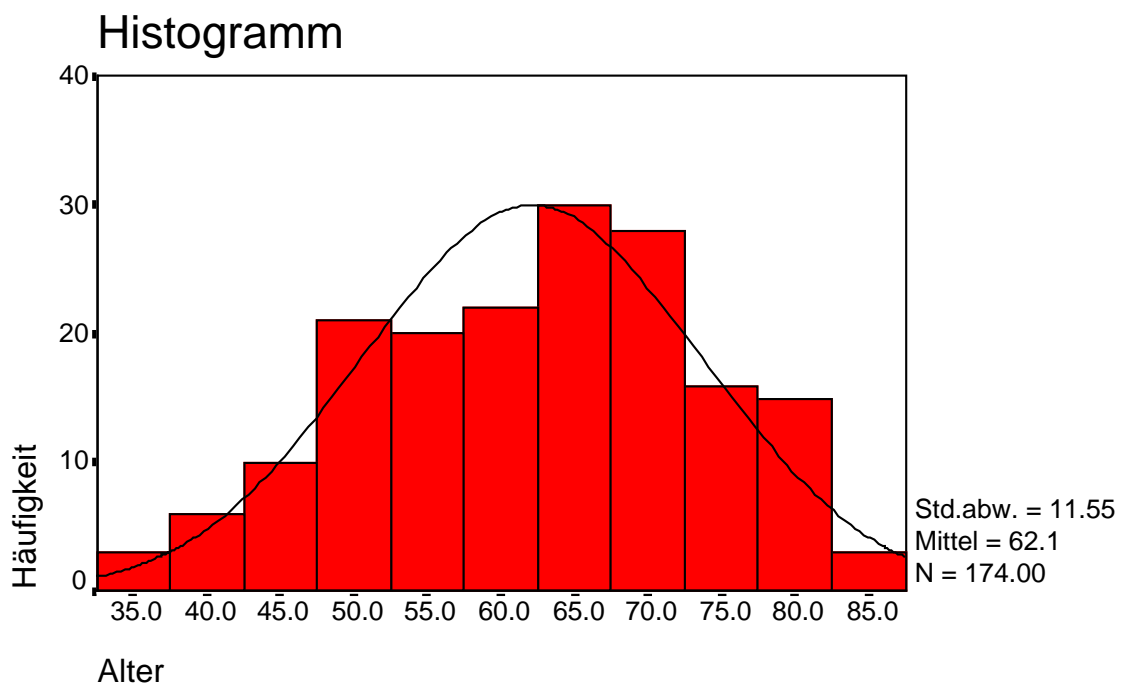


Statistische Datenanalyse mit SPSS für Windows



SPSS[®]
Real Stats. Real Easy.™

Impressum:

Universität Osnabrück
- Rechenzentrum -
Autor: Dipl.-Math. Frank Elsner
Albrechtstraße 28
D-49076 Osnabrück
E-Mail: Frank.Elsner@rz.uni-osnabrueck.de

Version: 1.7
Stand: 10.03.2003
WWW: <http://www.rz.uni-osnabrueck.de> >Skripte und Tutorials Online

Inhaltsverzeichnis

Einleitung	1
Voraussetzungen und Zielsetzung	1
SPSS Version	1
Typografische Konventionen	2
Namenskonventionen für Dateien	2
Hersteller	3
Weiterführende Literatur	3
Miete von SPSS für Windows für Studenten und Mitarbeiter	3
Überblick über SPSS für Windows	4
Funktionsumfang von SPSS für Windows	4
Typische Arbeitsschritte	4
Bedienen der Benutzeroberfläche	5
Online Hilfe	6
Definieren von Variablen und Erfassen von Beobachtungen	8
Fragebogen - Sonntagsumfrage	8
Überblick - SPSS Dateneditor	9
Vorgehensweise - Eingeben von Beobachtungen im Dateneditor	10
Zusammenfassung - Welche Informationen werden über eine Variable gespeichert?10	
Motivation - Einlesen von anderen Dateiformaten	10
Aufgaben	11
Berechnen neuer Variablen und Auswählen von Beobachtungen.....	12
Motivation – Hinzufügen oder oder Löschen von Beobachtungen/Variablen.....	12
Vorgehensweise - Berechnen neuer Variablen.....	12
Vorgehensweise - Filtern von Beobachtungen.....	14
Aufgaben	14
Arbeiten mit Datums-Variablen	16
Motivation - Darstellen von Datums- und Zeitangaben	16
Überblick – Funktionen für Variablen vom vom Datentyp DATUM.....	17
Vorgehensweise - Definieren von Variablen vom Datentyp DATUM	17
Vorgehensweise - Berechnen neuer Variablen vom Datentyp DATUM.....	17
Aufgaben	18
Aggregieren (Zusammenfassen) von Daten in eine neue SPSS Datei	19
Vorgehensweise – Zusammenfassen von Beobachtungen	19
Laden der aggregierten Daten	20
Aufgaben	20
Überblick über die deskriptive Statistik	21
Aufgaben der deskriptiven Statistik	21
Stichprobe und Grundgesamtheit.....	21
Messung von Variablen	22
Kenngrößen von Stichproben.....	23
Aufgaben	25
Erstellen von einfachen Tabellen und Berechnen von Kennzahlen.....	26
Überblick - Darstellen des Datenmaterials in tabellarischer Form und Berechnen von Kennzahlen	26
Vorgehensweise - Berechnen von Häufigkeiten	26
Vorgehensweise - Erstellen einer Kreuztabelle.....	27
Vorgehensweise - Berechnen von charakterisierenden Kennzahlen.....	28
Aufgaben	29
Erstellen von Diagrammen.....	30
Überblick - Visualisieren von Daten	30
Vorgehensweise - Erstellen eines einfachen Balkendiagramms.....	31
Vorgehensweise - Erstellen eines gruppierten Balkendiagramms	32
Vorgehensweise - Erstellen eines gestapelten Flächendiagramms.....	33
Vorgehensweise - Erstellen eines Histogramms (empirische Dichte).....	34
Vorgehensweise - Vergleichen von empirischen Verteilungen mit Hilfe von Boxplots35	
Vorgehensweise - Bearbeiten von Diagrammen	36
Aufgaben	36

Zufallsexperimente, Zufallsvariablen und Wahrscheinlichkeit	38
Zufallsexperiment und Wahrscheinlichkeit	38
Zufallsvariablen und ihre Verteilung	38
Aufgaben	39
Überblick über die mathematische Statistik	40
Ziehen von Rückschlüssen aus einer Stichprobe	40
Durchführen von Schätzungen und Hypothesentests	40
Einschränken der gesuchten theoretischen Verteilung auf eine Klasse (parametrische Tests)	41
Formulieren von Fragestellungen.....	42
Treffen von Entscheidungen anhand einer Entscheidungsregel.....	42
Aufgaben	44
Exploratives Analysieren von Daten und Überprüfen von Voraussetzungen	45
Motivation - Exploratives Analysieren von Daten und Überprüfen von Voraussetzungen für statistische Tests.....	45
Vorgehensweise - Testen auf Normalverteilung	45
Vorgehensweise - Testen auf Varianzhomogenität.....	47
Aufgaben	48
Berechnen eines Vertrauensbereiches (Konfidenz-Intervalls).....	49
Motivation - Interpretieren von Vertrauensbereichen	49
Vorgehensweise - Berechnen eines Vertrauensbereichs	49
Statistischer Hintergrund - Ableiten eines Vertrauensbereichs	50
Aufgaben	52
Testen der Unabhängigkeit von 2 Variablen.....	53
Motivation - Ableiten der Chi-Quadrat Testgröße.....	53
Vorgehensweise - Berechnen der Chi-Quadrat-Testgröße	53
Aufgaben	55
Berechnen von Korrelationskoeffizienten	56
Motivation - Festlegen eines Maßes für den linearen Zusammenhang	56
Vorgehensweise - Ermitteln des Korrelationskoeffizientens	56
Exkurs – nicht-lineare Zusammenhänge.....	58
Aufgaben	58
Approximieren von x-y-Punkten durch Geraden (lineare Regression).....	60
Motivation - Untersuchen eines möglichen linearen Zusammenhangs.....	60
Vorgehensweise - Durchführen einer linearen Regression.....	61
Statistischer Hintergrund - Bewerten der Güte eines Regressionsmodells.....	62
Exkurs: Vorgehensweise -Approximieren durch andere Kurven	63
Aufgaben	63
Vergleichen von 2 Gruppenmittelwerten (t-Test)	65
Motivation - Interpretieren von Unterschieden zwischen Gruppen.....	65
Vorgehensweise - Testen auf gleiche Erwartungswerte	65
Aufgaben	66
Vergleichen mehrerer Gruppenmittelwerte (Varianz-Analyse)	67
Motivation - Aufstellen eines ein-faktoriellen Modells.....	67
Vorgehensweise - Vergleichen von mehrern unabhängigen Stichproben.....	67
Exkurs: Motivation - Aufstellen eines mehr-faktoriellen Modells	69
Vorgehensweise - Durchführen einer 2-faktoriellen Varianzanalyse.....	69
Exkurs: Statistischer Hintergrund - Zurückführen der Varianz-Analyse auf ein lineares Modell.....	69
Aufgaben	69
Reduzieren der Variablenanzahl (Faktor-Analyse)	71
Motivation - Ermitteln von gemeinsamen Faktoren	71
Vorgehensweise - Durchführen einer Faktoren-Analyse.....	72
Stat. Hintergrund - Reduzieren der Variablenanzahl.....	75
Aufgaben	75
Zusammenfassen von Beobachtungen in Clustern (Cluster-Analyse)	77
Motivation - Zusammenfassen von Beobachtungen	77
Vorgehensweise - Ermitteln von hierarchisch geordneten Clustern.....	78
Aufgaben	82
Index	1

