkt (bzw. über den auf Ihrem Rechner für SPSS eingerichteten Menüpunkt)::

**„Start > Programme > SPSS > SPSS 16 (deutsch)“**

Wählen Sie dann im **SPSS Assistenten**, der verschiedene Möglichkeiten zur Nutzung von SPSS vorschlägt, die 2. Möglichkeit „Daten eingeben“ und deaktivieren Sie den Assistenten für die Zukunft, indem Sie „Dieses Dialogfeld nicht mehr anzeigen“ anklicken.



Beenden Sie den Dialog über die Schaltfläche „OK“.

Sie arbeiten in SPSS zu Beginn einer Sitzung im „**SPSS Dateneditor**“-Fenster, in dem Sie interaktiv Variablen und Daten eingeben und in das Sie auch gespeicherte SPSS Datendateien (<datei>.sav) laden können.

Die folgende Abbildung zeigt das leere „Dateneditor“-Fenster direkt nach dem Programm-Start.

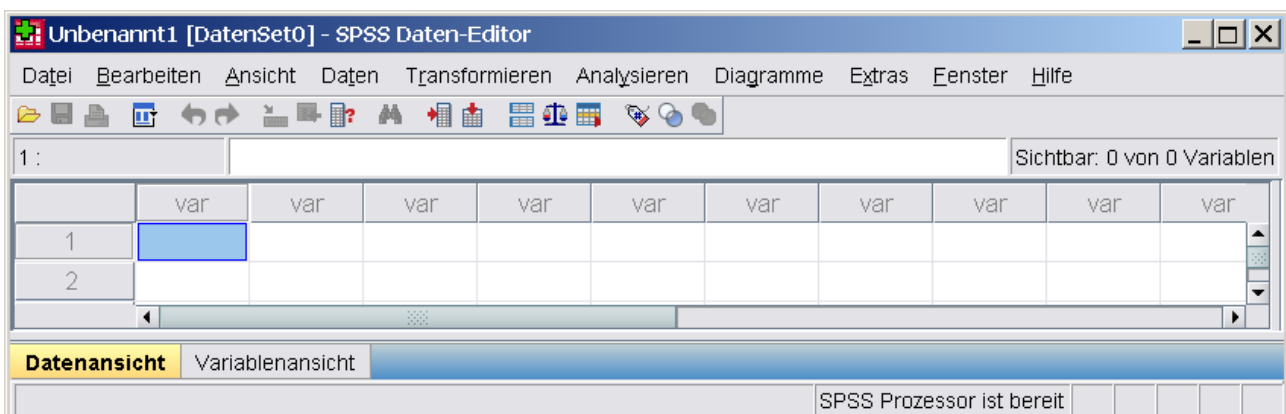


Abbildung 4.1: Daten-Editor (Datenansicht)

Wählen Sie den Reiter „**Variablenansicht**“ (unten links, 2. Position), um auf die Ansicht zum Definieren von Variablen umzuschalten:

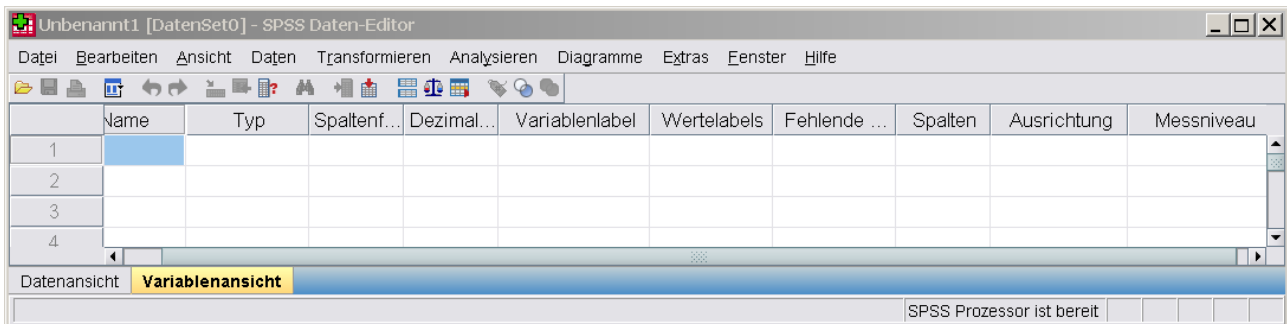


Abbildung 4.2: Daten-Editor (Variablenansicht)

## 4.3 Definieren von Variablen

In der Variablenansicht erstellen Sie für jede Variable eine Zeile mit den benötigten Eigenschaften. Zwingend notwendig sind die Eigenschaften **Name**, **Typ** und **Messniveau**, alle weiteren Eigenschaften können auf Voreinstellungen belassen werden.

Eigenschaft	Bedeutung	mögliche Werte	Beispiel
Name	Variablenname, voreingestellt sind die Namen: var0001, var002 usw.	maximal 8 Zeichen, 1. Zeichen Buchstabe, danach Buchstaben, Ziffern und einige Sonderzeichen wie \$ oder _	sex
Typ	Datentyp, voreingestellt sind Dezimalzahlen mit 8 Stellen, hiervon 2 Dezimalstellen	Zahl, Datum, Währung, Zeichenkette	Zahl
Spaltenformat	Länge der Anzeige		8
Dezimalstellen	Anzahl der Nachkommastellen bei Dezimalzahlen		0
Variablenlabel (variable labels)	sprechende Bezeichnung für den Variablennamen	max. 18 Zeichen, Leerstellen usw. erlaubt	Geschlecht
Wertelabel (value label)	sprechende Bezeichnung für einen Wert	Jeweils Paarbildung Wert=Label	1 = weiblich 2 = männlich 0 = keine Angabe
(benutzerdefinierte) fehlende Werte	Werte, die von SPSS als ungültig behandelt werden sollen	z.B. ein einzelner Wert wie Null oder ein Bereich von ungültigen Werten	0

Eigenschaft	Bedeutung	mögliche Werte	Beispiel
Meßniveau	Skalenniveau oder Rangordnung	metrisch, ordinal, nominal	nominal (keine Rangordnung)

Tabelle 2: Definition einer Variablen und deren Eigenschaften

Definieren Sie nun die Eigenschaften der Variablen „id“, „sex“, „alter“ und „partei“ bezogen auf den Fragebogen für die Sonntagsfrage. Die folgenden Schritte zeigen beispielhaft die Definition für die Variable „sex“.

Im ersten Schritt legen Sie den Namen und den Typ fest:

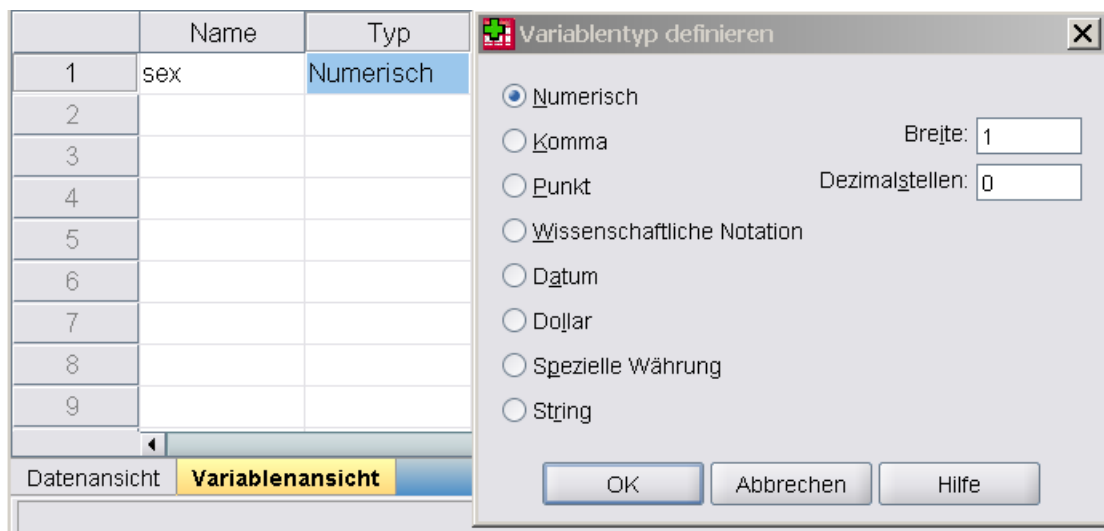


Abbildung 4.3: Variable sex, Typ: numerisch

Im 2. Schritt geben Sie die Paare „Wert und zugehöriges Wertelabel“ nacheinander ein und klicken jeweils auf „Hinzufügen“, um ein neues Paar zur Liste hinzuzufügen.

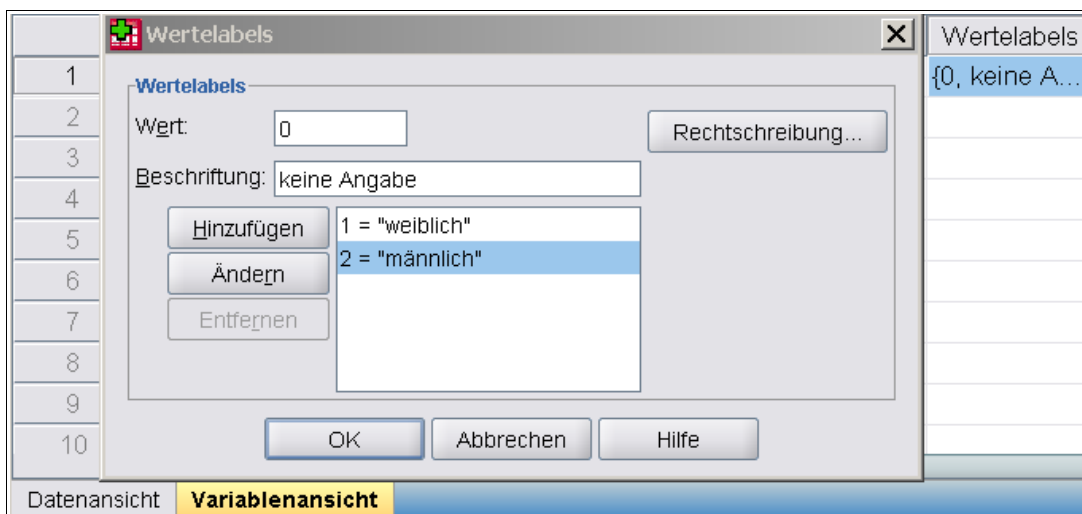
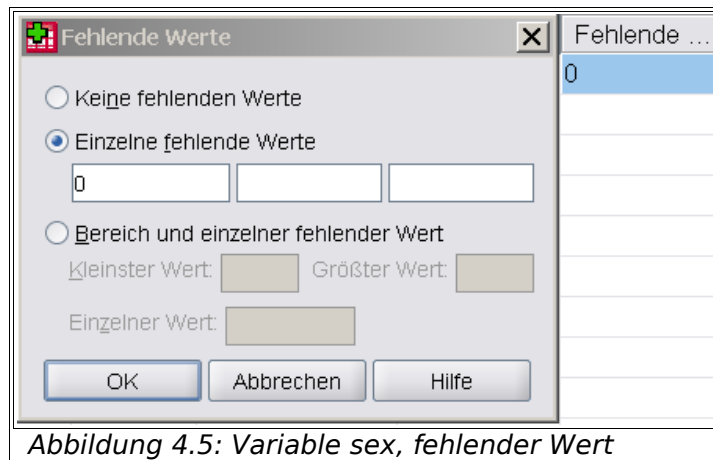
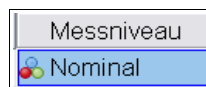


Abbildung 4.4: Variable sex: Wertelabel

Im 3. Schritt legen Sie den fehlenden Wert fest, d.h. SPSS interpretiert den Wert 0 im folgenden nicht als gültigen, sondern als fehlenden Wert:



Im 4. Schritt legen Sie „Nominal“ als Messniveau fest. Die anderen möglichen Messniveaus sind „metrisch“ und „nominal“ (Rang-Ordnung, geordnete Reihenfolge).



Im 5. Schritt legen Sie einen sprechenden Namen für die Variable fest, hier „Geschlecht“.

Definieren Sie nun die weiteren Variablen gemäß der Vorgaben im Fragebogen.

Name	Typ	Spaltenf...	Dezimal...	Variablenlabel	Wertelabels	Fehlende ...	Spalten	Ausrichtung	Messniveau
id	String	10	0	Fragebogen-ID	Keine	Keine	8	☰ Linksbündig	🌈 Nominal
sex	Numerisch	1	0	Geschlecht	{0, keine A...	0	8	☰ Rechtsbü...	🌈 Nominal
alter	Numerisch	3	0	Alter	Keine	-1	8	☰ Rechtsbü...	🔧 Metrisch
partei	Numerisch	3	0	Partei	{1, CDU/C...	-1	8	☰ Rechtsbü...	🌈 Nominal

Abbildung 4.6: Variablenansicht: Definition der Variablen

Kontrollieren Sie Ihre Eingaben über „Extras > Variablen“, hier als Beispiel die Anzeige für die Variable „Fragebogen-ID“:

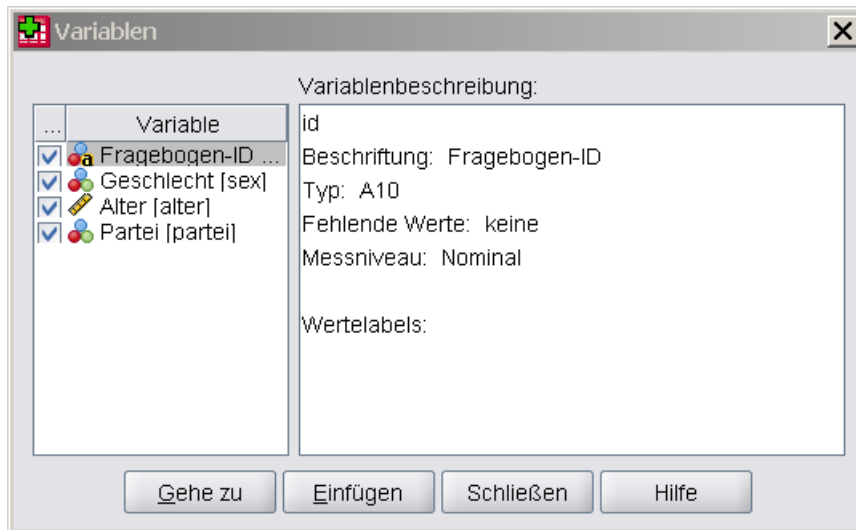


Abbildung 4.7: Extras > Variablen: Fragebogen-ID

Schalten Sie nun auf die Datenansicht um und geben Sie Werte für 5 (fiktive) Fragebögen ein:

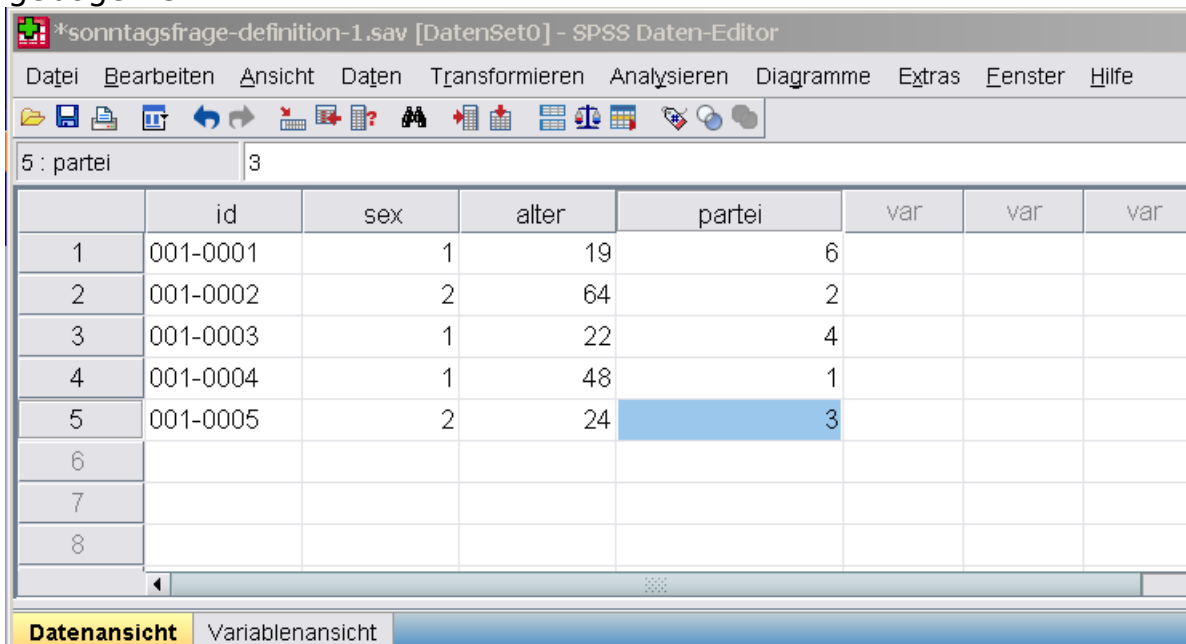


Abbildung 4.8: Datenansicht: 5 Beobachtungen

## 4.4 Begriffe

1. Die Daten sind in **Form einer Matrix** organisiert, die aus **Zeilen (Beobachtungen oder Fälle)** und **Spalten (Variablen)** besteht. Jeder Eintrag der tabellen- oder matrixartigen Arbeitsdatei stellt einen **Datenwert** (*data value*) dar. Ein Datenwert ist die kleinste Informationseinheit, die von **SPSS** verarbeitet werden kann.

2. Jede **Zeile** der Tabelle stellt eine **Beobachtung** (*observation*) dar; andere Bezeichnung: **Fall** (*case*). Eine Beobachtung setzt sich aus Informationen über ein Objekt oder eine Person zusammen. Die unterschiedlichen Informationen werden als **Variablen** (Eigenschaften oder Merkmale) bezeichnet. Der **Wertebereich** von Variablen können diskrete oder kontinuierliche Zahlenbereiche (numerische Werte) sein, Zeichenketten (*strings*, alphanumerische Werte) oder auch spezielle Wertebereiche wie z .B. Datum, Zeit oder Währung.
3. Variablen werden über **Variablennamen** bezeichnet, die in einer Tabelle typischerweise als Titelzeile (Spaltenüberschrift) verwendet werden.
4. Die Datenansicht enthält 5 **Beobachtungen** (oder **Fälle**) für diesen Fragebogen. In der aktuellen Datenansicht wird die Kodierung, also der Wert (*value*) und nicht das Werte-Etikett (*value label*) angezeigt. Der Wert 1 für Geschlecht („sex“) steht stellvertretend für „weiblich“, der Wert 2 für Partei („partei“) steht stellvertretend für „FDP“.

## 4.5 Anzeigen der Werte-Etiketten (Werte-Label)

Klicken Sie auf den Menüpunkt „Ansicht > Wertelabel anzeigen“, um sich die Werte-Label (auch als Etiketten bezeichnet) statt der Werte anzeigen zu lassen:

id	sex	alter	partei
001-0001	weiblich	19	Republikaner
001-0002	männlich	64	FDP
001-0003	weiblich	22	Grüne/Bündnis90
001-0004	weiblich	48	CDU/CSU
001-0005	männlich	24	SPD

Abbildung 4.9: Datenansicht: Werte-Etiketten statt Werte

## 4.6 Speichern von Variablen- und Datenansicht

Sorgen Sie nun im letzten Schritt dafür, die Definition der Variablen und die erfaßten Datenwerte permanent in eine SPSS Datendatei abzuspeichern. Verwenden Sie hierzu den Menüpunkt „Datei > Speichern“. Verwenden Sie den Namen „sonntagsfrage-01.sav“. Diese SPSS Datendatei kann in späteren SPSS Sitzungen von Ihnen oder anderen Bearbeitern erneut geöffnet und bearbeitet werden.

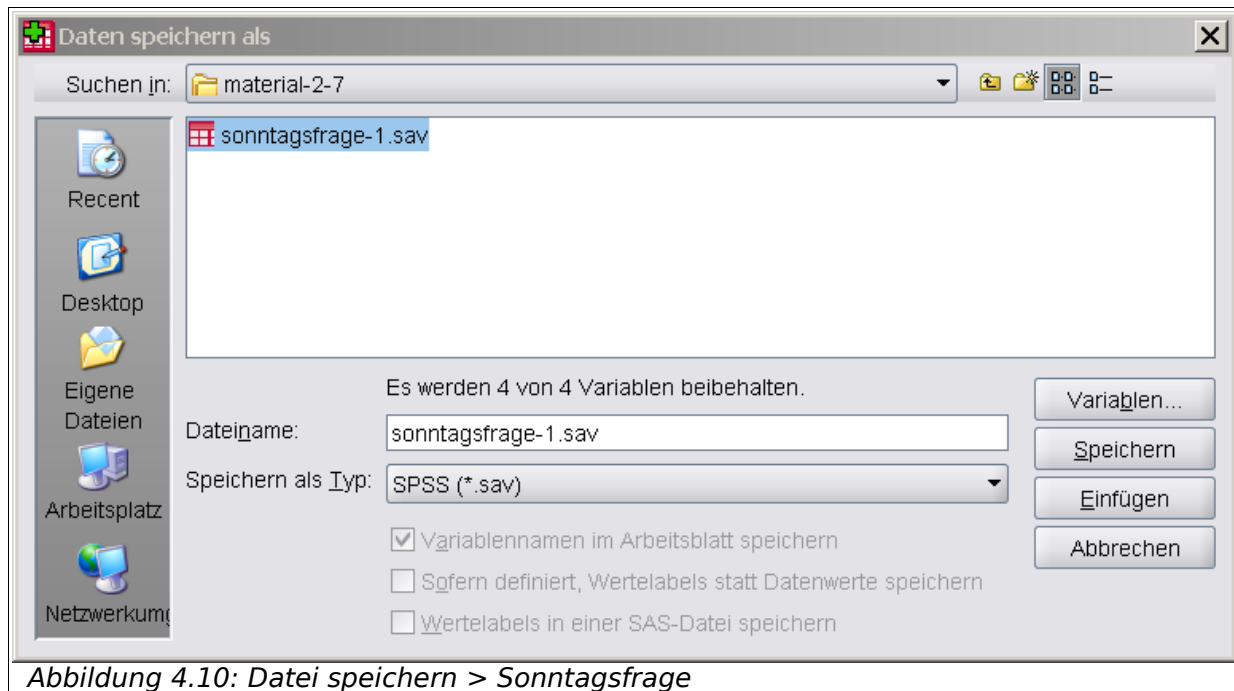


Abbildung 4.10: Datei speichern > Sonntagsfrage

## 4.7 Übungen

### Hinweis

Die erste Übung lautet immer (also auch für alle folgenden Kapitel), das aktuelle Kapitel mit SPSS in wesentlichen Zügen am eigenen Rechner nachzuarbeiten.

1. Erstellen Sie ein eigenes Verzeichnis für diesen SPSS Kurs. Verwenden Sie im PC-Raum des Rechenzentrums das Verzeichnis „Z:\SPSS-Kurs“ und speichern Sie im folgenden alle SPSS Datendateien und sonstige Dateien in diesem Verzeichnis.
2. Erstellen Sie die SPSS Datendatei „meine-sonntagsfrage-01.sav“ mit den Variablen „id, sex, alter, partei“ in Anlehnung an die Beschreibung im aktuellen Kapitel.
3. Geben Sie weitere Beobachtungen in der Datenansicht ein, wobei Sie einige Zellen für numerische und Zeichenketten (strings, alpha-numerische Werte) einfach freilassen (systembedingte fehlende Werte). Wie werden systembedingte fehlende Werte für Zahlen und Zeichenketten dargestellt?
4. Laden Sie die Datendatei „fragebogen-1.sav“ aus dem Bereich „Material“ (siehe oben) in das Verzeichnis für den SPSS Kurs herunter. Öffnen Sie diese SPSS Datendatei.

## 5. [zur Diskussion]

Welche Gründe sprechen allgemein für die Vergabe von Werte-Etiketten für Variablennamen und Werte?

**Hinweise:**

Fassen Sie z.B. die Möglichkeit in Betracht, daß jeweils eine Auswertung für englisch- und deutschsprachiges Publikum benötigt wird

## 6. [zusätzlich]

Ergänzen Sie den Fragebogen (und die SPSS Datei) um weitere Fragen (Variablen) wie zum Beispiel: „Welche Partei/Koalition sollte regieren (Bürgerlich, Ampel, Jamaika, ...)?“, „Wer sollte Kanzler/in werden?“, „Welche Themen werden den Wahlkampf bestimmen (mit Mehrfachauswahl)?“. Nutzen Sie hierzu ggf. auch die Textdatei „fragebogen-01.rtf“.

Beispiel zu Aufgabe 6:

Name der Variablen	Text / Hinweise	Kodierung (Werte-Etiketten)	Antwort (kodiert)
koalition	Welche ...	1 (grosse Koalition, CDU/CSU+SPD) 2 (CDU/CSU+FDP) ... 0 (keine Angabe)	<input type="checkbox"/>

## 4.8 Einlesen aus externen Quellen

Sie können das Datenmaterial auch außerhalb von **SPSS** in einem Tabellenkalkulationsprogramm wie **Microsoft Excel** oder **OpenOffice Calc** bzw. in einem Datenbankprogramm wie **Microsoft Access** erfassen und in dem spezifischen Format dieses Programmes abspeichern.

Im Anschluß können dieses Datenmaterial in **SPSS** mit einem Importfilter einlesen und dann das in eine SPSS Datendatei abspeichern.

Einen Überblick über die möglichen Import-Formate liefern die Menüpunkte „**Datei > Öffnen**“ bzw. „**Datei > Datenbankzugriff**“ und „**Daten > Textdaten lesen**“.

Abhängig davon, in welcher Form Daten bereits vorliegen, kann es Vor- und Nachteile mit sich ziehen, mit externen Quellen zu arbeiten. Als Faustregel kann gelten, daß es für kleine Projekte mit nur einem Bearbeiter nur eine SPSS Datendatei als Quelle geben sollte, die regelmäßig auf USB-Stick, CD mit einer Versions- und oder Datums-Historie (Stichwort: Subversion) gesichert wird.

# 5 Berechnen neuer Variablen

In diesem Kapitel werden Methoden beschrieben, um in **SPSS** neue Variablen aus den bestehenden Variablen zu erzeugen

**Zielsetzung**

Ich möchte meine Beobachtungen anhand einer berechneten Variable "Altersgruppe" in Alterklassen einteilen. Ich habe keine Lust, diese Variable selbst zu berechnen ...

## 5.1 Beobachtungen und Variablen

Eine Datendatei ist aus **Variablen** (Spalten, vertikal) und **Beobachtungen** (Zeilen, horizontal) zusammengesetzt. Hieraus ergibt sich insgesamt ein rechteckiges (oder matrixartiges) Schema. Sie können eine Datendatei deshalb grundsätzlich auf 2 Arten erweitern (oder verkleinern):

<b>Hinzufügen von Variablen</b> (hier: <b>x</b> neue Spalte)	<b>Hinzufügen oder Filtern von Beobachtungen</b> (hier: <b>x</b> neue Zeile)
<b>x</b>           <b>x</b>           <b>x</b>           <b>x</b>	                           <b>x</b>   <b>x</b>   <b>x</b>   <b>x</b>   <b>x</b>

Tabelle 5.1 :Beobachtungen und Variablen

Darüberhinaus ist es in SPSS auch möglich, einzelne Zeilen anhand vorgegebener Bedingungen vorübergehend "auszufiltern"; d.h. auszublenden, aber nicht permanent zu löschen.

## 5.2 Berechnen einer Altersgruppe

Im folgenden Beispiel berechnen Sie auf Grundlage der Datendatei „sonntagsfrage-01.sav“ eine neue Variable *g\_alter* (Altersgruppe). Wählen Sie hierzu **„Transformieren > Berechnen“**. Im folgenden Dialog-Fenster tragen Sie oben links den Variablen-Namen ein und oben rechts die „Rechenvorschrift“, um die neue Variable (zeilenweise) zu berechnen:

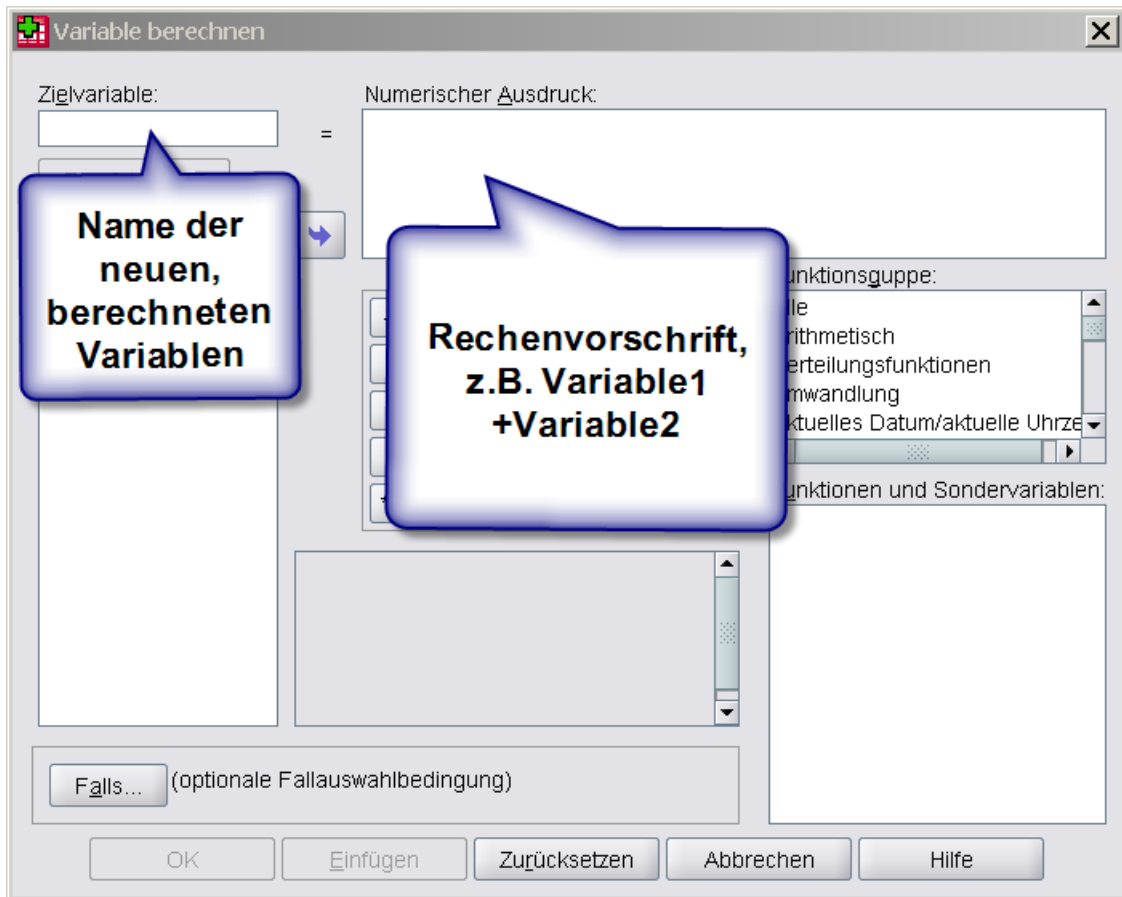


Abbildung 5.1: Transformieren > Berechnen

Geben Sie links oben den Namen der Zielvariablen ein , hier: `g_alter`, und nach dem Gleichheitszeichen auf der rechten Seite den Ausdruck, über den der Wert der Zielvariablen für jede Beobachtung festgelegt werden soll. Klicken Sie dann auf die Schaltfläche „OK“, um die neue Variable zu erstellen.

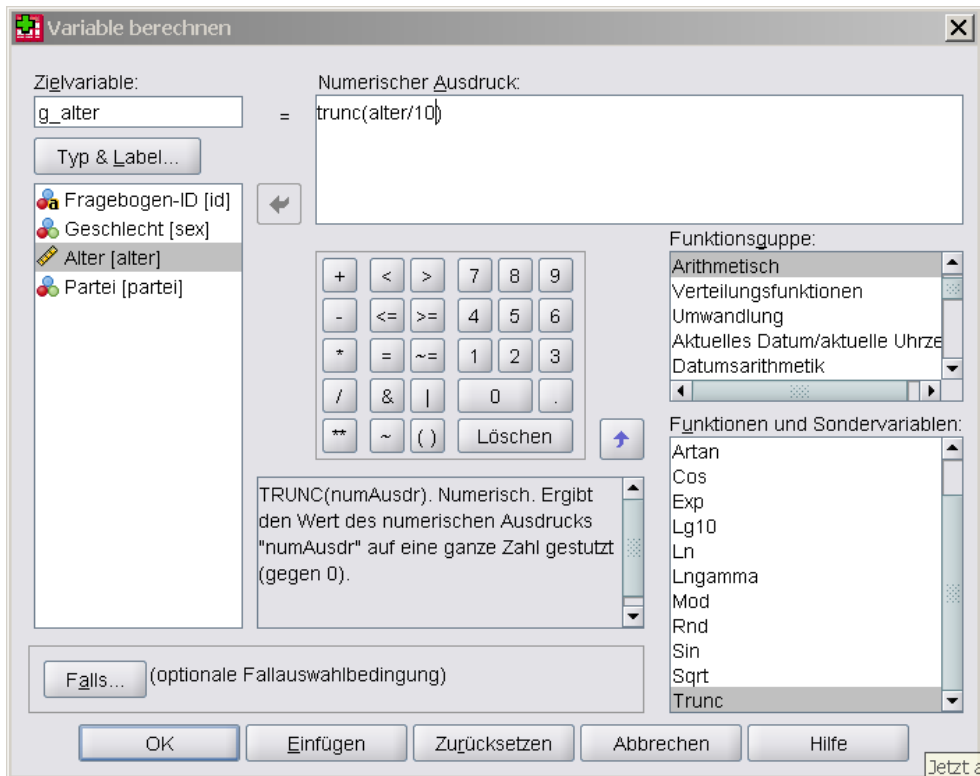


Abbildung 5.2: Numerischer Ausdruck: Abschneiden (truncate)

Die Variable `g_alter` wird im Beispiel berechnet auf Basis der Variablen `alter`. Diese Variable wird zunächst durch 10 geteilt, danach werden die Dezimalstellen abgeschnitten (`trunc` für `truncate`).

`trunc(alter/10)`; Beispiel:  $19/10=1,9$  `trunc(1,9) = 1`

**SPSS** berechnet die neue Variable `g_alter` und zeigt sie sofort im Dateneditor an. Nach Änderung der Anzahl der Dezimalstellen auf Null ergibt sich folgende Datenansicht:

	id	sex	alter	partei	g_alter
1	001-0001	weiblich	19	Republikaner	1
2	001-0002	männlich	64	FDP	6
3	001-0003	weiblich	22	Grüne/Bündnis90	2
4	001-0004	weiblich	48	CDU/CSU	4
5	001-0005	männlich	24	SPD	2
6	001-0006	weiblich	39	CDU/CSU	3
7	001-0007	weiblich	26	Grüne/Bündnis90	2
8	001-0008	männlich	34	FDP	3
9	001-0009	männlich	20	SPD	2
10	001-0010	weiblich	87	Sonstige	8
11	001-0011	männlich	14	FDP	1
12	001-0012	männlich	33	CDU/CSU	3
13	001-0013	weiblich	67	SPD	6
14	001-0014	weiblich	28	FDP	2
15	001-0015	männlich	22	CDU/CSU	2
16	001-0016	weiblich	19	SPD	1

Abbildung 5.3: Neue Variable Altersgruppe

## 5.3 Erläuterungen

1. Es stehen Ihnen eine Vielzahl von arithmetischen Operatoren und logischen Ausdrücken sowie eine Vielzahl von Funktionen zur Verfügung. Das Hilfesystem enthält detaillierte Informationen zum Thema „Berechnen neuer Variablen“.
2. Die Werte für die neue Variable werden auf Grundlage der aktuellen Datenwerte für bereits vorhandene Beobachtungen berechnet, bei später hinzugefügten oder auch bei modifizierten Beobachtungen wird KEINE (!) automatische Aktualisierung durchgeführt. Deshalb sollten berechnete Variablen erst NACH vollständiger Eingabe aller Beobachtungen erzeugt werden.

## 5.4 Übungen

1. Laden Sie die SPSS Datendatei „broca-01.sav“. Berechnen Sie als neue Variable den sogenannten **Broca-Index**  $bi=100 \cdot gewicht / (groesse-100)$  aus den Variablen `gewicht` (Gewicht in Kilogramm) und `groesse` (Körpergröße in Zentimeter).
2. Berechnen Sie als neue Variable den sogenannten **Body-Mass-Index** (BMI). Der Wert berechnet sich gemäß folgender Formel, wobei vorher die Größe in Meter umgerechnet werden muß:

$$bmi = gewicht / (groesse\_in\_m * groesse\_in\_m)$$

## 3. [zusätzlich]

Berechnen Sie für die Datendatei „broca-01.sav“ als neue Variable zusätzlich die Gruppe „bmi\_gruppe“ gemäß der Tabelle „BMI – Einteilung in Gruppen“.

## 4. [zusätzlich]

Definieren Sie in „sonntagsfrage.sav“ über **Transformieren > Berechnen** die folgende neue Variable `konsum_gruppe` für die Einteilung in Altersgruppen von 0-12, 13-19, 20-39, 40+ und 65+.

Werte- Etikett für bmi_gr	Werte für bmi_gruppe	Bereich für bmi
untergewichtig	-1	Kleiner als 18.5
normal	0	18.5 bis 25
übergewichtig	1	25 bis 30
stark übergewichtig	2	Größer als 30

Tabelle 5.2 : BMI: Einteilung in Gruppen

### Hinweise zu den Aufgaben 3 und 4

Eine UND-Verknüpfung wird im SPSS Formel-Editor durch den Operator `&` definiert, Beispiel: `(0 < bmi) & (bmi < 18.5)`; eine ODER-Verknüpfung durch den Operator `|`, Beispiel: `(alter < 20) | (alter > 60)`.

Das Ergebnis eines Vergleiches (oder allgemeiner eines logischen Ausdrucks) ist in **SPSS** entweder die Zahl 0 (Ausdruck ist falsch) oder die Zahl 1 (Ausdruck ist wahr). Deshalb kann mit logischen Ausdrücken wie mit Zahlen gerechnet werden:

```
bmi_gruppe = -1 * ((0 < bmi) & (bmi < 18.5)) +
             0 * (...)
             1 * ((25.0 <= bmi) & (bmi < 30)) +
             2 * (30 <= bmi)
```

```
konsum_gruppe = 1 * (( 0 < alter) & (alter <= 12)) +
                2 * ((12 < alter) & (alter <= 19)) +
                ...
```

## 6 Auswählen von Beobachtungen

Sie können die folgenden Auswertungen auf eine Auswahl (Teilmenge) von Beobachtungen einschränken.

### Zielsetzung

Ich möchte meine Untersuchung auf einer Teilmenge aller Beobachtungen durchführen, und ich möchte die Teilmengen variieren können.