Einleitung

In diesem Kapitel wird ein Überblick über Zielsetzung und Aufbau des Skriptes gegeben sowie einige Hinweise auf weiterführende Literatur.

Voraussetzungen und Zielsetzung

Dieses Skript wendet sich an Benutzer, die mit **SPSS für Windows** (im folgenden mit **SPSS** bezeichnet) menügeführt statistische Datenanalysen durchführen wollen. Der Schwerpunkt liegt auf **der statistischen Absicherung von Aussagen** über das untersuchte Datenmaterial mit Unterstützung durch SPSS - und weniger auf der rein tabellarischen oder grafischen Darstellung des Datenmaterials.

Grundlegende Kenntnisse über die grafische Benutzeroberfläche Microsoft **Windows** werden vorausgesetzt wie auch grundlegende wahrscheinlichkeitstheoretische und statistische Kenntnisse.

Es handelt sich weder um ein Lehrbuch über Windows noch über Wahrscheinlichkeitstheorie und Statistik. Dieses Skript ist als Begleitmaterial zu einem Kurs des Rechenzentrums konzipiert worden und ist deshalb nur mit Einschränkungen zum Selbststudium geeignet. Der Autor empfiehlt, alle im Text behandelten Beispiele im direkten Zusammenspiel mit SPSS auszuprobieren und zumindest einige der Übungen zu bearbeiten.

Der Autor hat sich entschlossen, vom typisch deutschen Lehrbuchstil abzuweichen und immer den Leser persönlich anzusprechen ("Starten Sie das Programm ...").

In diesem Handbuch werden folgende Fragen behandelt:

1. **Wie können Sie vorhandene Daten in SPSS einlesen bzw. Daten direkt im SPSS Dateneditor erfassen?**
2. **Wie können Sie Daten tabellarisch und grafisch darstellen, um sich einen Überblick über das Datenmaterial zu verschaffen und Anregungen für weiterführende Analysen zu erhalten?**
3. **Wie können Sie beschreibende Statistiken der Stichprobe wie z.B. Mittelwert, empirische Varianz, emp. Standardabweichung und Spannweite berechnen?**
4. **Wie können Sie Stichproben mit statistischen Verfahren untersuchen und die Ergebnisse interpretieren (t-Test, Faktor-Analyse, Varianz-Analyse, ...)?**

Hierzu wird in jedem Kapitel zunächst anhand eines Fallbeispiels im 1. Schritt eine kurze motivierende Einführung in den statistischen Hintergrund gegeben. Im Anschluß wird im 2. Schritt die prinzipielle Vorgehensweise mit Hilfe des Fallbeispiels erläutert, um das entsprechende statistische Verfahren mit SPSS menügeführt durchzuführen und die Ergebnisse zu interpretieren. Gelegentlich wird der statistische Hintergrund in einem 3. Schritt ausführlicher dargestellt, um die Interpretation der Ergebnisse fundierter begründen zu können. Jedes Kapitel enthält als letzten Schritt Übungen zur Vertiefung des Stoffes. Alle im Handbuch genannten Dateien stehen maschinenlesbar zur Verfügung (siehe Impressum, dort Hinweis auf WWW Server). Anspruchsvollere Übungen sind durch einen (*) oder zwei (**) Sterne gekennzeichnet.

SPSS Version

Als Grundlage der Beschreibung dient **SPSS für Windows**, Version 11. Die meisten Menüpunkte sind allerdings auch in den Vorgängerversionen ab Version 9.0 vorhanden. Das beim Rechenzentrum verwendete SPSS enthält folgende Module, die für die Universität Osnabrück lizenziert sind:

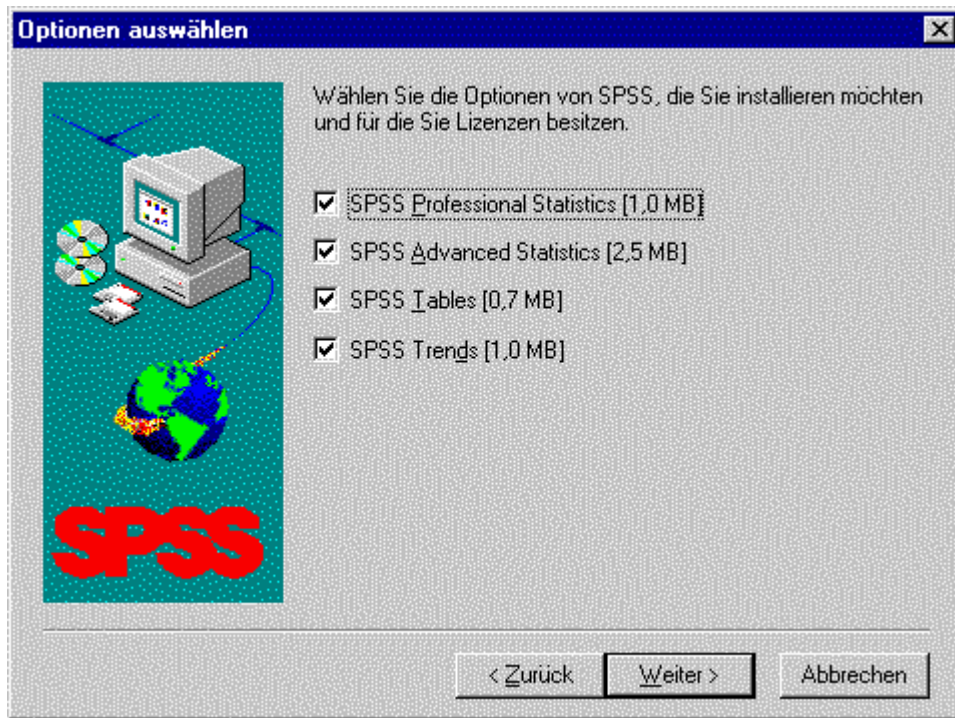


Abb./Tab. 1: SPSS Module

Beachten Sie bitte, daß andere Universitäten oder andere Firmen ggf. weitere oder auch weniger Module anbieten, so daß ggf. weitere Menüpunkte oder auch weniger Menüpunkte zur Verfügung stehen.

Typografische Konventionen

Fettschrift Fettschrift bezeichnet Definitionen, Programme und wichtige Textpassagen.
Beispiel: Starten Sie **Word für Windows** ...

Kursivschrift Fette Kursivschrift bezeichnet englische Fachausdrücke oder Menüpunkte.
Beispiel: Wählen Sie *Datei -> Speichern*.

Courier Schreibmaschinenschrift bezeichnet Dateinamen und Datenwerte.
Beispiel: Speichern Sie die Arbeitsdatei unter dem Namen wahl1.sav.

Courier Fette Schreibmaschinenschrift bezeichnet Variablennamen und Vorlagen.
Beispiel: Berechnen Sie die Variable **alter** ...

Namenskonventionen für Dateien

Während der Arbeit mit SPSS können Sie u.a. folgende Typen von Dateien lesen, erstellen, bearbeiten und speichern:

Dateiendung	Bedeutung	Beispiel
* .spo	SPSS Navigator Datei	wahl.spo
* .sav	permanente SPSS Arbeitsdatei	wahl.sav
* .sps	SPSS Syntaxdatei (Programmdatei)	wahl.sps

Tabelle/Abbildung 2: Dateiformate

Hersteller

Weitere Informationen zu SPSS finden Sie beim Hersteller:

SPSS GmbH Software - Rosenheimer Straße 30 - 81669 München - Tel.: (0 89) 48 90 74-0
Fax: (0 89) 448 31 15 - Internet: <http://www.spss.com>

Weiterführende Literatur

In diesem Handbuch werden einige statistische Themengebiete nur angeschnitten, zum anderen ist die Beschreibungen des Menüsystems sehr kurz gehalten. Abhängig von Ihren konkreten Aufgaben benötigen Sie weiterführende Literatur zur Statistik oder Informationen aus der SPSS Dokumentation oder dem integrierten SPSS Hilfesystem. Die Handbücher zum SPSS in gedruckter Form können in einigen Buchhandlungen erworben werden. Eine mögliche (Online) Bestelladresse lautet:

BSB Bücherdistribution

<http://www.bsb.de>

Geben Sie dort unter Suche das Stichwort „SPSS“ ein, um alle SPSS Handbücher mit Preisen aufzulisten.

Miete von SPSS für Windows für Studenten und Mitarbeiter

(Nur für Univ. Osnabrück!)

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

Nähere Informationen finden Sie unter: <http://www.rz.uni-osnabrueck.de>

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.

Funktionsumfang von SPSS für Windows

SPSS für Windows ist ein modular auf-gebautes Statistik-Analyse-System. Es besteht aus dem Basis-System (Base System), das bereits das komplette Daten- und Dateimanagement, sämtliche Grafiktypen und eine breite Palette an statistischen Funktionen umfaßt. Diverse Zusatzmodule erweitern die statistische Leistungsfähigkeit des Base Systems. (zitiert aus einem SPSS Werbeprospekt)

SPSS¹ ist ein umfassendes Programmsystem zum Verwalten und zum statistischen Auswerten von Datenmaterial. Sie können Daten erfassen oder einlesen, Daten bearbeiten, tabellarische Berichte, Diagramme und Plots erzeugen, Kennzahlen berechnen und sowohl einfache als auch komplexe statistische Verfahren auf das Datenmaterial anwenden.

SPSS versetzt Sie in die Lage, die genannten Aufgaben ohne Kenntnis einer Programmiersprache durchzuführen. Grundlage hierfür ist die grafische Benutzeroberfläche von SPSS, die auf einem Menüsystem mit nachgeordneten Dialogboxen sowie einem integrierten Dateneditor beruht.

Sie teilen **SPSS** durch Auswahl eines Menüpunktes im Menüsystem und durch Eingabe von Informationen in nachgeschalteten Dialogboxen mit, welche Aktionen mit den Daten durchgeführt werden sollen.

Für Kenner der SPSS Programmiersprache besteht auch weiterhin die Möglichkeit, SPSS Programme in ein Syntax-Fenster zu laden und aus dem Syntax-Fenster heraus auszuführen. Innerhalb des Menüsystems ermöglicht Ihm die Aktionsschaltfläche Befehl darüberhinaus, die über das Menüsystem ausgewählten Kommandos in das Syntax-Fenster zu übertragen. Es ist in dieser Form z.B. möglich, über das Menüsystem ein "Grundgerüst" (Prototyp) eines SPSS Programmes zu erstellen, das im Anschluß im Syntax-Fenster individuell ergänzt werden kann.

Auf diese Art und Weise können Sie die Vorteile einer grafischen Benutzeroberfläche (u.a. Auswahl über Menüs, Point&Click, Online Hilfe) und die Vorteile einer Programmiersprache (u.a. Reproduzierbarkeit von Ergebnissen bzw. Durchführen von gleichartigen Analysen auf mehreren Datensätzen) verbinden.

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. Eine gelungenen Überblick liefert das auf der SPSS CD mitgelieferte Dokument **SPSS Survey Tips** ([survtips.pdf](#)).

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 Daten-Werte (*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 auf Festplatte
3. (Optional) **Transformieren** der Daten in eine zweckmäßigere Form bzw. **Erzeugen** von neuen Variablen bzw. **Aggregieren** oder **Verschmelzen** von Daten

¹ SPSS war ursprünglich eine Abkürzung für *Statistical Package for the Social Sciences*. Um die Sozialwissenschaften nicht als einziges Anwendungsgebiet auszuzeichnen, soll die Abkürzung in neuerer Zeit für *Superior Performance Software System* stehen.

Überblick über SPSS für Windows

4. (Optional)- **Auswählen** von Fällen und/oder Variablen für die folgende Analysen
5. Tabellarisches **Darstellen** des Datenmaterials und **Berechnen** von Kennzahlen zur Vorbereitung von statistischen Analysen
6. Grafisches Darstellen (**Visualisieren**) der Daten zur Vorbereitung von statistischen Analysen
7. **Analysieren** der Daten mit Verfahren der mathematischen Statistik

Bedienen der Benutzeroberfläche

Die zuvor genannte Vorgehensweise spiegelt sich direkt in der Anordnung der Menüpunkte im SPSS Hauptfenster wider.

Starten Sie SPSS und wählen Sie im Dialog den Punkt: "Eingeben von Daten" aus. SPSS meldet sich dann mit folgendem Hauptfenster:

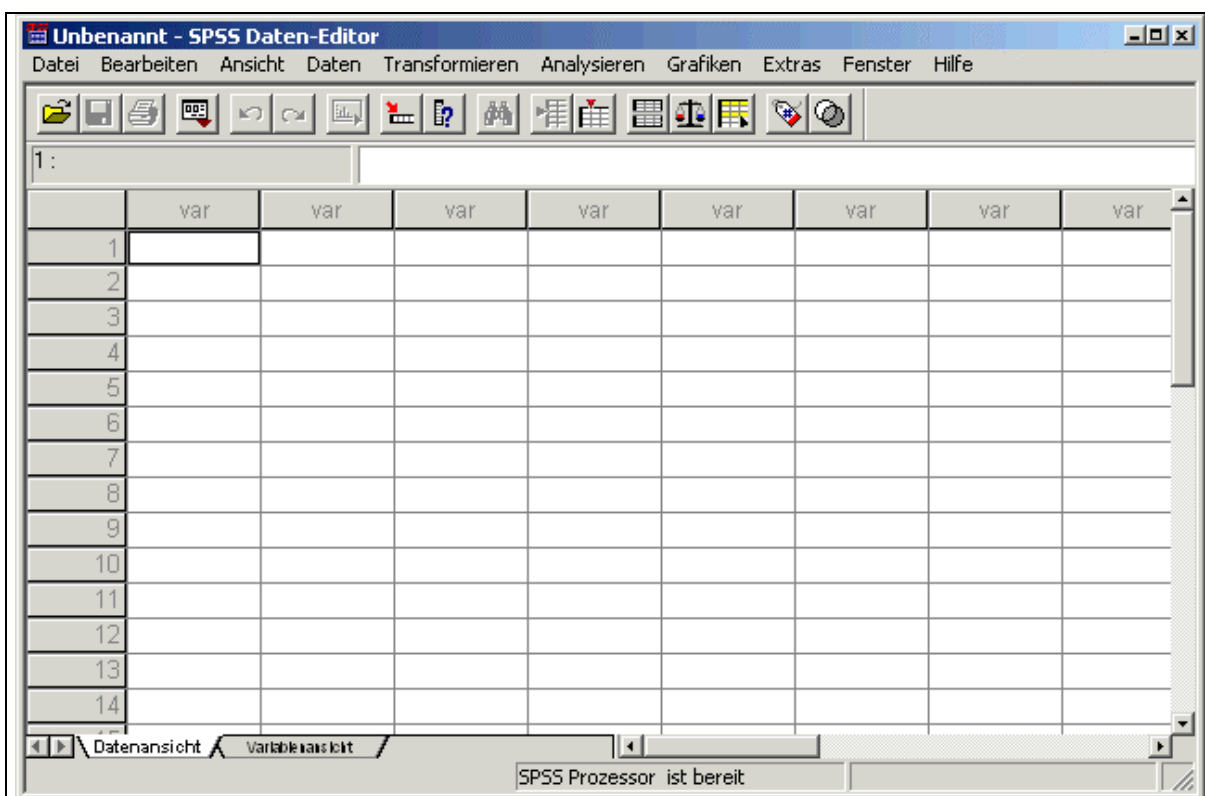


Abb./Tab. 3: Hauptfenster

Sie arbeiten in SPSS zu Beginn einer Sitzung zunächst im Hauptfenster und geben dort Daten ein bzw. laden dort eine bereits vorher erzeugte Arbeitsdatei. Abhängig von den von Ihnen ausgewählten Menüpunkten öffnet SPSS weitere Fenster, um dort z.B. tabellarische Ergebnisse oder Grafiken anzuzeigen oder die Möglichkeit zur Bearbeitung von Grafiken zu geben. Der Wechsel zwischen den Fenstern erfolgt über den Menüpunkt **Fenster**.