### 6.1 Filtern von Beobachtungen

Im folgenden Beispiel führen Sie in der SPSS Datendatei „sonntagsfrage-01.sav“ eine Auswahl (Filterung) nach Frauen (Bedingung: „sex=1“) durch.

Wählen Sie zunächst „**Daten** > **Fälle auswählen**“.

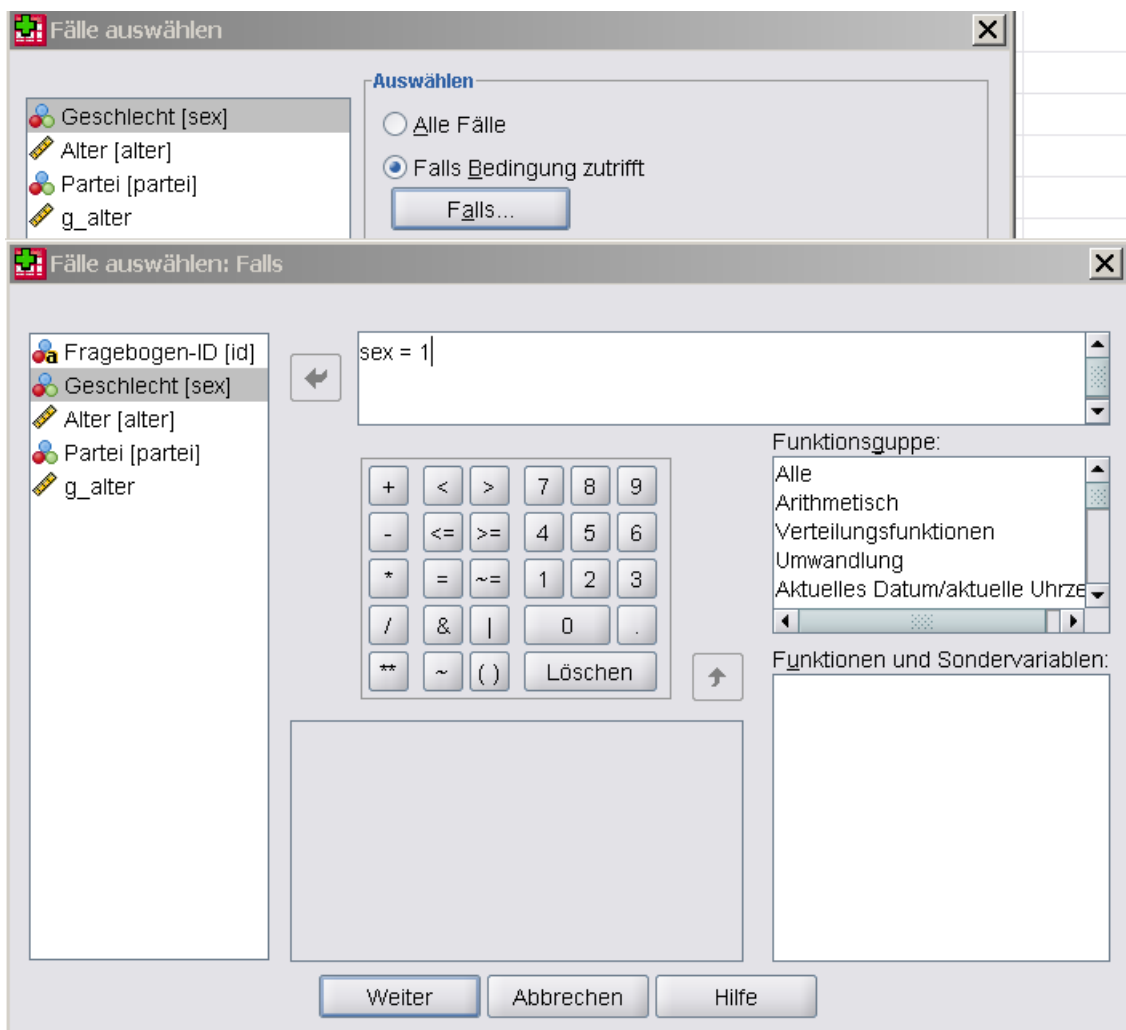


Abbildung 6.1: Daten > Fälle auswählen

Wählen Sie das Auswahlfeld „Falls Bedingung zutrifft“ und klicken Sie dann auf die Schaltfläche „Falls ...“. Tragen Sie dann als „Filter“ (Auswahlkriterium) ein: „sex=1“. Klicken Sie auf „Weiter“ und dann auf die Schaltfläche „OK“, um den Filter anzuwenden.

Im „Dateneditor“-Fenster erscheint nun eine neue Variable `filter_$` mit den möglichen Werten 0 (für: ausgeschlossen/Not Selected) und 1 (für: beteiligt/Selected).

Alle ausgefilterten Beobachtungen sind zusätzlich im „Dateneditor“-Fenster am linken Rand durch eine durchgestrichene Nummer gekennzeichnet (hier: 2,5 usw.). Ausgefilterte Beobachtungen werden in den folgenden Auswertungen nicht berücksichtigt.

	id	sex	alter	partei	g_alter	filter_\$
1	001-0001	weiblich	19	Republikaner	1	Selected
<del>2</del>	001-0002	männlich	64	FDP	6	Not Selected
3	001-0003	weiblich	22	Grüne/Bündnis90	2	Selected
4	001-0004	weiblich	48	CDU/CSU	4	Selected
<del>5</del>	001-0005	männlich	24	SPD	2	Not Selected
6	001-0006	weiblich	39	CDU/CSU	3	Selected
7	001-0007	weiblich	26	Grüne/Bündnis90	2	Selected
<del>8</del>	001-0008	männlich	34	FDP	3	Not Selected
<del>9</del>	001-0009	männlich	20	SPD	2	Not Selected
10	001-0010	weiblich	87	Sonstige	8	Selected
<del>11</del>	001-0011	männlich	14	FDP	1	Not Selected
<del>12</del>	001-0012	männlich	33	CDU/CSU	3	Not Selected
13	001-0013	weiblich	67	SPD	6	Selected
14	001-0014	weiblich	28	FDP	2	Selected
<del>15</del>	001-0015	männlich	22	CDU/CSU	2	Not Selected
16	001-0016	weiblich	19	SPD	1	Selected

Abbildung 6.2: Datenansicht: aktiver Filter `filter_$`

## 6.2 Übungen

1. Wählen Sie der SPSS Datendatei „sonntagsfrage-01.sav“ nacheinander die Fälle (Beobachtungen) aus, für die folgende Bedingungen gelten:

- Die befragte Person (P) ist zwischen 40 und 60 Jahre alt.
- P ist weiblich und älter als 60.
- P ist älter als 25, männlich und würde die FDP wählen.
- P würde CDU/CSU oder FDP wählen.

Kontrollieren Sie jeweils in der Arbeitsdatei, ob die Filter-Variable korrekt gesetzt worden ist.

2. [zusätzlich]  
Wie können Sie eine Filter-Variable vor dem Überschreiben durch die nächste Filter-Regel „retten“?
3. [zusätzlich]  
Was spricht dagegen, Beobachtungen zu löschen anstatt sie nur zu filtern?

## 7 Arbeiten mit Kalenderdaten

In diesem Kapitel wird speziell auf den Datentyp Datum eingegangen.

In meiner Untersuchung benötige ich den zeitlichen Abstand zwischen zwei Messungen als neue Variable. Muß ich im Kalender nachschlagen oder kann ich Zeitdifferenzen von **SPSS** berechnen lassen ?

### 7.1 Definieren von Variablen mit Datentyp DATUM

Sie benötigen Datums- oder Zeitangabe, um ein Datum, z.B. das Datum einer Messung, in eine SPSS Arbeitsdatei aufnehmen zu können. Falls eine Datendatei mehrere Datums-Variablen enthält, können Sie auch zeitliche Abstände zwischen ihnen berechnen.

Die Festlegung des Datumsformats erfolgt bei der Definition einer Variablen in der Auswahlliste für die Eigenschaft Typ.

Die Darstellung läßt sich durch Auswahl eines Musters steuern, hierbei steht die Abkürzung „tt“ (00-31) für Tag in, die Abkürzung „mmm“ für Monat als Buchstaben-Abkürzung ( JAN, FEB, ..., DEC), „mm“ für Monat als Zahl (01-12) und „jj“ für eine 2-stellige Jahreszahl (09) bzw. „jjjj“ für eine 4-stellige Jahreszahl (2009).

Erstellen Sie über „Datei > Daten > Neu“ eine neue SPSS Datendatei „ferien-01.sav“. Wählen Sie den Reiter "Variablenansicht" und definieren Sie eine Variable „beginn“ vom Datentyp „Datum“ und mit dem Muster „tt.mm.jjjj“ (Beispiel: 04.01.2009):

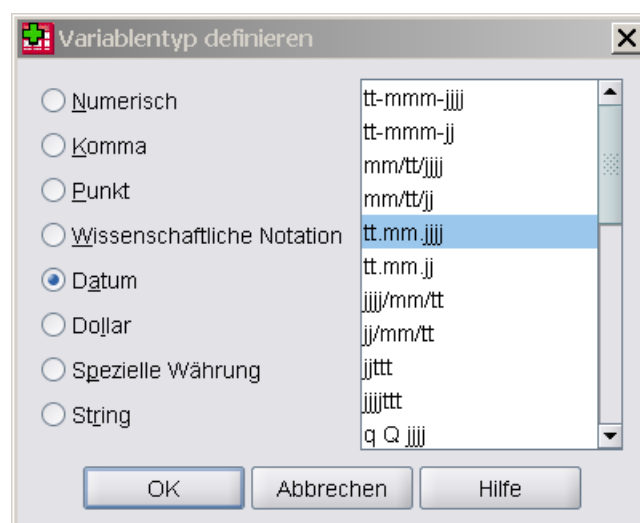


Abbildung 7.1: Variable definieren > Datum

Definieren Sie eine weitere Variable „ende“ vom Datentyp Datum und eine Variable „ferien“ vom Datentyp „String“ („Zeichenkette“) der Länge 32.

Geben Sie dann die Ferientermine für Niedersachsen im Jahr 1995, wie folgt ein:

	beginn	ende	ferien
1	01.01.1995	07.01.1995	Weihnachtsferien
2	03.04.1995	19.04.1995	Osterferien
3	06.06.1995	06.06.1995	Pfingstferien
4	22.06.1995	02.08.1995	Sommerferien
5	02.10.1995	14.10.1995	Herbstferien
6	23.12.1995	31.12.1995	Weihnachtsferien

Abbildung 7.2: Datenansicht: Ferientermine Niedersachsen 1995

## 7.2 Erläuterungen

Eine Datumsangabe wird von SPSS intern als Anzahl von Sekunden seit dem **14. Oktober 1582**, also ab einem willkürlich gewählten Zeitnullpunkt, abgelegt, d.h. als numerischer Wert in Sekunden.

<----- ----->
14.10.1582 (willkürlich gewählter Datums-Nullpunkt)
0 [Sekunden] interne Darstellung in SPSS
<----- ----->
01.02.1995
interne Speicherung in SPSS: 13010976000 [Sekunden]
<----- ----->
01.06.1966
interne Speicherung in SPSS: 12100924800 [Sekunden]

Tabelle 7.1 : internes Datums-Format

Die Differenz zwischen den Datums-Werten 01.06.1966 und dem 01.02.1995 [in Sekunden] läßt sich in **SPSS** direkt als Differenz der entsprechenden Werte berechnen. Unter Beachtung elementarer Umrechnungen (60 Sekunden = 1 Minute, 60 Minuten = 1 Stunde, 24 Stunden = 1 Tag) läßt sich hieraus auch die Differenz zwischen 2 Datumswerten in Tagen berechnen.

## 7.3 Funktionen für den Datentyp Datum

Für Variablen des Datentyps `DATUM` stehen spezielle Funktionen zur Verfügung, mit denen Sie u.a. Datums- oder Zeitdifferenzen berechnen oder Wochentage

bestimmen können. Die folgende Tabelle enthält eine kleine Auswahl an Funktionen:

<b>Funktion</b>	<b>Berechnet ...</b>	<b>Beispiel</b>
CTIME.DAYS (date)	Anzahl Tage seit dem 14.10.1582, benötigt als Argument eine Datums-Variable	CTIME.DAYS(14.10.1582)
YRMODA (year, month, day)	Anzahl Tage seit dem 14.10.1582, benötigt als Argument 3 numerische Variablen	YRMODA(1582,10,14)

Tabelle 7.2 : Funktionen für Datums-Variablen

Es folgt eine Tabelle mit Geburts- und Todes-Datum einiger berühmter Persönlichkeiten:

<b>Name</b>	<b>Geburtsdatum</b>	<b>Todesdatum</b>
Newton, Isaac	25. Dezember 1642	20. März 1726
Leibniz, Gottfried	21. Juni 1646	14. November 1716
Einstein, Albert	14. März 1879	18. April 1955

Tabelle 7.3: Lebensdaten

<http://de.wikipedia.org/wiki/Leibniz>

[http://de.wikipedia.org/wiki/Isaac\\_Newton](http://de.wikipedia.org/wiki/Isaac_Newton)

<http://de.wikipedia.org/wiki/Einstein>

Im folgenden Beispiel sind entsprechende Variablen und Werte in der SPSS Datendatei "persoenlichkeiten-01.sav" eingetragen:

Name	Vorname	Geburtsdatum	Todesdatum
Einstein	Albert	14.03.1879	18.04.1955
Newton	Isaac	25.12.1642	20.03.1726
Leibniz	Gottfried	21.06.1646	14.11.1716

Abbildung 7.3: Lebensdaten

Berechnen Sie mit Hilfe der genannten Funktionen das Lebensalter der Persönlichkeiten in Tagen (Alter\_T) und in Jahren (Alter\_J) und speichern Sie in der SPSS Datendatei "persoenlichkeiten-02.sav":

Alter\_T = CTIME.DAYS(Todesdatum) – CTIME.DAYS(Geburtsdatum)

Alter\_J = DATEDIFF(Todesdatum,Geburtsdatum,'years')

Name	Vorname	Geburtsdatum	Todesdatum	Alter_T	Alter_J
Einstein	Albert	14.03.1879	18.04.1955	27793	76
Newton	Isaac	25.12.1642	20.03.1726	30400	83
Leibniz	Gottfried	21.06.1646	14.11.1716	25713	70

Abbildung 7.4: Lebensalter

## 7.4 Übungen

- Schalten Sie in der SPSS Datendatei "ferien-01.sav" zwischen mindesten 3 verschiedenen Mustern für das Datum um; Beispiele: 2009-01-01, 1. Januar 2009, 01.01.09.
- Berechnen Sie für die SPSS Datendatei „ferien-01.sav“ in der neuen Variablen `diff_tag` die Anzahl der Ferientage für die jeweiligen Ferien und erläutern Sie, weshalb die folgende Formel verwendet werden sollte. Betrachten Sie insbesondere (fiktiv) Ferien, die nur einen Tag dauern.

`diff_tag = CTIME.DAYS(ende) - CTIME.DAYS(beginn) + 1`

- [zusätzlich]  
Ergänzen Sie in der SPSS Datendatei "persoenlichkeiten-02.sav" einige weitere Persönlichkeiten (Napoleon, Galilei, Shakespeare, Goethe, Schiller etc.) und berechnen Sie deren Lebensalter.

## 7.5 Berechnen von Durchschnittswerten

Im folgenden Beispiel werten Sie Daten über das Betanken eines Autos aus. Die Daten sind in der SPSS Datendatei „fahrtenbuch-01.sav“ festgehalten und zwar das Datum der Betankung in den drei Variablen `tag`, `monat`, `jahr`, der jeweilige Kilometerstand (`kmstand`) und die getankte Benzinmenge (`liter`) bei jeder Betankung.

	tag	monat	jahr	kmstand	liter
1	16	12	1992	20580	60,3
2	23	12	1992	21250	57,4
3	4	1	1993	21874	56,6
4	17	1	1993	22476	56,3
5	28	1	1993	22954	45,4
6	12	2	1993	23450	48,6
7	27	2	1993	24020	57,0
8	14	3	1993	24611	56,7

Abbildung 7.5: Fahrtenbuch

Berechnen Sie die neuen Variablen `km_pro_tag` (durchschnittlich gefahrene Kilometer pro Tag) und `verbrauch` (durchschnittlicher Benzinverbrauch auf 100 km) zwischen 2 Betankungen anhand der folgenden Tabelle:

Berechnung	Bedeutung
<code>ntage=yrmoda(jahr, monat, tag)</code>	<code>ntage</code> ist die Anzahl Tage seit dem 14.10.1582.
<code>diff_tage=ntage-lag(ntage)</code>	<code>diff_tage</code> ist die Differenz zwischen zwei aufeinanderfolgenden Betankungen in Tagen. <code>lag(var)</code> liefert den Wert der Variablen <code>var</code> für die vorhergehende (!) Beobachtung.
<code>diff_km=kmstand-lag(kmstand)</code>	<code>diff_km</code> ist die Differenz zwischen zwei aufeinanderfolgenden Kilometerständen.
<code>verbrauch</code>	<code>verbrauch</code> ist der Verbrauch zwischen zwei Betankungen (in liter / 100 km).
<code>km_pro_tag</code>	<code>km_pro_tag</code> ist die durchschnittlich zurückgelegte Strecke pro Tag.

Tabelle 7.4 : berechnete Variablen

Die SPSS Arbeitsdatei enthält nach erfolgreicher Berechnung folgende neuen Variablen und zugehörige Werte:

tag	monat	jahr	kmstand	liter	ntage	diff_tage	diff_km	verbrauch
16	12	1992	20580	60,3	149813	.	.	.
23	12	1992	21250	57,4	149820	7	670,00	8,57
4	1	1993	21874	56,6	149832	12	624,00	9,07
17	1	1993	22476	56,3	149845	13	602,00	9,35
28	1	1993	22954	45,4	149856	11	478,00	9,50
12	2	1993	23450	48,6	149871	15	496,00	9,80
27	2	1993	24020	57,0	149886	15	570,00	10,00
14	3	1993	24611	56,7	149901	15	591,00	9,59

Abbildung 7.6: Fahrtenbuch: Durchschnittswerte

## 7.6 Übungen

- Berechnen Sie in der SPSS Arbeitsdatei „fahrtenbuch-01.sav“ den Verbrauch zwischen 2 Betankungen über die folgende Formel:  
 $\text{verbrauch} = \text{liter} * 100 / \text{diff\_km}$
- Berechnen Sie die durchschnittliche tägliche Fahrstrecke über die folgende Formel:  
 $\text{km\_pro\_tag} = \text{diff\_km} / \text{diff\_tage}$
- [zusätzlich]  
Berechnen Sie in der SPSS Arbeitsdatei „ferien-01.sav“ in der Variablen „wartezeit“

jeweils den Abstand zwischen Beginn der Schulzeit und Anfang der Ferien bei aufeinanderfolgenden Ferienterminen, also z.B. die Anzahl der Tage zwischen Ende der Sommerferien und Beginn der Herbstferien; also die „Wartezeit auf die nächsten Ferien“:

**wartezeit = CTIME.DAYS(beginn) - CTIME.DAYS(LAG(ende)) - 1**

4. Erklären Sie die oben verwendete Formel zur Berechnung der Wartezeit. Betrachten Sie insbesondere (fiktiv) direkt aufeinanderfolgende Ferientermine.
5. [zusätzlich]  
Berechnen Sie, wieviele Jahre Überlappung es zwischen den Lebensdaten von Leibniz und Newton gab.
6. [zusätzlich]  
Berechnen Sie den Zeitraum zwischen dem Tod von Newton und der Geburt von Einstein.
7. [Projekt]  
Ordnen Sie die Lebensdaten der Persönlichkeiten auf einer Zeitleiste an.

## 8 Überblick über die deskriptive Statistik

In diesem Kapitel werden wichtige Begriffe aus der beschreibenden Statistik erläutert. Dieses Kapitel dient als Auffrischung und kann ggf. überschlagen werden.

Ich entsinne mich, daß ich in einführenden Veranstaltungen zur Wahrscheinlichkeitsrechnung und Statistik etwas über Median, Mittelwert, Varianz und Standardabweichung gehört habe ...

### 8.1 Aufgaben der deskriptiven Statistik

Die **deskriptive Statistik** (oder beschreibende Statistik) befaßt sich mit der tabellarischen und grafischen Darstellung von Daten sowie mit der Zusammenfassung (Verdichtung, Aggregation) von Daten mit Hilfe neuer Variablen oder mit Hilfe charakteristischer Kenngrößen (wie z.B. Lage- bzw. Streumaße). Sie dient damit als Ausgangspunkt für die mathematische oder analytische Statistik, da sie Hinweise auf grundlegende Zusammenhänge im Datenmaterial liefert.

Der Untersuchungsgegenstand der deskriptiven Statistik sind **Beobachtungen** von zufälligen und nicht-zufälligen **Variablen (Merkmalen oder Eigenschaften)** von Objekten oder Personen. Zufällige Variablen können im Anschluß mit Verfahren der mathematischen Statistik analysiert werden, um z.B. statistisch abgesicherte Aussagen über Zusammenhänge zwischen einzelnen Variablen ableiten zu können.

### 8.2 Tabellen und Diagramme

In der deskriptiven Statistik werden Variablen und Beziehungen zwischen Variablen u.a. mit folgenden Tabellen und grafischen Hilfsmitteln dargestellt:

- Häufigkeitstabelle
- Kreuztabelle
- Histogramm
- Streudiagramm (*scatterplot*)

In der Regel werden von einem Objekt oder einer Person mehrere Variablen gleichzeitig beobachtet, so daß eine Beobachtung aus mehreren Variablen besteht. Die Beobachtung  $\mathbf{x}$  wird dann als **vektoriell** oder **multivariat** bezeichnet. Ein häufiger Spezialfall sind **bivariate** Variablen bzw. Auswertungen,

die sich auf genau zwei Variablen beziehen; wichtige Beispiele sind Korrelation und lineare Regression. Eine Beobachtung, die aus nur einer Variablen besteht, bzw. eine Analyse, die nur eine Variable berücksichtigt, wird als **univariat** bezeichnet. Hieraus leiten sich die Begriffe univariate, bivariate und multivariate Statistik ab.

Zum Beispiel könnten für eine Person gleichzeitig die zufälligen Variablen Größe, Gewicht und Alter sowie die gruppierende Variable Geschlecht beobachtet werden; d.h. die Beobachtung einer Person setzt sich aus vier Datenwerten für 4 Variablen zusammen:

$$\mathbf{x} = (\text{Größe}, \text{Gewicht}, \text{Alter}, \text{Geschlecht})$$

## 8.3 Stichprobe und Grundgesamtheit

Eine **Stichprobe**  $S$  (*sample*) ist eine Auswahl von **Beobachtungen** (*observations*) aus einer **Grundgesamtheit** oder **Population**  $P$  (*population*). In der Terminologie der Wahrscheinlichkeitsrechnung besteht eine Stichprobe aus Realisierungen von zufälligen, also nicht deterministisch vorhersagbaren Variablen (Zufallsvariablen).

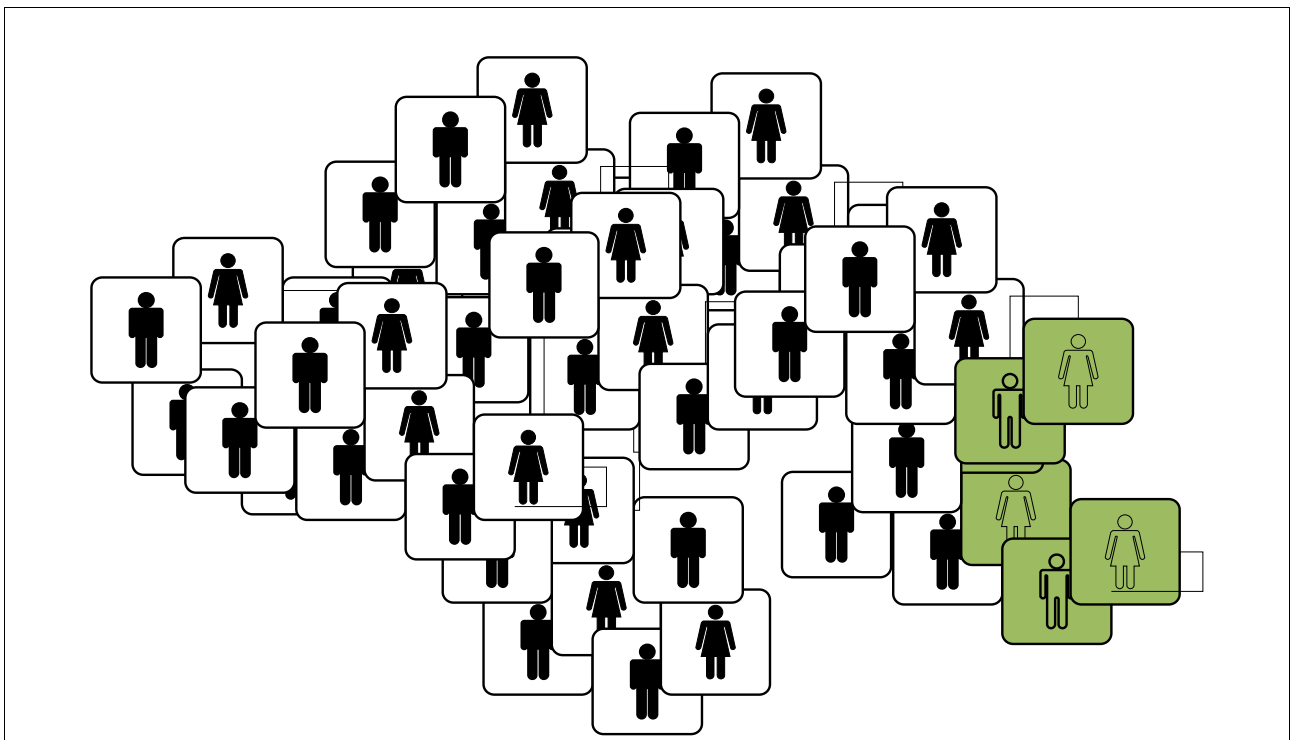


Tabelle 8.1: Grundgesamtheit und Stichprobe (grün)

## 8.4 Auswahlmechanismen

Eine Stichprobe erhebt den Anspruch, im Kleinen das Verhalten der Grundgesamtheit widerzuspiegeln, also **repräsentativ** zu sein. Hierzu müssen Auswahlmechanismen festgelegt werden, die eine repräsentative Auswahl der Stichprobe aus der Grundgesamtheit garantieren. Eine geeignete Auswahl erfolgt z.B. durch eine Einteilung der Grundgesamtheit in Klassen und einer anschließenden zufällige Auswahl von Repräsentanten aus jeder Klasse (mehrstufige Zufallsauswahl).

Eine **Stichprobe**  $S$  wird im Gegensatz zu einer **Gesamterhebung** aus folgenden **Gründen** durchgeführt:

1. Eine **Gesamterhebung**, d.h. die Untersuchung der Grundgesamtheit, ist prinzipiell unmöglich, scheitert am politischen Widerstand, zerstört die untersuchten Objekte, ist zu teuer oder dauert zu lange.
2. Aus vorherigen Untersuchungen ist bekannt, daß sich das Verhalten der Grundgesamtheit zufriedenstellend über Beobachtungen von ausgewählten Repräsentanten beschreiben läßt.

## 8.5 Kennzeichen einer Stichprobe

Eine Stichprobe  $S$  ist u.a. gekennzeichnet durch:

- Erhebungszeitraum (einmalig, periodisch)
- Grundgesamtheit (Umfang, Verteilungsannahme)
- Auswahlverfahren (einstufig, mehrstufig)
- Erhebungsverfahren (Interview, Fragebogen, physikalische Messung)
- Umfang (Anzahl der Beobachtungen und der Variablen pro Beobachtung)
- Skalenbereich und Genauigkeit der Messung

Bekannte **Beispiele** für Stichproben sind:

- repräsentative Umfragen vor Bundestagswahlen
- Untersuchung des Wahlverhaltens durch Kennzeichnung von Wahlzetteln
- zufälliges Entnehmen von Wasserproben
- Testen einzelner Produkte in der Qualitätskontrolle
- Markieren einzelner Vögel oder anderer Tiere

- Messen der Nutzung von Fernsehkanälen in ausgesuchten Haushalten

## 8.6 Messung von Variablen

Variablen lassen sich einteilen in **nominale**, **ordinale** und **metrische** Variablen:

Eine Variable ist **nominal-skaliert** (andere Bezeichnungen: klassifizierend, kategorisierend, gruppenbildend), wenn sie die Stichprobe in unterschiedliche "Kategorien", "Teilmengen", "Klassen" oder "Gruppen" einteilt, wobei die möglichen Werte (Ausprägungen) der Variablen keine offensichtliche (Rang-) Ordnung aufweisen.

Beispiele sind die Variable Geschlecht, die die Stichprobe in die Gruppen "Männer" und "Frauen" zerlegt, oder die Variable Haarfarbe, die eine Stichprobe nach Haarfarbe untergliedert. Ein Spezialfall sind **dichotom-skalierte** Variablen mit genau 2 möglichen Werten wie zum Beispiel „Ja/Nein“-Antworten.

Bei einer **nominalen Variablen** werden die möglichen Werte zur Abkürzung häufig, mehr oder weniger willkürlich auf Zahlen abgebildet. Die Zahl hat dabei keine andere Bedeutung als das sie stellvertretend für eine Zeichenkette steht.

Zum Beispiel können die Bundesländer mit Zahlen von 1 bis 16 durchnummeriert werden, wobei die Zahl 1 für das Bundesland Berlin steht, 2 für Sachsen usw.

Es ist unsinnig, für nominal gemessene Variablen einen Mittelwert zu bilden oder ein Histogramm zu erzeugen.

Eine Variable ist **ordinal-skaliert** (geordnet, mit Rangfolge), wenn ihr Wertebereich aus Zahlen besteht, zwischen denen eine offensichtliche Reihenfolge oder (Rang-) Ordnung existiert. Beispiele sind Schulnoten oder Sympathie-Werte für Politiker.

Bei einer **ordinalen Variablen** gibt es eine auf- oder absteigende Ordnung zwischen den möglichen Werten. Es kommt dabei aber nur auf die Reihenfolge und nicht auf die Abstände zwischen den Werten an.

Zum Beispiel legen Schulnoten oder andere Bewertungen zwischen 1 und 6 eine Reihenfolge fest, aber die Abstände haben keine gleichbleibende Bedeutung. Ähnlich kann ein Gesundheitszustand mit "gesund", "leicht erkrankt" und "krank" bewertet werden, die Abstände zwischen den möglichen Ausprägungen haben jedoch keine gleichbleibende Bedeutung.

Bei einer **metrischen Variablen** gibt es eine offensichtliche Ordnung und zusätzlich besitzen die Abstände zwischen möglichen Werten eine gleichbleibende Bedeutung.

Zum Beispiel ist ein im Jahr 1992 geborenes Kind 3 Jahre älter als ein 1995 geborenes Kind, dies ist wiederum 3 Jahre älter als ein 1998 geborenes Kind. Die Absolutwerte (Jahreszahlen) sind hier allerdings nicht von Bedeutung, da der Nullpunkt der Skala willkürlich gewählt ist.

## 8.7 Meßniveau

Bei der **Messung** von Variablen sind zusammengefaßt folgende Wertebereiche (Skalen) und Bedeutungen der möglichen Werte zu unterscheiden,

<b>Typ der Messung</b>	<b>Wertebereich (Menge oder Intervall)</b>
nominale Messung	{a,b, ..., }, {0,1,2,...,n} oder ähnlich, keine Ordnung, nur zur Klassifikation verwendbar
ordinale Messung	{0,1,2,...,n}, [a,b] oder ähnlich, (Rang-) Ordnung
metrische Messung	{0,1,2,...,n}, [a,b] oder ähnlich, Ordnung UND Abstände aussagekräftig (Stichwort: Lineal)

Tabelle 8.2 : Skalen-Niveaus

Viele Verfahren der deskriptiven oder mathematischen Statistik können nur angewendet werden, wenn gewisse Voraussetzungen hinsichtlich der Skalierung (*levels of measurement*) vorliegen.

## 8.8 Kenngrößen von Stichproben

Eine **Stichprobe**  $S = (x_1, \dots, x_n)$  setzt sich aus n **Beobachtungen** (andere Bezeichnung: **Fall**)  $x_1, \dots, x_n$  zusammen, wobei jede Beobachtung selbst aus mehreren Variablen bestehen kann. Im folgenden Abschnitt soll zur Vereinfachung jede Beobachtung nur aus einer Variablen bestehen (univariate Statistik).

Oftmals sind nicht die einzelnen beobachteten Werte interessant, sondern **Kenngrößen** oder **Maßzahlen** (*statistics*), die einen Überblick über die gesamte Stichprobe vermitteln. So können Sie zum Beispiel eine Stichprobe "verdichten", indem Sie nur den kleinsten, den größten Wert und den Mittelwert der Stichprobe betrachten. Mit jeder Verdichtung ist allerdings grundsätzlich ein Informationsverlust - bezogen auf das vollständige Datenmaterial - verbunden.

Wichtige **Kenngrößen** der Stichprobe sind für eine numerische Variable im folgenden aufgelistet.

Bekannt sind die Bezeichnungen **Lagemaße** für Mittelwert und empirischen Median und **Streuemaße** für empirische Varianz und empirische Standardabweichung der Stichprobe. Der Zusatz "empirisch" soll jeweils verdeutlichen, daß es sich um eine (bekannte), also meßbare Kenngröße der Stichprobe

handelt, die sich von der (unbekannten) Kenngröße der Grundgesamtheit unterscheidet.

Der **empirische Mittelwert** (*empirical mean*)  $\bar{x}$  der Stichprobe  $S = (x_1, \dots, x_n)$  beschreibt den gemittelten beobachteten Wert (arithmetisches Mittel) und ist definiert als:

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \dots + x_n)}{n}$$

Die **empirische Varianz** (*emp. variance*)  $s^2$  der Stichprobe  $S = (x_1, \dots, x_n)$  beschreibt die mittlere quadratische Abweichung vom empirischen Mittelwert und ist definiert als:

$$s^2 = \frac{((x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2)}{(n-1)}$$

Die **empirische Standardabweichung** (*emp. standard deviation, stddev*)  $s$  der Stichprobe  $S$  ist definiert als die Wurzel aus der empirischen Varianz. Sie hat den Vorteil, die selbe Einheit wie die Variable zu besitzen:

$$s = \sqrt{s^2}$$

Die **geordnete Stichprobe** (*ordered sample*) enthält die nach aufsteigender Reihenfolge geordneten Werte  $x_1, \dots, x_n$  der Stichprobe  $S$ . Mit  $x_{(1)}$  wird der kleinste, mit  $x_{(n)}$  der größte Wert bezeichnet, mit  $x_{(2)}$  der zweitkleinste usw.:

$$S_{\text{sorted}} = (x_{(1)}, x_{(2)}, \dots, x_{(n-1)}, x_{(n)})$$

Das **Minimum** und das **Maximum** der Stichprobe  $S$  bezeichnen den kleinsten und den größten beobachteten Wert und sind definiert als:

$$\min(S) = \min(x_1, \dots, x_n) = x_{(1)}$$

$$\max(S) = \max(x_1, \dots, x_n) = x_{(n)}$$

Die **Spannweite** (*range*) der Stichprobe  $S$  beschreibt die Differenz zwischen dem größten Wert (Maximum) und dem kleinsten Wert (Minimum) in der Stichprobe und ist definiert als:

$$\text{range}(S) = \max(S) - \min(S) = x_{(n)} - x_{(1)}$$

Der **empirische Median** (*emp. median*) der Stichprobe  $S$  ist definiert als der mittlere Wert in der geordneten Stichprobe (bzw. als das arithmetische Mittel der mittleren Werte für eine gerade Anzahl von Beobachtungen). Der Median trennt die Stichprobe in 2 Hälften; d.h. die Hälfte der Beobachtungen liegt oberhalb des Medians und die andere Hälfte liegt unterhalb des Medians. Er ist für eine ungerade Anzahl  $n$  definiert als:

$$\text{med}(S) = x_{(n/2)}$$

Ein **25%-Quantil** der Stichprobe  $S$  trennt die Stichprobe in 2 Hälften; wobei 25% der Beobachtungen unterhalb liegen und 75% oberhalb. Es ist für eine ungerade Anzahl  $n$  definiert als:

$$\text{25\%-Quantil}(S) = x_{(n/4)}$$

Die **relative Häufigkeit** (*relative frequency*) beschreibt die Anzahl der Beobachtungen  $x_i$ , die gleich einem vorgegebenen Wert  $k$  sind, geteilt durch die Gesamtanzahl  $n$  aller Beobachtungen und ist definiert als:

$$\begin{aligned} h(x) &= (\text{Anzahl beobachtete Werte } x_i = k) / n \\ &= \#(x_i = k) / n \end{aligned}$$

Die **empirische Verteilungsfunktion** (*empirical distribution function*) der Stichprobe  $S$  beschreibt die die kumulierten (aufsummierten) relativen Häufigkeiten und ist definiert als:

$$\begin{aligned} F^{\wedge}(x) &= (\text{Anzahl beobachtete Werte } x_i \leq x) / n \\ &= \#(x_i \leq x) / n \end{aligned}$$

$$\text{Es gilt immer: } 0 = F^{\wedge}(x_{(1)}) < \dots < F^{\wedge}(x_{(n)}) = 1$$

## 8.9 Übungen

1. Welche Skalierung (Wertebereich, Maßeinheit, Messniveau, ggf. Meßgenauigkeit) würden Sie für folgende Variablen wählen?
  - Sympathiewerte für Politiker
  - Schulnoten
  - Europäische Staaten
  - Bundesländer
  - Zeitmessung für 50km-Skilanglauf
  - Zeitmessung für 100-m-Lauf
2. [zusätzlich]  
Nennen Sie weitere Beispiele für Variablen und geeignete Skalierungen. Gehen Sie dabei auch auf das Problem der Fragebogen-Gestaltung ein (Alterangabe im Detail oder in Altersgruppen, analog für Familieneinkommen, Bildungsstand, Krankheiten oder andere sensible Daten).
3. Berechnen Sie (per Hand) einige der oben genannten Kenngrößen für die folgende kleine Stichprobe  $S=(1,5,3)$ , die die Augenzahlen bei 3 zufälligen Würfelwürfen darstellen soll.



4. [zusätzlich]  
Welche weiteren Kenngrößen einer Stichprobe kennen Sie?

## 9 Erstellen von einfachen Tabellen

In diesem Kapitel werden anknüpfend an das Kapitel über die deskriptive Statistik einige Möglichkeiten zum tabellarischen Darstellen von Variablen vorgestellt.

Ich will mir zunächst einen tabellarischen Überblick über meine Daten verschaffen, um Ideen zu erhalten, welche Zusammenhänge existieren könnten ...

### 9.1 Berechnen von Häufigkeiten

Im folgenden Beispiel führen Sie für die Variable „partei“ aus „sonntagsfrage-01.sav“ eine Häufigkeitsauszählung durch. Wählen Sie „Analysieren > Deskriptive Statistiken > Häufigkeiten“.

Klicken Sie dann links in der Liste aller Variablen auf „Partei“ und danach auf den Pfeil nach rechts, um die Variable in die Liste der Variablen zur Bearbeitung aufzunehmen:

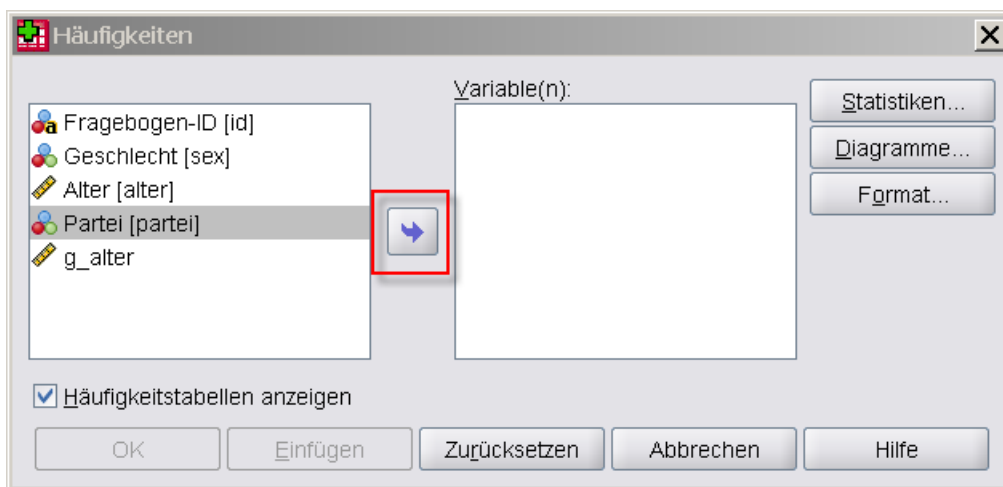


Abbildung 9.1: Analysieren > Deskriptive Statistiken > Häufigkeiten

Klicken Sie dann auf die Schaltfläche "OK", um die Häufigkeiten für die ausgewählten Variablen zu berechnen.

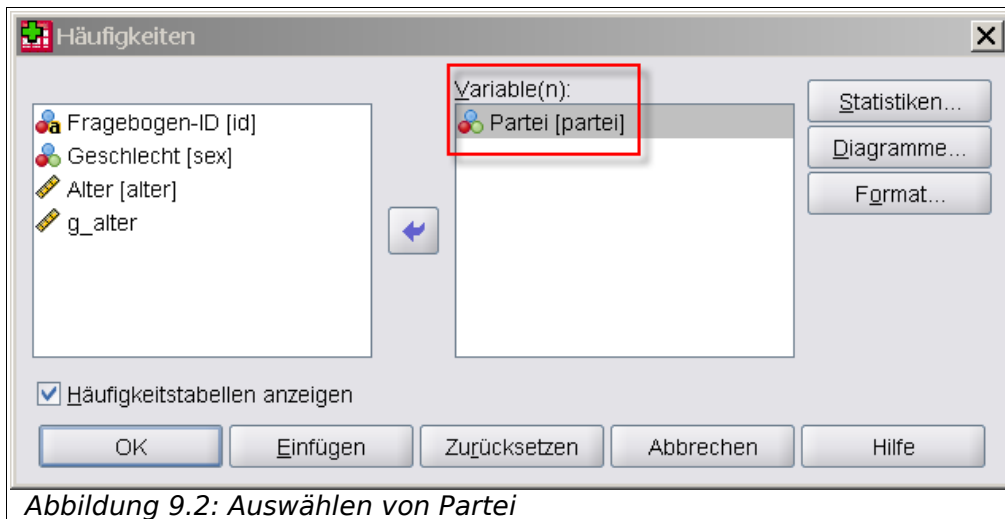


Abbildung 9.2: Auswählen von Partei

Aufgrund der gewählten Einstellungen erhalten Sie in einem neuen „SPSS Viewer“-Fenster das Ergebnis angezeigt:

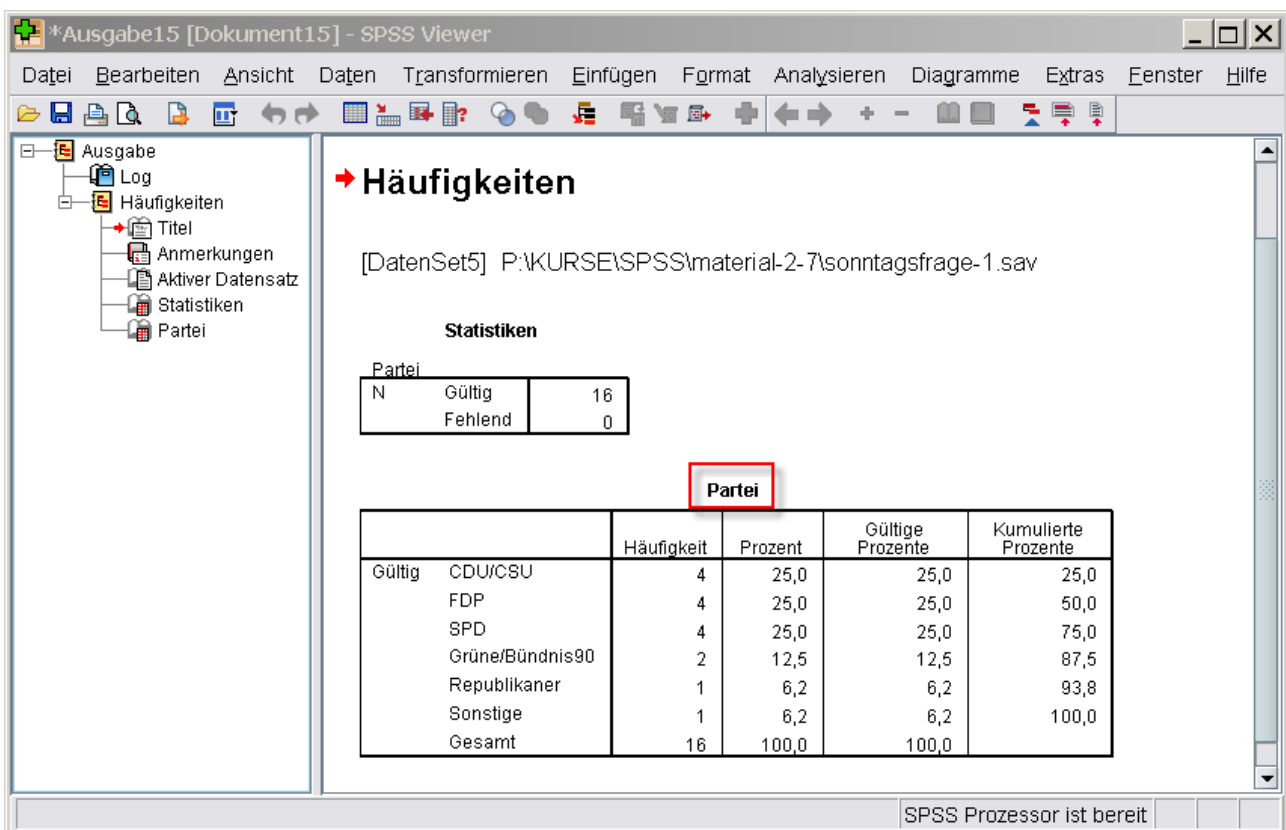


Abbildung 9.3: SPSS Viewer Fenster: Häufigkeiten

In diesem Beispiel gibt es keine fehlenden Werte.

## 9.2 Erstellen einer Kreuztabelle

Im folgenden Beispiel führen Sie für die Variablen **partei** (Partei) und **sex** (Geschlecht) eine Häufigkeitsauszählung in einer Kreuztabelle durch.

Wählen Sie hierzu „*Analysieren > Deskriptive Statistiken > Kreuztabellen*“. Wählen Sie dann aus der Liste der Variablen die zu bearbeitenden Variablen aus.

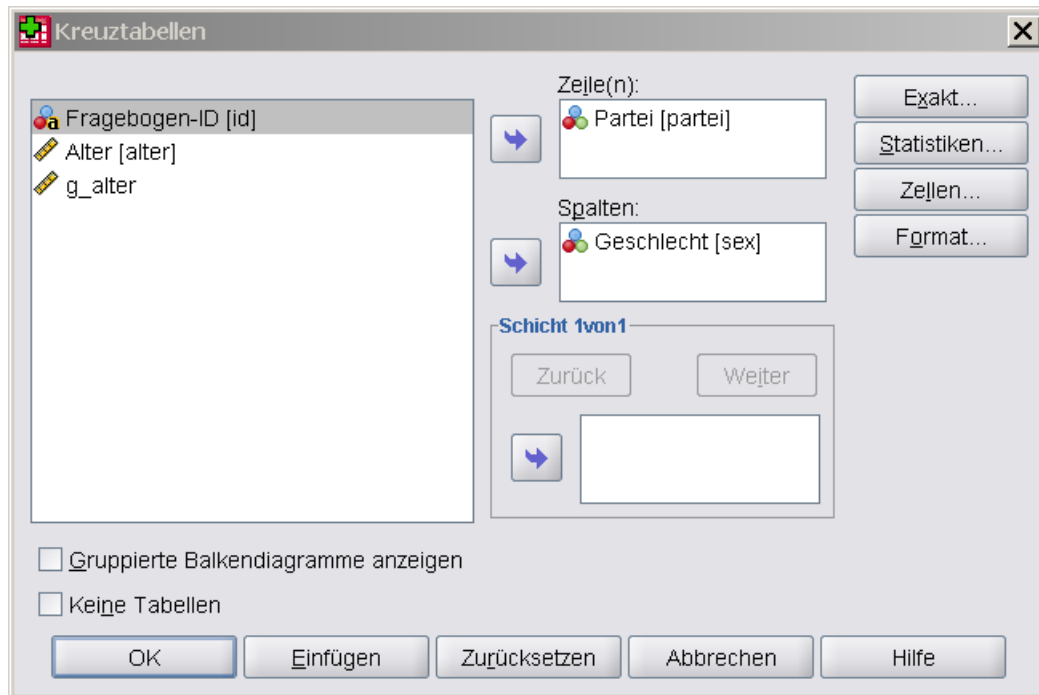


Abbildung 9.4: Kreuztabelle: Zeilen Partei, Spalten Geschlecht

Die Häufigkeiten werden im „SPSS Viewer“-Fenster, gruppiert nach Männern und Frauen angezeigt:

The screenshot shows the SPSS Viewer window with the following content:

**Kreuztabellen**

[DatenSet5] P:\KURSE\SPSS\material-2-7\sonntagsfrage-1.sav

**Verarbeitete Fälle**

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
Partei * Geschlecht	16	100,0%	0	,0%	16	100,0%

**Partei \* Geschlecht Kreuztabelle**

Anzahl		Geschlecht		
		weiblich	männlich	Gesamt
Partei	CDU/CSU	2	2	4
	FDP	1	3	4
	SPD	2	2	4
	Grüne/Bündnis90	2	0	2
	Republikaner	1	0	1
	Sonstige	1	0	1
	Gesamt	9	7	16

SPSS Prozessor ist bereit

Abbildung 9.5: SPSS Viewer Fenster: Kreuztabellen

## 9.3 Übungen

1. Führen Sie für die SPSS Datendatei "schueler.sav" (Noten einer Klasse in einzelnen Fächern) Häufigkeitsauszählungen für die Variablen „deutsch“ und „physik“ durch.
2. Führen Sie die Häufigkeitsauszählungen erneut aus, diesmal in einer Kreuztabelle, getrennt nach Jungen und Mädchen.
3. Fügen Sie einen (fiktiven) Datensatz für einen Schüler hinzu, für den keine Noten in Deutsch und Physik vorliegen, und führen Sie die Berechnungen erneut durch. Wo ergeben sich Unterschiede?
4. In der SPSS Datendatei „sonntagsfrage-03.sav“ ist offensichtlich ein Eingabefehler, welcher? Wie können Sie diesen Fehler erkennen und beseitigen?
5. Erzeugen Sie für die SPSS Datendatei „sonntagsfrage-01.sav“ eine neue Variable "a\_gruppe" mit der Einteilung 1=„bis 25“, 2=„25-40“ und 3=„40-65“ und 4=„über 65“. Erstellen Sie eine Kreuztabelle mit "agruppe/partei" (Partei: horizontal und Altersgruppe: vertikal).
6. Führen Sie für die SPSS Datendatei „sonntagsfrage-01.sav“ eine Kreuztabelle mit "agruppe/partei" und der zusätzlichen Gruppierung (Schichtung) nach der Variablen „sex“ durch.

# 10 Berechnen von Kennzahlen

In diesem Kapitel werden anknüpfend an das Kapitel über die deskriptive Statistik einige Möglichkeiten zum Verdichten der Informationen über eine Stichprobe vorgestellt.

Ich will mir zunächst einen Überblick über die Kennzahlen meiner Daten verschaffen, um Zusammenhänge, ungewöhnliche Werte, ggf. auch Eingabefehler, zu entdecken ...

## 10.1 Berechnen einfacher Kennzahlen

Im folgenden Beispiel berechnen Sie für die numerische Variable „groesse“ aus „broca-01.sav“ einige für die Stichprobe charakteristische Kennzahlen (Lage- und Streumaße) wie z.B. den Mittelwert und die Standardabweichung:

Wählen Sie hierzu „Analysieren > Deskriptive Statistiken > Häufigkeiten“. Wählen Sie dann aus der Liste der Variablen die zu bearbeitenden Variablen aus, hier die Variable „groesse“.

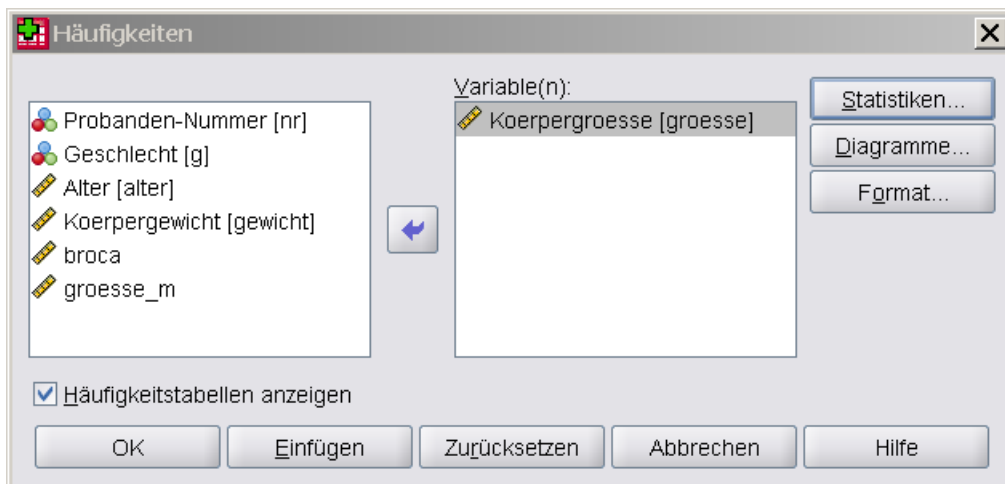


Abbildung 10.1: Häufigkeiten: Variable groesse

Fordern Sie dann über die Schaltfläche „Statistik“ gezielt die gewünschten Kennzahlen an:

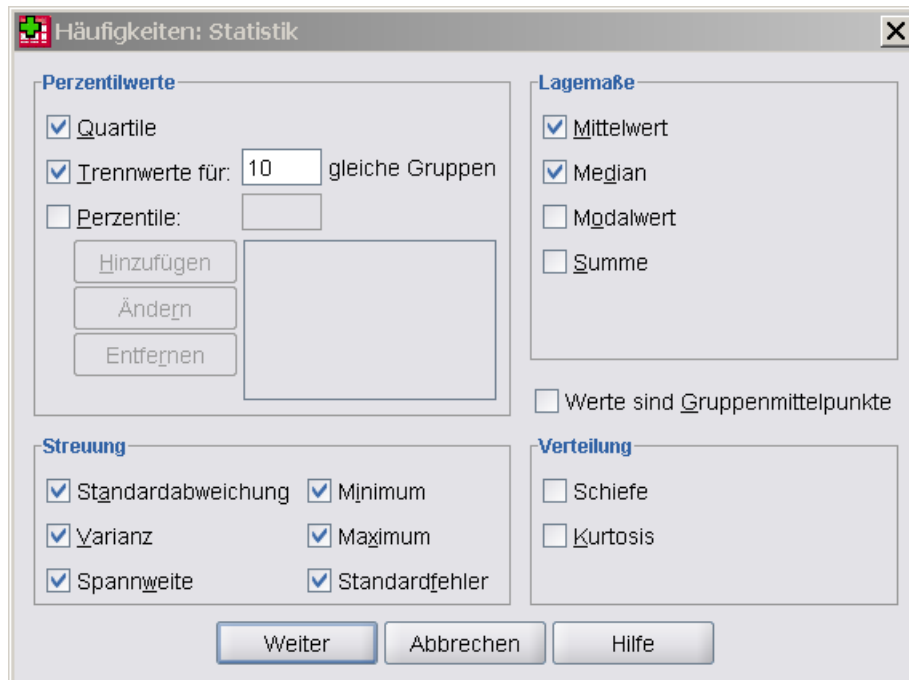


Abbildung 10.2: Häufigkeiten: Zusatzmenü: Statistiken

Klicken Sie im Untermenü „Statistik“ auf „Weiter“ und im Menü „Häufigkeiten“ auf „OK“. Die Ergebnisse werden im „SPSS Viewer“-Fenster angezeigt:

Statistiken		
Koerpergroesse		
N	Gültig	174
	Fehlend	0
Mittelwert		165,17
Standardfehler des Mittelwertes		,612
Median		165,00
Standardabweichung		8,079
Varianz		65,273
Spannweite		40
Minimum		145
Maximum		185
Perzentile		
	10	155,00
	20	158,00
	25	159,75
	30	160,00
	40	163,00
	50	165,00
	60	167,00
	70	170,00
	75	171,00
	80	173,00
	90	176,00

Abbildung 10.3: SPSS Viewer Fenster: Häufigkeiten: Statistiken

Die Bedeutung der Kennzahlen für die Stichprobe ist zum Teil im Kapitel über die deskriptive Statistik erläutert worden; Informationen über Kennzahlen können Sie aber auch über das Hilfesystem abrufen oder in guten Lehrbüchern zur Statistik nachlesen.

## 10.2 Übungen

1. Berechnen Sie für die SPSS Datendatei „*schueler.sav*“ für die Variable **physik** die Kennzahlen Mittelwert, Minimum, Maximum, Spannweite, Varianz, Standardabweichung und Median.
2. Wie können Sie die Klasse hinsichtlich der Variablen „*physik*“ in 3 gleich-grosse Leistungsgruppen einteilen (stark, mittel, schwach)? Welche Schüler, identifiziert anhand der Variablen „*nr*“ befinden sich in der Gruppe „mittel“?
3. Berechnen Sie die Kennzahlen Mittelwert, Minimum, Maximum, Spannweite, Varianz, Standardabweichung und Median für die Variable „**physik**“ getrennt nach Mädchen und Jungen.

4. [zusätzlich]  
Ermitteln Sie bezogen auf die SPSS Datendatei „`schueler.sav`“ in welchem Fach es jeweils die größte Spannweite, die größte Varianz und den größten bzw. kleinsten Mittelwert gibt. Welche Schlüsse würden Sie als Klassenlehrer aus diesen Ergebnissen ziehen?
5. In der Datendatei „`sonntagsfrage-03.sav`“ ist offensichtlich ein Eingabefehler, welcher? Wie können Sie diesen Fehler entdecken und beseitigen?
7. [zusätzlich]  
Was ist die **Schiefte** einer empirischen Verteilung? Kann eine Stichprobe mit "extrem schiefer" empirischer Verteilung approximativ normal-verteilt sein?
8. [zusätzlich]  
Was ist die **Kurtosis** einer empirischen Verteilung?
9. [zusätzlich]  
Welche weiteren Kennzahlen können berechnet werden?

# 11 Erstellen von Diagrammen

In diesem Kapitel werden Methoden zum grafischen Darstellen von Variablen vorgestellt. Die visuelle Darstellung dient als Ergänzung zur tabellarischen Darstellung und hilft häufig, Verhältnisse und absolute Werte zu verdeutlichen und ggf. auch interessante Zusammenhänge im Datenmaterial zu entdecken, getreu dem Motto:

*1 picture is worth a 1000 words.*

## 11.1 Visualisieren von Datenmaterial

Im Bereich der Diagrammerstellung gibt es zwischen den SPSS Versionen 15 und 16 signifikante Unterschiede. Die folgenden Beschreibungen beziehen sich auf die Version 16 – nach Ansicht des Autors ist diese Version ein Rückschritt in Sachen Verständlichkeit und Bedienbarkeit gegenüber Version 15.

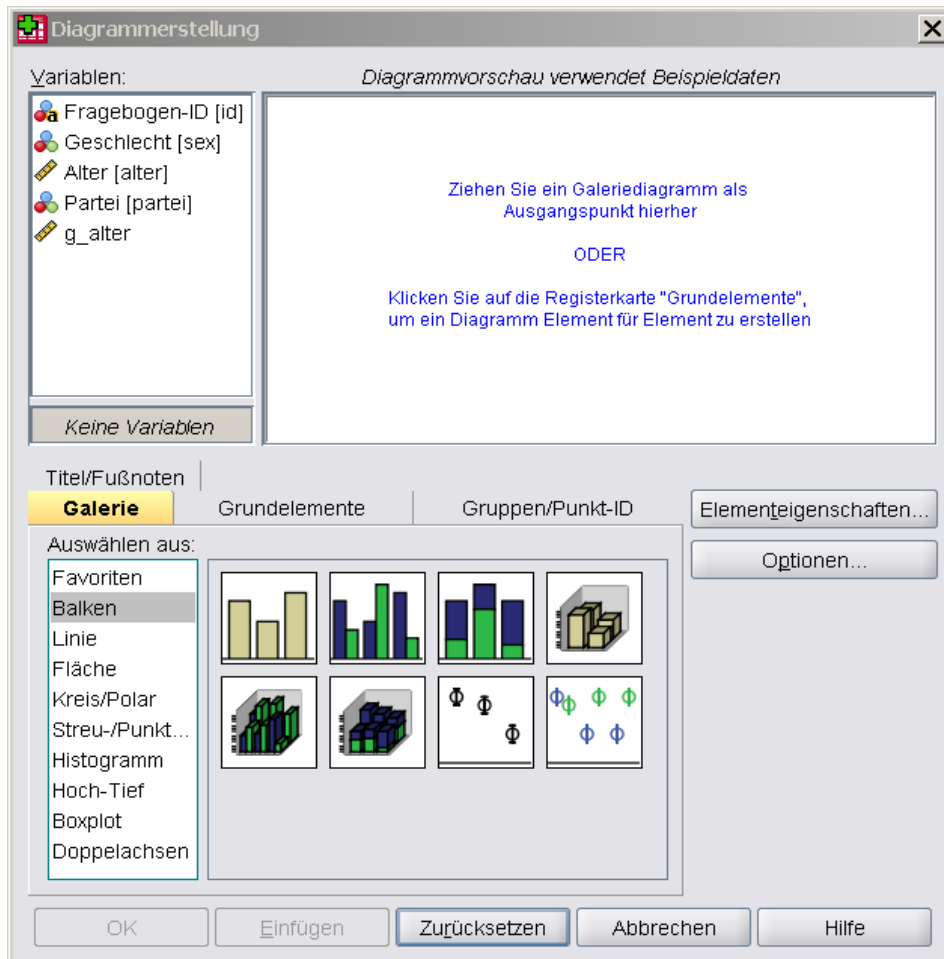


Abbildung 11.1: Diagramme > Diagrammerstellung (Version 16)

## 11.2 Erstellen eines einfachen Balkendiagramms

Stellen Sie die Häufigkeit der Kategorien, d.h. die unterschiedlichen beobachteten Werte, der Variablen `partei` aus „sonntagsfrage-01.sav“ als Balkendiagramm dar.

Wählen Sie „Diagramme > Veraltete Dialogfelder > Balken“.

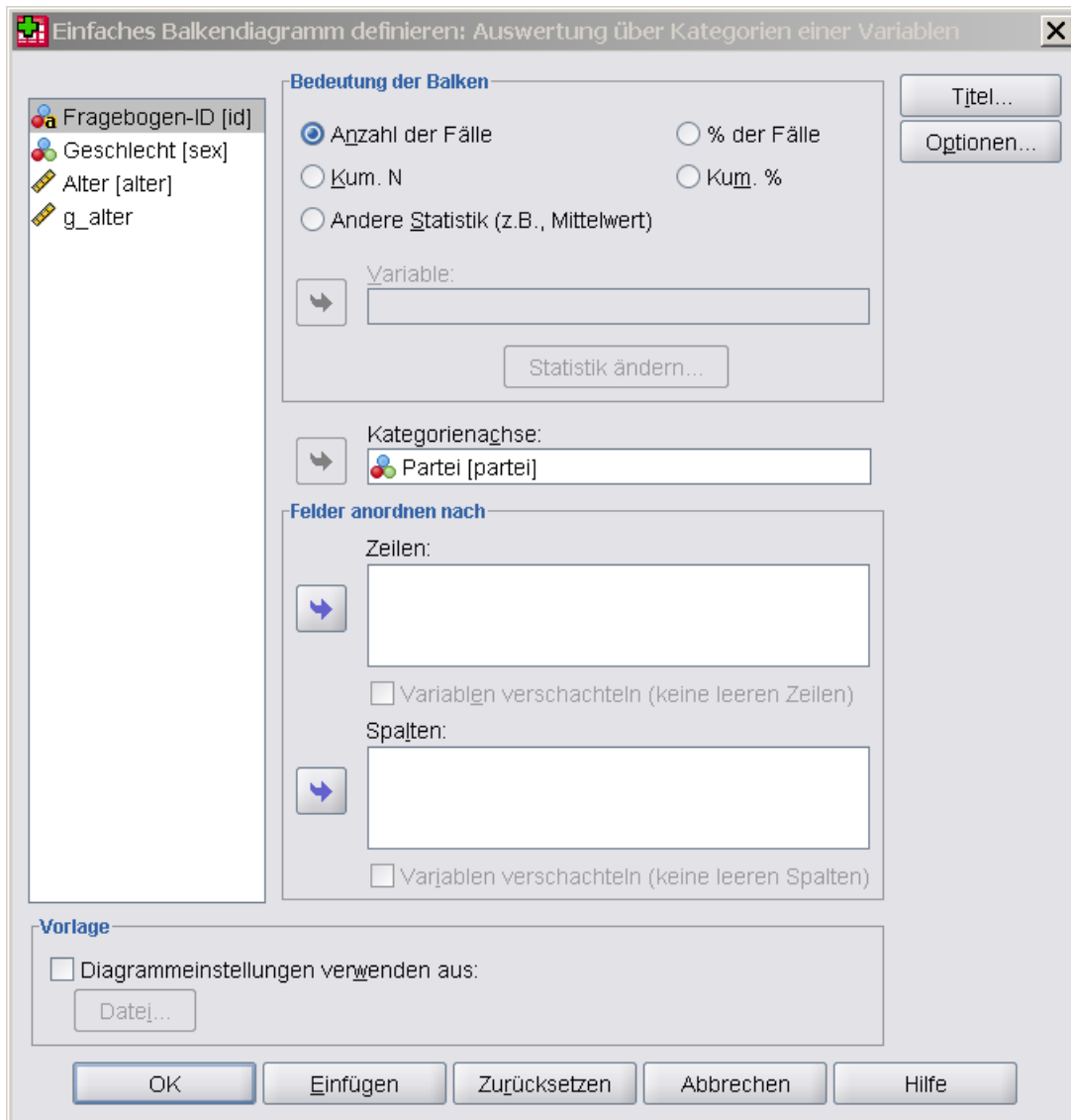


Abbildung 11.2: Diagramme > Veraltete Dialogfelder > Balken

Wählen Sie für die Kategorien-Achse (x-Achse) die Variable „partei“. Die x-Achse repräsentiert hiermit Kategorien, also die unterschiedlichen möglichen Werte (Ausprägungen) einer Variablen. Wählen Sie „Anzahl der Fälle“ als darzustellende Größe in y-Richtung aus; d.h. die Höhe eines Balkens repräsentiert die Anzahl der Beobachtungen für den an der x-Achse angezeigten Wert.

Klicken Sie abschließend auf die Schaltfläche „OK“. Die Ergebnisse werden im „SPSS Viewer“-Fenster angezeigt und sind hier bereits nachbearbeitet:

## Diagramm

[DatenSet4] P:\KURSE\SPSS\material-2-7\sonntagsfrage-1.sav

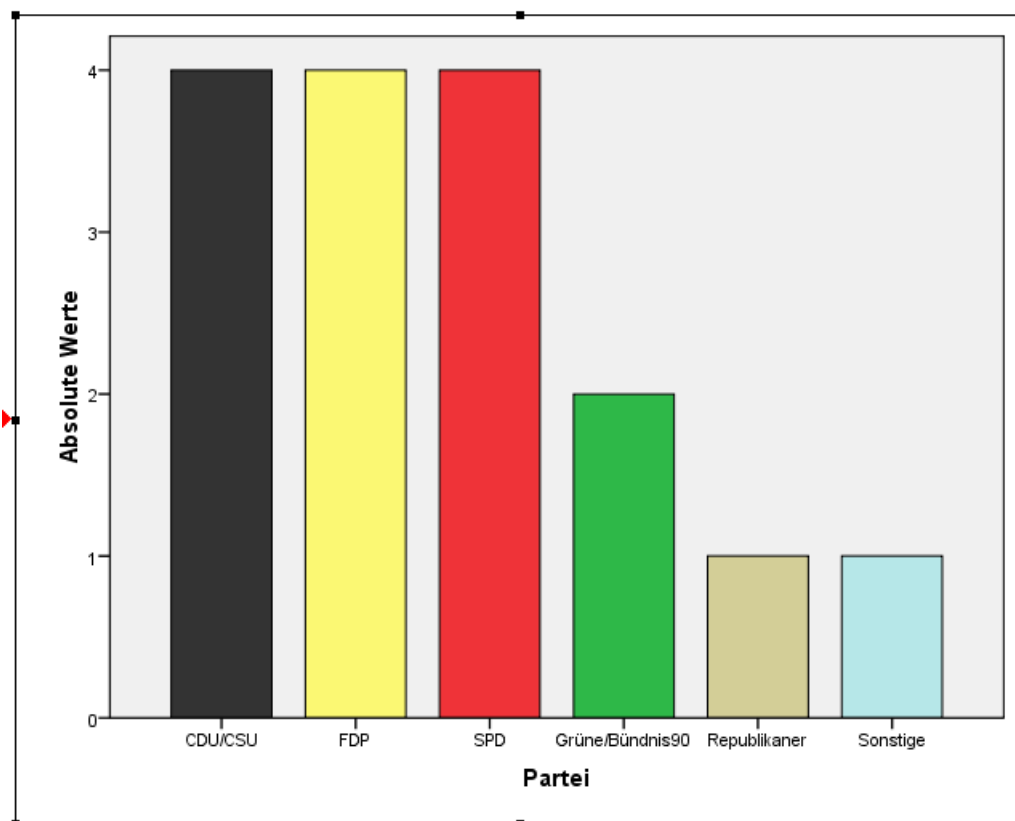
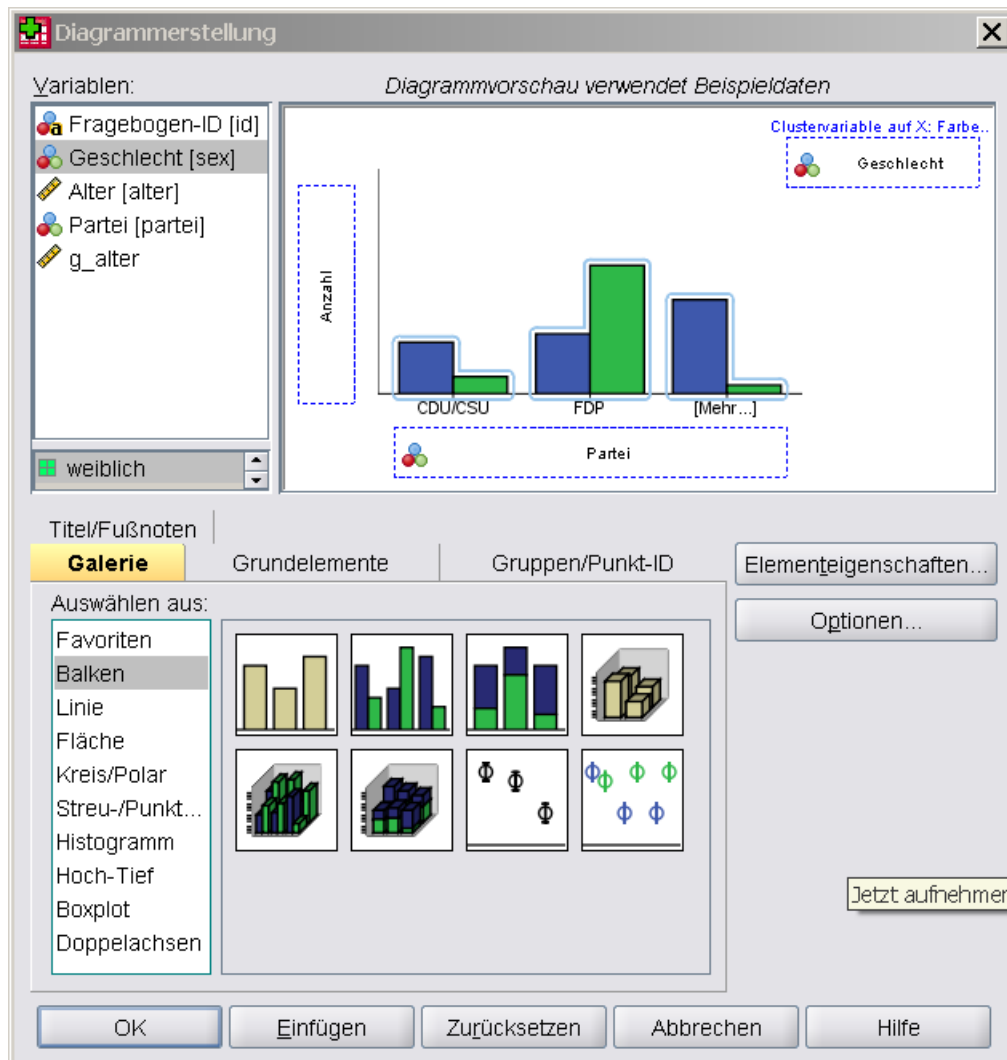


Abbildung 11.3: Diagramm: Balkendiagramm (nachbearbeitet)

Das Balkendiagramm zeigt die Anzahl der Beobachtungen für jede Kategorie der Variablen `partei` an. Es gibt zum Beispiel jeweils insgesamt 4 Personen, die CDU/CSU, FDP oder SPD wählen würden.

## 11.3 Erstellen eines gruppierten Balkendiagramms

Stellen Sie die Häufigkeit der Kategorien von der Variablen `partei` als Balkendiagramm dar, wobei Sie eine Untergliederung nach Geschlecht (`sex`) vornehmen.



Klicken Sie abschließend auf die Schaltfläche „OK“. Die Ergebnisse werden im „SPSS Viewer“-Fenster angezeigt.

Das gruppierte Balkendiagramm zeigt die Anzahl der Beobachtungen für jede Kategorie der Variablen „partei“, gruppiert nach der Variablen „sex“ an.

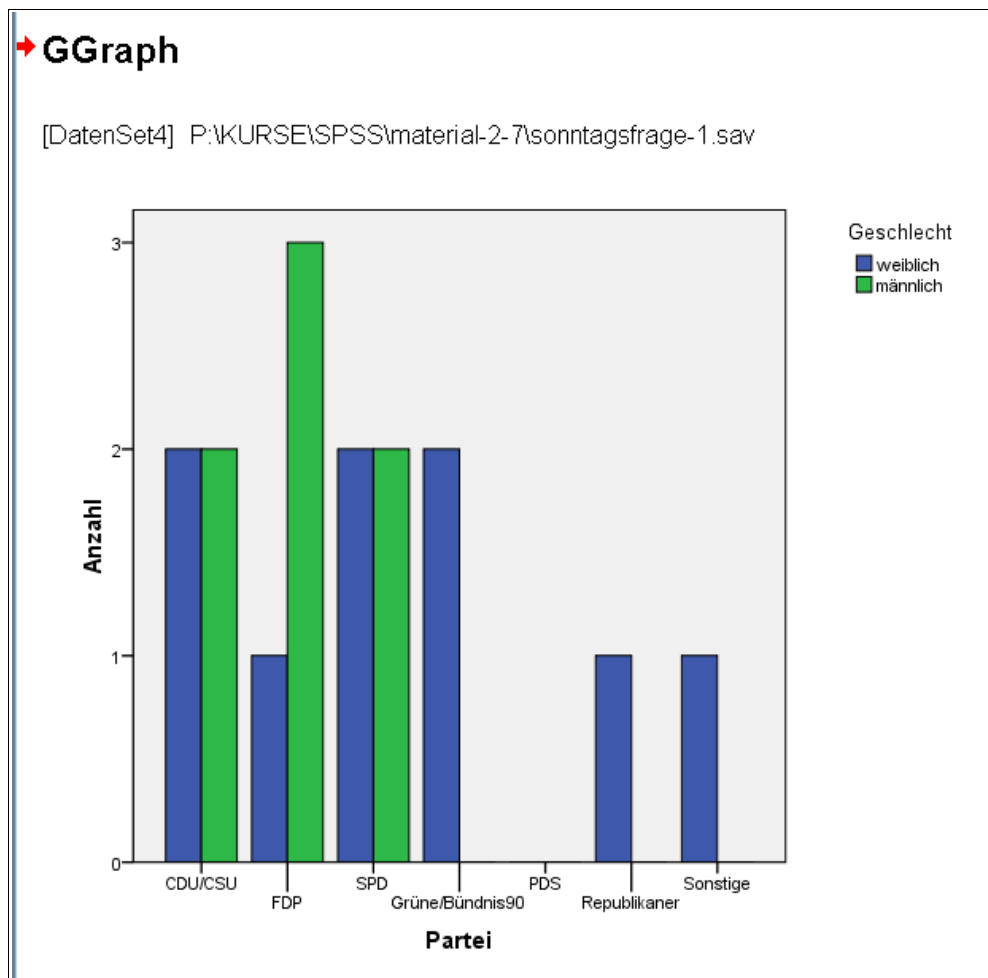


Abbildung 11.4: Diagramme: gestapeltes Balkendiagramm

## 11.4 Übungen

- Erzeugen Sie folgende Diagramme für die SPSS Datendatei „sonntagsfrage-01.sav“:
  - Tortendiagramm für **partei**
  - gestapeltes Balkendiagramm für **partei** mit Gruppierung nach einer Altersgruppe, zum Beispiel nach „g\_alter“.
- Exportieren Sie das Tortendiagramm aus Übung 1a in das Format PNG (Portable Network Graphic) oder „WMF“ (Windows Metafile) und fügen sie das Diagramm in ein Word- bzw. OpenOffice-Dokument ein - oder in eine Powerpoint Präsentation

3. Vergleichen Sie mit einem Kopieren/Einfügen über die Windows Zwischenablage.
4. zusätzlich:  
Experimentieren Sie mit dem „3D-Effekt“ (siehe folgende Abbildung).