Neben dem Hauptfenster ist das Viewer-Fenster von besonderer Bedeutung, in dem alle Ergebnisse in einer baumartigen Struktur angezeigt werden:

Überblick über SPSS für Windows

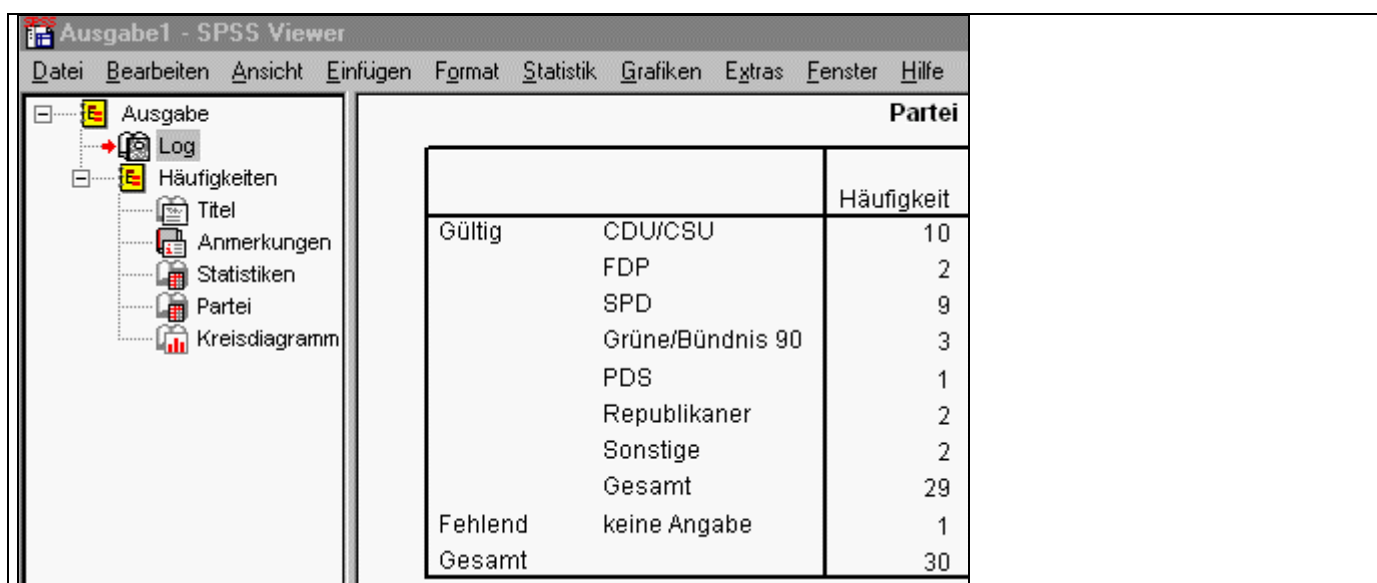


Abb./Tab. 4: Viewer oder Navigator Fenster (Ausschnitt)

Online Hilfe

Die Online Hilfe, die über den Menüpunkt *Hilfe* erreichbar ist, liefert ausführliche Informationen zu SPSS:

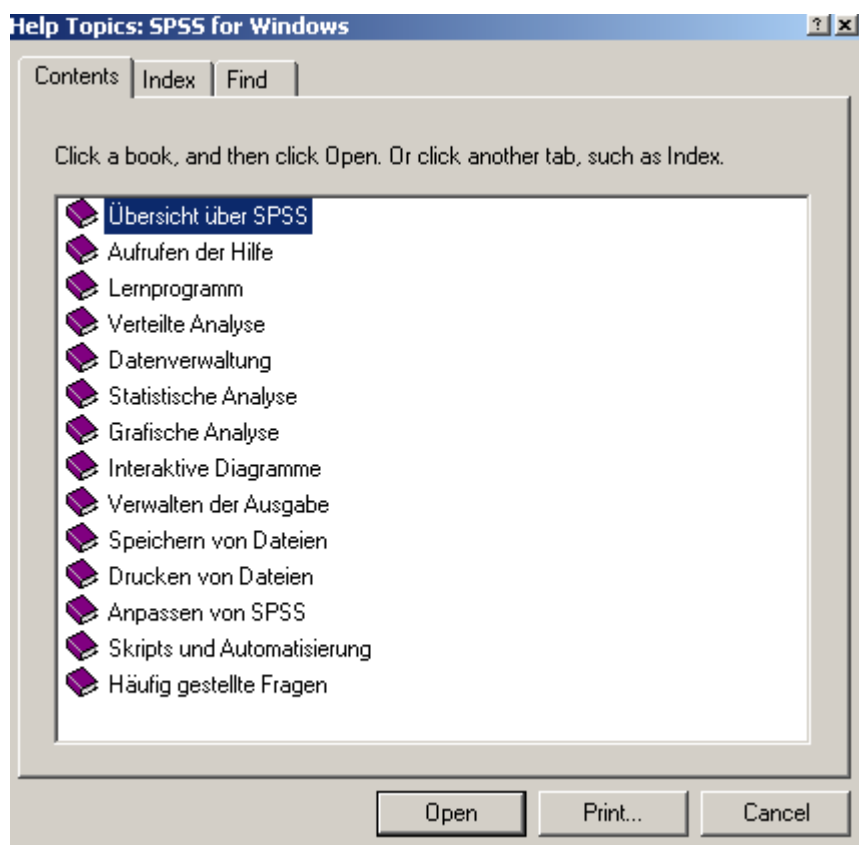


Abb./Tab. 5: Hilfe

Sie können dort u.a. auch ein Lernprogramm aufrufen. Eine ebenfalls gute Einführung liefert die auf der SPSS CD mitgelieferte **Tour durch SPSS**.

Aufgaben

Überblick über SPSS für Windows

1. Machen Sie sich mit der Benutzeroberfläche von SPSS für Windows vertraut, indem Sie zunächst **SPSS für Windows** durch Anklicken des zugehörigen Piktogrammes oder über **Start > Programm > ... > SPSS für Windows** aufrufen.
2. Wählen Sie verschiedene Menüpunkte aus.
3. Vergleichen Sie die Benutzeroberfläche von SPSS mit der von anderen Windows Programmen. Welche Gemeinsamkeiten bzw. Unterschiede können Sie feststellen?
4. Machen Sie sich mit dem Online Hilfesystem von SPSS vertraut. (Der entsprechende Menüpunkt lautet: *Hilfe*)
5. Beenden Sie das Programm über **Datei -> Beenden**.

Definieren von Variablen und Erfassen von Beobachtungen

In diesem Kapitel stehen der SPSS Dateneditor und der Speichern und Laden einer SPSS Arbeitsdatei im Mittelpunkt.

Meine Diplomarbeit basiert auf einem Fragebogen, den ich an 400 Firmen verschickt habe. 250 Fragebögen stehen nun als Rücklauf zur Verfügung. Wie geht es nun weiter ...