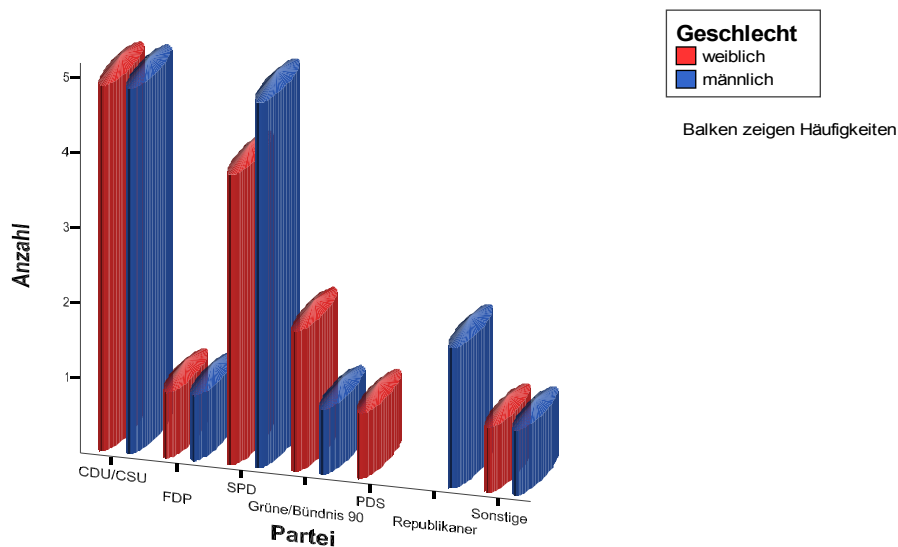


Abbildung 11.5 : 3D-Effekt für Diagramme

## 11.5 Erstellen eines Flächendiagramms

Die folgenden Diagramme beruhen auf Datenmaterial über die Entwicklung von Studentenzahlen im Fach Informatik in den Jahren von 1975 bis 1993.

Die SPSS Datendatei „studiengang-informatik-01.sav“ enthält u.a. folgende Variablen:

Variable	Bedeutung
JAHR	Jahr (von 1975-1993)
STUD_GES	Studenten gesamt
STUD_W	davon: Studenten weiblich
STUD_M	davon: Studenten männlich

Erzeugen Sie ein Flächendiagramm, um die zeitliche Entwicklung der Studentenzahlen zu verdeutlichen. Wählen Sie hierzu „Grafiken Interaktiv > Fläche“.

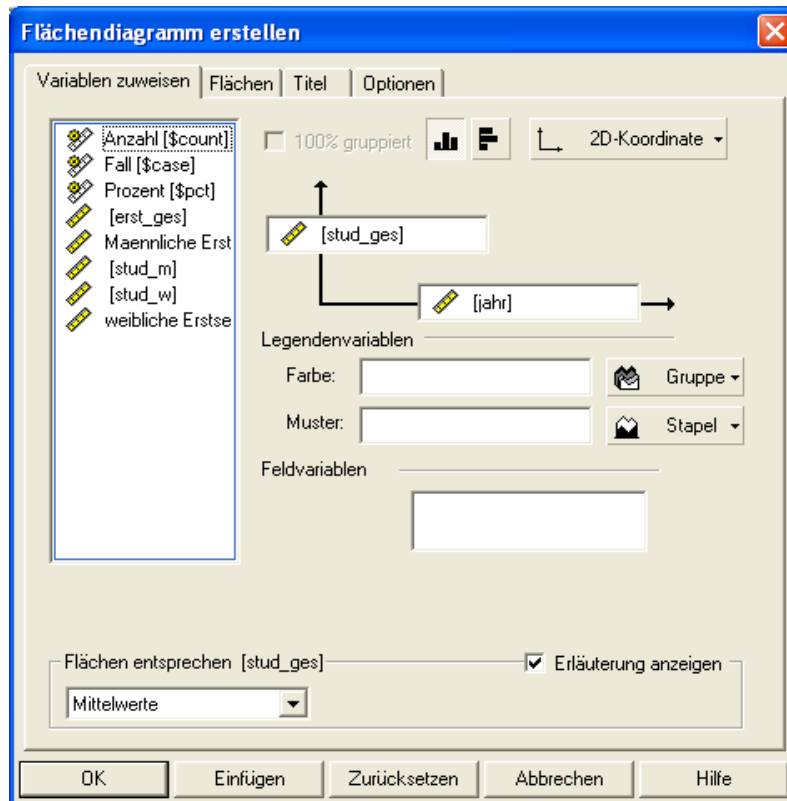


Abbildung 11.6 : Flächendiagramm erstellen

Klicken Sie auf die Schaltfläche „OK“. Die Ergebnisse werden im „SPSS Viewer“-Fenster angezeigt.

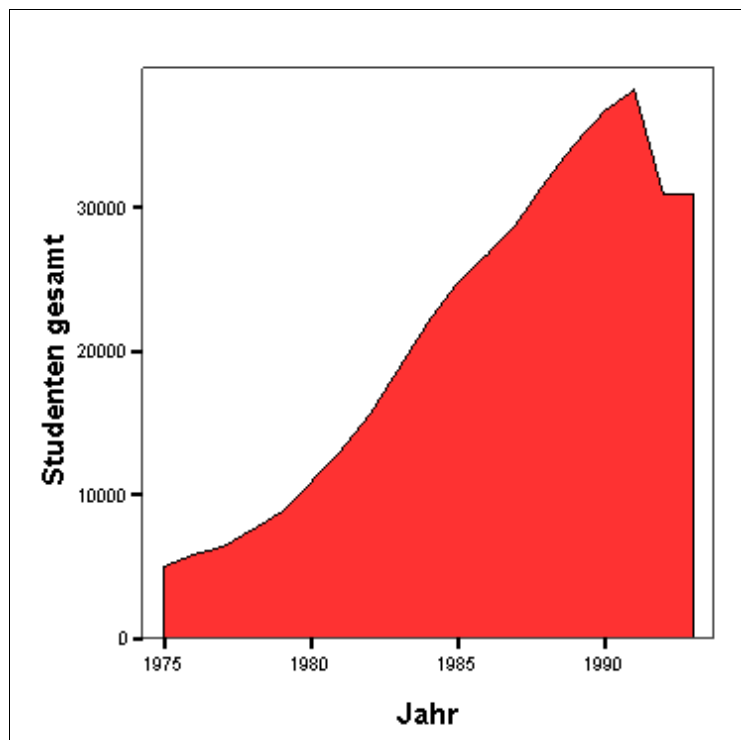


Abbildung 11.7: Flächendiagramm

Das Flächendiagramm zeigt die Werte einzelner Beobachtungen. Statt der laufenden Nummer der Beobachtung wurde die Variable `jahr` zur Beschriftung der x-Achse ausgewählt.

## 11.6 Erstellen eines gestapelten Flächendiagramms

Im folgenden Beispiel wird das Flächendiagramm verändert, indem weibliche und männliche Studenten getrennt ausgewiesen werden.

Wählen Sie „Grafik > Flächendiagramm“.



Abbildung 11.8: Gestapeltes Flächendiagramm

Wählen Sie die Variablen aus, die gestapelt (aufsummiert) dargestellt werden sollen, hier `stud_m` (Anzahl männlicher Studenten) und `stud_w` (Anzahl weiblicher Studenten).

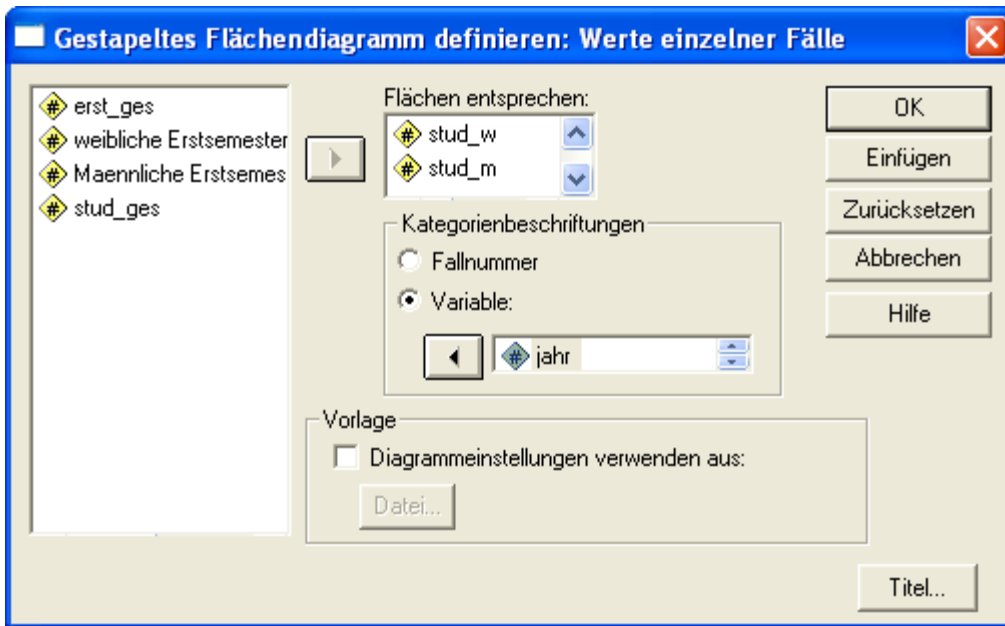


Abbildung 11.9: Gestapeltes Flächendiagramm definieren

Klicken Sie auf „OK“. Die Ergebnisse werden im „Viewer“-Fenster angezeigt.

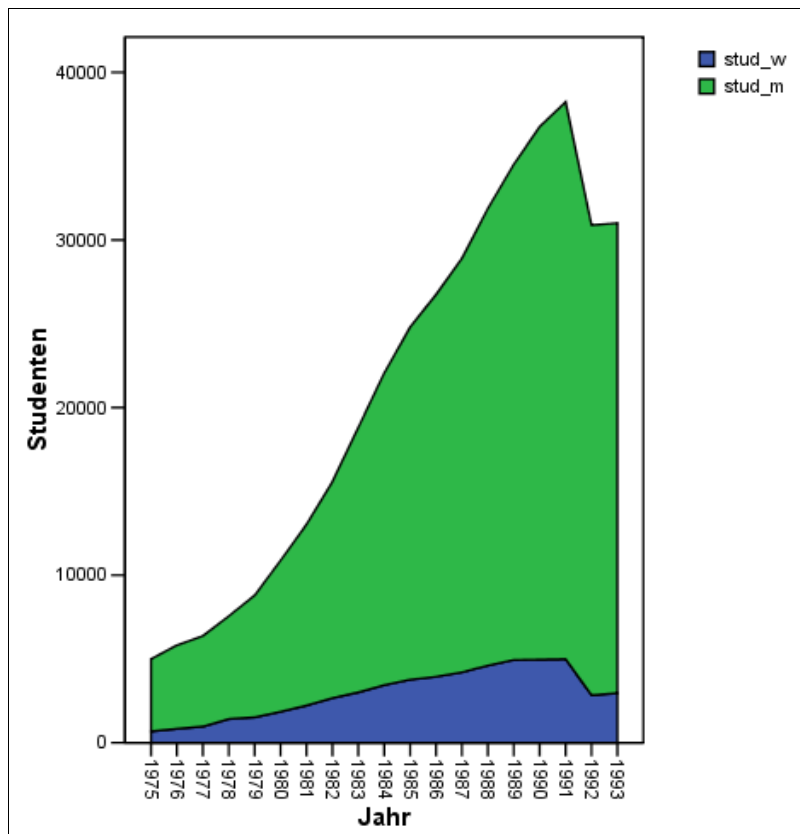


Abbildung 11.10: Gestapeltes Flächendiagramm

## 11.7 Erstellen eines Histogramms

Im folgenden Beispiel erstellen Sie für die Variable **groesse** aus der SPSS Datendatei „broca-01.sav“ ein Histogramm (Häufigkeitsverteilung) mit überlagerter Normalverteilungskurve. Wählen Sie „*Grafiken > Interaktiv > Histogramm*“. Wählen Sie die Variablen aus, für die Histogramme erstellt werden sollen.

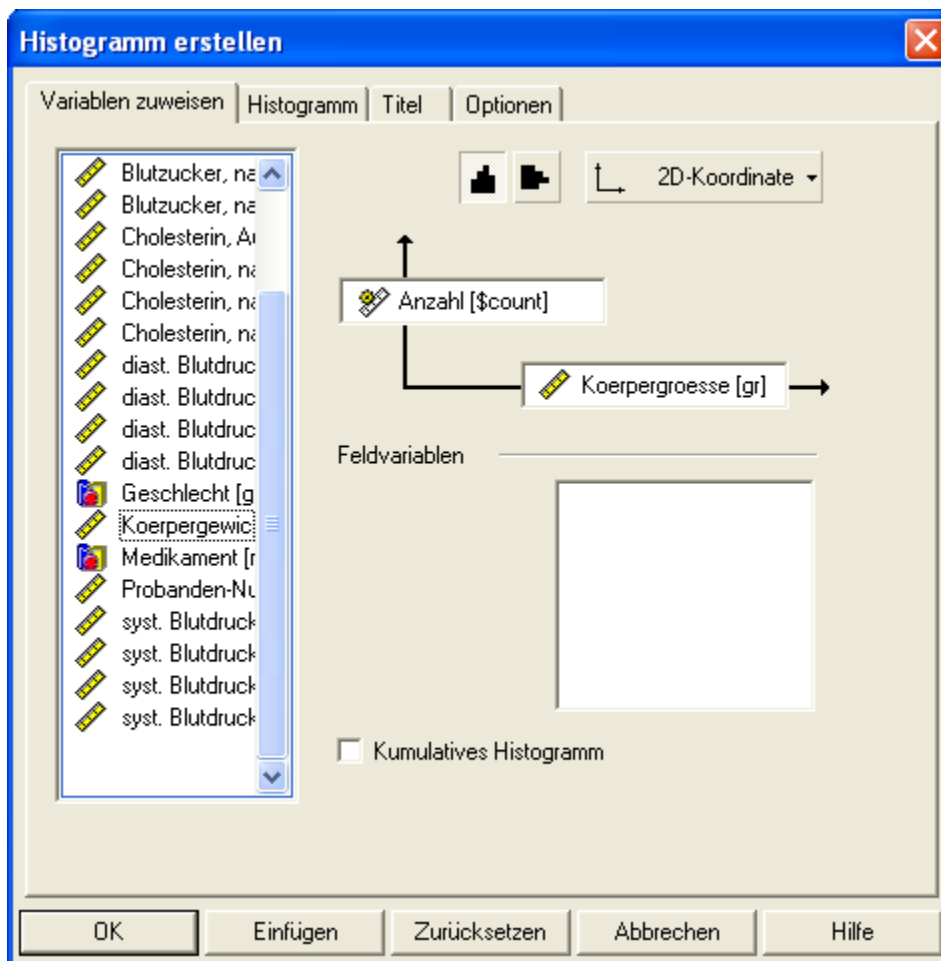


Abbildung 11.11

Fordern Sie durch Ankreuzen der entsprechenden Checkbox zusätzlich eine eingezeichnete Normalverteilungskurve an.



Abbildung 11.12

Klicken Sie auf die Schaltfläche „OK“. Die Ergebnisse werden im „SPSS Viewer“-Fenster angezeigt.

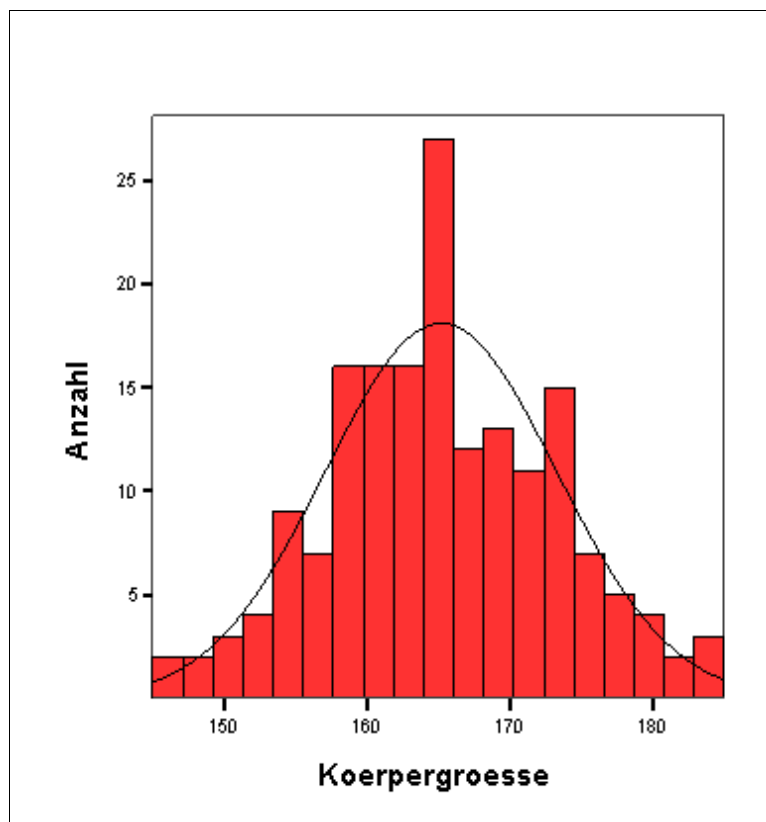


Abbildung 11.13: Histogramm

Das Histogramm zeigt an, wieviele Beobachtungen in die vorgegebenen Intervalle fallen. Sie haben die Möglichkeit, die Anzahl der Intervalle zu verändern und damit das Histogramm zu verändern.

Die überlagerte Normalverteilungskurve gibt Anlaß zur Vermutung, daß die Stichprobe annähernd normal-verteilt ist. Die Normalverteilungskurve wird mit Hilfe von Schätzwerten der Stichprobe für die Parameter der Normalverteilung konstruiert. Die benötigten Parameter Erwartungswert  $\mu$  und Varianz  $\sigma^2$ , die die Normalverteilung eindeutig bestimmen, werden hierbei durch den Mittelwert und die empirische Varianz der Stichprobe geschätzt.

## 11.8 Übungen

1. Erstellen Sie ein Histogramm mit überlagerter Normalverteilungskurve für die Variable „Gewicht“ („gewicht“) aus der SPSS Datendatei „broca-01.sav“.
2. Überprüfen Sie, ob tatsächlich der Mittelwert der Stichprobe als Schätzwert für den Parameter der überlagerten Normalverteilung verwendet wird.

3. Erstellen Sie ein gestapeltes Flächendiagramm für männliche und weibliche Erstsemester aus der SPSS Datendatei „studiengang-informatik-01.sav“.

## 11.9 Vergleichen von empirischen Verteilungen

Ein **Box- und Whisker-Plot** enthält folgende Informationen:

Symbol	Bezeichnung	Bedeutung
*	obere Extremwerte	Werte, die weiter als 3 Boxlängen oberhalb vom 75%-Quantil liegen
o	obere Ausreißer	Werte, die weiter als 1.5 Boxlängen oberhalb vom 75%-Quantil liegen
-----	größter "normaler" Wert	größter beobachteter Wert, der noch kein Ausreißer ist (nicht zu verwechseln mit MAX)
	Verbindungsline	
+----+	75% Quantil	Wert, der größer ist als 75% aller beobachteten Werte (Die Box enthält entsprechend 50% aller Werte.)
+---	50% Quantil, Median	Wert, der größer ist als 50% aller beobachteten Werte
+----+	25% Quantil	Wert, der größer ist als 25% aller beobachteten Werte
	Verbindungsline	
-----	kleinster "normaler" Wert	größter beobachteter Wert, der noch kein Ausreißer ist (nicht zu verwechseln mit MIN)
o	untere Ausreißer	Werte, die weiter als 1.5 Boxlängen unterhalb vom 25%-Quantil liegen
*	untere Extremwerte	Werte, die weiter als 3 Boxlängen unterhalb vom 25%-Quantil liegen

Im folgenden Beispiel betrachten Sie die Variable „**groesse**“ aus der SPSS Datendatei „broca-01.sav“.

Vergleichen Sie die Verteilung der Körpergröße mit Hilfe eines Box- und Whisker-Plots. Wählen Sie „Grafik > Interaktiv > Boxplot“. Wählen Sie als darzustellende Variable „Körpergröße“ und als Gruppierung „Geschlecht“. Die Box- und Whisker-Plots werden dann gruppiert nach den Kategorien von „Geschlecht“, d.h. nach „männlich“ und „weiblich“, getrennt dargestellt.

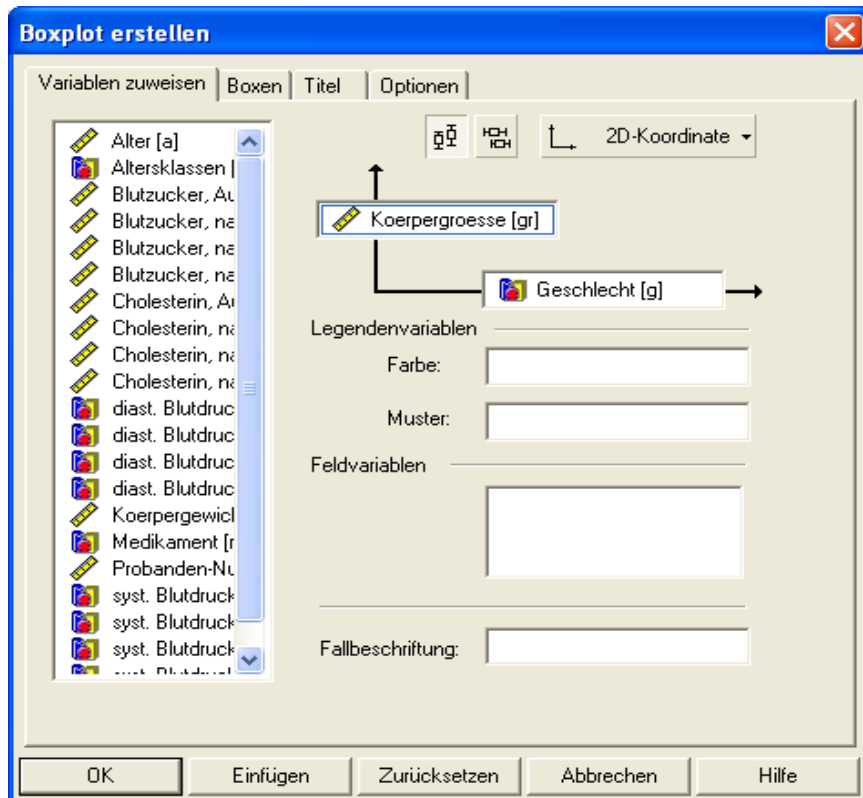


Abbildung 11.14

Klicken Sie auf „OK“. Die Ergebnisse werden im „SPSS Viewer“-Fenster angezeigt.

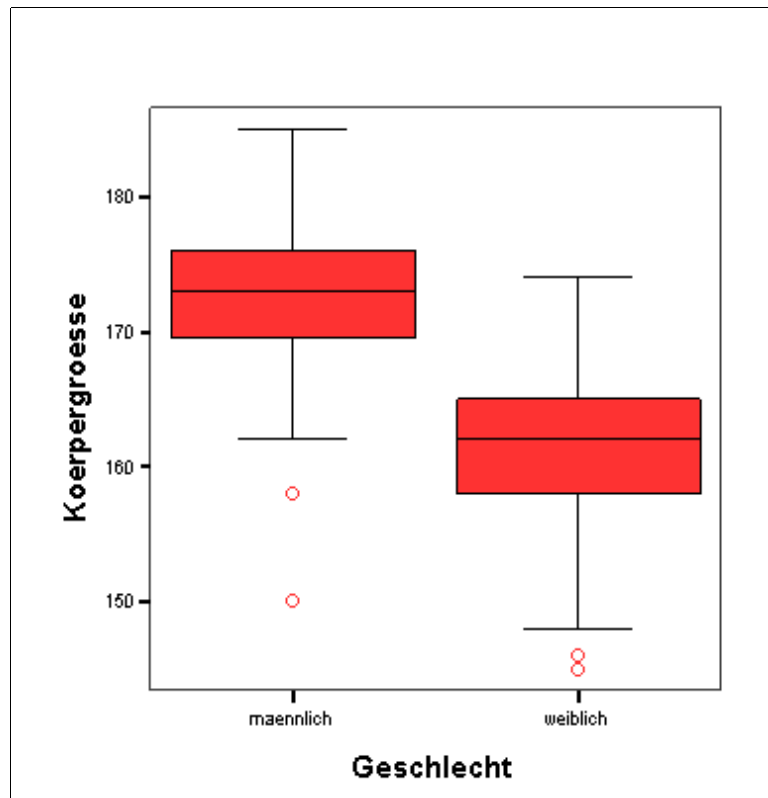


Abbildung 11.15

Die Boxplots sind leicht gegeneinander verschoben, d.h. die untersuchten Männer waren im Mittel größer als die Frauen. Es gibt Ausreißer, aber keine Extremwerte.

## 11.10 Bearbeiten von Diagrammen

Sie können Grafiken, die Sie im „SPSS Viewer“-Fenster betrachten, individuell nachbearbeiten.

Klicken Sie hierzu auf das gewünschte Diagramm. Das aktuell angezeigte Diagramm wird nun in den Diagramm-Editor geladen, der zahlreiche Werkzeuge zur Nachbearbeitung zur Verfügung stellt.

Aufgrund der Vielzahl der Möglichkeiten wird an dieser Stelle nur auf die grundsätzliche Vorgehensweise verwiesen:

1. Markieren Sie durch Doppelklicken das Element des Diagrammes, das bearbeitet werden soll (z.B. Achsenbeschriftung, Linie, Überschrift, Legende).
2. Wählen Sie neue Einstellungen in der Dialogbox aus und klicken Sie auf „OK“, um die neuen Einstellungen wirksam werden zu lassen

## 11.11 Übungen

1. Erstellen Sie einen Box- und Whisker-Plot die Variable „Gewicht“ (gewicht) aus „broca-01.sav“, einmal gesamt und einmal getrennt nach Männern und Frauen.
2. [zusätzlich]  
Das Statistische Bundesamt bietet in der Genesis Online-Datenbank, erreichbar über „<https://www-genesis.destatis.de/genesis/online/logon>“, die Möglichkeit zum Download von diversen statistischen Daten, u.a. im Format Excel.  
Laden Sie die folgende Tabelle im Format Excel herunter:  
Tabelle 21311-0003; Studierende: Deutschland, Semester, Nationalität, Geschlecht, Studienfach; eingeschränkt auf das WS 2006/2007  
Sie finden diese Tabelle auch als Excel-Datei „studiengang-ws0607.xls“ oder als SPSS Datendatei „studiengang-ws0607.sav“ im Bereich Material.
3. [zusätzlich]  
Bereinigen Sie die Excel-Datei „studiengang-ws0607.xls“ für den Import nach SPSS.
4. Erzeugen Sie für die SPSS Datendatei „studiengang-ws0607.sav“ eine neue Variable „D\_Gesamt“ mit der Gesamtzahl der männlichen und weiblichen deutschen Studenten. Stellen Sie die 10 Studiengänge mit den meisten Studenten in einem Diagramm dar. Vergleichen Sie mit den 10 Studiengängen, die für Männer und Frauen getrennt am beliebtesten sind.
5. [zusätzlich]  
Experimentieren Sie mit Hoch-/Tief-Diagrammen (Beispiel: Aktien, Fieberkurven) oder Fehlerbalken-Diagrammen (Beispiel: Meßfehler in physikalischen Experimenten). Verwenden Sie hierzu zum Beispiel die Kurse der „Hypo Real Estate“ (WKN: 802770 ISIN: DE0008027707).
6. [zusätzlich]  
Verschaffen Sie sich einen Überblick über die weiteren verfügbaren Diagrammtypen.
7. Speichern Sie abschließend das Sitzungsprotokoll.

# 12 Zufallsexperimente, Zufallsvariablen und Wahrscheinlichkeit

In diesem Kapitel wird erläutert, wie Ergebnissen von Zufallsexperimenten Wahrscheinlichkeiten zugeordnet werden. Dieses Kapitel dient zur Auffrischung und kann ggf. überschlagen werden.

Ich habe immer Mühe, mir unter dem Begriff „Wahrscheinlichkeit“ etwas vorzustellen.

## 12.1 Zufallsexperiment und Wahrscheinlichkeit

Ein Vorgang oder Versuch, dessen Durchführung "zufällig" zu genau einem von mehreren möglichen Ergebnissen führt, wird als **Zufallsexperiment** oder **Zufallsvorgang** bezeichnet.

Klassische Beispiele für Zufallsexperimente stammen aus der Welt der Spiele wie das Werfen eines Würfels, das Ziehen von Losen in einer Lotterie oder das Ziehen von Spielkarten beim Poker.

Das Ziehen von Karten beim Poker wird in der gewöhnungsbedürftigen Sprache der **Kombinatorik** („Kunst des Zählens“) bezeichnet als *„Auswahl einer Stichprobe aus einer Grundgesamtheit, bei der  $m$  Objekte "zufällig" ohne Zurücklegen aus der Grundgesamtheit mit  $n$  Objekten ausgewählt werden“*.

Das Zufallsexperiment besitzt eine Ergebnismenge  $\Omega$  (Menge der möglichen Ergebnisse). Die **Wahrscheinlichkeit  $P$**  ist eine Abbildung von  $\Omega$  in das Intervall  $[0,1]$ , die jedem (Elementar-) **Ergebnis**  $\omega$  des Zufallsexperimentes eine positive Zahl  $p$ , die Wahrscheinlichkeit für das Eintreten dieses Ergebnisses, zuordnet.

$$P(\omega) = p$$

Die Abbildung  $P$  hat die folgenden Eigenschaften:

$$P(\{\omega\}) > 0$$

$$P(\Omega) = 1$$

$$P(\{\omega_1, \omega_2\}) = P(\{\omega_1\}) + P(\{\omega_2\})$$

Die Abbildung  $P$ , die Elementar-Ergebnissen "Wahrscheinlichkeiten" zuweist, wird aufgrund einer mathematischen Theorie, anhand von plausiblen Annahmen, von Erfahrungswerten oder von Schätzungen aufgestellt.

Eine gute intuitive Vorstellung sieht  $P(\omega)$  als "stabilisierte" relative Häufigkeit für das Ergebnis  $\omega$  bei einer sehr großen Anzahl von gleichartigen Versuchen.

## 12.2 Zufallsvariablen und ihre Verteilung

Bei einem Zufallsexperiment seien die Ergebnisse  $\Omega = \{w_1, \dots, w_n\}$  möglich. Jedem Ergebnis sei durch die weitere Abbildung  $X$  eine reelle Zahl zugeordnet. Die Abbildung  $X$  heißt **Zufallsvariable**, die möglichen Werte von  $X$  ergeben den **Wertebereich** von  $X$ . (Zufalls-) **Variablen** sind die beobachtbaren **Merkmale** oder **Eigenschaften** von Objekten oder Personen, die in einem Zufallsexperiment ausgewählt werden.

Bei einem Zufallsexperiment interessieren oft nicht die elementaren Ergebnisse, sondern eine vom elementaren Ergebnis  $w$  abgeleitete (Zufalls-) Variable  $X(w)$ .

Bei einem Angelwettbewerb interessieren zum Beispiel nicht die einzelnen gefangenen Fische, sondern nur die Anzahl (oder das Gesamtgewicht) aller gefangenen Fische oder auch einfach nur der schwerste gefangene Fisch.

Die (Wahrscheinlichkeits-) **Verteilung**  $P^X$  einer Zufallsvariablen  $X$  wird über die Wahrscheinlichkeit der „Urbilder“ in der Ergebnismenge  $\Omega$  definiert; d.h. die Summe aller Wahrscheinlichkeiten für elementare Ergebnisse, die zu einem Wert  $k$  von  $X$  führen. Die Verteilung von  $X$ ,  $P^X$ , gibt also Auskunft darüber, mit welcher Wahrscheinlichkeit die Zufallsvariable  $X$  einen bestimmten Wert  $k$  aus dem Wertebereich annimmt.

## 12.3 Vorher und Nachher

Beachten Sie, daß Sie nur **vor der Durchführung** des Zufallsexperiments Aussagen über die **möglichen Werte** und **deren Wahrscheinlichkeiten** treffen können, während **nach der Durchführung** genau ein **beobachteter** (realisierter) **Wert** zur Verfügung steht.

Dieser Sachverhalt wird im folgenden durch folgende Notation verdeutlicht:

(Zufalls-) Variablen werden immer mit Großbuchstaben (gebräuchlich sind die Buchstaben  $X, Y, Z$ ) bezeichnet, während die möglichen oder tatsächlich beobachteten Werte für eine (Zufalls-) Variable mit Kleinbuchstaben (gebräuchlich sind die Buchstaben  $x, y, z$ ) bezeichnet werden:

<b>Zufallsexperiment</b>	
Vorher (vor der Durchführung):	Nachher (nach der Durchführung):
Wahrscheinlichkeiten für mögliche Ergebnisse	Realisierung eines Ergebnisses
Zufallsvariable $X$ , mögliche Werte $x_1, x_2, x_3, \dots$	beobachteter Wert von $X$ , z.B. $x_3$
$P(X=x_3) = p_3$ ; $0 < p_3 < 1$	

Bei einer Durchführung eines Zufallsexperimentes ist also sehr wohl möglich, daß der Wert  $x_3$  mit der kleinsten Wahrscheinlichkeit  $p_3$  beobachtet wird. Nur **bei häufiger Wiederholung** eines Zufallsexperiments ist zu erwarten, daß Werte mit großen Wahrscheinlichkeiten auch häufiger eintreten (Gesetz der großen Zahlen).

## 12.4 Aufgaben

1. Kennen Sie Spiele, die auf Zufallsexperimenten beruhen?
2. Viele Spiele haben Elemente von „Glück und Zufall“ und „Strategie“. Spielen Sie lieber Schach oder Blackjack?
3. Wie lautet die Verteilung der Augenzahl beim Zufallsexperiment "Werfen von 2 echten Würfeln"? (siehe unten)
4. [zusätzlich]  
Zeichnen Sie die Verteilungsfunktion und berechnen Sie einige Kenngrößen der Verteilung (Grundgesamtheit).
5. [zusätzlich]  
Führen Sie das Zufallsexperiment nun 10-mal mit 2 echten Würfeln durch und vergleichen Sie die empirische und die tatsächliche Verteilungsfunktion. Berechnen Sie einige Kenngrößen Ihrer Stichprobe und vergleichen Sie mit den korrespondierenden Kenngrößen der Grundgesamtheit.
6. [zusätzlich]  
Kennen Sie den Begriff **Gauß'sche Glockenkurve**? Unter welchem Namen ist die zugehörige Verteilung bekannt? Worin besteht die besondere Bedeutung dieser Verteilung? Der "alte" 10-DM-Schein enthält sowohl eine Grafik der Dichtefunktion wie auch die recht komplizierte explizite Formel der Dichtefunktion (siehe unten).
7. [zusätzlich]  
Wie lautet der zentrale Grenzwertsatz? Wie ist er zu interpretieren?
8. [zusätzlich]  
Welche weiteren wichtigen diskreten und kontinuierlichen Verteilungen sind Ihnen bekannt und wie lauten jeweils die Verteilungsfunktionen?



Abbildung 12.1: 10 DM Schein, Vorderseite (Quelle: Wikipedia)

	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

Tabelle 12.1: Würfelwurf mit 2 echten Würfeln

# 13 Überblick über die mathematische Statistik

In diesem Kapitel werden die wesentlichen Prinzipien der mathematischen Statistik behandelt. Dieses Kapitel dient zur Auffrischung und kann ggf. überschlagen werden.

## 13.1 Ziehen von Rückschlüssen aus einer Stichprobe

In der mathematischen oder analytischen Statistik werden **Verfahren** entwickelt und angewendet, um anhand einer Stichprobe, d.h. Des Auswählen einer Teilmenge der Grundgesamtheit, **Rückschlüsse** oder **Folgerungen** (*statistical inference*) für die Grundgesamtheit ziehen zu können. Es wurde bereits sehr früh aus naheliegenden Gründen der Versuch unternommen, aus repräsentativen Beobachtungsdaten Gesetzmäßigkeiten abzuleiten, die über den Beobachtungszeitraum und -ort hinaus gültig waren.

Historisch gesehen entwickelte sich die Statistik mit dem aufkommenden Absolutismus und der Einführung merkantilistischer Wirtschaftsordnungen zu einem Hilfsmittel zentralistischer Staatspolitik. So wurden in Preußen erste statistische Erhebungen unterschiedlichen Umfangs in der Zeit des Großen Kurfürsten (1620 bis 1688) durchgeführt. Die Ergebnisse blieben aber meist in der Verwaltung und waren Interessierten kaum zugänglich, stellten also geheimes Herrschaftswissen dar, das es vor den Untertanen und konkurrierenden Staaten zu schützen galt. Sie beschäftigte sich hauptsächlich mit Tauf-, Heirats- und Sterberegistern, um Geschlechtsverhältnis, Altersaufbau und Sterblichkeit der Bevölkerung abzuschätzen.

**Stichprobe** und **Grundgesamtheit** lassen sich u.a. durch folgende, korrespondierende Größen beschreiben:

Grundgesamtheit (diskrete Verteilung)	Stichprobe S
Anzahl Elemente $m$	Anzahl Beobachtungen $N$
Verteilung, Wahrscheinlichkeit für bestimmte Werte $P(X=k)$	Relative Häufigkeit <sup>1</sup> von bestimmten Werten $h(k)=\#(x_i=k)/n$
Verteilungsfunktion von X $F(k)=P(X\leq k)$	Empirische Verteilungsfunktion von S $F^*(k)=\#(x_i\leq k)/n$

1 Das Zeichen # dient zur Abkürzung für Anzahl, z.B. #(Augenzahl=5) steht für: Anzahl der Würfe, bei denen die gewürfelte Augenzahl 5 beträgt.

<b>Erwartungswert</b> $\mu = E[X] = k_1 P(X=k_1) + \dots$	<b>Mittelwert von S</b> $\bar{x} = (k_1 + \dots) / n$
<b>Varianz</b> $\sigma^2 = \sigma_{XX} = E[(X-\mu)^2] = (k_1 - \mu)^2 P(X=k_1) + \dots$	<b>Empirische Varianz von S</b> $s^2 = s_{XX} = (k_1 - \bar{x})^2 + \dots$
<b>Standardabweichung</b> $\sigma = \sqrt{\text{Varianz}}$	<b>Empirische Standardabweichung von S</b> $s = \sqrt{\text{Empirische Varianz}}$
<b>Median</b> $m = F^{-1}(0.5)$	<b>Empirischer Median von S</b> $m^* = x_{(n/2)}$
<b>Kovarianz</b> $\rho_{XY} = \text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$	<b>Empirische Kovarianz</b> $s_{XY} = ((k_1 - \bar{x})(l_1 - \bar{y}) + \dots) / (s_{XX} s_{YY})$

Grundsätzlich gilt, daß die beobachteten **Kennzahlen der Stichprobe** nicht mit den entsprechenden (unbekannten) **Kennzahlen der Grundgesamtheit** übereinstimmen.

Die mathematische Statistik stellt jedoch Verfahren zur Verfügung, um auf Grundlage der Stichprobe **plausible Schätzungen** für die Grundgesamtheit abzugeben oder um **Tests** über bestimmte Aussagen zu (unbekannten) Kennzahlen der Grundgesamtheit durchzuführen.

## 13.2 Durchführen von Schätzungen und Hypothesentests

Viele Verfahren der mathematischen Statistik lassen sich stark reduziert auf folgende Fragestellung zurückführen:

Welche Aussage über eine **unbekannte** Kennzahl (wahrer Parameter) der **Grundgesamtheit** kann aufgrund der Beobachtung der korrespondierenden **realisierten** (empirischen, beobachteten, bekannten) Kennzahl der **Stichprobe** gemacht werden?

Entgegen einer weitverbreiteten Meinung bedeutet mathematische Statistik nicht (oder nur in sehr geringem Maße) Sammeln und tabellarisches Zusammenstellen (evtl. auch Manipulation?) von Unmengen an Zahlenmaterial [... es gibt die Notlüge, die gemeine Lüge und die Statistik ...], sondern die **Entwicklung und Begründung von Verfahren** zur Auswertung von zufallsabhängigen

gen Beobachtungsdaten, **mit denen sich "vernünftige" Entscheidungen bei ungewisser Sachlage treffen lassen**<sup>2</sup>.

Vernünftig heißt in diesem Zusammenhang, daß die Sicherheit, mit der ein statistisches Verfahren zu einer richtigen Entscheidung führt, vertrauensstärkend ist. Ein Verfahren hat eine Sicherheit (Erfolgswahrscheinlichkeit, Konfidenz-Niveau) von z.B. 0.95, wenn es im Mittel in 95 von 100 Durchführungen zu einer richtigen Entscheidung führt, und entsprechend eine Irrtumswahrscheinlichkeit von 0.05; d.h. Im Mittel führen 5 von 100 Durchführungen zu einer falschen Entscheidung.

Wichtig ist neben der statistischen Signifikanz natürlich auch die praktische Relevanz einer Aussage bzw. die Auswirkungen ("Verlust"), die eine "falsche" Entscheidung nach sich zieht.

So kann im privaten Bereich eine allzu hohe Erwartung an die Trinkfreudigkeit der Party-Gäste dazu führen, daß der nach der Party verbleibende Getränkevorrat für den Rest des Jahres für den eigenen Konsum ausreicht, eine zu niedrige Erwartung an die Trinkfreudigkeit kann leicht zu schlechter Stimmung führen.

### 13.3 Einschränken der gesuchten theoretischen Verteilung auf eine Klasse (parametrische Tests)

Bei konkreten Problemen liegen oft genaue oder gewisse Kenntnisse hinsichtlich der "Rahmenbedingungen" eines Zufallsexperimentes vor (z.B. bei einer Lotterie: "n-malige Stichprobenentnahme ohne Zurücklegen von Kugeln"), so daß die Menge aller in Frage kommenden theoretischen Verteilungen auf eine Klasse von Verteilungen eingeschränkt werden kann.

In diesem Fall spricht man von einer **Verteilungsannahme**, d.h. der Einschränkung auf eine **Klasse von Verteilungen**, in der sich die einzelnen Verteilungen nur noch durch unterschiedliche **Kenngößen** wie Lage- oder Streumaße (z.B. Erwartungswert und Varianz) unterscheiden. Die einfachere Aufgabe besteht in diesem Fall nun darin, mit Hilfe eines statistischen Verfahrens gesicherte Aussagen über die unbekanntenn Kennzahlen zu erhalten, die die gesuchte theoretische Verteilung vollständig charakterisieren.

Diese Aufgabenstellung ist weitaus einfacher, als aus der unendlichen Anzahl aller möglichen theoretischen Verteilungen eine „passende“ auszuwählen. Aus der anderen Bezeichnung **Parameter** für Kenngröße oder Maßzahl leitet sich der Begriff **parametrische Statistik** für diesen Bereich von statistischen Fragestellungen ab. Entsprechend gehören Fragestellungen, bei denen keine Verteilungsannahmen gemacht werden, zur **nicht-parametrischen Statistik**.

Viele der bekannten statistischen Verfahren setzen weiter einschränkend voraus, daß die beobachteten Zufallsvariablen **unabhängig** sind und daß die Ver-

---

<sup>2</sup> Da ungewisse Sachlagen eigentlich den Normalfall im Leben darstellen, treffen wir mehr oder weniger bewußt sehr häufig statistisch motivierte Entscheidungen.

teilung der Grundgesamtheit (wenigstens approximativ) eine **Normalverteilung** ist (**Normalverteilungsannahme**).

Der wohl wichtigste Satz der Statistik, der zentrale Grenzwertsatz der Statistik, besagt, daß der Mittelwert einer Stichprobe approximativ normalverteilt ist. Für große Stichprobenumfänge ist also die Normalverteilungsannahme häufig gerechtfertigt.

## 13.4 Formulieren von Fragestellungen

Die möglichen statistischen Fragestellungen sollen am folgenden einfachen Beispiel erläutert werden:

Beim 100-maligen Werfen eines Würfels mit den beobachteten Augensummen  $x_1, \dots, x_{100}$  interessiere der unbekannte Erwartungswert  $\mu$  der gewürfelten Augenzahl. Bei einem "echten" Würfel berechnet sich der Erwartungswert  $\mu$  aus Symmetriegründen zu 3.5, aber vielleicht ist der Würfel manipuliert!

1. Welcher **Schätzwert**  $T(x_1, \dots, x_n)$  für den Parameter (Erwartungswert)  $\mu$  kann aus der Stichprobe  $S=(x_1, \dots, x_n)$  abgeleitet werden?  
(*Punkt-Schätzung*)
2. Welcher **Schätzwert für ein Intervall**  $[a, b] = [Cl_L(x_1, \dots, x_n), Cl_R(x_1, \dots, x_n)]$ , das den unbekanntem wahren Parameter (Erwartungswert)  $\mu$  mit vorgegebener Sicherheit enthält, kann aus der Stichprobe  $S=(x_1, \dots, x_n)$  abgeleitet werden?  
(*Vertrauensbereich- oder Konfidenz-Intervall-Schätzung*)
3. Wie kann aufgrund der Stichprobe  $S=(x_1, \dots, x_n)$  eine begründete Entscheidung gegeben werden, ob die **Null-Hypothese** ' $\mu = 3.5$ ' angenommen oder abgelehnt werden soll? Wie groß sind die Fehler 1. Art  $\alpha$  (Annahme der Hypothese, obwohl sie falsch ist) und 2. Art  $\beta$  (Ablehnung der Hypothese, obwohl sie wahr ist)?  
(*Hypothesen-Test*)

## 13.5 Treffen von Entscheidungen anhand einer Entscheidungsregel

Sie treffen **nach** Durchführen eines Hypothesen-Tests eine Entscheidung über die Annahme oder Ablehnung der Null-Hypothese  $H$ . Ihre Entscheidung ist, abhängig vom gewählten statistischen Verfahren, mit einer gewissen Wahrscheinlichkeit  $(1-(\alpha+\beta))$  "richtig" und mit einer gewissen Wahrscheinlichkeit  $(\alpha+\beta)$  "falsch". Sie können bei den meisten Verfahren sogar nur eine Aussage darüber treffen, mit welcher Wahrscheinlichkeit  $\alpha$  (Fehler 1. Art) Sie die Hypothese irrtümlicherweise verworfen haben<sup>3</sup>.

<sup>3</sup> In der Mathematik sind die Anforderungen weitaus anspruchsvoller: Sie müssen eine Behauptung allgemeingültig beweisen, eine Anhäufung von Datenmaterial gilt nicht als Beweis. So ist z.B. die Behauptung, daß 24 durch alle Zahlen teilbar ist, mit den Zahlen 1,2,3,4,6,12

Die Entscheidungsregeln für statistische Verfahren zum Hypothesentest haben sämtlich folgende Form:

Falls der anhand der Stichprobe S realisierte Wert $t$ der Testgröße $T$	größer ist als ein von Ihnen vorgegebener kritischer Wert $c$	wird die Null-Hypothese $H$ von Ihnen	abgelehnt.
...	kleiner ist ...	...	nicht abgelehnt.

oder prägnanter formuliert:

**Falls  $t > c$ , dann: Ablehnen**

**Falls  $t \leq c$ , dann: Annehmen**

Beim Hypothesentest gibt es ein Dilemma besonderer Art; denn es können zwei verschiedene Typen von falschen Entscheidungen auftreten. Die folgende Tabelle zeigt die möglichen Kombinationen von **Wahrheit** und **Entscheidung**:

<b>Wahrheit/ Ihre Entscheidung</b>	<b>Hypothese ist wahr.</b>	<b>Hypothese ist falsch.</b>
<b>Hypothese wird angenommen.</b>	Richtige Entscheidung	Falsche Entscheidung Fehler 2. Art $\beta$
<b>Hypothese wird abgelehnt.</b>	Falsche Entscheidung Fehler 1. Art $\alpha$	Richtige Entscheidung

Es ist unmöglich, ein statistisches Verfahren zu konstruieren, mit dem **beide** Fehlerarten **gleichzeitig** minimiert werden können. Es ist allerdings häufig möglich, bei **vorgegebenem** Fehler 1. Art ein Verfahren mit minimalem Fehler 2. Art zu konstruieren (z.B. *Maximum Likelihood* Verfahren).

Ein **Hypothesen-Test** basiert zusammengefaßt auf einer **Null-Hypothese**  $H$ , einer **Testgröße**  $T$  zum Überprüfen der Null-Hypothese und einem **kritischem Wert**  $c$ , der den Annahme- bzw. Ablehnungsbereich für die Null-Hypothese trennt und damit die Entscheidungsregel zur Annahme bzw. Ablehnung der Null-Hypothese festlegt. Jedem kritischem Wert  $c$  ist eindeutig eine Irrtumswahrscheinlichkeit 1. Art  $\alpha$  und entsprechend ein **Konfidenz-Niveau**  $(1-\alpha)$  zugeordnet.

---

zu belegen. Ein einziges Gegenbeispiel reicht allerdings aus, um eine Behauptung zu widerlegen, im Beispiel ist 24 z.B. nicht durch 5 teilbar.

Annahmereich	Ablehnungsbereich
$P(T < c) = 1 - \alpha$	$P(T > c) = \alpha$
Trennung zwischen den Bereichen	

Tabelle 13.1 : Annahme- und Ablehnungsbereich für einen Hypothesen-Test

Es besteht folgender Zusammenhang zwischen der Verteilung der Testgröße  $T$ , dem kritischen Wert  $c$  (Beginn des Ablehnungsbereiches) und der Irrtumswahrscheinlichkeit  $\alpha(c)$ :

$P(T(X_1, \dots, X_n) > c) = \alpha(c)$  ist monoton fallend im kritischen Wert  $c$ , d.h. je größer der durch  $c$  festgelegte Annahmereich für die Null-Hypothese  $H$  wird, desto größer wird das Konfidenz-Niveau  $(1 - \alpha)$  und desto kleiner wird die Irrtumswahrscheinlichkeit  $\alpha$  (Wahrscheinlichkeit, die Hypothese fälschlicherweise abzulehnen)<sup>4</sup>. Umgekehrt gilt: Je größer  $\alpha$  gewählt wird, desto größer wird der Ablehnungsbereich und desto größer wird die Irrtumswahrscheinlichkeit dafür, die Hypothese fälschlicherweise abzulehnen.

## 13.6 Entscheidungsregel

SPSS setzt **automatisch** den beobachteten (in der Stichprobe realisierten) Wert  $t$  der Teststatistik  $T$  als kritischen Wert  $c$  ein und berechnet die zugehörige Irrtumswahrscheinlichkeit  $\mathbf{p}$ .

Sie brauchen nun nur die von Ihnen gewünschte oder von anderen Personen vorgegebene Irrtumswahrscheinlichkeit  $\alpha$  (z.B. 0.01) mit der von SPSS berechneten Irrtumswahrscheinlichkeit  $\mathbf{p}$  zu vergleichen und entscheiden nun folgendermaßen:

- Ist  $\mathbf{p}$  (von SPSS berechnet)  $< \alpha$  (von Ihnen gewünscht), sollten Sie die Null-Hypothese ablehnen.
- Ist  $\mathbf{p}$  (von SPSS berechnet)  $> \alpha$  (von Ihnen gewünscht), sollten Sie die Null-Hypothese annehmen, oder sich die Frage stellen, ob Sie auch eine größere Irrtumswahrscheinlichkeit  $\alpha$  akzeptieren wollen, um die Null-Hypothese ablehnen zu können.

Die Vorgehensweise des "normalen" Statistikers ist übrigens genau umgekehrt, denn bei Vorgabe von  $\alpha=0.01$  o.ä. berechnet er hieraus den kritischen Wert  $c$ .

Die von SPSS vorgegebene Wahrscheinlichkeit  $\mathbf{p}$  ist zu interpretieren als die **minimale** Irrtumswahrscheinlichkeit, bei der die Null-Hypothese  $H$  noch abgelehnt werden kann. SPSS kann nicht wissen, welche Irrtumswahrscheinlichkeit

<sup>4</sup> Allerdings wird die Wahrscheinlichkeit für einen Fehler 2. Art immer größer, nämlich die Hypothese  $H$  anzunehmen, obwohl sie falsch ist.

$\alpha$  Sie ansetzen möchten und berechnet deshalb die **minimal** zulässige Irrtumswahrscheinlichkeit, die zur Ablehnung der Hypothese führen kann.

## 13.7 Übungen

1. Was halten Sie davon, den (unbekannten) Erwartungswert im obigen Beispiel des 100-fachen Würfelwurfes durch folgende Punktschätzer  $T(X_1, \dots, X_n)$  zu schätzen:

- $T_1$ : Schätzwert ist Ergebnis des 1. Würfelwurfes
- $T_2$ : Schätzwert ist Mittelwert von 1. und letztem Würfelwurf
- $T_3$ : Schätzwert ist Median aller Würfelwürfe
- $T_4$ : Schätzwert ist 3.5, unabhängig davon, was gewürfelt wurde
- Ihr Schätzwert ?

### Hinweise:

Die Aufgabe eines Statistikers besteht u.a. darin, möglichst effiziente Verfahren zu entwickeln, die bei „geringer“ Stichprobenanzahl möglichst „optimale“ Ergebnisse liefern. Als Anwender brauchen Sie sich nur ein „passendes“ Verfahren aussuchen und sich aufgrund Ihres Datenmaterials und eines vernünftigen Signifikanzniveaus  $\alpha$  die Antwort (Annahme/Ablehnung) von **SPSS** berechnen lassen.

2. [Ergänzung]

Kriterien für einen "guten" Punktschätzer  $T(X_1, \dots, X_n)$  sind z.B. Erwartungstreue sowie Konsistenz. Interpretieren Sie diese Kriterien.

### Hinweise:

- $E(T) = \mu$
- $\text{Var}(T)$  konvergiert gegen 0 bei wachsendem Stichprobenumfang

3. [Zur Diskussion]

Nennen Sie "Alltagssituationen", in denen Sie oder andere Punkt- oder Bereichsschätzungen vornehmen und entsprechende Entscheidungen treffen.

### Hinweise:

Verbrauch an Lebensmitteln am Wochenende, Dimensionierung von Parkhäusern, Planung von Zugkapazitäten, ...

4. [Zur Diskussion]

Erläutern Sie mit eigenen Worten, welche Probleme sich bei einem Hypothesentest ergeben.

5. [Zur Diskussion]

Wie würden Sie die Irrtumswahrscheinlichkeit  $\alpha$  festlegen

- für einen genetischen Test ("genetischer Fingerabdruck"), der in einem Vergewaltigungs- und Mordprozeß zur Urteilsfindung herangezogen werden soll,
- für eine Marketing-Untersuchung,
- für den Nachweis der Wirksamkeit eines Medikamentes als Befürworter/Gegner des Medikamentes?

## 6. [Zur Diskussion]

Interpretieren Sie folgende statistische Grundweisheit für Konfidenz-Intervalle:

"Sichere Aussagen sind unscharf, scharfe Aussagen sind unsicher."

**Hinweise:**

Welcher Zusammenhang besteht zwischen Irrtums-Wahrscheinlichkeit und Länge von des Konfidenz-Intervalls?

## 14 Berechnen eines Vertrauensbereiches (Konfidenz-Intervalls)

In diesem Kapitel wird für eine metrische Variable ein Vertrauensbereich für den unbekanntem Erwartungswert berechnet und das gewonnene Ergebnis interpretiert. Die ausführliche Behandlung des statistischen Hintergrundes soll einen Brückenschlag zwischen "Alltagswissen" und mathematischer Statistik herstellen.

Mit welcher Wahrscheinlichkeit kann ich damit rechnen, daß das Packungsgewicht von Pralinen in einem bestimmten Intervall (z.B. 500 g +/- 5%) liegt, wenn ich bei einer Stichprobe 30 Packungen kontrolliere?

### 14.1 Interpretieren von Vertrauensbereichen

Sie können für eine metrische Variable einen **Vertrauensbereich für den Erwartungswert** berechnen. Der Erwartungswert ist eine (Ihnen unbekannt) Kenngröße der Grundgesamtheit, das arithmetische Mittel eine (Ihnen bekannte) Kenngröße der Stichprobe.

Ein **Vertrauensbereich (Konfidenz-Intervall)** enthält einen unbekanntem Parameter  $\mu$ , hier den Erwartungswert der Verteilung, mit einer **Sicherheit (Konfidenz-Niveau)** von z.B. 95% und entsprechend einer **Irrtumswahrscheinlichkeit**  $\alpha$  von 5%.

Das **Konfidenz-Niveau** ( $1-\alpha$ ), hier  $\alpha=5\%$ , ist folgendermaßen zu interpretieren:

Falls Sie das ausgewählte Verfahren 100-mal durchführen würden - was Sie aber aufgrund von Geld- und Zeitmangel nicht tun - erhalten Sie im Mittel 95-mal einen **Vertrauensbereich**, der den unbekanntem Parameter tatsächlich enthält, allerdings auch 5-mal einen Vertrauensbereich, der ihn nicht enthält. Da Sie aber nur eine und nicht 100 Untersuchungen durchführen, kann Ihre aktuelle Untersuchung also zu den 5 von 100 Untersuchungen gehören, bei denen das Verfahren einen "falschen" Vertrauensbereich liefert, also einen Vertrauensbereich, der den wahren Parameter  $\mu$  **nicht** enthält.

Bei einer **Schätzung** aufgrund einer Stichprobe bleibt also immer ein **Risiko**, das Sie nur mit einer **Gesamterhebung** (Stichprobe = Grundgesamtheit) ausschließen können.

Es ist daher falsch zu behaupten, daß das berechnete Konfidenz-Intervall den wahren Parameter enthält. Das berechnete Konfidenz-Intervall enthält den wahren Parameter vielmehr mit einer berechneten Wahrscheinlichkeit.

## 14.2 Berechnen eines Vertrauensbereichs

Im folgenden Beispiel berechnen Sie für die Variable **gewicht** aus der SPSS Datendatei „hypertonie-01.sav“ ein 95%-Vertrauensbereich (**Konfidenz-Intervall**, *confidence interval*, CI).

Wählen Sie *Analysieren > Deskriptive Statistiken > Explorative Datenanalyse*. und wählen Sie dort die Variable „gewicht“.

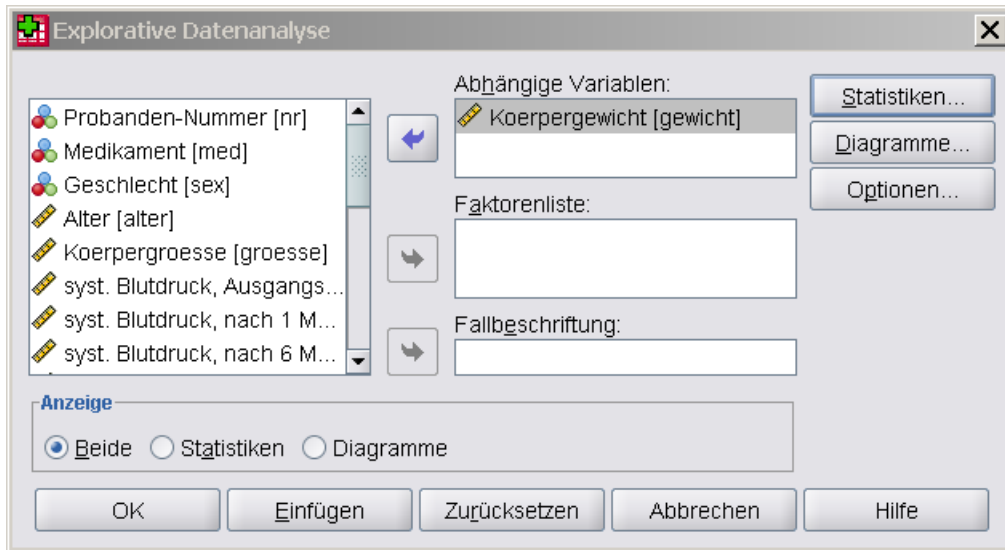


Abbildung 14.1: Explorative Datenanalyse

Geben Sie über die Schaltfläche „Statistiken“ das gewünschte Konfidenz-Niveau für den Mittelwert in Prozent ein, hier 95%. Die Irrtumswahrscheinlichkeit berechnet sich damit zu:  $\alpha = (1 - 95\%) = 5\% = 0.05$ .

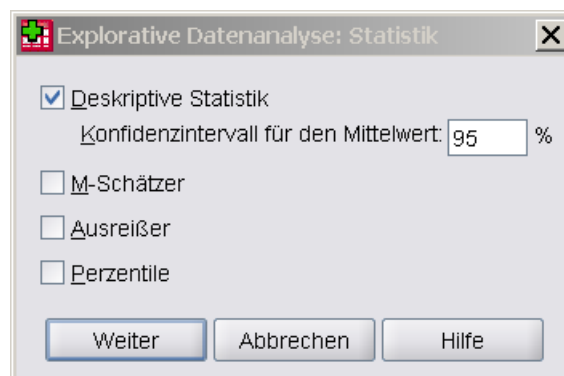
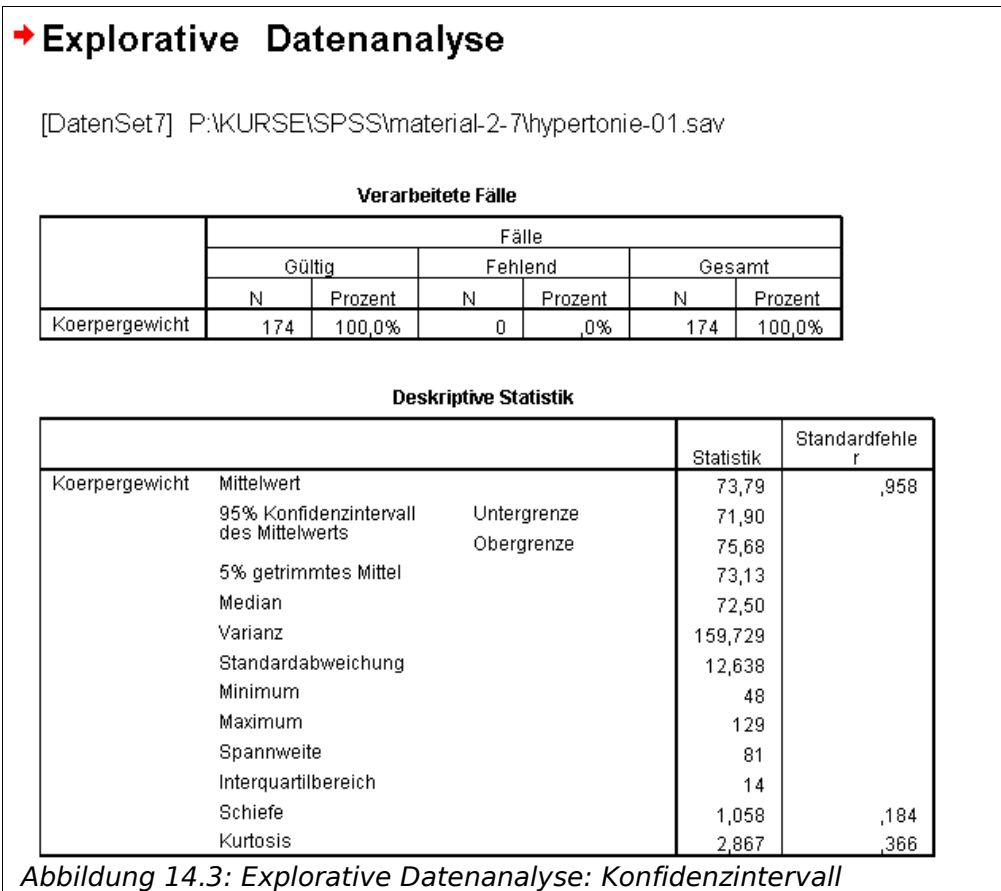


Abbildung 14.2: Konfidenzintervall (Signifikanz-Niveau)

Klicken Sie auf die Schaltfläche „OK“. Die Ergebnisse werden im „SPSS Viewer“-Fenster angezeigt.



Der 95%-Vertrauensbereich für den unbekanntem Erwartungswert lautet demnach:  $CI = [71.90, 75.68]$  oder anders formuliert:  $71.90 < \mu < 75.68$ . Dieser Vertrauensbereich enthält den unbekanntem Erwartungswert  $\mu$  mit einer Irrtumswahrscheinlichkeit von 5% - falls die Stichprobe die Grundgesamtheit angemessen repräsentiert.

Ein Blick auf das zugehörige Histogramm läßt diese Berechnung plausibel erscheinen:

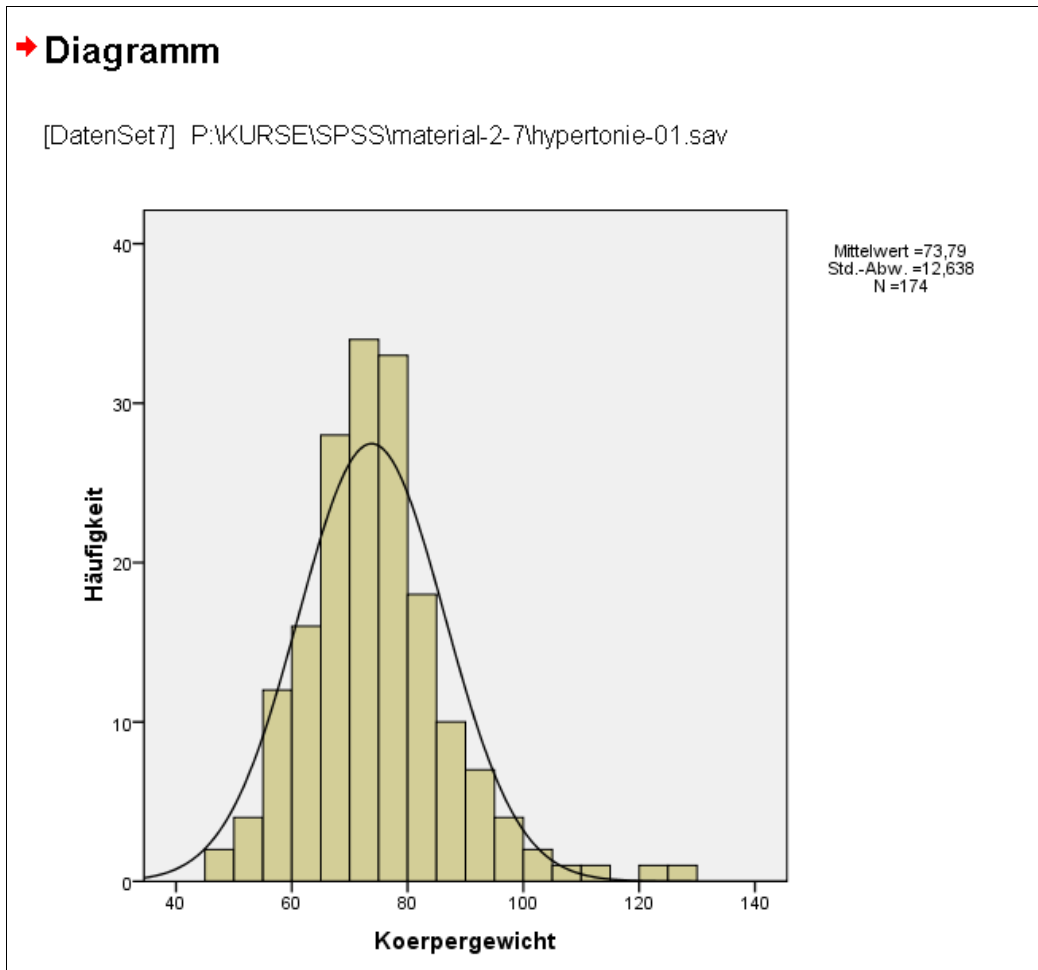


Abbildung 14.4: Histogramm

## 14.3 Ableiten der Formel für den Vertrauensbereich

Es sei  $\bar{X}$  der arithmetische Mittelwert der  $n$  Beobachtungen in der Stichprobe. Dann gilt unter der Voraussetzung, daß alle  $X_i$  identisch und normalverteilt<sup>5</sup> sind (in Anlehnung zum zentralen Grenzwertsatz) für beliebige Werte  $c$ :

$$(1) \quad n^{1/2}(\bar{X} - \mu) / \hat{\sigma}$$

ist verteilt nach der tabellierten Verteilungsfunktion  $t(x; n-1)$  der Student'schen **t-Verteilung** mit  $(n-1)$  Freiheitsgraden, anders formuliert:

$$(2) \quad P(-c \leq n^{1/2} (\bar{X} - \mu) / \hat{\sigma} \leq c) = t(c, n-1) - t(-c; n-1)$$

<sup>5</sup> Hier macht sich eine angenehme Eigenschaft von unabhängigen normal-verteilten Zufallsvariablen bemerkbar, daß deren Summe selbst wieder normal-verteilt ist.

Die Formel (2) läßt sich nach dem Erwartungswert  $\mu$  umstellen:

$$(3) P( \bar{X} - c \hat{\sigma} n^{-1/2} \leq \mu \leq \bar{X} + c \hat{\sigma} n^{-1/2} ) = t(c;n-1) - t(-c;n-1)$$

Sei nun im folgenden die Irrtumswahrscheinlichkeit  $\alpha$  für das Verfahren folgendermaßen (von Ihnen kraft eigener Willkür oder bestimmter Vorgaben) festgelegt:

$$(4) \alpha = 0.05$$

Damit der Erwartungswert  $\mu$  mit einer Wahrscheinlichkeit von  $(1-\alpha)$  im Konfidenz-Intervall liegt, muß die rechte Seite den Wert  $(1-\alpha)$  ergeben:

$$(5) t(c;n-1) - t(-c;n-1) = 1 - \alpha = 0.95$$

Nach einigen Umformungen läßt sich hieraus der kritische Wert  $c$  bestimmen:

$$(6) \quad \begin{aligned} t(c;n-1) - t(-c;n-1) &= t(c;n-1) - (1 - t(c;n-1)) \\ 2 t(c;n-1) - 1 &= 1 - \alpha \\ t(c;n-1) &= 1 - \alpha/2 \\ c &= t^{-1}(1-\alpha/2;n-1) \\ &= (1-\alpha/2)\text{-Quantil der t-Verteilung mit } (n-1) \text{ df (Freiheitsgraden)} \end{aligned}$$

Diese kritischen Werte  $c$  mußten Sie früher in Tabellen nachschlagen, heute berechnet SPSS diesen Wert. Die "zufällige" Länge  $L$  des Konfidenz-Intervalls beträgt:

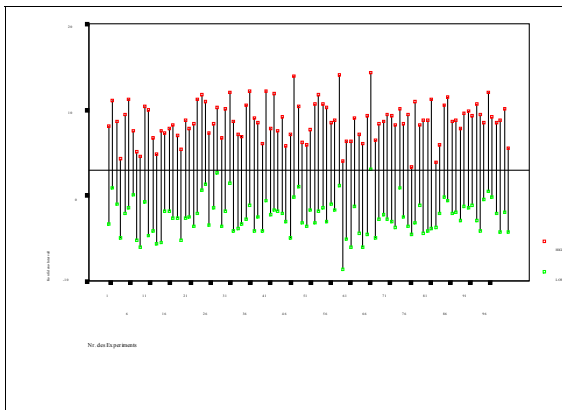
$$(7) L = 2 c s / n^{1/2}$$

Hierbei ist der kritische Wert  $c$  (durch Vorgabe von  $\alpha$ ) bereits vor Beginn der Untersuchung vorgegeben,  $n$  ist der Stichprobenumfang und  $s$  (emp. Standardabweichung) ist ein "zufälliger" Wert, der erst nach Durchführung der Untersuchung bekannt ist. Insgesamt handelt es sich also bei  $L$  um eine **vor** Durchführung der Untersuchung unbekannte Länge.

Beachten Sie, daß Sie selbst das Konfidenz-Niveau und damit den kritischen Wert  $c$  festlegen. Sie "erkaufen" sich eine größere Sicherheit Ihrer Schätzung (kleines  $\alpha$ ) durch ein längeres Konfidenzintervall  $L$  und umgekehrt erhalten Sie bei kleinerer Sicherheit ein kürzeres Konfidenzintervall  $L$ .

Sie können das Konfidenz-Intervall natürlich dadurch verkleinern, indem Sie die Stichprobe vergrößern – und damit die Untersuchung verteuern. Wegen des Faktors  $n^{1/2}$  im Nenner verkleinert die Länge des Konfidenzintervalls  $L$  mit wachsendem  $n$  – aber nur mit der Wurzel, im Grenzfall wird aus dem Intervall ein Punkt.

Die folgende Abbildung zeigt für 100 Durchführungen<sup>6</sup> eines Experiments die jeweils ermittelten 99%-Konfidenz-Intervalle (Experiment: Ziehen einer Stichprobe mit 25 Beobachtungen aus einer Grundgesamtheit mit Normalverteilung  $\mu=3$  und  $\sigma=10$ ).



An der Referenzlinie  $y=3$  ist erkennbar, daß der wahre Parameter  $\mu=3$  bei 99 von 100 Durchführungen im 99%-Konfidenz-Intervall enthalten ist und bei genau einer Durchführung nicht.

Gut erkennbar ist auch die jeweils unterschiedliche Länge und Lage der Konfidenz-Intervalle.

Das Diagramm ist realisiert als ein Hoch-Tief-Diagramm)

## 14.4 Übungen

1. Berechnen Sie einen 99%-Vertrauensbereich für den Erwartungswert von **bz0** (Blutzucker) aus der SPSS Datendatei „hypertonie-01.sav“. Nennen Sie eine sinnvolle Grundgesamtheit. Halten Sie einen Rückschluß auf die Gesamtbevölkerung für sinnvoll?
2. [zusätzlich]  
Führen Sie nun die Berechnung aus Übung 1 analog für die Irrtumswahrscheinlichkeiten  $\alpha=1\%$ ,  $2\%$ ,  $3\%$ ,  $4\%$ ,  $5\%$ ,  $10\%$  und  $20\%$  durch und vergleichen Sie die Länge und Lage der Konfidenzintervalle tabellarisch und grafisch. Erklären Sie, weshalb „große“ Konfidenz-Intervalle „sicher“ und „kleine“ entsprechend „unsicher“ sind.

<sup>6</sup> Die 100 Ziehungen sind über den SPSS Zufallszahlengenerator mit  $RV.NORMAL(3.0,10.0)$  realisiert.

## 15 Testen der Unabhängigkeit

In diesem Kapitel wird der **Chi-Quadrat-Test** zum Überprüfen der Unabhängigkeit von 2 kategorial-skalierten Variablen X und Y behandelt. Grundlage für den Test ist eine  $r \times s$  **Kontingenztafel** (r Kategorien von X und s Kategorien von Y), in der neben den berechneten und zusätzlich die „erwarteten“ Häufigkeiten eingetragen werden.

Sterben Raucher häufiger an Krebs? Werden farbige Kriminelle härter bestraft als weiße? Sind Intelligenz-Quotient und Geschlecht voneinander unabhängig ...

Solche und ähnliche sinnige und unsinnige Fragestellungen werden im ersten Ansatz mit Unabhängigkeits-Test behandelt.

### 15.1 Berechnen der Chi-Quadrat-Testgröße

Im folgenden Beispiel untersuchen Sie aggregiertes Datenmaterial über die Religionszugehörigkeit von Braut und Bräutigam bei Eheschließungen in Köln im Jahr 1970 aus der SPSS Arbeitsdatei „heirat.sav“.

Es soll die **Null-Hypothese H** überprüft werden, daß die Religionszugehörigkeit der Braut (X=braut) und die Religionszugehörigkeit des Bräutigams (Y=braeutigam) keinen Einfluß auf das Zustandekommen einer Eheschließung hat.

H: X=**braut** und Y=**braeutigam** sind unabhängig.

Gewichten Sie zunächst die Beobachtungen mit der Variablen **anzahl1** (Anzahl der Hochzeiten pro Kombination der Religionszugehörigkeit von Braut und Bräutigam). Wählen Sie hierzu „Daten > Fälle gewichten“

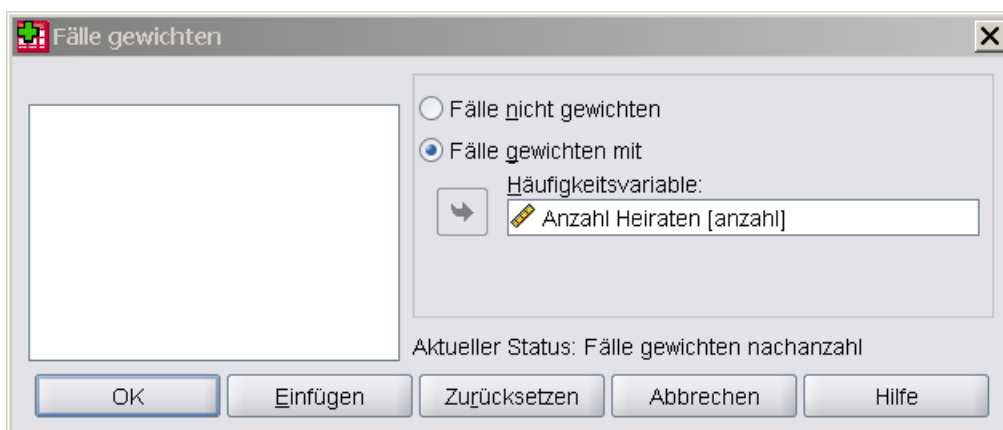


Abbildung 15.1: Daten > Fälle gewichten

Wählen Sie dann „*Analysieren > Deskriptive Statistiken > Kreuztabellen*“. Wählen Sie die Variablen für die Kreuztabulation aus, hier: **braut** und **brautigam**.