Fragebogen - Sonntagsumfrage

In diesem Kapitel wird der folgende Fragebogen "Sonntagsfrage" bearbeitet:

Variable	Frage	Kodierung (Werte-Etiketten)	Antwort (kodiert)
nr	Laufende Nummer des Fragebogens <i>(wird vom Interviewer ausgefüllt)</i>	<i>nnn</i>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
w_o	Befindet sich Ihr Wohnort in den alten oder neuen Bundesländern?	2 (5 neue Bundesländer, "Osten") 1 (alte Bundesländer, "Westen") 0 (keine Angabe)	<input type="checkbox"/>
sex	Geschlecht <i>(wird vom Interviewer ausgefüllt)</i>	1 (weiblich) 2 (männlich) 0	<input type="checkbox"/>
alter	Alter	<i>nnn</i> -1 (keine Angabe)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
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) 0 (keine Angabe)	<input type="checkbox"/>

Tabelle/Abb. 6: Variablen

Die folgende Tabelle enthält 4 Beobachtungen (oder Fälle) für diesen Fragenbogen, wobei hier die Kodierung und (nicht die Werte-Etiketten) angezeigt werden:

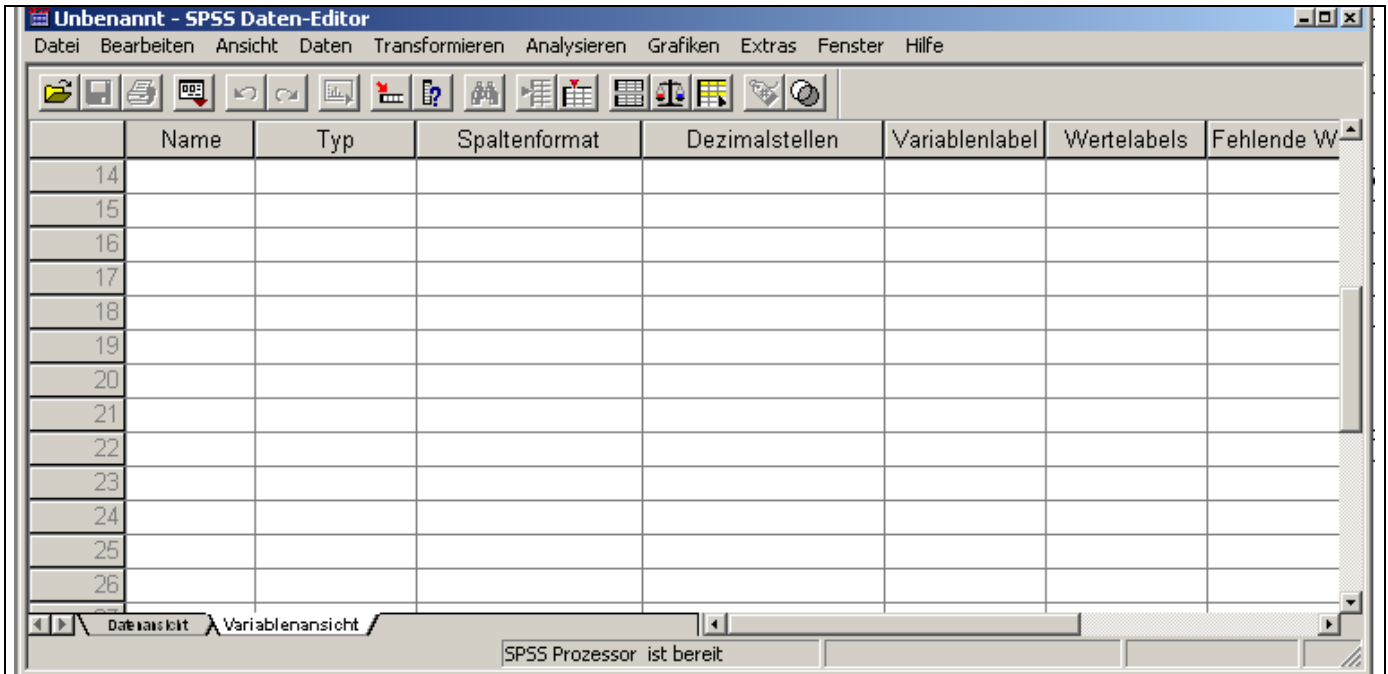
nr	alter	sex	w_o	partei
1	45	1	1	1
2	22	2	1	3
3	19	2	2	3
4	42	1	2	1

Tabelle/Abb. 7: Beobachtungen aus wahl.sav (Ausschnitt)

Überblick - SPSS Dateneditor

Der SPSS Dateneditor ist ein arbeitsblattähnliches Fenster mit Zeilen und Spalten zum Erfassen von neuem Datenmaterial in eine SPSS Arbeitsdatei und zum Laden, Bearbeiten und Speichern von vorhandenen SPSS Arbeitsdateien.

Definieren Sie im ersten Schritt die benötigten Variablen. Wählen Sie hierzu durch Klicken auf die entsprechende Reiter-Schaltfläche die Ansicht "Variablenansicht".



Tabelle/Abb. 8: Variablenansicht

Geben Sie nun zeilenweise für jede Variable die gewünschte Definition ein:

1. **Variablenname**
voreingestellt sind die Namen `var0001`, usw.;
zu verwenden sind: maximal 8 Zeichen, 1. Zeichen Buchstabe, danach Buchstaben, Ziffern und einige Sonderzeichen wie \$ oder _
2. **Typ**
voreingestellt sind Dezimalzahlen mit 8 Stellen, hiervon 2 Dezimalstellen;
z.B. Zahl, Datum, Währung, Zeichenkette
3. **Datenformat des Typs**
z.B. 8.2 für numerische Variablen steht für 8 Zeichen Breite und 2 Nachkommastellen und ermöglicht die Darstellung von Zahlen im Bereich von -99999.99 bis 99999.99
4. **(benutzerdefinierte) fehlende Werte**
z.B. ein einzelner Wert wie Null oder ein Bereich von ungültigen Werten²
5. **Labels (Etikett für den Variablenname)**
sprechende Bezeichnung, *variable label*, (z.B. `Lebensalter` statt `alter`)
6. **Labels (Etiketten für einzelne Werte)**
sprechende Bezeichnungen, *value labels*;
z.B. "ungenügend" für 6 oder "weiblich" für 1; die Werte werden in einem kleinen Dialogfenster eingegeben und dann mit "Hinzufügen" übernommen
7. **Spaltenformat (Spaltenbreite und Spaltenausrichtung bei der Ausgabe)**
z.B. 8 Zeichen, rechtsbündig

² Für numerische Variablen ist der Punkt (.) der systemdefinierte fehlende Wert. Für Zeichenketten existiert kein systemdefinierter fehlender Wert, da jede Zeichenkette (auch die leere Zeichenkette) prinzipiell einen gültigen Datenwert darstellen könnte. Falls Sie mit benutzerdefinierten fehlenden Werten arbeiten wollen, empfiehlt sich daher grundsätzlich eine numerische Kodierung.

Definieren von Variablen und Erfassen von Beobachtungen

8. Meßniveau

metrisch, ordinal, nominal

Für die Variable `sex` ergibt sich folgender Eintrag:

	Name	Typ	Spaltenformat	Dezimalstellen	Variablenlabel	Wertelabels	Fehlende Wert	Spalten	Ausrichtung	Meßniveau
1	sex	Numeri	1	0	Geschlecht	{1, maennlich}	0	8	Rechts	Nominal

Tabelle/Abb. 9: Definition einer Variablen

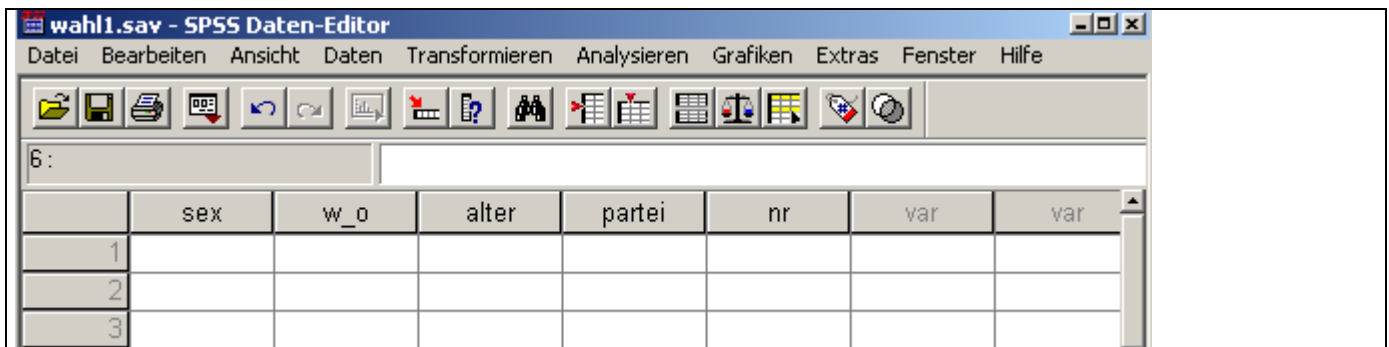
Speichern Sie sämtliche Informationen über den Menüpunkt **Datei > Speichern** in die Datei `wahl1.sav`; d.h. die Definition der Variablen (und später auch die erfaßten Datenwerte) aus dem temporären Arbeitsblatt werden in eine permanente SPSS Arbeitsdatei `wahl1.sav` auf Festplatte gespeichert, damit Sie sie auch in späteren SPSS Sitzungen erneut verwenden können.