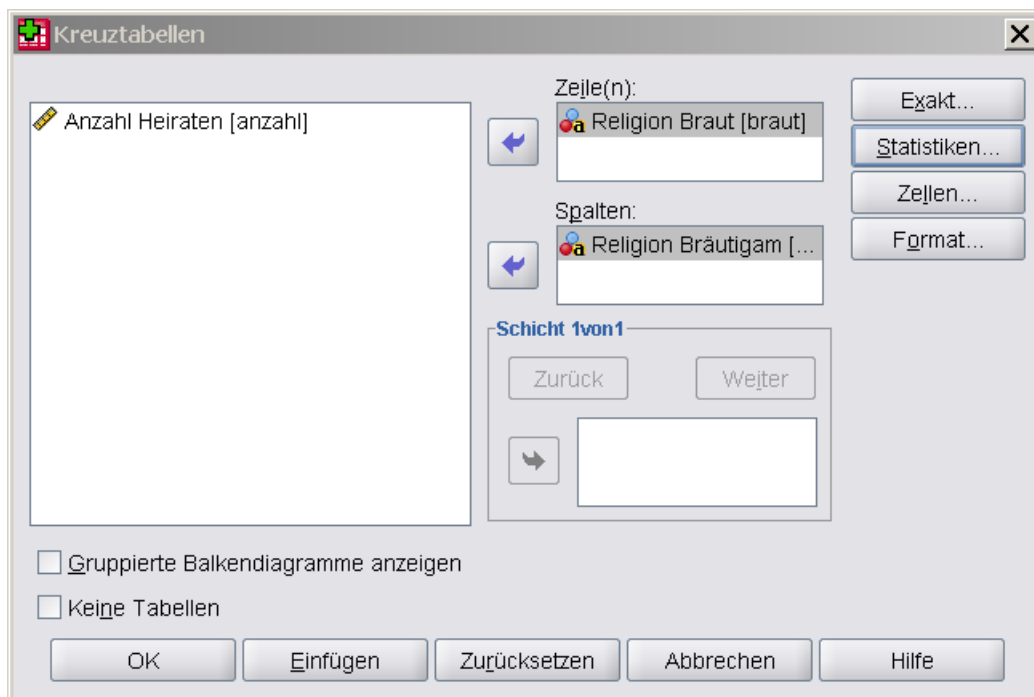
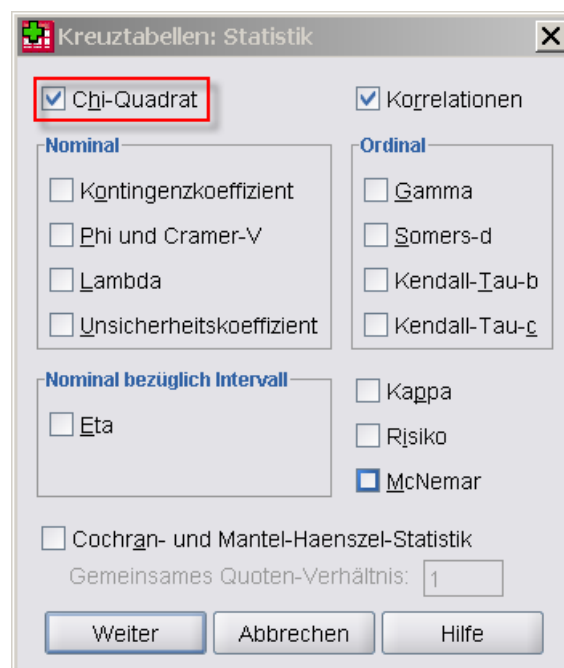


Abbildung 15.2: Kreuztabellen

Fordern Sie über die Schaltfläche „Statistiken“ einen Chi-Quadrat-Test“ an:



Fordern Sie über die Schaltfläche „Zellen“ zusätzlich die erwarteten Werte an:



Abbildung 15.3: Inhalte der Kreuztabelle

Klicken Sie auf die Schaltfläche „OK“. Die Ergebnisse werden im „SPSS Viewer“-Fenster angezeigt.

**Religion Braut \* Religion Bräutigam Kreuztabelle**

		Religion Bräutigam				Gesamt	
		ev	ohneB	rk	sonstB		
Religion Braut	ev	Anzahl	784	122	1100	40	2046
		Erwartete Anzahl	608,4	105,1	1244,2	88,4	2046,0
		% von Religion Braut	38,3%	6,0%	53,8%	2,0%	100,0%
		% von Religion Bräutigam	38,5%	34,7%	26,4%	13,5%	29,9%
		% der Gesamtzahl	11,4%	1,8%	16,0%	,6%	29,9%
	Residuen	175,6	16,9	-144,2	-48,4		
ohneB		Anzahl	47	78	56	14	195
		Erwartete Anzahl	58,0	10,0	118,6	8,4	195,0
		% von Religion Braut	24,1%	40,0%	28,7%	7,2%	100,0%
		% von Religion Bräutigam	2,3%	22,2%	1,3%	4,7%	2,8%
		% der Gesamtzahl	,7%	1,1%	,8%	,2%	2,8%
	Residuen	-11,0	68,0	-62,6	5,6		
rk		Anzahl	1193	152	2987	90	4422
		Erwartete Anzahl	1314,9	227,1	2689,1	191,0	4422,0
		% von Religion Braut	27,0%	3,4%	67,5%	2,0%	100,0%
		% von Religion Bräutigam	58,5%	43,2%	71,7%	30,4%	64,5%
		% der Gesamtzahl	17,4%	2,2%	43,6%	1,3%	64,5%
	Residuen	-121,9	-75,1	297,9	-101,0		
sonstB		Anzahl	14	0	25	152	191
		Erwartete Anzahl	56,8	9,8	116,1	8,2	191,0
		% von Religion Braut	7,3%	,0%	13,1%	79,6%	100,0%
		% von Religion Bräutigam	,7%	,0%	,6%	51,4%	2,8%
		% der Gesamtzahl	,2%	,0%	,4%	2,2%	2,8%
	Residuen	-42,8	-9,8	-91,1	143,8		
Gesamt		Anzahl	2038	352	4168	296	6854
		Erwartete Anzahl	2038,0	352,0	4168,0	296,0	6854,0
		% von Religion Braut	29,7%	5,1%	60,8%	4,3%	100,0%
		% von Religion Bräutigam	100,0%	100,0%	100,0%	100,0%	100,0%
		% der Gesamtzahl	29,7%	5,1%	60,8%	4,3%	100,0%

Abbildung 15.4: Kreuztabelle

Ein erster unschuldiger Vergleich der tatsächlich beobachteten mit den erwarteten Häufigkeiten bei Unabhängigkeit zeigt, daß diese stark voneinander abweichen. Diese Vermutung läßt sich durch den Chi-Quadrat-Test nun auch statistisch absichern:

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	3166,034 <sup>a</sup>	9	,000
Likelihood-Quotient	1185,537	9	,000
Anzahl der gültigen Fälle	6854		

a. 0 Zellen (.0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 8,08.

Der Chi-Quadrat Test liefert einen verschwindend kleinen und deshalb auf Null abgerundeten Wert für die Irrtumswahrscheinlichkeit  $p < 0.00001$ . Es gibt also eine extrem kleine Irrtumswahrscheinlichkeit  $p$  für eine irrtümliche Ablehnung der Hypothese. Sie sollten die Null-Hypothese  $H$  deshalb ablehnen und damit verwerfen. Die Religionszugehörigkeit spielt demnach sehr wohl eine Rolle bei der Auswahl des Ehepartners.

## 15.2 Ableiten der Formel für die Chi-Quadrat Testgröße

Die Null-Hypothese der Unabhängigkeit zwischen zwei diskreter Zufallsvariablen  $X$  (Wertebereich  $W_1$ ) und  $Y$  (Wertebereich  $W_2$ ) wird folgendermaßen definiert:

Alle gemeinsamen Wahrscheinlichkeiten sind gleich dem Produkt der Einzelwahrscheinlichkeiten; d.h.

$$(1) H: p_{ij} = P(X=i, Y=j) = v_i u_j = P(X=i) P(Y=j)$$

für alle möglichen Kombinationen von  $i$  aus  $W_1$  und  $j$  aus  $W_2$

Beim Testen der Unabhängigkeit von zwei Variablen  $X$  und  $Y$  auf Grundlage einer Stichprobe mit  $n$  Beobachtungen werden zunächst die zugehörigen empirischen Werte berechnet, im folgenden gekennzeichnet durch  $\sim$ :

$$\begin{aligned}
 (2) \quad & p_{ij} \sim h_b = N_{ij} / n && \text{(relative Häufigkeit)} \\
 & v_i \sim a_i = N_{i.} / n \\
 & u_j \sim b_j = N_{.j} / n \\
 & h_e = a_i \cdot b_j
 \end{aligned}$$

Als Testgröße  $T$  (**Chi-Quadrat-Statistik**) wird nun die Summe der quadrierten Abweichungen  $(h_b - h_e)^2$  zwischen den erwarteten Häufigkeiten  $h_e$  und den beobachteten Häufigkeiten  $h_b$  verwendet, wobei jeder Summand geeignet normiert wird:

$$(3) T = \text{Summe über alle: } (h_b - h_e)^2 / h_e$$

Sofern diese Testgröße „große Werte“ annimmt, kann von zu großen Abweichungen ausgegangen werden; d.h. die Hypothese (1) trifft wahrscheinlich nicht zu.

## 15.3 Übungen

1. Untersuchen Sie für das Datenmaterial aus der Datendatei „*strafe.sav*“ (Untersuchung über die Art der Verurteilung von weißen und schwarzen Mördern in den USA) die Variablen „*strafe*“ (Urteil bei Mord (Zuchthaus oder Todesstrafe) und *hautfarbe* (Hautfarbe des Verurteilten) auf Unabhängigkeit. Die Gewichtung (*Daten > Fälle gewichten ...*) erfolgt über die Variable *anzahl*. Messen Sie dieser Untersuchung politische Bedeutung zu?
2. Überlegen, wie Sie obige Ergebnis für die SPSS Arbeitsdatei „*heirat.sav*“ begründen könnten. Untersuchen Sie insbesondere, wo es auffallend zu „wenig“ und wo es auffallend zu „viele“ Eheschließungen gibt.

### **Hinweise:**

Könnte das Ergebnis z.B. auf indirekte Zusammenhänge wie geografische oder soziale Gruppierungen zurückzuführen sein, die ihrerseits bei der Wahl des Ehepartners eine Rolle spielen?

# 16 Berechnen von Korrelationskoeffizienten

In diesem Kapitel wird der (Pearson-) **Korrelationskoeffizient** behandelt, der ein Maß für die lineare Abhängigkeit zwischen zwei metrischen Variablen liefert.

Häufig besteht die Vermutung, daß zwischen zwei Variablen ein linearer Zusammenhang besteht: Wächst die eine, wächst die andere, bzw. wächst die eine, fällt die andere. Wie „stark“ ist dieser lineare Zusammenhang?

## 16.1 Festlegen eines Maßes für den linearen Zusammenhang

Die Korrelation  $\rho$  zweier metrischer Zufallsvariablen  $X$  und  $Y$  berechnet sich über Erwartungswert und Varianz:

$$\text{Cov}(X,Y) = E[XY] - E[X] E[Y]$$

$$\rho = \text{Cov}(x,y) / (\sigma_{xx} \sigma_{yy})^{1/2}$$

Beispiele für  $X$  und  $Y$  wären die Schulnoten in den Fächern Mathematik und Physik, bei denen ein großer linearer Zusammenhang zu erwarten wäre.

Beim Berechnen der empirischen Korrelation  $\hat{\rho}=r$  auf Grundlage einer Stichprobe werden, wie nicht anders zu erwarten, die entsprechenden empirischen Werte verwendet.

Für Beobachtungspunkte, die "ungefähr" auf einer steigenden Geraden liegen, ergibt der empirische Korrelationskoeffizient  $r$  einen Wert, der "ungefähr" bei 1 liegt, für solche auf einer fallenden Geraden einen Wert "ungefähr" bei (-1) und z.B. für stark streuende einen Wert bei 0. Variablen mit einem **positiven Korrelationskoeffizienten** heißen positiv korreliert, in diesem Fall wächst  $Y$  mit  $X$ , Variablen mit negativen Korrelationskoeffizienten heißen negativ korreliert, in diesem Fall fällt  $Y$  mit wachsendem  $X$ .

Abhängig vom Betrag des empirischen Korrelationskoeffizientens  $r$  sind folgende Aussagen üblich:

$r$	Bewertung	Formulierung
$0.0 <  r  \leq 0.2$	sehr gering	Es besteht ein sehr geringer linearer Zusammenhang zwischen den Variablen $X$ und $Y$ .
$0.2 <  r  \leq 0.5$	gering	... geringer ...
$0.5 <  r  \leq 0.7$	mittel	... mittelgroßer ...
$0.7 <  r  \leq 0.9$	hoch	... hoher ...
$0.9 <  r  \leq 1.0$	sehr hoch	... sehr hoher ...

## 16.2 Ermitteln des Korrelationskoeffizientens

Im folgenden Beispiel berechnen Sie für die Noten aus der SPSS Datendatei „schueler.sav“ empirische Korrelationskoeffizienten.

Tragen Sie zunächst **mathe** (Mathematik), **physik** (Physik), **deutsch** (Deutsch) und **latein** (Latein) in einem mehrfachen x-y-Streudiagramm (*Scatterplot Matrix*) gegeneinander auf. Wählen Sie „Grafiken -> Streudiagramm > Einfach“.

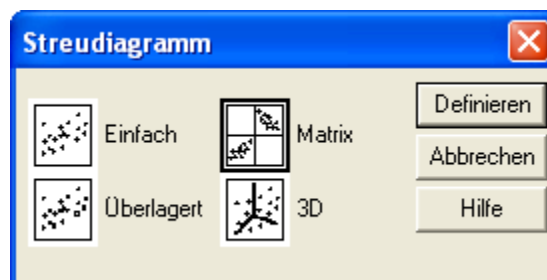


Abbildung 16.1

Wählen Sie alle Variablen aus, für die paarweise Streudiagramme erzeugt werden sollen, hier die Fächer Deutsch, Mathematik, Physik und Latein.

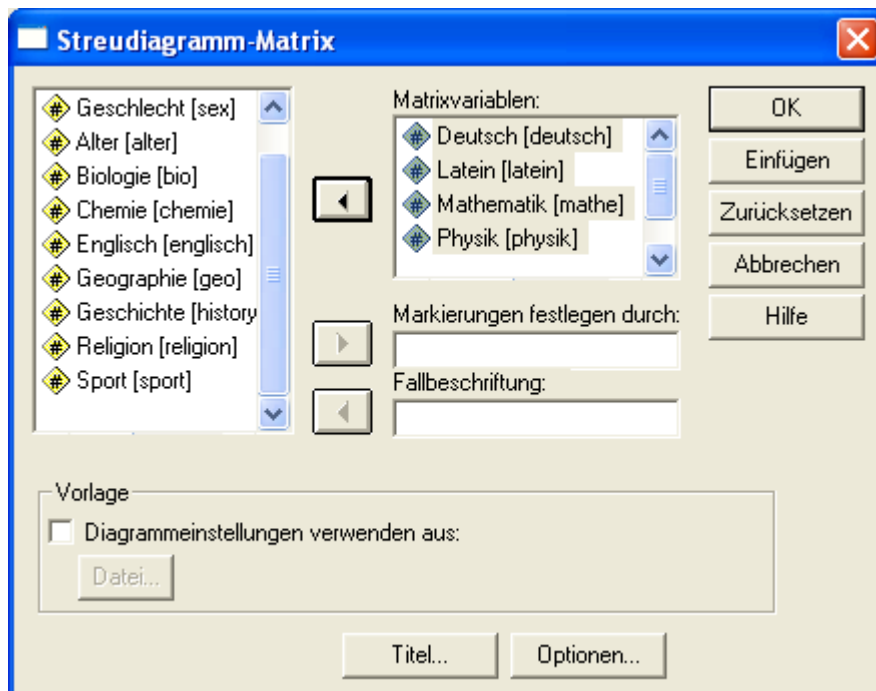


Abbildung 16.2

Klicken Sie auf „OK“. Die Ergebnisse werden im „Viewer“-Fenster angezeigt.

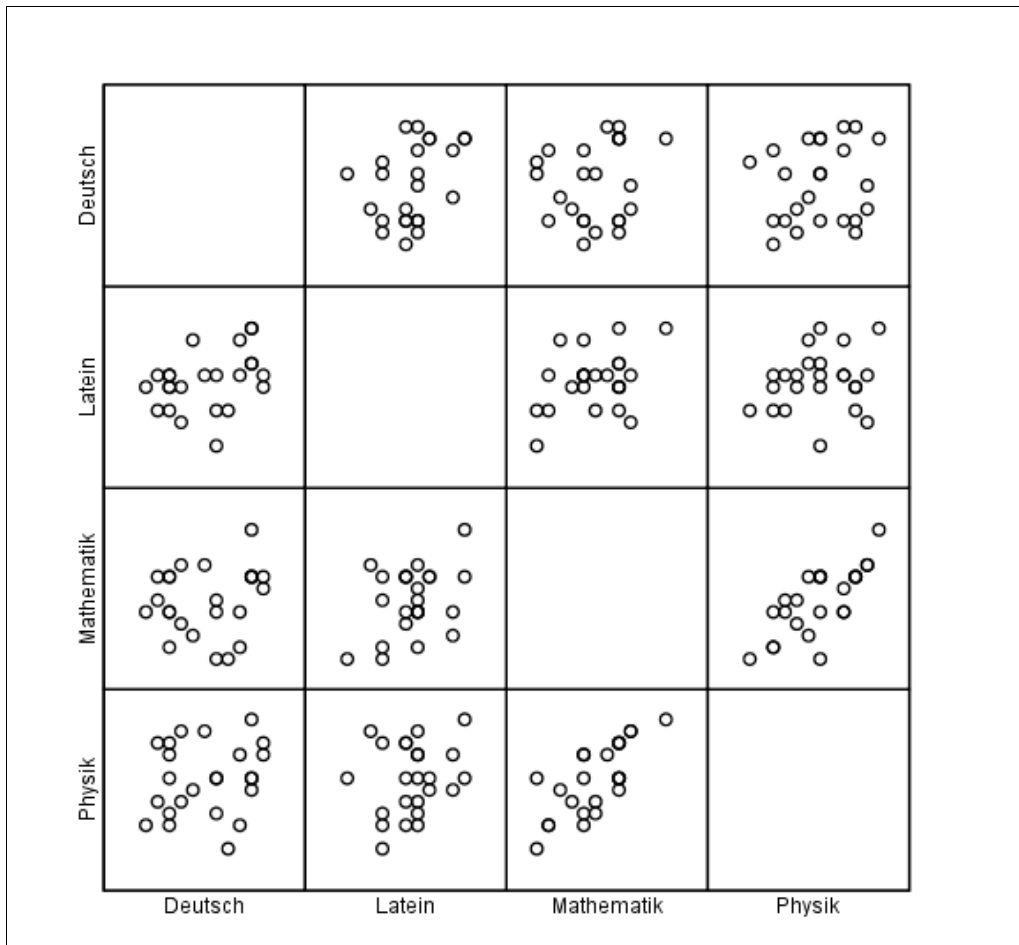


Abbildung 16.3

Die Streudiagramme in der Matrix legen nahe, daß nur für Mathematik und Physik eine hohe emp. Korrelation besteht. Prüfen Sie diese Vermutung im nächsten Schritt durch Berechnung der Korrelationskoeffizienten. Wählen Sie hierzu „Analysieren > Korrelation-> Bivariat“. Wählen Sie alle Variablen aus, für die der emp. Korrelationskoeffizient berechnet werden soll, hier die Fächer Deutsch, Mathematik, Physik und Latein.



Abbildung 16.4 Bivariate Korrelation

Klicken Sie auf die Schaltfläche „OK“. Die Ergebnisse werden im „SPSS Viewer“-Fenster angezeigt.

Die berechneten empirischen Korrelationskoeffizienten bestätigen das vermutete Ergebnis, daß nur zwischen Mathematik und Physik eine signifikant große Korrelation besteht.

## 16.3 Übungen

1. Nehmen Sie in die Untersuchung der Korrelation für die Datendatei „*schueler.sav*“ zusätzlich die Schulfächer Biologie und Chemie auf.
2. Der sozio-ökonomische Status (*socioeconomic status*, **SES**) einer Person werde auf einer Skala von 11 (niedrig) bis 77 (hoch) gemessen. SES ist dabei ein (nicht weiter definierter) Index für schulische und berufliche Qualifikation. Untersuchen Sie für die fiktiven (!) Daten aus der Arbeitsdatei *ses.sav*, inwieweit der SES von Vätern im Alter von 45 Jahren (*vater*) mit dem SES ihrer Söhne (*sohn*) korreliert, wobei der SES der Söhne ebenfalls im Alter von 45 Jahren ermittelt wird (also eine Generation später). Interpretieren Sie Ihr Ergebnis auch unter Zuhilfenahme eines Streudiagramms von *sohn* (y-Achse) und *vater* (x-Achse)

### Hinweise:

Unterscheiden Sie zwischen Familien mit niedrigem, mittlerem und hohem SES. Beachten Sie, daß SES nach oben und unten beschränkt ist.

3. (Diskussion)  
Wie würden Sie SES definieren?
4. (Diskussion)  
Für ordinal-skalierte Variablen wie in SES empfiehlt sich anstelle des **Pearson** der **Spearman Rang-Korrelationskoeffizient**, da er sich auf die rang-transformierten und nicht auf die ursprünglichen Werte bezieht. Führen Sie die Untersuchung aus der vorherigen Aufgabe erneut mit dem

Spearman Rang-Korrelationskoeffizienten  $r_s$  durch. Vergleichen Sie die Streudiagramme der ursprünglichen und der rang-transformierten Variablen und die gemessenen Korrelationskoeffizienten  $r$  und  $r_s$ .

## 17 Approximieren von x-y-Punkten durch Geraden (lineare Regression)

In diesem Kapitel wird die **lineare Regression** behandelt, bei der durch eine Menge von x-y-Beobachtungspunkten eine "möglichst optimale" Gerade gelegt werden soll.

Wie stark steigt der Cholestinspiegel bei erhöhter Fettaufnahme? Wie groß ist der Umsatzzuwachs, den eine Firma bei Verdoppelung der Werbeausgaben erwarten kann? Der Korrelationskoeffizient lag nahe bei Eins – wie sieht denn nun der lineare Zusammenhang genau aus?

### 17.1 Untersuchen eines möglichen linearen Zusammenhangs

Während Ihnen die Korrelation nur einen einzigen Wert zur Beschreibung einer linearen Abhängigkeit, nämlich den Korrelationskoeffizienten  $r$  liefert, dient die lineare Regression zur genaueren Modellierung eines vermuteten linearen Zusammenhangs.

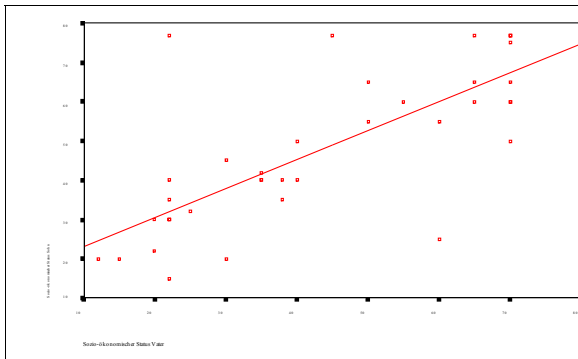
Bei einer Regression postulieren Sie allgemein ein Modell, bei dem eine abhängige Variable  $Y$  **funktional** von einer unabhängigen Variablen  $X$  abgeleitet wird, im Fall der **linearen Regression** wird einschränkend ein lineares Modell vermutet.

Sie untersuchen mit der linearen Regression die Null-Hypothese  $H_0$ , daß sich die Variablen  $Y$  und  $X$  in der Form  $Y = mX + b + Z$ , also in Form einer Geradengleichung, darstellen lassen. Dabei sind  $m$  und  $b$  feste, aber unbekannte Parameter und  $Z$  ist ein "zufälliger Fehler", z.B. ein physikalischer Meßfehler oder ein individueller Störeffekt.

$$H_0: Y = mX + b + Z$$

Die Regressions-Gerade  $g(x)$  erfüllt unter allen möglichen Geraden die Minimaleigenschaft, daß die Summe der vertikalen Abstandsquadrate zwischen den Beobachtungspunkten  $P_1=(x_1, y_1), \dots, P_n=(x_n, y_n)$  und der Geraden  $g(x)$  so klein wie möglich (**Gauß'sche Methode der kleinsten Quadrate**).

Der für  $X$  erwartete Wert von  $Y$  auf der Regressions-Geraden wird im folgenden als  $\hat{Y}$  ( $Y$  erwartet oder  $Y$  geschätzt) bezeichnet.



Die eingezeichnete Regressions-Gerade minimiert die Summe der vertikalen Abstandsquadrate.

Für jeden Wert  $x_i$  ist  $g(x_i)$  der für  $y_i$  erwartete Wert auf der Regressions-Geraden, d.h.  $\hat{y}_i = g(x_i)$ .

Bei der linearen Regression wird folgende Terminologie verwendet:

Bezeichnung	Bedeutung
$Y = mX + b + Z$	Modellgleichung, hier: lineares Modell (postuliert)
$g(x) = mx + b$	Gleichung der Regressionsgeraden (aufgrund der Modellannahme berechnet, optimale Anpassung nach der Gauß'schen Methode der kleinsten Quadrate)
$Y$	abhängige oder erklärte Variable
$y_i$	beobachtete Werte von Y in der Stichprobe
$\hat{y}_i$	geschätzte Werte (aufgrund der Modellannahme berechnet)
$X$	unabhängige oder erklärende Variable (Regressor)
$x_i$	beobachtete Werte von X in der Stichprobe
$b$	Schnittpunkt der Regressionsgeraden mit der horizontalen Achse (aufgrund der Modellannahme berechnet)
$m$	Steigung der Regressionsgeraden oder Koeffizient vor $x$ (aufgrund der Modellannahme berechnet)
$Z$	Residuum oder zufälliger Fehler (postuliert)
$z_i$	errechnete Werte für jede Beobachtung in der Stichprobe (berechnet als Differenz zwischen erwarteten Wert $\hat{y}_i$ und $y_i$ )

## 17.2 Durchführen einer linearen Regression

Im folgenden Beispiel untersuchen Sie in der SPSS Datendatei „buecher.sav“ die Variablen „anzahl“ (Anzahl verkaufter Bücher) und „jahr“ in Hinblick auf einen linearen Zusammenhang. Wählen Sie „Analysieren > Regression > Linear“.

Wählen Sie die abhängige Variable, hier **anzahl** und die unabhängigen Variablen, hier **jahr**, aus.

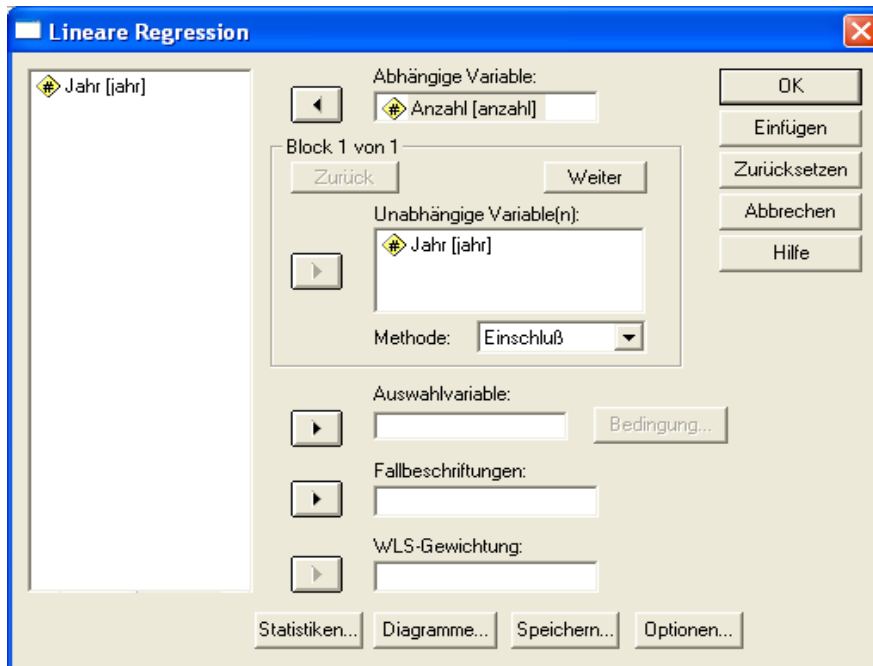


Abbildung 17.1

Fordern Sie über die Schaltfläche „Statistiken“ die Schätzwerte für die Gera-  
den-Gleichung an:

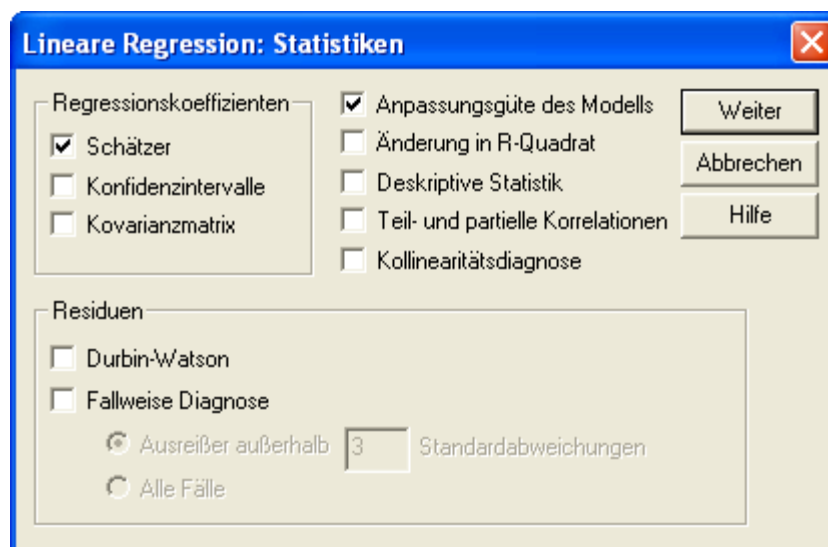


Abbildung 17.2

Forden Sie zusätzliche Variablen an für die erwarteten Wert  $\hat{y}_i$  und die Abwei-  
chungen (Residuen) zur Regressionsgeraden:

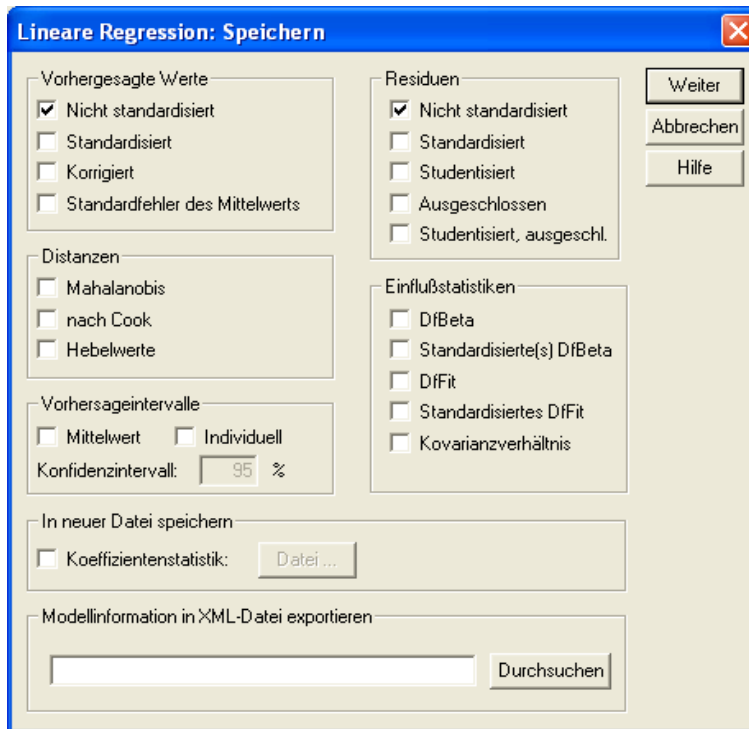


Abbildung 17.3

Klicken Sie auf „OK“. Die Ergebnisse werden im „Viewer“-Fenster angezeigt.

Variable B  
 JAHR 1583.6 = m  
 (Constant) -3081413.9 = b ->  $g(x) = 1593.6 x + (-3081413.9)$

**Koeffizienten<sup>a</sup>**

Modell	Nicht standardisierte Koeffizienten	Standardisierte Koeffizienten	T	Signifikanz
1 (Konstante)	-3081414	218574,437		
Jahr	1583,687	110,586	,938	14,321

a. Abhängige Variable: Anzahl

Die Gleichung der Regressions-Geraden  $g(x)$  lautet:

$$g(x) = 1593.6 x - 3081413.9$$

Der berechnete Wert der Testgröße  $R^2$  (Güte des Modells) liegt bei **.87987**, das lineare Modell liefert demnach eine hervorragende Anpassung (siehe auch unten).

## 17.3 Visualisieren der linearen Regression

Kontrollieren Sie das vorige Ergebnis, indem Sie die Regressionsgerade in einem Streudiagramm einzeichnen. Wählen Sie hierzu „Grafik > Interaktiv > Streudiagramm“.

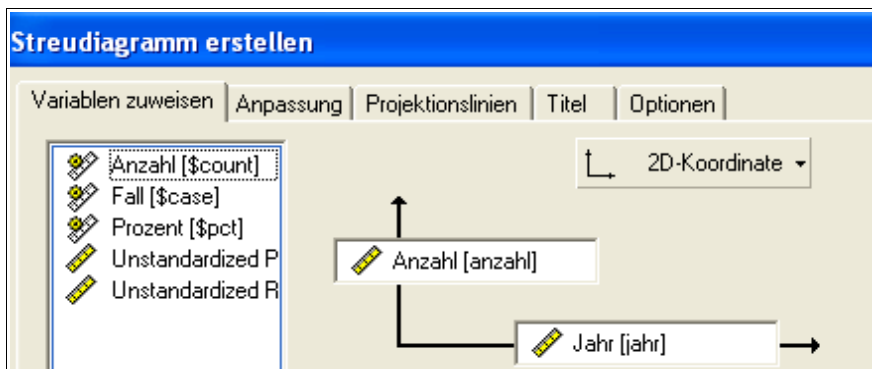


Abbildung 17.4

Fordern Sie über die Schaltfläche „Anpassung“ eine Regressionsgerade an:

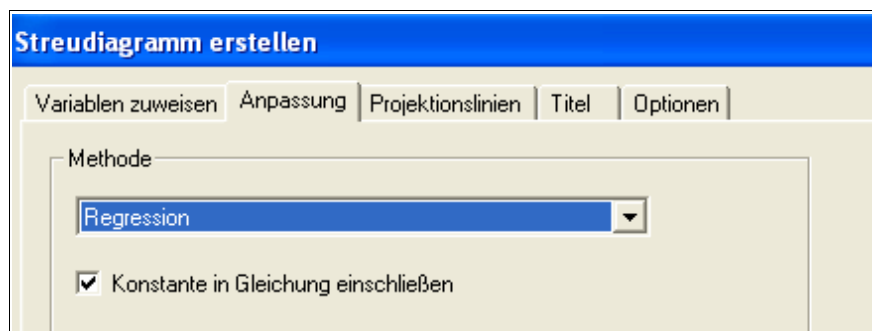


Abbildung 17.5

Klicken Sie auf „OK“. Die Ergebnisse werden im „Viewer“-Fenster angezeigt.

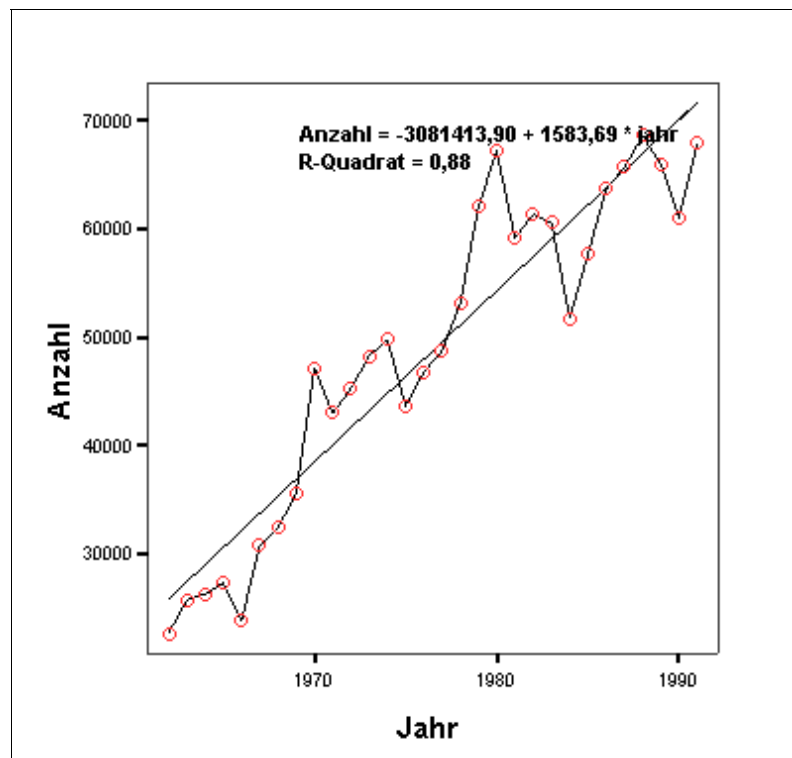


Abbildung 17.6

Es ist deutlich erkennbar, daß die lineare Approximation mit der Regressionsgeraden den generellen Trend (Wachstum) gut erfaßt, aber nicht sensibel auf Schwankungen reagiert. Eine Zeitreihen-Analyse würde voraussichtlich eine bessere Approximation liefern.

## 17.4 Bewerten der Güte eines Regressionsmodells

Bei der linearen Regression berechnen Sie ausgehend von den Beobachtungspunkten  $P_1=(x_1,y_1), \dots, P_n=(x_n,y_n)$  der Stichprobe  $S$  Schätzwerte  $m$  und  $b$  für eine Gerade (Regressions-Gerade), die "möglichst optimal" durch die Beobachtungspunkte  $P_1, \dots, P_n$  verläuft. Da es nur für  $n=2$  eine eindeutig bestimmte Gerade gibt (2 Punkte bestimmen eindeutig eine Gerade), können Sie das Problem für 3 und mehr Beobachtungspunkte ( $n > 3$ ) nicht eindeutig lösen.

Sie fordern nun vielmehr, daß die Regressions-Gerade  $g(x) = m \cdot x + b$  folgende **Minimaleigenschaft** erfüllt (**Gauß'sche Methode der kleinsten Quadrate**):

$$(1) \text{ SSE} = [y_1 - (m \cdot x_1 + b)]^2 + \dots + [y_n - (m \cdot x_n + b)]^2$$

$$\leq [y_1 - (m x_1 + b)]^2 + \dots + [y_n - (m x_n + b)]^2$$

(minimal für alle Variationen von  $b$  und  $m$ )

SSE: *Sum of Squares Errors or Residual*

Die Schätzwerte  $\hat{b}$  (Achsenabschnitt, *intercept*) und  $\hat{m}$  (Koeffizient vor der unabhängigen Variablen, *slope*) berechnen Sie durch Lösen der Minimierungsaufgabe (1). Die verbleibende Abweichung der Beobachtungspunkte zur Geraden (**SSE**) drückt das Verhalten von Y aus, daß sich nicht durch das Modell  $Y=mX+b$  erklären läßt, sondern allein vom Fehler Z abhängt.

Die Güte der Modellgleichung  $Y=mX+b$  überprüfen Sie nun folgendermaßen (Normierende Faktoren werden im folgenden vernachlässigt!):

Die Varianz des Modells (SSM, *Sum of Squares Model*) beschreibt die Abweichung des Mittelwertes  $\bar{y}$  von der Regressionsgeraden (Abstandsmaß: ebenfalls quadrierter vertikaler Abstand):

$$(2) \quad \mathbf{SSM} = [\bar{y} - (m x_1 + b)]^2 + [\bar{y} - (m x_2 + b)]^2 + \dots + [\bar{y} - (m x_n + b)]^2$$

Die gesamte Quadratsumme der Abweichungen der abhängigen Variablen Y von ihrem Mittelwert  $\bar{Y}$  (SSY) läßt sich in zwei Summanden aufspalten, in SSE und SSM:

$$(3) \quad \mathbf{SSY} = \mathbf{SSE} + \mathbf{SSM}$$

Das Verhältnis<sup>7</sup> F zwischen SSM und zu SSE dient Ihnen als Maß, wie groß die emp. Varianz des Modells SSM im Vergleich zur emp. Varianz des Fehlers SSE ist; d.h. wie "gut" das Modell die Varianz der abhängigen Variablen erklärt:

$$(4) \quad F = (SSM/1) / (SSE/(n-2))$$

ist verteilt nach  $\mathbf{F(x;1,n-2)}$   
 $\mathbf{F(x;1,n-2)}$  ist die Fisher-Verteilung mit (1,n-2) Freiheitsgraden

Je größer F ist, desto „mehr“ Varianzanteil wird durch das lineare Modell "erklärt" und desto weniger Varianzanteil muß durch den (an sich störenden) Term SSE<sup>8</sup> erklärt werden.

Eine ähnliche Größe  $R^2$  beschreibt den Quotienten aus SSM und SSY.

$$(5) \quad R^2 = SSM / SSY$$

Für  $R^2$  "nahe bei 1" erklärt das lineare Modell  $Y=mX+b$  einen Großteil der gesamten empirischen Varianz von Y, während der Fehler Z nur unwesentlich zur Va-

---

7 Die einzelnen Größen sind nicht aussagekräftig, da Sie z.B. bei 100 Beobachtungen mit kleinen Abweichungen vom Mittelwert einen ähnlich großen Wert für SSE erhalten könnten wie bei 10 Beobachtungen mit großen Abweichungen. Nur der Quotient aus SSE und SSY bzw. aus SSM und SSE ist aussagekräftig, da diese die beiden Größen zueinander ins Verhältnis setzen.

8 Optimal wäre  $SSE=0$ . In diesem Fall besteht ein genauer linearer Zusammenhang. Je größer SSE wird, desto größer sind die störenden Einflüsse.

rianz beiträgt (vgl. (3)). Sie können die Testgröße F auch für einen formalen Hypothesentest verwenden, da die Verteilung von F bekannt ist.

## 17.5 Übungen

1. Führen Sie für das Datenmaterial aus der SPSS Arbeitsdatei „*umwelt.sav*.“ eine lineare Regression für den zeitlichen Ablauf von Umweltstraftaten durch.  
Verwenden Sie hierzu für die y-Achse (abhängige Variable) jeweils die Variablen **ua** (umweltgefährdende Abfallbeseitigung) und **gv** (Gewässerverunreinigung) und für die x-Achse (unabhängige Variable) die Variable **jahr**.
2. Zusätzlich:  
Erzeugen Sie für die SPSS Arbeitsdatei „*buecher.sav*“ auf Grundlage der Variablen **anzahl** und der neuen Variablen **pre\_1** (erwarteter Wert) überlagerte Streudiagramme mit den Beobachtungspunkten und der Regressionsgeraden und verbinden Sie die Punkte durch eine Spline-Interpolation.
3. (Ergänzung)  
Welche Prognosen können Sie aus den linearen Modellen aus Aufgabe (1) für das Jahr 2000 ablesen (*forecasting*) und inwieweit würden Sie den Prognosen vertrauen?

**Hinweise:**

$$g(x) = mx + b, x = 2000$$

## 18 Vergleichen von 2 Gruppenmittelwerten (t-Test)

In diesem Kapitel werden Verfahren vorgestellt, um die beobachteten arithmetischen Mittelwerte zweier Gruppen miteinander zu vergleichen und zu entscheiden, ob ein Unterschied zwischen den beiden Gruppen zufällig zu erklären ist oder als **signifikant** einzustufen ist.

Wie unterscheiden sich 2 verschiedene Behandlungsverfahren in ihrer Wirksamkeit? Verdienen Männer mehr als Frauen?

### 18.1 Interpretieren von Unterschieden zwischen Gruppen

Ein häufig auftretendes statistisches Problem ist der Vergleich von zwei Stichproben hinsichtlich einer gemeinsam beobachteten Variablen  $X$ . Durch Einteilung einer Stichprobe in zwei oder mehrere Gruppen  $G_1, \dots, G_m$  können Sie ebenfalls "Teil-Stichproben" erzeugen, die miteinander verglichen werden können.

Im folgenden Beispiel wird die Variable „Behandlungserfolg“ (=“Senkung des Blutdruckes bei Bluthochdruck-Patienten“), betrachtet, die für 2 Präparate gemessen und verglichen werden kann.

Sie wollen die **Null-Hypothese  $H$**  untersuchen, daß die Grundgesamtheiten, aus denen die Gruppen stammen, den selben Erwartungswert besitzen, so daß der Unterschied zwischen den beobachteten Gruppenmittelwerten zufällig entstanden ist .

$$H: \begin{array}{l} \mu_1 = \mu_2 \\ \mu_1 = E(X) \text{ für 1. Gruppe, } \mu_2 : E(X) \text{ für 2. Gruppe} \end{array}$$

Die Alternative  $A$  besagt, daß der Unterschied zwischen den Gruppenmittelwerten zu groß (**signifikant**) ist, um sich zufällig aus den Unterschieden zwischen Individuen erklären zu lassen, sondern nur systematisch durch unterschiedliche Erwartungswerte erklärt werden kann.

### 18.2 Testen auf gleiche Erwartungswerte

Im folgenden Beispiel werden für die SPSS Datendatei „hypertonie-01.sav“ die beiden fiktiven Medikamente **alphasan** (**med=1**) und **betasan** (**med=2**) hinsichtlich der Wirkung bei der Senkung des Blutdrucks während einer 1-monatigen Behandlung (**diff=rrs1-rrs0**) untersucht.

Definieren Sie zunächst eine neue Variable **diff=rrs1-rrs0** (Behandlungserfolg durch Absenkung des Blutdrucks).

Die Null-Hypothese  $H_0$  lautet, daß die Erwartungswerte von  $\text{diff}$  für die Gruppen, die durch Medikament 1 bzw. 2 festgelegt werden, übereinstimmen. Führen Sie zur statistischen Absicherung Ihrer Vermutung, daß die Mittelwerte signifikant unterschiedlich sind, einen statistischen Test durch. Wählen Sie „Analysieren > Mittelwertvergleiche > t-Test für unabhängige Stichproben“. Wählen Sie die zu analysierende Variable aus, hier  $\text{diff}$ ,

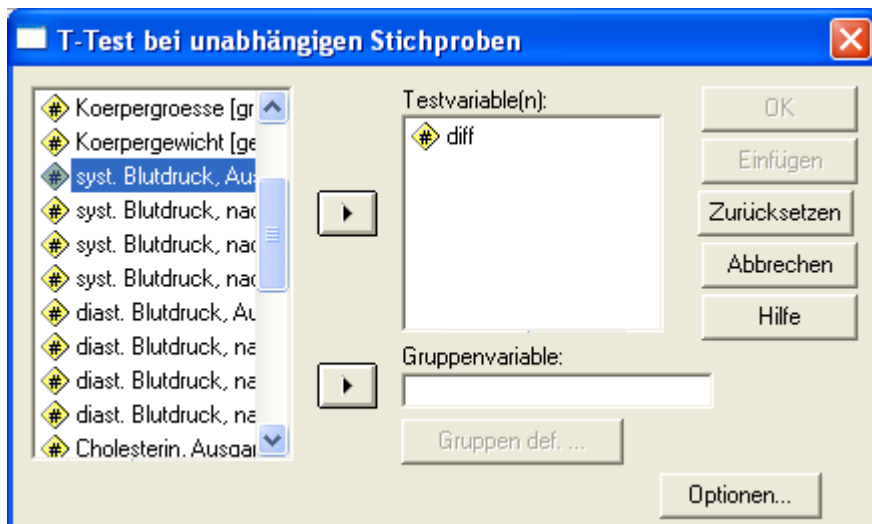


Abbildung 18.1

Verwenden Sie als Variable, nach der gruppiert werden soll, die Variable  $\text{med}$ . Zusätzlich müssen unter „Gruppen definieren“ die beiden Werte eingegeben werden, nach denen die Gruppen unterschieden werden, hier:  $\text{med}=1$  und  $\text{med}=2$ .

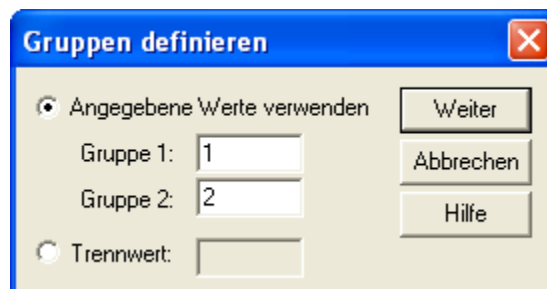


Abbildung 18.2

Klicken Sie auf „OK“. Die Ergebnisse werden im „Viewer“-Fenster angezeigt.

T-Test für die Mittelwertgleichheit						
T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	95% Konfidenzintervall der Differenz	
					Untere	Obere
-2,792	172	,006	-5,9195	2,1200	-10,1041	-1,7350
-2,792	170,218	,006	-5,9195	2,1200	-10,1044	-1,7347

Ihre Null-Hypothese  $H_0$  lautet, daß der Erfolg der Medikamente in Hinblick auf Blutdrucksenkung gleich ist.

Der t-Test liefert Ihnen den Wert der Teststatistik (*t-value*) und die zugehörige Irrtumswahrscheinlichkeit  $p$ . Die Irrtumswahrscheinlichkeit  $\alpha$ , die Null-Hypothese  $H_0$  fälschlicherweise abzulehnen, obwohl sie wahr ist, können Sie bis zum Wert  $p=0.006$  wählen. Die Null-Hypothese  $H_0$  sollte dementsprechend abgelehnt werden.

Der Unterschied zwischen den beobachteten Mittelwerten ist also zu signifikant, um nur allein auf zufällige Schwankungen zurückgeführt werden zu können.

## 18.3 Übungen

- Führen Sie einen t-Test durch für die Variable **physik** (Abiturnote einer Klasse in Physik) aus der SPSS Arbeitsdatei „schueler.sav“, wobei Sie nach **sex** (Geschlecht) unterscheiden.
- (zusätzlich:  
Vergleichen Sie mit einem nicht-parametrischen Test wie z.B. dem **Mann-Whitney U-Test**, der nicht die arithmetischen Mittelwerte, sondern die Ränge der Gruppen, miteinander vergleicht.  
**Hinweise:**  
Der U-Test sollte eingesetzt werden, wenn die Voraussetzungen für den t-Test – welche? - nicht erfüllt sind. Welchen Einfluß haben jeweils Ausreißer auf das Testergebnis (Stichwort: Robustheit)?

## 19 Vergleichen mehrerer Gruppenmittelwerte (Varianz-Analyse)

In diesem Kapitel wird die ein-faktorielle Varianz-Analyse **ANOVA** (*Analysis of Variance*) für eine abhängige Variable als Methode zum Vergleichen von drei und mehr Gruppenmittelwerten behandelt. Auf die mehr-faktorielle Varianz-Analyse für eine abhängige Variable wird abschließend kurz eingegangen.

*Nimmt die Merkfähigkeit mit dem Alter ab? Der t-Test kann nur für 2 Gruppen durchgeführt werden, ich habe aber 3 Gruppen ...*

### 19.1 Aufstellen eines ein-faktoriellen Modells

Die Unterteilung einer Stichprobe in  $m$  Gruppen  $G_1, \dots, G_m$  nehmen Sie wie bereits im vorherigen Kapitel beschrieben über eine kategoriale Variable  $X$  (auch als Faktor oder Gruppenvariable bezeichnet) mit  $m$  ( $m > 2$ ) unterschiedlichen Werten (Kategorien) vor. Jede der  $m$  Gruppen enthalte  $n_i$  Beobachtungen,  $i=1, \dots, m$ .

Im folgenden Beispiel soll die Merkfähigkeit in Abhängigkeit vom Alter untersucht werden. Sie wollen nun zunächst die Null-Hypothese  $H$  überprüfen, daß alle Gruppen den selben Erwartungswert besitzen (zusammengesetzte Hypothese, da nicht ein Vergleich, sondern viele Vergleiche durchgeführt werden müssen, bei  $n=4$  z.B. 6 Vergleiche):

$$H: \mu_1 = \mu_2 = \dots = \mu_m$$

$$E(X|G_1) = \mu_1 \text{ usw.}$$

Bei der **Varianz-Analyse** zerlegen Sie die gesamte Varianz einer Variablen  $Y$  in einen Anteil **SSM**, der auf den Unterschieden zwischen Erwartungswerten unterschiedlicher Gruppen beruht (*Between-Groups-Variance*) und einen Anteil **SSE**, der auf der Unterschiedlichkeit von Individuen innerhalb einer Gruppe beruht (*Within-Group-Variance*).

Aufgrund dieser Zerlegung der gesamten Varianz in zwei Bestandteile können Sie nun anhand der Größenverhältnisse entscheiden, ob die Gruppen hinsichtlich ihrer Erwartungswerte als "gleich" (SSM klein im Vergleich zu SSE) oder als

signifikant "unterschiedlich" (SSE klein im Vergleich zu SSM) zu betrachten sind<sup>9</sup>.

Falls die Null-Hypothese nicht zutrifft, muß im nachhinein (**a-posteriori** oder **post-hoc**) untersucht werden, welche paarweisen Unterschiede zwischen Gruppen signifikant sind und welche allein durch die Variabilität von individuellen Beobachtungen, also zufällig, zu erklären sind.

Hierzu ordnen Sie die Mittelwerte in aufsteigender Reihenfolge und überprüfen jeweils benachbarte Gruppen auf signifikante Unterschiede:

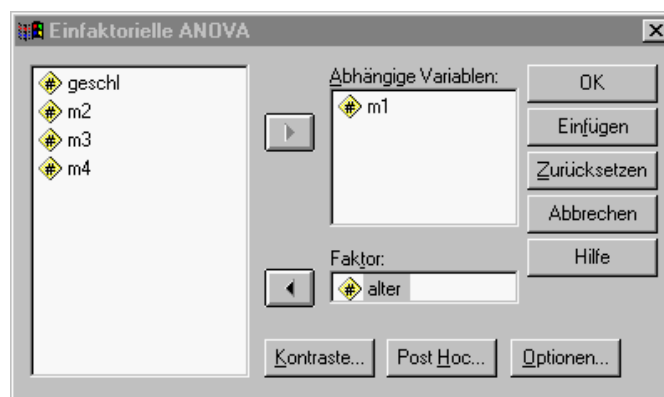
**post-hoc:**  $H: \mu_{(1)} = \mu_{(2)}$  oder Alternative:  $\mu_{(1)} < \mu_{(2)}$  usw.

Bei aufsteigender Sortierung ergibt sich typischerweise eine Einteilung in homogene Mengen (subsets), wobei innerhalb der Menge keine signifikanten Unterschiede bestehen, sondern nur solche zwischen den Mengen.

## 19.2 Vergleichen von mehreren unabhängigen Stichproben

Im folgenden Beispiel verwenden Sie einen Merkfähigkeitstest in der Datendatei `varana.sav`, bei dem die Merkfähigkeit `m1` untersucht werden soll. Als Gruppenvariable dient die Variable `alter` (Altersklasse) mit 3 unterschiedlichen Ausprägungen. Wählen Sie „Analysieren > Mittelwerte vergleichen > Ein-faktorielle ANOVA“.

Wählen Sie eine abhängige Variable aus, deren Varianz analysiert werden soll, und einen Faktor, der für die Gruppenaufteilung verwendet wird, hier: `m1` und `alter`.



<sup>9</sup> Als aufmerksamer Leser werden Sie die Bezeichnungen SSE und SSM wiedererkannt haben. Lineare Regression und Varianzanalyse sind beide auf das allgemeine lineare Modell (*generalized linear model*) zurückführbar.

Verwenden Sie die Aktionsschaltfläche Post-hoc, um die Analyse um Post-Hoc Tests zu erweitern – Sie vermuten, daß die Erwartungswerte signifikant unterschiedlich sind ...



Klicken Sie auf „OK“. Die Ergebnisse werden im „Viewer“-Fenster angezeigt.

ANOVA					
M1					
	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	142,929	2	71,465	31,917	,000
Innerhalb der Gruppen	53,737	24	2,239		
Gesamt	196,667	26			

Die Varianz-Analyse liefert Ihnen die Irrtumswahrscheinlichkeit  $p = .0001$  für den Wert  $t = 31,917$  der Testgröße F. Die Null-Hypothese sollte abgelehnt werden, d.h. es liegt **mindestens ein signifikanter Unterschied** zwischen den Erwartungswerten der Gruppen vor.

Die nachgeschaltete (**a-posteriori, post-hoc**) Betrachtung der Mittelwerte verdeutlicht, daß die Merkfähigkeit signifikant mit dem Alter abnimmt - traurig, aber wahr.

Mehrfachvergleiche						
Abhängige Variable: M1						
Bonferroni						
(I) ALTER	(J) ALTER	Mittlere Differenz (I-J)	Standardfehler	Signifikanz	95%-Konfidenzintervall	
					Untergrenze	Obergrenze
bis 30 Jahre	31 - 50 Jahre	1,22	,754	,354	-,72	3,16
	ueber 50 Jahre	5,27*	,723	,000	3,41	7,13
31 - 50 Jahre	bis 30 Jahre	-1,22	,754	,354	-3,16	,72
	ueber 50 Jahre	4,05*	,673	,000	2,32	5,78
ueber 50 Jahre	bis 30 Jahre	-5,27*	,723	,000	-7,13	-3,41
	31 - 50 Jahre	-4,05*	,673	,000	-5,78	-2,32

\*. Die mittlere Differenz ist auf der Stufe .05 signifikant.

### 19.3 Aufstellen eines mehr-faktoriellen Modells

Die Unterteilung einer Stichprobe in ein **mehr-faktorielles Design** nehmen Sie über mehrere kategoriale Variablen  $X_1, X_2, \dots, X_k$  (Faktoren) mit jeweils  $m_i$  unterschiedlichen Ausprägungen (Faktor-Stufen oder kurz Stufen) vor. Sie erhalten auf diese Art und Weise z.B. bei 2 Faktoren  $X_1$  und  $X_2$  ein Design mit  $m_1 \times m_2$  unterschiedlichen Zellen.

Das Problem bei der k-faktoriellen Varianzanalyse sind u.a. die möglichen **Wechselwirkungen (Interaktionsterme)** zwischen den Faktoren. Optimal ist eine Auswahl von schwach korrelierten Faktoren, die nur vernachlässigbar kleine Interaktionsterme hervorrufen. Es könnte deshalb sinnvoll sein, wenn Sie vor einer Varianz-Analyse eine Faktoren-Analyse durchzuführen.

### 19.4 Zurückführen der Varianz-Analyse auf ein lineares Modell

Bei entsprechender Zuordnung läßt sich die **Varianz-Analyse** auf ein **lineares Modell** der Form  $Y=X\beta+Z$  (Matrix-Schreibweise) abbilden, im einfachsten Fall lautet die Null-Hypothese  $Y=b+Z$  (identische Erwartungswerte für alle Zellen).

Damit sind alle Ausführungen bzgl. SSE und SSM, die zum linearen Modell erfolgten, auf die Varianz-Analyse übertragbar.

### 19.5 Übungen

1. Führen Sie eine einfache Varianzanalyse für die Variable `cpitn` (Behandlungsbedürftigkeit des Gebisses) aus der Arbeitsdatei `zahn.sav` jeweils mit den Variablen `alter` (Alter), `g` (Geschlecht), `s` (Schulabschluß), `pu` (Putzhäu-

figkeit), **zb** (Wechsel der Zahnbürste) und **beruf** als mögliche Faktoren durch.

**Hinweise:**

Wie lauten Ihre Hypothesen? Sollten die Hypothesen verworfen oder nicht verworfen werden? Welche Mittelwerte sind jeweils signifikant unterschiedlich (**post-hoc** Tests)?

2. zusätzlich:

Welche Voraussetzungen sollten **vor** einer Varianz-Analyse überprüft werden und welche Möglichkeiten gibt es hierzu? (Stichworte: Normalverteilung, Varianzhomogenität)

3. zusätzlich:

Besteht für das Einkommen ein signifikanter Unterschied bzgl. der Schulbildung? Verwenden Sie passende Variablen für Einkommen und Schulbildung aus der Arbeitsdatei `allbus90.sav`, um diese Fragestellung zu untersuchen. Welche anderen Faktoren könnten einen Einfluß haben?

## 20 Reduzieren der Variablenanzahl (Faktor-Analyse)

In diesem Kapitel wird die **Faktor-Analyse** behandelt, bei der Sie versuchen, eine i.d.R. große Anzahl von Variablen auf wenige, nicht direkt beobachtbare Einflußgrößen (**Faktoren**) zurückzuführen. Die Faktor-Analyse setzt – zumindest zum tieferen Verständnis - grundlegende Kenntnisse der Matrix-Algebra voraus.

*Was ist Liebe, Intelligenz, Kreativität, Qualifikation, Ausländerfeindlichkeit, ...*

### 20.1 Ermitteln von gemeinsamen Faktoren

Ausgangspunkt einer Faktor-Analyse ist eine Untersuchung mit einer Vielzahl von Variablen, bei denen nicht a-priori bekannt ist, ob in und welcher Weise sie miteinander verbunden sind. Gesucht werden bei der Faktor-Analyse sogenannte „**Hintergrund-Variablen**“ wie zum Beispiel „Kreativität“, „Qualifikation“, „Allgemeine oder sprachliche Intelligenz“, die im Rahmen der Faktor-Analyse als **Faktoren** bezeichnet werden. Besonders bei einer großer ("unüberschaubaren") Anzahl von Variablen besteht der Wunsch, diese auf einige grundlegende, allerdings von Ihnen nicht direkt beobachtbare oder quantifizierbare Hintergrund-Variablen oder Faktoren zurückzuführen.

Zum Beispiel stellt das "Lebensgefühl in einem Wohngebiet" einen Faktor dar, der sich nicht direkt messen läßt, aber sicher in einem starken Zusammenhang (hohe Korrelation) mit Variablen wie "Zahl der hinzugezogenen/fortgezogenen Personen", "Wohndauer und Umzugshäufigkeit im Wohngebiet", "Zufriedenheit mit der Infrastruktur", "Anzahl Kindergärten", „Anzahl Seniorenwohnheime“, "Altersstruktur" usw. steht.

Sie verwenden die **Faktor-Analyse** als ein mathematisches (nicht statistisches) Verfahren, um auf Grundlage der Korrelationsmatrix von beobachteten Variablen auf neue, nicht direkt beobachtete Variablen (Faktoren) zu schließen. Eine erfolgreiche Faktor-Analyse zeichnet sich dadurch aus, daß Sie disjunkte Mengen<sup>10</sup> von **Variablen mit hoher Korrelation** zu einem gemeinsamen Faktor zusammenzufassen **und** diese Faktoren hinsichtlich ihrer realen Bedeutung interpretieren können.

Eine Faktor-Analyse besteht aus folgenden Schritten, bei der Sie nur bei den fett markierten Schritte Entscheidungen treffen bzw. eine Interpretation vornehmen müssen, die restlichen, eher technischen Schritte werden automatisch von SPSS durchgeführt:

<sup>10</sup> Mengen sind disjunkt, wenn ihre Schnittmenge leer ist. Eine Zerlegung in disjunkte Teilmengen ist ein Ziel der Faktorenanalyse, das nicht immer vollständig erreicht werden kann.

1. **Auswählen** von Variablen
2. Normieren von Variablen  
(z-Transformation,  $z_i = (x_i - \bar{x})/s$ , d.h. die transformierten Variablen haben Mittelwert=0 und emp. Standardabweichung  $s=1$ )
3. Berechnen der Korrelationsmatrix für die normierten Variablen
4. Berechnen der Eigenwerte der Korrelationsmatrix  
(Hauptkomponenten-Analyse)
5. **Festlegen der Anzahl der Faktoren**  
(subjektive Entscheidung, z.B. Kriterium: Eigenwert > 1)
6. Ggf. Rotieren des Koordinatensystems der Eigenvektoren  
(z.B. orthogonale Rotation mit Varimax-Methode oder schiefwinkliges Rotieren)
7. Berechnen der Koeffizienten der Eigenvektoren (Faktor-Ladungen) für jede Variable
8. **Zuordnen der Variablen zu Faktoren** (möglichst eindeutig, d.h. Einteilung in disjunkte Mengen)
9. **Interpretieren der Eigenvektoren (Faktoren)**  
(manchmal unvermeidliches Ergebnis:  
Abbrechen der Faktor-Analyse, da keine sinnvolle Interpretation der Faktoren möglich ist!)  
im günstigen Fall:
10. Hinzufügen der neuen Variablen (Faktorwerte) zur Arbeitsdatei

## 20.2 Durchführen einer Faktoren-Analyse

Im folgenden führen Sie eine Faktor-Analyse für die demoskopischen und klimarelevanten Variablen **sb** (Anteil der Stadtbevölkerung), **lem** (Lebenserwartung der Männer), **lew** (Lebenserwartung der Frauen), **ks** (Kindersterblichkeit), **so** (Anzahl Sonnenscheintage pro Jahr), **nt** (Anzahl Regentage pro Jahr), **tjan** (Tagestemperatur im Januar), **tjul** (Tagestemperatur im Juli) aus der Datendatei `europa.sav` durch. Die Variable **land** enthält eine Kurzbezeichnung für das jeweilige Land.

Berechnen Sie zunächst die Korrelationsmatrix:

Correlation Matrix								
	KS	LEM	LEW	NT	SB	SO	TJAN	TJUL
KS	1.00000							
LEM	-.75208	1.00000						
LEW	-.85207	.88887	1.00000					
NT	-.64264	.38860	.44709	1.00000				
SB	-.71494	.55289	.68082	.53841	1.00000			
SO	.57481	-.43154	-.42041	-.72078	-.48714	1.00000		
TJAN	.22960	-.20682	-.16388	-.38540	-.12580	.69510	1.00000	
TJUL	.70231	-.54736	-.60219	-.84244	-.60951	.72690	.41945	1.00000

Tabelle 20.1: Korrelation

Wählen Sie „Analysieren -> Dimensionsreduktion -> Faktorenanalyse“. Wählen Sie die Variablen aus, die in eine mit Standardeinstellungen durchgeführte Faktor-Analyse einbezogen werden sollen, hier: ) **sb**, **lem**, **lew**, **ks**, **so**, **nt**, **tjan** und **tjul**.

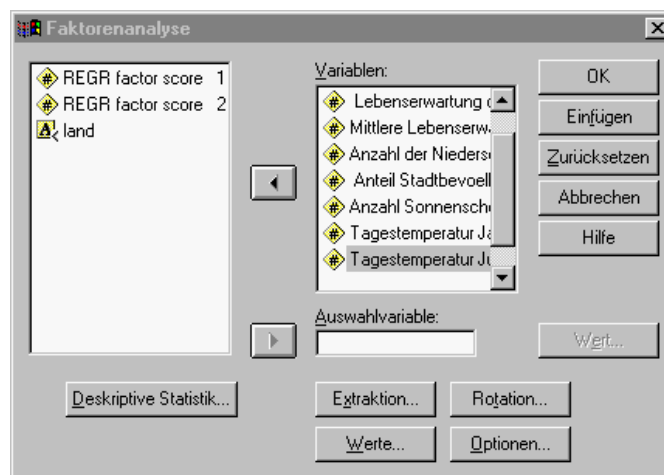
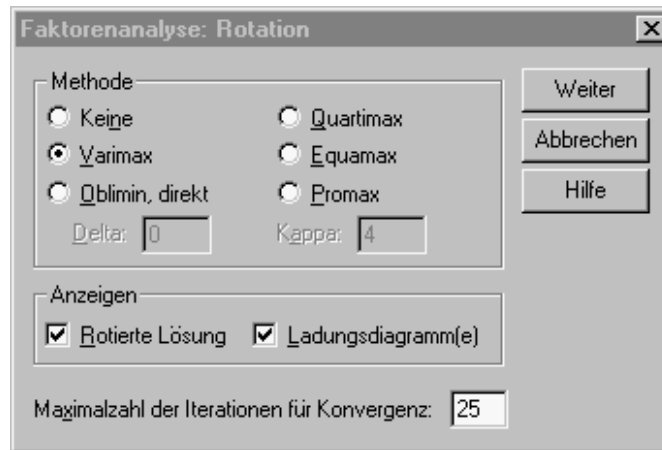


Abbildung 20.1 : Faktoren

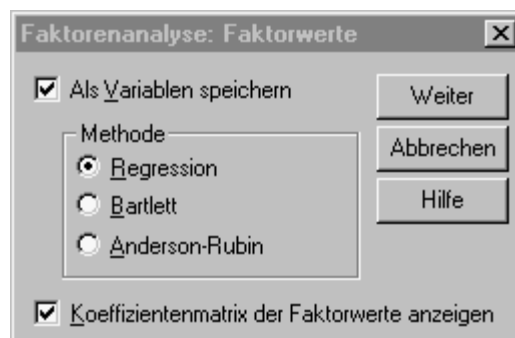
Die Anzahl der Faktoren wird über die Größe der Eigenwerte festgelegt (Standard-Einstellung).



Es soll eine Rotation nach der Varimax-Methode erfolgen, um möglichst optimale Faktoren zu finden.



Die Faktorwerte sollen in die Arbeitsdatei aufgenommen werden (Koeffizienten der einzelnen Beobachtungen bezüglich der Faktoren als neue Variablen).



Klicken Sie auf „OK“. Die Ergebnisse werden im „Viewer“-Fenster angezeigt.

Die Faktor-Analyse liefert u.a. die **Eigenwerte** (*eigen values*) der Hauptkomponenten-Analyse und die **Faktor-Ladungen** (*factor scores*) nach einer Rotation des Koordinatensystems:

Komponente	Anfängliche Eigenwerte		
	Gesamt	% der Varianz	Kumulierte %
1	4,944	61,801	61,801
2	1,408	17,604	79,406

Abbildung 20.2 : Eigenwerte und Anteil an Gesamtvarianz

Zwei der Eigenwerte sind größer als 1 und werden aufgrund der verwendeten Einstellung in die folgenden Berechnungen aufgenommen, alle weiteren Eigenwerte werden ignoriert. Die Variable `factor1` gehört zum 1. (größten) Eigenwert, `factor2` zum 2. (zweit-größten) Eigenwert.

	Komponente	
	1	2
Kindersterblichkeit bei 1000 Geburten	-,875	,325
Lebenserwartung der Maenner	,866	-,138
Mittlere Lebenserwartung der Frauen	,940	-,127
Anzahl der Niederschlagstage pro Jahr	,462	-,719
Anteil Stadtbevoelkerung	,780	-,245
Anzahl Sonnenscheinstunden pro Jahr	-,334	,874
Tagestemperatur Januar	5,572E-02	,852
Tagestemperatur Juli	-,594	,675

Extraktionsmethode: Hauptkomponentenanalyse.  
Rotationsmethode: Varimax mit Kaiser-Normalisierung.  
a. Die Rotation ist in 3 Iterationen konvergiert.

Abbildung 20.3 : Faktorladungen (factor score coefficients)

Die von SPSS vorgenommene „Rotation“ (Drehung des Koordinatensystems“) verfolgt das Ziel, jede Variable auf nur genau einen Faktor "hochzuladen".

Im vorliegenden Beispiel können Sie die Variablen tatsächlich sinnvoll in 2 disjunkte Mengen aufteilen. Die 1. Menge mit den Variablen `sb` (Anteil der Stadtbevölkerung), `lem` (Lebenserwartung der Männer), `lew` (Lebenserwartung der Frauen), `ks` (Kindersterblichkeit) lädt hoch auf den 1. Faktor, und die 2. Menge mit den Variablen `so` (Anzahl Sonnenscheintage pro Jahr), `nt` (Anzahl Regentage pro Jahr) und `tjan` (Tagestemperatur im Januar) auf den 2. Faktor <sup>11</sup>.

<sup>11</sup> Eine Interpretation der Faktoren ist einfacher, wenn alle einem Faktor zugeordneten Variablen zum Faktor positiv korreliert sind. So ist z.B. ein hoher Wert für `lebw` ein als positiv zu bewertendes Merkmal, ein hoher Wert für `ks` ein als negativ zu wertendes Merkmal. Die Variablen sollten deshalb transformiert werden, z.B. `ks2=100-ks`.

## 20.3 Interpretieren der Faktoren in einem Streudiagramm

Im folgenden Beispiel ist die Faktor-Analyse gemäß Aufgabe (1) der Übungen durchgeführt worden. Erstellen Sie ein Streudiagramm für die beiden Faktoren über „*Grafiken Interaktiv* > *Streudiagramm*“. Tragen Sie die beiden Faktoren gegeneinander auf und wählen Sie die Variable **land** als Fallbeschriftung.

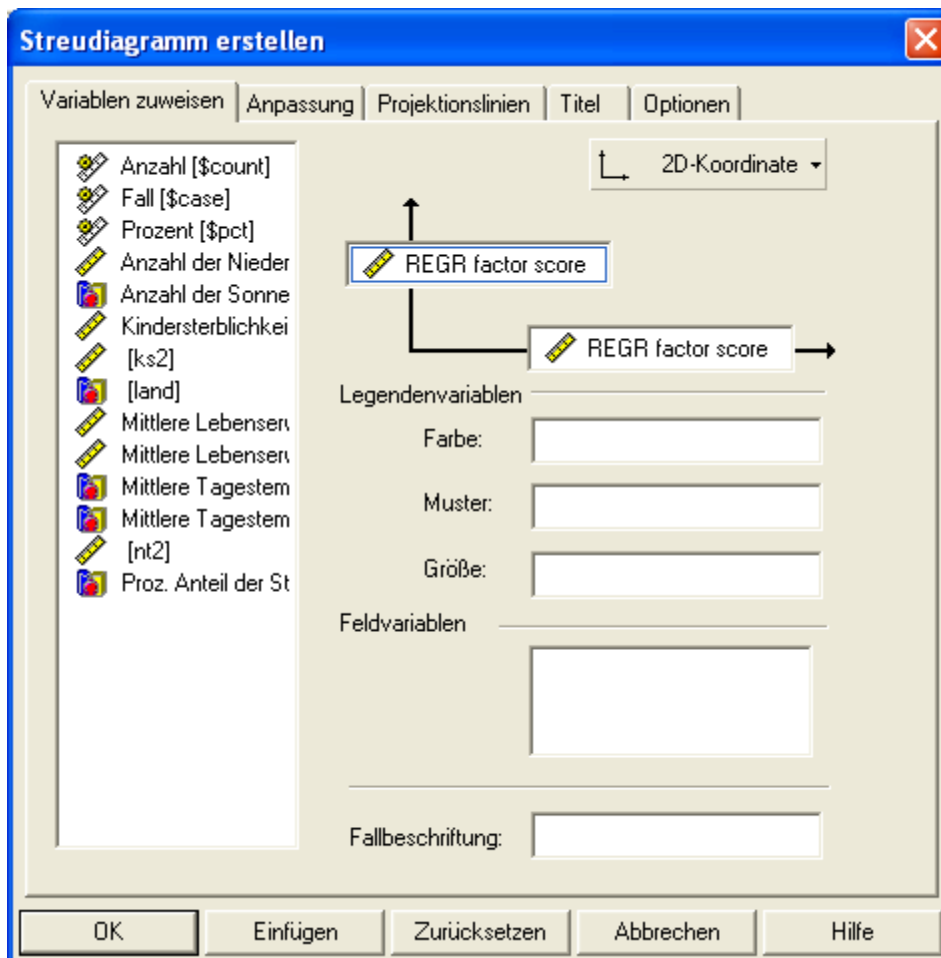


Abbildung 20.4

Klicken Sie auf „OK“. Die Ergebnisse werden im „Viewer“-Fenster angezeigt.



Der 1. Faktor (vertikal) ist ein Maß für "Lebensdauer", der 2. Faktor (horizontal) ein Maß für "Klima". Es fällt nun nicht mehr schwer, die europäischen Länder nach diesen Kriterien zu klassifizieren —und die Hauptreise-Länder auf der Klima-Skala (Faktor 2) ganz links zu finden.

## 20.4 Reduzieren der Variablenanzahl

Bei der Faktor-Analyse sollen  $p$  von Ihnen ausgewählte Variablen  $X_1$  bis  $X_p$  als Linearkombinationen von  $k$  (zunächst unbekannt) Faktoren  $F_1$  bis  $F_k$  ( $k < p$ , Reduktion der Anzahl!) dargestellt werden<sup>12</sup>:

$$(1) \quad X_i = A_{i1} F_1 + \dots + A_{ik} F_k + U_i$$

Hierbei sind  $F_1$  bis  $F_k$  neue Variablen (*common factors*, gemeinsame Faktoren) und  $A_{i1}$  bis  $A_{ik}$  die Koeffizienten der einzelnen Faktoren für die Variable  $X_i$ .  $U_i$  ist

<sup>12</sup> Es sollte  $k < p$  gelten, denn sonst führt dieser Ansatz nicht zu einer Reduktion der Variablenanzahl.

der Anteil von  $X_i$ , der sich nicht auf gemeinsame Faktoren zurückführen läßt (*unique factor*).

Die Faktoren  $F_1 - F_k$  lassen sich nun ihrerseits durch die Variablen  $X_1 - X_p$  ausdrücken:

$$(2) \quad F_j = W_{j1} X_1 + \dots + W_{jp} X_p$$

Hierbei werden die Koeffizienten  $W_{jk}$  als Faktor-Ladungen (*factor score coefficients*) bezeichnet.