Laden Sie die zuvor gespeicherte Datei über **Datei > Öffnen > Daten**.

Vorgehensweise - Eingeben von Beobachtungen im Dateneditor

Geben Sie nun im zweiten Schritt die Daten für die zuvor definierten Variablen ein.

Wechseln Sie hierzu in die Datenansicht. Geben Sie direkt im SPSS Dateneditor (fiktive) Datenwerte für die ersten 4 Beobachtungen ein.



	sex	w_o	alter	partei	nr	var	var
1	1						
2							
3							

Abb./Tab. 10 : Datenwerte (Ausschnitt)

Zusammenfassung - Welche Informationen werden über eine Variable gespeichert?

Die Daten sind in Form einer Matrix organisiert, die aus Zeilen (Beobachtungen) 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.

Jede **Zeile** der Tabelle stellt eine **Beobachtung** (*observation*) oder einen **Fall** (*case*) dar. 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. Variablen werden über **Variablennamen** bezeichnet, die in einer Tabelle typischerweise als Titelzeile (Spaltenüberschrift) verwendet werden.

Motivation - Einlesen von anderen Dateiformaten

Sie können das Datenmaterial in einem Tabellenkalkulationsprogramm wie **Excel für Windows** oder einem Datenbankprogramm wie **dBase** oder **Access** erfassen und in dem spezifischen Format des Programmes abspeichern.

Definieren von Variablen und Erfassen von Beobachtungen

Sie können nun mit SPSS Datenmaterial in vielen Datei-Formaten einlesen (importieren) und im Anschluß in das SPSS Format (.sav Format) abspeichern.

Einen Überblick über die möglichen Import-Formate liefert der Menüpunkt **Datei > Öffnen** bzw. **Datei > Datenbankzugriff** und dort die diversen Unterpunkte.

Aufgaben

1. Definieren Sie die Variablen für die Arbeitsdatei `wahl1.sav` wie im Fragebogen vorgegeben, erfassen Sie 4 (frei erfundene!) Beobachtungen im SPSS Dateneditor und speichern Sie die Arbeitsdatei unter dem Namen `wahl1.sav` ab
2. Geben Sie weitere Beobachtungen ein, wobei Sie einige Zellen für numerische und alphanumerische Werte freilassen (systembedingte fehlende Werte). Wie werden fehlende Werte dargestellt?
Zeigen Sie kodierte Werte (Zahlen) an und alternativ Werte-Label (Zeichenketten) an.
Hinweis: Menüpunkt: *Ansicht > Wertelabels*.
3. (*) Definieren Sie die Variablen `partei_1` bis `partei_5` (`partei_1` entspricht CDU/CSU usw.). Definieren Sie für diese Variablen Werte-Etiketten für Sympathiewerte ("Schulnoten") von 1 bis 6 mit den Etiketten "sehr gut" bis "ungenugend", -1 als benutzerdefinierter fehlender Wert) als Bewertung für die Parteien.
4. Welche Gründe sprechen für die Vergabe von Etiketten für Variablennamen und Werte?
Hinweis: Fassen Sie z.B. die Möglichkeit in Betracht, daß jeweils eine Auswertung für englisch- und deutschsprachiges Publikum benötigt wird und skizzieren Sie die Vorgehensweise bei "harter Kodierung" von z.B. Ja/Nein und "weicher numerischer Kodierung" mit 0/1 und Verwendung von Etiketten wie z.B. "Yes/No" oder "Ja/Nein".
5. (**) Erzeugen Sie eine Access Datenbank `sonntagsfrage.mdb` mit einer entsprechenden Tabelle `fragebogen` und lesen Sie die Daten über die ODBC Schnittstelle ein (Menüpunkt: **Datei > Datenbankzugriff**).

Berechnen neuer Variablen und Auswählen von Beobachtungen

In diesem Kapitel werden Methoden beschrieben, um neue Variablen zu erzeugen bzw. um die folgenden Auswertungen auf eine Auswahl von Beobachtungen einzuschränken. Das Aggregieren und Verschmelzen von Daten wird in einem der folgenden Kapitel behandelt.

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

Motivation – Hinzufügen oder oder Löschen von Beobachtungen/Variablen

Sie setzen eine SPSS Arbeitsdatei aus **Variablen** (Spalten, vertikal) und **Beobachtungen** (Zeilen, horizontal) zusammen, die insgesamt ein rechteckiges oder matrixartiges Schema ergeben. Sie können eine SPSS Arbeitsdatei deshalb grundsätzlich auf 2 Arten erweitern (oder verkleinern):

Hinzufügen von Variablen (hier: x neue Spalte)	Hinzufügen von Beobachtungen (hier: x neue Zeile)
<pre> x x x x </pre>	<pre> x x x x x </pre>

Abb./Tab. 11: Erweitern einer Datenmatrix

Darüberhinaus ist es möglich, einzelne Zeilen anhand vorgegebener Bedingungen zu löschen bzw. "auszufiltern".

Beispiele: Hinzufügen von Variablen:

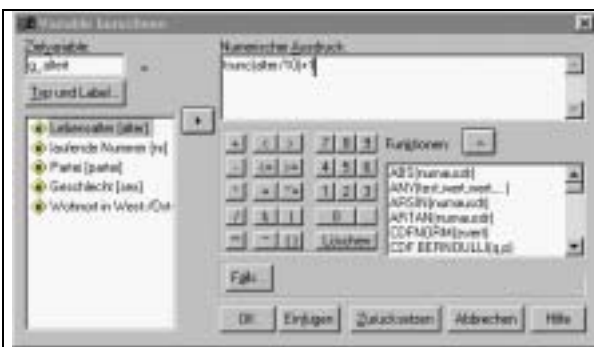
- Sie beziehen eine neue Variablen in die Untersuchung ein, die sich aus den vorhandenen Variablen ableiten lassen. Als Beispiel für eine abgeleitete Variable dient der **Broca-Index**, der sich als Quotient aus Körpergewicht und Normalgewicht in Prozent berechnet.

Beispiele: Hinzufügen oder Filtern/Löschen von Beobachtungen:

- Sie führen eine **Bedingung (Filter, Selektion)** für Beobachtungen ein, die an den weiteren statistischen Auswertungen teilnehmen sollen. Als Beispiel für eine Bedingung dient der Wertebereich "Alter zwischen 20 und 40" einer Variablen `alter`.

Vorgehensweise - Berechnen neuer Variablen

Im folgenden Beispiel berechnen Sie auf Grundlage der Arbeitsdatei `wah1.dat` eine neue Variable `g_alter` (Altersgruppe). Wählen Sie **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.

Hierzu stehen Ihnen arithmetische Operatoren, logische Ausdrücke und Funktionen zur Verfügung, hier soll folgender Ausdruck verwendet werden:

```
trunc(alter/10)+1
```

Berechnen neuer Variablen und Auswählen von Beobachtungen

SPSS berechnet eine neue Variable `g_alter`, die sofort im Dateneditor angezeigt wird. Ärgerlicherweise werden nur Datenwerte für die vorhandenen (!) Beobachtungen berechnet, bei später hinzugefügten oder auch bei modifizierten Beobachtungen wird nicht mehr automatisch eine Aktualisierung durchgeführt – berechnete Variablen sollten also möglichst nach vollständiger Eingabe aller Beobachtungen erzeugt werden!