Sie stehen nun vor der Aufgabe, die Variablen so in (disjunkte) Teilmengen  $M_1, M_2, \dots, M_k$  aufzuteilen, daß Variablen in einer Teilmenge  $M_i$  hohe Faktorladungen für den Faktor  $F_i$  tragen und nur möglichst geringe Faktorladungen für die anderen Faktoren.

Die Aufteilung der Variablen in disjunkte Teilmengen; d.h. die Zuordnung von Variablen zu Faktoren, ist nicht eindeutig lösbar und wird i.d.R. am besten vom **Varimax Verfahren** gelöst. Darüberhinaus besteht das Problem, daß Sie die im voraus nicht bekannten und nur **mathematisch**, d.h. nicht aus der eigentlichen Problemstellung, abgeleiteten Faktoren im nachhinein **bezogen auf die ursprüngliche Problemstellung** interpretieren müssen. Falls Ihnen eine derartige Interpretation nicht möglich ist, ist die Faktor-Analyse als gescheitert anzusehen.

## 20.5 Übungen

1. Transformieren Sie die Variablen aus der Arbeitsdatei `europa.sav` in neue Variablen, so daß für diese neuen Variablen nur positive Korrelationen zu "hochgeladenen" Faktoren bestehen. Führen Sie nun erneut eine Faktor-Analyse durch. Interpretieren Sie die neuen Faktoren hinsichtlich ihrer "realen" Bedeutung.  
**Hinweise:**  
Die Variablen `ks` und `nt` mit „negativer“ Bedeutung sollten in Variablen mit „positiver“ Bedeutung transformiert werden. Wählen Sie hierzu: `ks2=100-ks`, `nt2=365-nt`.
2. Führen Sie für eine Untersuchung über die Einstellung zu Ausländern in der Arbeitsdatei `ausland.sav` eine Faktor-Analyse für die Variablen `a01` bis `a15` durch und interpretieren Sie die ermittelten Faktoren. Welche Variablen (Antworten) laden auf genau einen Faktor hoch?  
Die Variablen `a01` bis `a15` repräsentieren die Antworten auf folgende Fragen (auf einer Skala von 1="Völlige Ablehnung" bis 7 = "Vollständige Zustimmung").

mung"):

- a01: Die Integration der Ausländer muß verbessert werden.
- a02: Das Flüchtlingselement muß gemindert werden.
- a03: Deutsches Geld sollte für deutsche Belange ausgegeben werden.
- a04: Deutschland ist nicht das Sozialamt der Welt.
- a05: Ein gutes Miteinander ist anzustreben.
- a06: Das Asylrecht ist einzuschränken.
- a07: Die Deutschen werden zur Minderheit.
- a08: Das Asylrecht ist europaweit zu schützen.
- a09: Die Ausländerfeindlichkeit schadet der deutschen Wirtschaft.
- a10: Wohnraum sollte zuerst für Deutsche geschaffen werden.
- a11: Wir sind auch Ausländer, fast überall.
- a12: Multikulturell bedeutet multikriminell.
- a13: Das Boot ist voll. a14: Ausländer raus.
- a15: Ausländerintegration ist Völkermord.

**Hinweise:**

Wählen Sie 3 Faktoren aus. Eine mögliche Aufteilung in 3 Gruppen ist z.B. (1,12,13,15,3,4,7), (5,7), (6,14).

3. zusätzlich:

In einer Studie zu Frühgeburten in `fruehgeb.sav` werden die Faktoren *allgemeine Intelligenz* (AI) und *sprachliche Intelligenz* (SI) vermutet. Versuchen Sie, hierfür eine Faktoren-Analyse durchzuführen.

## 21 Exploratives Analysieren von Daten

In diesem Kapitel werden Möglichkeiten behandelt, wie das Datenmaterial tabellarisch und grafisch dargestellt werden kann und wie die Hypothesen **Normalverteilung** und **Varianzhomogenität** in einer vorgeschalteten Untersuchung überprüft werden können. Normalverteilung und Varianzhomogenität werden bei vielen statistischen Verfahren als Voraussetzungen gefordert.

### 21.1 Exploratives Analysieren von Daten

Mit der Prozedur "Explorative Datenanalyse" werden Auswertungsstatistiken und grafische Darstellungen für alle Fälle oder für separate Fallgruppen erzeugt. Es kann viele Gründe für die Verwendung der Prozedur "Explorative Datenanalyse" geben:

*Sichten von Daten, Erkennen von Ausreißern, Beschreibung, Überprüfung der Annahmen und Charakterisieren der Unterschiede zwischen Teilgrundgesamtheiten (Fallgruppen)*

Beim Sichten der Daten können Sie ungewöhnliche Werte, Extremwerte, Lücken in den Daten oder andere Auffälligkeiten erkennen. Durch die explorative Datenanalyse können Sie sich vergewissern, ob die für die Datenanalyse vorgesehenen statistischen Methoden geeignet sind. Die Untersuchung kann ergeben, daß Sie die Daten transformieren müssen, falls die Methode eine Normalverteilung erfordert. Sie können sich statt dessen auch für die Verwendung nichtparametrischer Tests entscheiden. (aus dem SPSS Hilfesystem)

Wichtige Voraussetzungen, die Sie häufig vor der Durchführung von statistischen Verfahren absichern müssen, sind **Normalverteilung** und **Varianzhomogenität**.

Zur Absicherung dieser Voraussetzungen dienen u.a. die Teststatistik von Kolmogorov-Smirnov (**Test auf Normalverteilung**) und die Teststatistik von Levene (**Test auf Varianzhomogenität**).

Die Ergebnisse der genannten Tests beeinflussen die Auswahl weiterer statistischer Verfahren, da diese häufig genau diese Voraussetzungen an die Stichprobe stellen. Falls die Voraussetzungen nicht erfüllt sind, können Sie z.B. auf nicht-parametrische Verfahren zurückgreifen.

### 21.2 Testen auf Normalverteilung

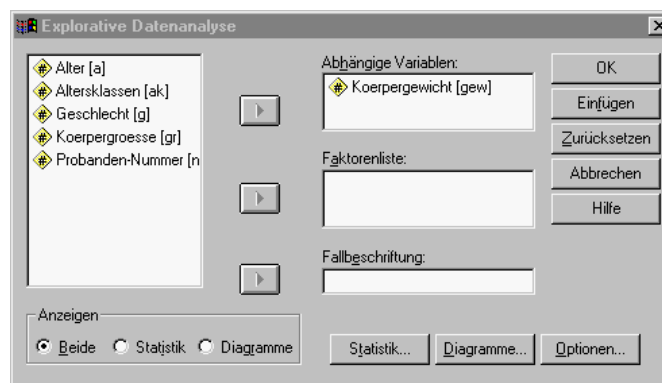
Die Testgröße des **Kolmogorov-Smirnov Tests** mißt die Abweichungen zwischen der vermuteten theoretischen und der tatsächlich beobachteten empirischen Verteilungsfunktion von  $X$ . Die Null-Hypothese  $H$  lautet entsprechend: Die unbekannte theoretische Verteilungsfunktion  $F_x(t)$  ist eine Normalverteilung

$N(t)$  mit Erwartungswert  $\bar{x}$  und Varianz (Schätzungen für unbekannte Parameter!).

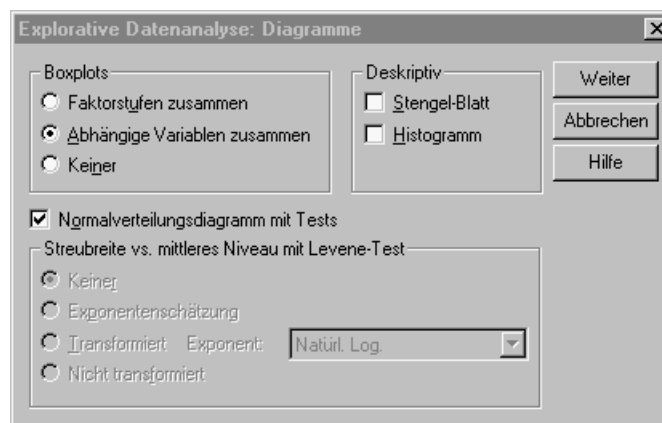
Hypothese:

Die Variable  $X$  ist normalverteilt;  
d.h.  $F_x(t) = N(t; \bar{x}, s^2)$

Im folgenden Beispiel untersuchen Sie die Variable **gew** aus der Arbeitsdatei „hyper.sav“ auf Normalverteilung. Wählen Sie „Analysieren > Deskriptive Statistiken > Explorative Datenanalyse“. Wählen Sie die Variable **gew** als abhängige Variable



Aktivieren Sie nun *Diagramme* und nehmen Sie dort weitere Einstellungen vor. Kreuzen Sie hier Normalverteilungsplots mit Tests an.



Klicken Sie auf „OK“. Die Ergebnisse werden im „Viewer“-Fenster angezeigt.

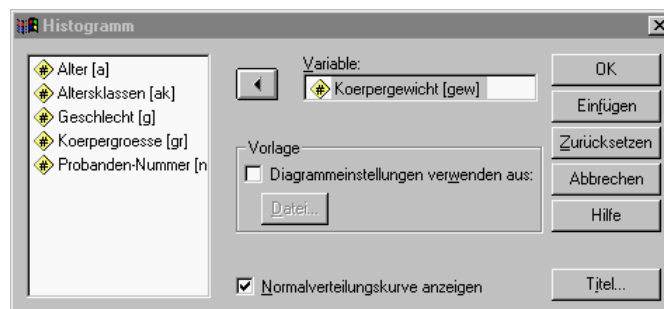
Tests auf Normalverteilung			
	Kolmogorov-Smirnov <sup>a</sup>		
	Statistik	df	Signifikanz
Koerpergewicht	,082	174	,006

a. Signifikanzkorrektur nach Lilliefors

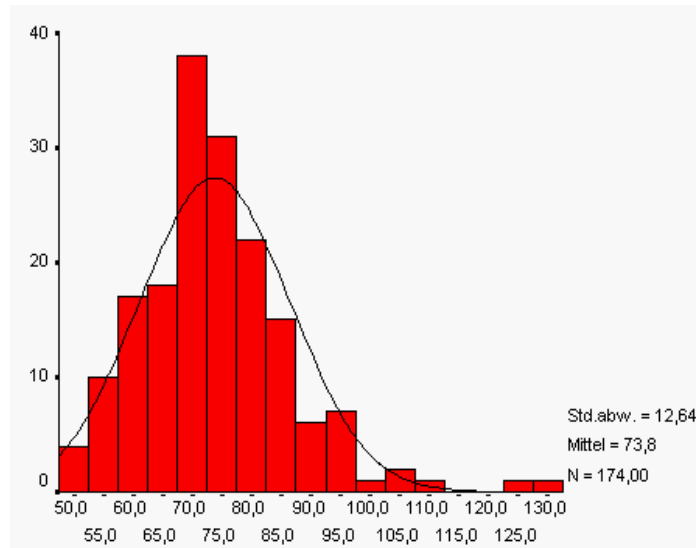
Die explorative Datenanalyse liefert Ihnen einen Wert  $t=0.082$  für die Kolmogorov-Smirnov-Testgröße K-S (Normalverteilung). Die zugehörige Irrtumswahrscheinlichkeit beträgt  $p=0.006$ .

Die Null-Hypothese sollte deswegen abgelehnt werden; d.h. es handelt sich vermutlich nicht um eine Normalverteilung.

Fordern Sie ein Histogramm mit überlagerter Normalverteilungskurve an. Wählen Sie „*Grafik > Interaktiv > Histogramm*“.



Klicken Sie auf „OK“. Die Ergebnisse werden im „Viewer“-Fenster angezeigt.



Die Ablehnung der Null-Hypothese wird auch visuell durch ein Histogramm mit eingezeichneter Normalverteilungskurve für die Variable **gew** unterstützt. Ein Histogramm dient hier als visuelles Hilfsmittel zur Überprüfung der Normalverteilungsannahme.

## 21.3 Testen auf Varianzhomogenität

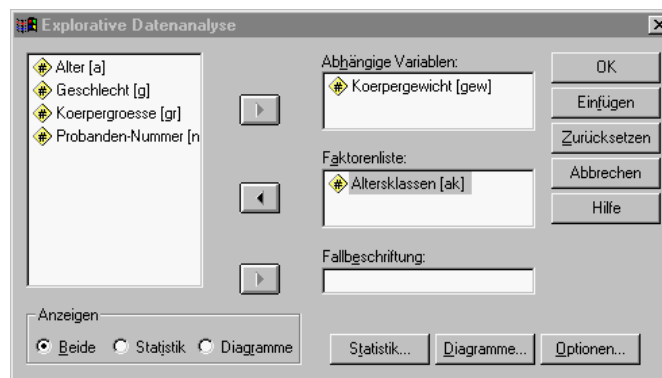
Die Testgröße des **Levene-Tests** mißt die Unterschiedlichkeit der Standardabweichungen einer Variablen X in unterschiedlichen Gruppen G1, G2, ..., Gm. Die Null-Hypothese H lautet entsprechend :Die Varianz von X ist in allen Gruppen gleich.

Hypothese:

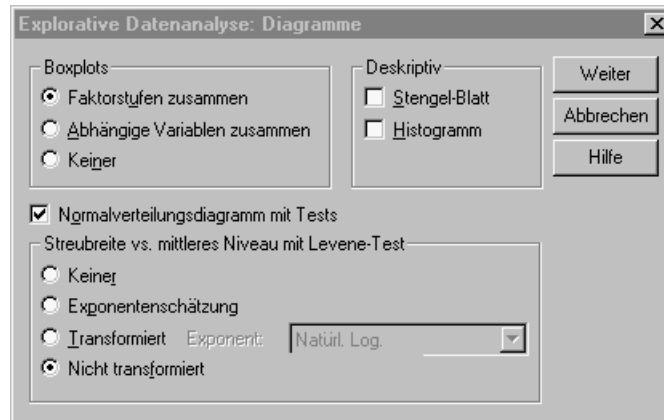
Die Varianz von X ist in allen Gruppen gleich;  
d.h.  $\text{Var}(X|G1) = \text{Var}(X|G2) = \dots = \text{Var}(X|Gm)$

Im folgenden Beispiel untersuchen Sie die Variable **gew** aus der Arbeitsdatei `broca.sav` bzgl. **ak** (Altersklasse) auf Varianzhomogenität. Wählen Sie „Analysieren > Deskriptive Statistiken > Explorative Datenanalyse“.

Wählen Sie zunächst die Variable **gew** als abhängige und die Variable **ak** (Altersklasse) als Faktor (unabhängige Variable).



Aktivieren Sie nun *Diagramme* und nehmen Sie dort weitere Einstellungen vor. Fordern Sie *Boxplots* an, die für jede Altersklasse (d.h. gruppiert nach **ak**) getrennt erstellt werden und nebeneinander angezeigt werden. Kreuzen Sie Normalverteilungsplots mit Tests an und zusätzlich ein Histogramm. Fordern Sie für die nicht transformierten Beobachtungen den **Levene-Test** an.



Klicken Sie auf „OK“. Die Ergebnisse werden im „Viewer“-Fenster angezeigt.

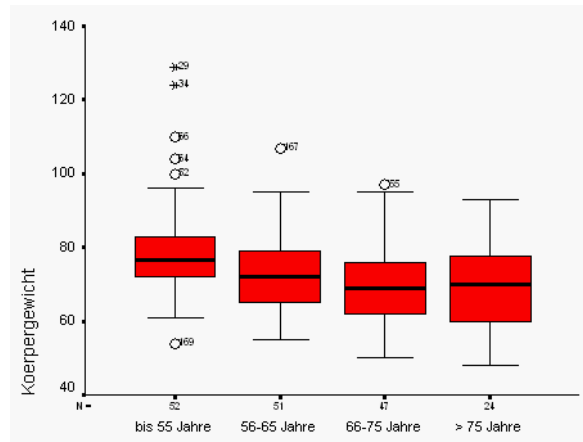
		Levene-Statistik	df1	df2	Signifikanz
Körpergewicht	Basiert auf dem Mittelwert	,313	3	170	,816
	Basiert auf dem Median	,145	3	170	,933
	Basierend auf dem Median und mit angepaßten df	,145	3	136,657	,933
	Basiert auf dem getrimmten Mittel	,221	3	170	,882

Die explorative Datenanalyse liefert Ihnen für den gemessenen Wert  $t$  der **Levene Testgröße (Varianzhomogenität)** eine zugehörige Irrtumswahrscheinlichkeit  $p=0,816$ . Dieser Wert ist folgendermaßen zu interpretieren:

Die Null-Hypothese  $H_0$ , daß die Varianzen der Gruppen gleich sind, kann nur mit einer Irrtumswahrscheinlichkeit von  $p=0,816$  abgelehnt werden, d.h. nur wenn eine Irrtumswahrscheinlichkeit von  $\alpha=0,9020$  verwendet wird, befindet sich der beobachtete Wert  $t$  der Testgröße  $T$  im Ablehnungsbereich der Null-Hypothese.

Da eine derartig hohe Irrtumswahrscheinlichkeit nicht zu rechtfertigen ist, kann die Null-Hypothese nicht verworfen werden; d.h. die Null-Hypothese ist sinnvoll und sollte aufrechterhalten werden. Es kann also von homogenen Varianzen ausgegangen werden.

Die Beibehaltung der Null-Hypothese wird auch visuell durch die nebeneinander gezeichneten **Boxplots** mit den Altersklassen als Kategorien unterstützt, die sich nur wenig voneinander unterscheiden.



## 21.4 Übungen

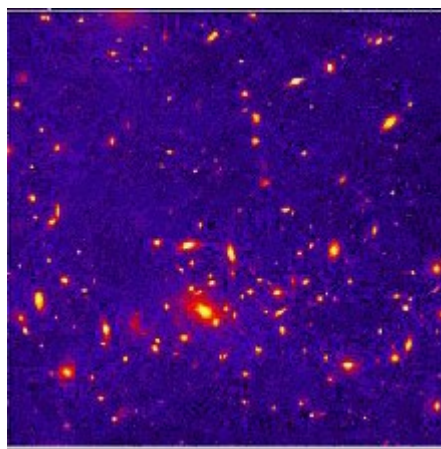
1. Erzeugen Sie für die Arbeitsdatei `hyper.sav` Boxplots für die Variable `gr` nach `g` (Geschlecht) gruppiert.
2. Führen Sie nun den **Levine Test** durch, um die Varianzhomogenität der beiden Gruppen Männer und Frauen zu testen. Entspricht das Ergebnis Ihrer visuellen Vorstellung, die durch die Boxplots erzeugt wird?  
Formulieren Sie Ihre Interpretation der Ergebnisse ähnlich wie im Skript.
3. Zusätzlich:  
Überprüfen Sie, ob für die Variable `physik` (Note im Fach Physik) aus der Arbeitsdatei `schueler.sav` die Normalverteilungsannahme gerechtfertigt ist. Sehen Sie prinzipiell Probleme aufgrund der Skalierung der Variablen?  
**Hinweise:**  
Normalverteilte Zufallsvariablen können (zumindest theoretisch) alle Werte zwischen Minus-Unendlich und Plus-Unendlich annehmen. Überprüfen Sie, ob die Körpergröße, Variable `gr`, (annähernd) normalverteilt ist.  
Testen Sie nochmals getrennt für Männer und Frauen. Formulieren Sie Ihre Interpretation ähnlich wie im Skript.

## 22 Zusammenfassen von Beobachtungen in Clustern (Cluster-Analyse)

In diesem Kapitel wird die **Cluster-Analyse** behandelt, bei der "dicht zusammenliegende" Beobachtungen ("Pulks" oder "Anhäufungen") nach einem mathematisch definierten Verfahren zu **Clustern** (Haufen) zusammengefaßt werden.

*Siehst Du 1000 Sternlein stehen ...*

( <http://aitzu3.ait.physik.uni-tuebingen.de/~stuhli/html/astro.html> )



### 22.1 Zusammenfassen von Beobachtungen

Aufgrund der grafischen Darstellung von intervall-skalierten Beobachtungen in Streudiagrammen erkennen Sie häufig visuell auffällige Ansammlungen, Haufen oder Punktwolken von „dicht zusammenliegenden Beobachtungen“.

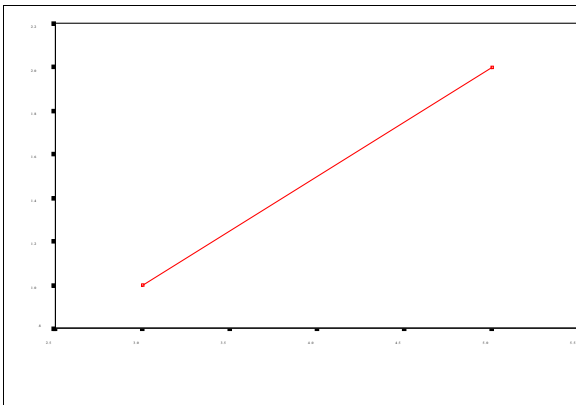
Die **Cluster-Analyse** liefert Ihnen Verfahren, um derartige Cluster (Haufen, Wolken, Pulks) nach mathematisch nachvollziehbaren (und nicht "intuitiven") Kriterien identifizieren zu können. Der Begriff Cluster ist zunächst für 3-dimensionale Sternenansammlungen verwendet und dann auf allgemeine (n-dimensionale) Streudiagramme übertragen worden.

Die Beobachtungen  $P_1, P_2$  usw. **innerhalb** eines Clusters sollen untereinander einen "kleineren" Abstand oder eine "größere Nähe" haben als Beobachtungen in **unterschiedlichen** Clustern, wobei der "Abstand" oder die "Nähe" zwischen zwei Beobachtungen und der "Abstand" oder die "Nähe" zwischen zwei Clustern mathematisch auf unterschiedliche Art und Weise definiert werden können, wodurch sich unterschiedliche Möglichkeiten zur Aufteilung der Beobachtungen in Cluster ergeben.

Ein häufig benutztes Abstandsmaß zwischen 2 Beobachtungspunkten  $P_1$  und  $P_2$  ist der **euklidische Abstand**. Bei 2 Variablen läßt sich der Abstand geome-

trisch deuten als Länge der Strecke zwischen den Punkten  $P_1=(x_1,y_1)$  und  $P_2=(x_2,y_2)$  (**Satz des Pythagoras**). Bei Beobachtungen mit n Variablen (n-dimensionale Beobachtungen) wird entsprechend der n-dimensionale euklidische Abstand verwendet.

$$D = \text{Abstand\_Zwischen\_Punkten}(P_1,P_2) = [(x_1-x_2)^2+(y_1-y_2)^2]^{1/2}$$



Das Streudiagramm enthält 2 Beobachtungspunkte  $P_1=(3,1)$  und  $P_2=(5,2)$  mit eingezeichnetem euklidischen Abstand.

Der euklidische Abstand d beträgt:

$$d = [(3-5)^2+(1-2)^2]^{1/2} = 25$$

Der Abstand zwischen zwei Clustern  $C_1$  und  $C_2$  läßt sich z.B. über folgende zwei Abstandsmaße definieren, die sich als **mittlerer Abstand** bzw. als **maximaler Abstand** zwischen Beobachtungspunkten aus unterschiedlichen Clustern ergeben:

1. **mittlerer Abstand:**

$\text{Abstand\_Zwischen\_Cluster}(C1,C2) = \text{Summe ueber alle Punk-Kombinationen: Abstand}(P_i,P_j) / (n * m)$

$P_i$  aus  $C_1$  und  $P_j$  aus  $C_2$ , n Beobachtungen in  $C_1$  und m Beobachtungen in  $C_2$

2. **maximaler Abstand:**

$\text{Abstand\_Zwischen\_Cluster}(C1,C2) = \text{Maximum} \{ \text{Abstand}(P_i,P_j), i \text{ aus } C_1, j \text{ aus } C_2 \}$

Die meisten Verfahren der Cluster-Analyse, die sich u.a. durch die Wahl der Abstandsmaße unterscheiden, suchen möglichst "kugelförmige Gebilde" im n-dimensionalen Raum, die sie zu einem Cluster zusammenfassen. Sie scheitern dementsprechend z.B. bei langgestreckten, unregelmäßigen oder überlappenden Formen, die aufgrund der verwendeten Abstandsmaße nicht als Cluster erkannt werden können.

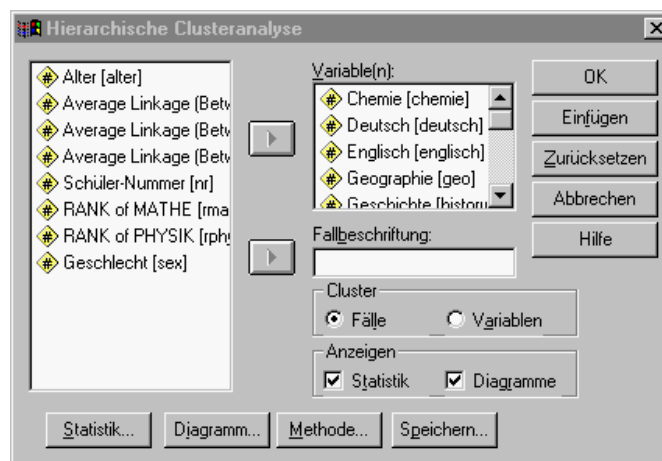
Unterscheidungsmerkmale für Verfahren zur Cluster-Analyse sind ferner der Zeitpunkt, zu dem die Anzahl der zu bildenden Cluster festgelegt wird und die Anzahl der gebildeten Cluster. Die Anzahl der Cluster kann entweder im voraus fest vorgegeben werden, im voraus auf einen Wertebereich eingegrenzt werden oder im nachhinein individuell auf eine Anzahl festgelegt werden, die gute Interpretationsmöglichkeiten bietet.

Wählen Sie ein hierarchisches Verfahren, wenn Sie erst im nachhinein die Anzahl der Cluster festlegen wollen. Sie können nun sukzessive für jede Aufteilung der Beobachtungen in  $n=2,3,4,5,6, \dots$  Cluster überprüfen, ob eine "ausgesagte" Interpretation möglich ist.

## 22.2 Ermitteln von hierarchisch geordneten Clustern

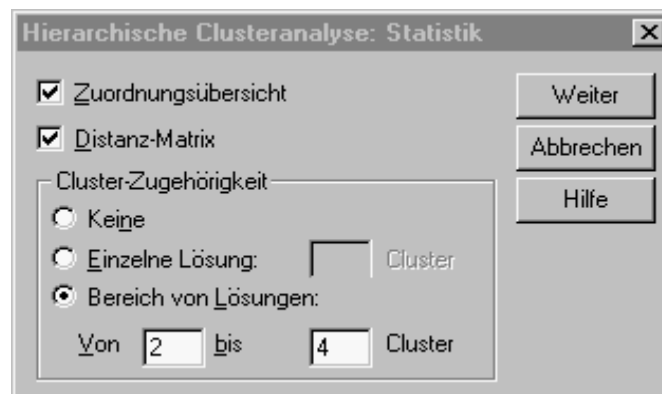
Im folgenden Beispiel führen Sie eine **hierarchische Cluster-Analyse** für die Arbeitsdatei `schueler.sav` durch.

Wählen Sie „Analysieren > Klassifizieren > Hierarchische Cluster“. Wählen Sie die Variablen aus, die als Berechnungsgrundlage für die Cluster-Analyse dienen sollen, hier die Faktorwerte aus der Faktorenanalyse. Nehmen Sie weitere Einstellungen vor über die Aktionsschaltflächen vor, hier: Statistiken, Diagramme und Speichern.

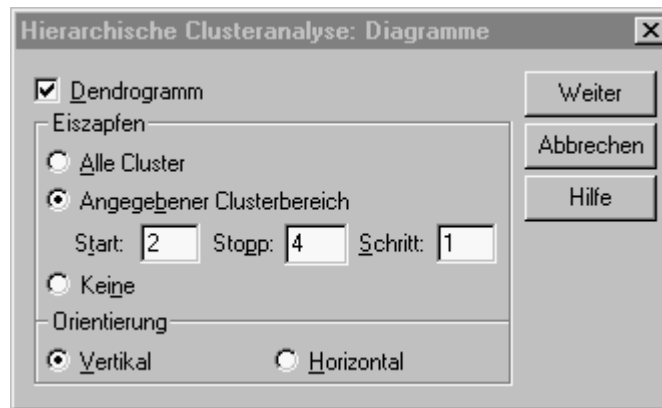


Wählen Sie insgesamt 3 Aufteilungen in disjunkte Cluster (2, 3, 4 Cluster).

Sie können sich im Anschluß jede Aufteilung ansehen und entscheiden, welche Aufteilung Ihnen optimal erscheint.

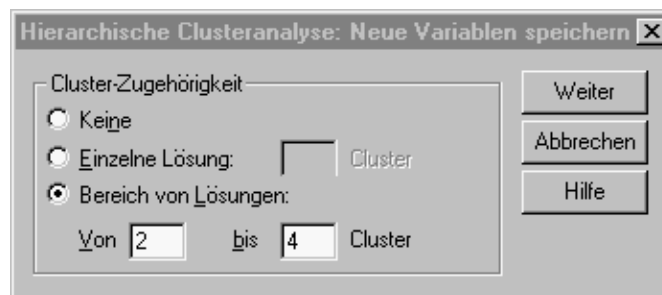


Fordern Sie diverse Diagramme an.



Erzeugen Sie neue Variablen `CLUSn_1`, die die Cluster-Zugehörigkeit der einzelnen Beobachtungen bei Vorgabe von  $n=2, 3, 4$  Clustern repräsentieren.

Die Variable `CLUS3_1` enthält z.B. (bei Aufteilung in insgesamt 3 Cluster) für jede Beobachtung die Cluster-Nummer (1, 2 oder 3), dem die Beobachtung zugeordnet wurde.



Die oben angeforderte Cluster-Analyse liefert eine sehr umfangreiche Ausgabe, u.a. eine **Abstandsmatrix** für die Beobachtungen, einen **Ablauf der Ver-**

**schmelzung** (Fusionierungsschema, *Agglomeration Schedule*), d.h. die Reihenfolge, in der Beobachtungen bzw. bereits vorhandene Cluster zu neuen Clustern verschmolzen werden und ein **Dendogramm**.

Fall	1	2	3	4	5	6
1		133,000	99,000	96,000	243,000	164,000
2	133,000		96,000	201,000	162,000	155,000
3	99,000	96,000		111,000	120,000	89,000
4	96,000	201,000	111,000		155,000	114,000
5	243,000	162,000	120,000	155,000		55,000
6	164,000	155,000	89,000	114,000	55,000	

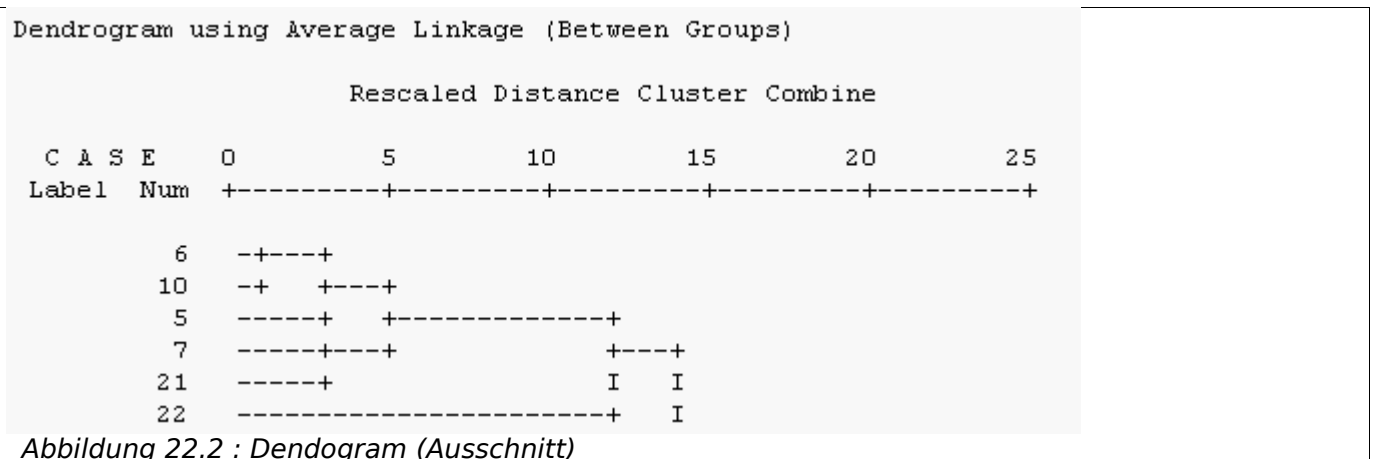
Abbildung 22.1 : Euklidische Abstandsmatrix zwischen den Beobachtungen (Ausschnitt)

Zuordnungsübersicht						
Schritt	Zusammengeführte Cluster		Koeffizienten	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	6	10	42,000	0	0	8
2	15	16	44,000	0	0	16
3	11	12	46,000	0	0	9
4	3	13	49,000	0	0	11

Tabelle 22.1 : Reihenfolge der Fusionierung und jeweiliger Cluster-Abstand vor der Fusionierung (Ausschnitt)

Im 1. Schritt (*stage 1*) wird die Beobachtung Nr. 6 mit der Beobachtung Nr. 10 zu einem Cluster verschmolzen, der mit der Nummer des ersten Elementes, hier 6, gekennzeichnet wird. Der euklidische Abstand zwischen Nr. 6 und Nr. 10 beträgt 42; d.h. die Summe aller quadrierten Abstände ist 42.

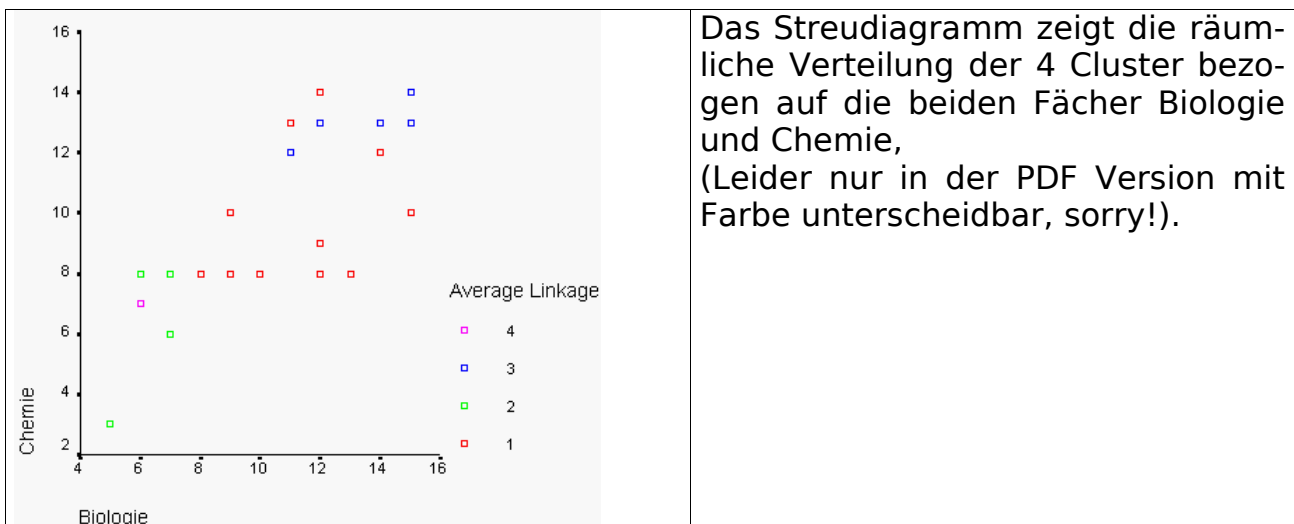
Sie sollten die Fusionierung abbrechen, wenn die die "Größe" (Durchmesser) der Cluster im Vergleich zu den "Abständen" zwischen den Clustern sprunghaft anwächst.

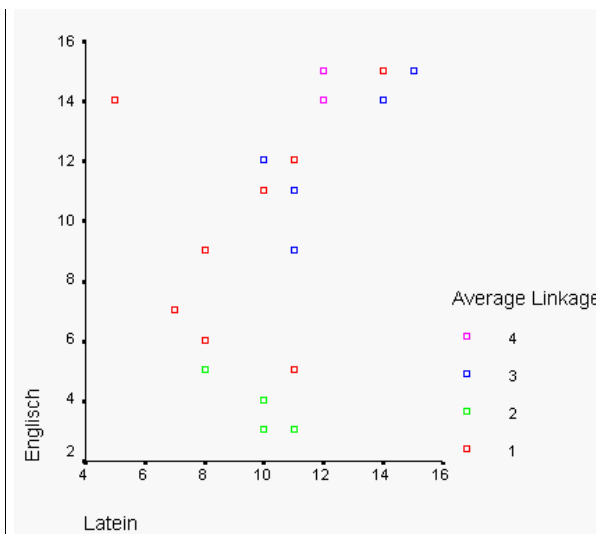


Das Dendrogramm zeigt Ihnen grafisch (von "oben nach unten gelesen") wie sukzessive Beobachtungen und dann Cluster miteinander verschmolzen werden. Dendrogramm und Fusionierungsschema (siehe zuvor) ergänzen sich, in dem sie numerische und grafische Informationen enthalten.

Sie können nun z.B. eine Sortierung nach der neuen Variable `CLU4_1` (Zugehörigkeit der Beobachtung zu einem von insgesamt 4 Clustern) durchführen und einige Kombinationen von 2 Variablen mit der Cluster-Zugehörigkeit als Beschriftung in einem Streudiagramm darstellen. Beachten Sie bitte, daß die Cluster-Analyse auf allen (!) Fächern beruht, während in den folgenden Diagrammen immer nur einige Fächer berücksichtigt werden!

Grafiken > Streudiagramm > Einfach





Das Streudiagramm zeigt die räumliche Verteilung der 4 Cluster bezogen auf die beiden Fächer Englisch und Latein.

## 22.3 Übungen

1. Führen Sie eine **Cluster-Analyse** (2-5 Cluster) für die Länder aus der Arbeitsdatei `europa.sav` durch, versuchen Sie eine optimale Clusteranzahl festzulegen und geben Sie in der von Ihnen gewählten Aufteilung jedem Cluster einen aussagekräftigen Titel.
2. zusätzlich:  
 Welche Auswirkungen haben unterschiedliche Wertebereiche bei den in der Cluster-Analyse beteiligten Variablen? Wie könnten alle Variablen "gleichermaßen" berücksichtigt werden?  
**Hinweise:**  
 Normierung auf einen einheitlichen Wertebereich [0,1] durch Z-Transformation.
4. zusätzlich:  
 Welche **a-posteriori** Auswertungen (d.h. Auswertungen im Anschluß an die Cluster-Analyse) und Grafiken halten Sie für sinnvoll?

### Hinweise:

Vergleich der Mittelwerte unterschiedlicher Cluster, räumliche Anordnung der Cluster und maximale vertikale bzw. horizontale Ausdehnung, für 2 Variablen: Berechnung (falls möglich!) von Trenn-Geraden, d.h. Einteilung der Ebene in Polygone, die Cluster voneinander trennen, oder Trenn-Kreisen, d.h. Einteilung der Ebene in nicht-überlappende Kreise, die jeweils einen Cluster enthalten.

# 23 Anhang

## 23.1 Bundestagswahlen