Das Hilfesystem enthält detaillierte Informationen zu den verfügbaren Funktionen, die zur Berechnung neuer Variablen verwendet werden können:

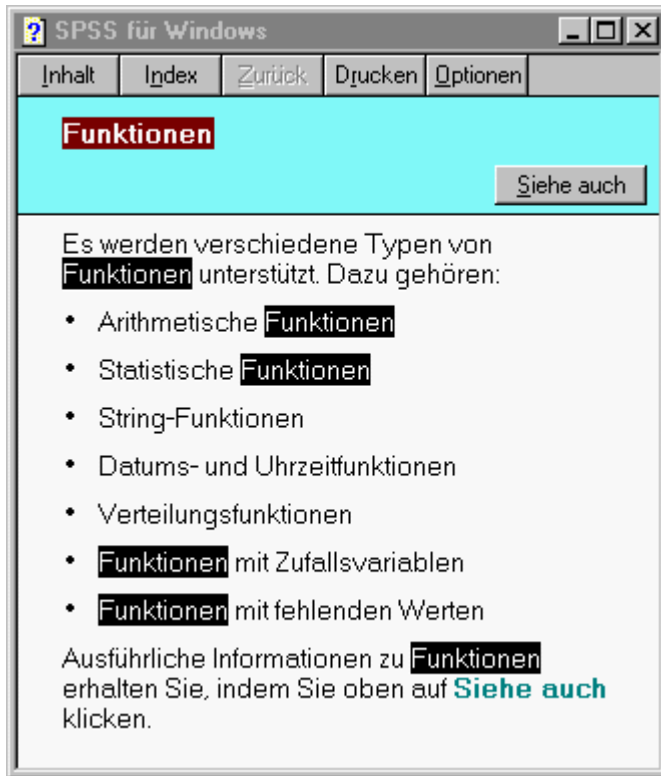


Abb./Tab. 12 : Funktionen



Es ist sinnvoll, als Variablen-Label einer berechneten Variablen den zugrundeliegenden numerischen Ausdruck zu verwenden.

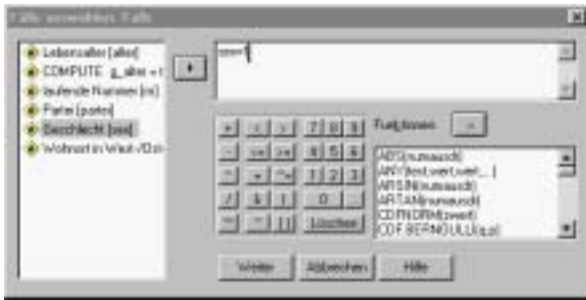
Somit ist es jederzeit möglich, für eine berechnete Variable die zugehörige Rechen-Vorschrift nachzuschlagen.

	sex	alter	partel	m	w_e	g_alter
1	1	45	1	1	1	COMPUTE g_alter = (sex(alter)/2)+1 (COMPU
2	2	22	3	2	1	3,00
3	2	19	3	3	1	2,00
4	1	42	1	4	1	5,00
5	2	34	4	5	1	4,00
6	1	72	2	6	1	8,00

Beim Anklicken der Spaltenüberschrift wird der zugehörige Label angezeigt.

Vorgehensweise - Filtern von Beobachtungen

Im folgenden Beispiel führen Sie in `wahl.sav` eine Filterung nach Frauen (Bedingung: „`sex=1`“) durch. Wählen Sie **Daten -> Fälle auswählen**.



Wählen Sie durch Angabe der Bedingung „`sex=1`“ nur die Beobachtungen aus, bei denen es sich bei der befragten Person um eine Frau handelt.

Im Dateneditor existiert nun eine neue Variable `filter_$` mit den Werten 0 (ausgeschlossen) und 1 (beteiligt).

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

Alle ausgefilterten, also in den folgenden Auswertungen nicht mehr berücksichtigten Beobachtungen sind im Dateneditor am linken Rand durch eine durchgestrichene Nummer gekennzeichnet (hier: 2,3,5).

Aufgaben

- Definieren Sie für die Datei `wahl.sav` über **Transformieren -> Berechnen** eine neue Variable `dekade`, die eine Altersgruppe "Dekade" repräsentiert.
Hinweis: Berechnen Sie `dekade` folgendermaßen mit der Rundungsfunktion `trunc` (*truncate*: abschneiden):

```
dekade = trunc(alter/10)
```

- Erläutern Sie das Ergebnis folgender Transformation in der Arbeitsdatei `wahl.sav` (neu berechnete Variable `gruppen`).

Hinweise:

Das Ergebnis eines Vergleiches ist entweder 0 für falsch oder 1 für wahr, d.h. falls eine Beobachtung für die Variable `a` den Wert 20 besitzt, liefert `a < 40` als Ergebnis 1. Eine UND-Verknüpfung wird durch `(...) & (...)` definiert, eine ODER-Verknüpfung durch `(...) | (...)`

```
gruppen= 1 * ((0 < alter) & (alter <= 20)) +
         2 * ((20 < alter) & (alter <= 40)) +
         3 * ((40 < alter) & (alter < 150))
```

- Selektieren Sie in der Arbeitsdatei `wahl.sav` Beobachtungen, für die folgende Bedingungen gelten:

- Befragte Person (P) ist zwischen 40 und 60 Jahre alt.
- P ist weiblich und älter als 60.
- P ist jünger als 25, männlich und würde die Republikaner wählen.
- P würde CDU/CSU oder FDP wählen.
- P wohnt in den alten Bundesländern.

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

Hinweis

: Eine UND-Verknüpfung wird durch `(...) & (...)` definiert, eine ODER-Verknüpfung durch `(...) | (...)`.

Berechnen neuer Variablen und Auswählen von Beobachtungen

4. Berechnen Sie für `broca.sav` den **Broca-Index** $bi=100*gew/(groesse-100)$ aus den Variable `gew` (Gewicht) und `gr` (Körpergröße).
5. (*) Berechnen Sie den Body-Mass-Index (BMI) in Ergänzung zum Broca-Index. Der Wert berechnet sich gemäß: $bmi = \text{Weight}(\text{kg}) / [\text{Height}(\text{m}) * \text{Height}(\text{m})]$. Berechnen Sie die Variable `bmi_gr` gemäß der folgenden Tabelle:

Werte- Etikett	Werte	bmi
untergewichtig	-1	unter 18.5
normal	0	18.5 - 24.9
übergewichtig	1	25.0 - 29.9
stark übergewichtig	2	30.0 und mehr

Arbeiten mit Datums-Variablen

In diesem Kapitel wird speziell auf den Datentyp `DATUM` eingegangen.

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

Motivation - Darstellen von Datums- und Zeitangaben

Sie benötigen Datums- oder Zeitangaben, um z.B. das Datum einer Messung in eine Arbeitsdatei aufzunehmen und dann zeitliche Abstände zu weiteren Messungen zu berechnen.

Eine Datumsangabe wird von SPSS intern in Sekunden seit dem **14. Oktober 1582** gespeichert, d.h. als numerischer Wert. Entsprechend wird eine Zeitangabe als numerischer Wert in Sekunden seit 0:00 Uhr intern gespeichert.

```

<-----|----->
14.10.1582 (willkürlich gewählter Nullpunkt)
0 [Sekunden] interne Darstellung in SPSS

<-----|----->
01.02.1995
interne Darstellung in SPSS: 13010976000 [Sekunden]

<-----|----->
01.06.1966
interne Darstellung in SPSS: 12100924800 [Sekunden]
    
```

Abb./Tab. 13 : Interne Darstellung eines Datums

Die Eingabe und Darstellung von Datums- und Zeitwerten vom Datentyp `DATUM` im SPSS Dateneditor, also in lesbarer Form) kann in verschiedenen Datenformaten erfolgen wie z.B.:

```

02.01.1995
2. Jan 1995
1995/1/2
    
```

Die **interne** Darstellung dieses Datums, nämlich in Sekunden seit dem Stichtag, für das Beispiel-Datum **13010976000**, ist immer gleich

Die Festlegung des Datumsformats erfolgt bei der Definition einer Variablen in der Auswahlliste, hierbei steht `tt` für Tag, `mmm` für Monat als Buchstaben-Abkürzung (`JAN`, `FEB`), `mm` für Monat als Zahl (01, 02) und `jj` für eine 2-stellige Jahreszahl und `jjjj` für eine 4-stellige Jahreszahl:

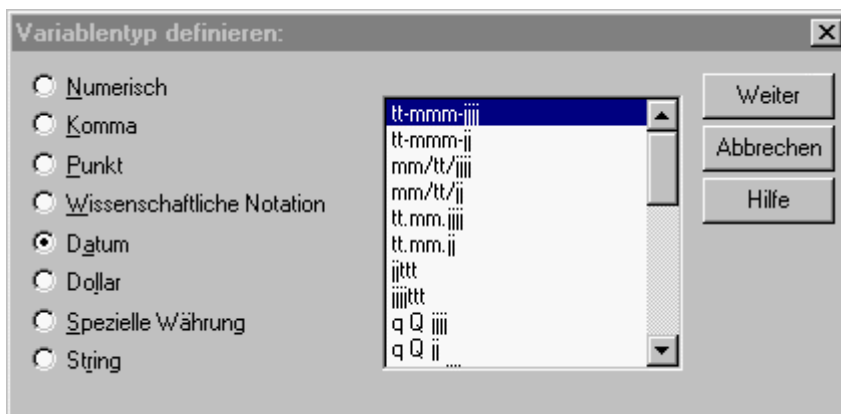


Abb./Tab. 14 : Auswahlliste für Darstellung eines Datums

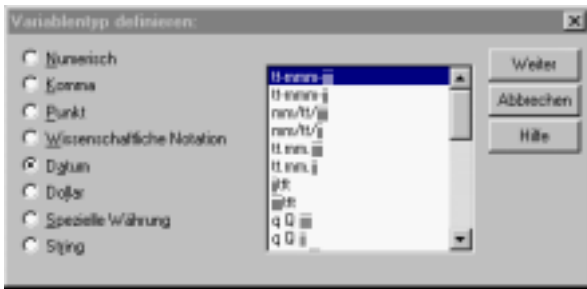
Überblick – Funktionen für Variablen vom Datentyp DATUM

Für Variablen des Datentyps `DATUM` stehen spezielle Funktionen zur Verfügung, mit denen Sie z.B. Datums- oder Zeitdifferenzen berechnen oder Wochentage bestimmen können. Die folgende Tabelle enthält eine kleine Auswahl:

Funktion	Berechnet ...	Beispiel
<code>CTIME.DAYS(date)</code>	Anzahl Tage seit dem 14.10.1582, benötigt als Argument eine Datums-Variable	<code>CTIME.DAYS(15.10.1582)=0</code>
<code>YRMODA(year,month,day)</code>	Anzahl Tage seit dem 14.10.1582, benötigt als Argument 3 numerische Variablen	<code>YRMODA(1582,10,15)=0</code>

Vorgehensweise - Definieren von Variablen vom Datentyp DATUM

In diesem Beispiel definieren Sie eine Variable mit dem Datentyp `DATUM`: Laden Sie die Arbeitsdatei `tank.sav`. Wählen Sie zunächst den Reiter "Variablenansicht" und definieren Sie dann eine neue Variable `datum_d` über **Daten > Variable einfügen**.



Wählen Sie aus der Liste der Datums- und Zeitformate das gewünschte Format aus, z.B. `tt.mm.yyyy` für Daten der Form `04.08.1984`. oder `hh:mm` für Zeitangaben der Form `13:19`.

Beachten Sie, daß 3-stellige Monatsnamen (mmm) auf Englisch bezeichnet werden müssen (also z.B. `Dec` statt `Dez`).

Wechseln Sie in die Datenansicht und geben Sie einige Werte für diese neue Variable ein.

Vorgehensweise - Berechnen neuer Variablen vom Datentyp DATUM

Im folgenden Beispiel werten Sie Daten über das Betanken eines Autos aus. Die Daten sind in einer Arbeitsdatei `tank.sav` festgehalten und zwar Datum (`tag`, `monat`, `jahr`), Kilometerstand (`kmstand`) und getankte Benzinmenge (`tank`) für jede Betankung.

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

Berechnung der neuen Variablen	Bedeutung
<code>ntage=yrmoda(jahr,monat,tag)</code>	<code>ntage</code> ist die Anzahl Tage seit dem 14.10.1582.
<code>difftage=ntage-lag(ntage,1)</code>	<code>difftage</code> ist die Differenz zwischen zwei aufeinanderfolgenden Betankungen in Tagen. <code>lag(var,1)</code> liefert den Wert der Variablen <code>var</code> für die vorhergehende (!) Beobachtung.
<code>diffkm=kmstand-lag(kmstand,1)</code>	<code>diffkm</code> ist die Differenz zwischen zwei aufeinanderfolgenden Kilometerständen.
<code>verbr=tank*100/diffkm</code>	<code>verbr</code> ist der Verbrauch zwischen zwei Betankungen (in liter / 100 km).
<code>kmtag=diffkm/difftage</code>	<code>kmtag</code> ist die durchschnittlich zurückgelegte Strecke pro Tag.

Abb./Tab: 15:Berechnungsformeln für neue Variablen

Die SPSS Arbeitsdatei enthält nun weitere Variablen und zugehörige Werte (Auszug):

tag	monat	jahr	kmstand	tank	Ntage	difftag e	diffkm	verbr	kmtag
16	12	1992	20580	60.3	149813
23	12	1992	21250	57.4	149820	7.00	670.00	8.57	95.71
4	1	1993	21874	56.6	149832	12.00	624.00	9.07	52.00

Abb./Tab. 16: Arbeitsdatei mit neuen Variablen (Ausschnitt)

Aufgaben

- Definieren Sie in einer leeren Arbeitdatei **nds.sav** die Variablen **beginn** und **ende** mit dem Datentyp **DATUM**, geben Sie jeweils Beginn und Ende der Schulferien in Niedersachsen im Jahr 1995 ein (siehe Tabelle unten) und berechnen Sie in der Variablen **diff1** die Anzahl Ferientage für jede Ferien :

$$\text{diff1} = \text{CTIME.DAYS}(\text{ende}) - \text{CTIME.DAYS}(\text{beginn}) + 1$$

- (*) Definieren Sie Beginn und Ende der Ferien über insgesamt 6 Variablen (jeweils Jahr, Monat, Tag) und berechnen Sie die Anzahl Ferientage in der Variablen **diff2** folgendermaßen:

$$\text{diff2} = \text{YRMODA}(\text{ende}_y, \text{ende}_m, \text{ende}_d) - \text{YRMODA}(\text{beg}_y, \text{beg}_m, \text{beg}_d) + 1$$

- (*) Berechnen Sie in der Variablen **diff3** 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 (ergo: die Nicht-Ferientage, natürlich mit Sonntagen und Feiertagen):

$$\text{diff3} = \text{CTIME.DAYS}(\text{beginn}) - \text{CTIME.DAYS}(\text{LAG}(\text{ende}, 1)) - 1$$

- (*) Kontrollieren Sie die Gesamtsumme (1995 war kein Schaltjahr, hatte also 365 Tage).

Tabelle:

Es galten für 1995 in Niedersachsen folgende Ferientermine (siehe hierzu auch die Arbeitdatei **ferien.sav**):

01-JAN-1995	07-JAN-1995	Weihnachten
03-APR-1995	19-APR-1995	Ostern
06-JUN-1995	06-JUN-1995	Pfingsten
22-JUN-1995	02-AUG-1995	Sommer
02-OCT-1995	14-OCT-1995	Herbst
23-DEC-1995	31-DEC-1995	Weihnachten

Aggregieren (Zusammenfassen) von Daten in eine neue SPSS Datei

In diesem Kapitel wird das Zusammenfassen von gleichartigen Beobachtungen zu einer einzigen Beobachtung behandelt.

Ich habe „zu viele Beobachtungen“ und möchte in einer neuen Arbeitsdatei nur noch mit aggregierten Daten arbeiten.

Vorgehensweise – Zusammenfassen von Beobachtungen

Im folgenden Beispiel wird eine Arbeitsdatei **angst.sav** behandelt, in der die (fiktive) Stärke von Angstzuständen von Personen über einen Zeitraum dokumentiert ist:

Nr. der Beobachtung	Person	Tag_Nr	Angstzustand
1	1	1	1
2	1	2	2
3	1	3	3
4	2	1	3
5	2	2	4
6	2	3	5
7	3	1	4
8	3	2	5
9	3	3	6

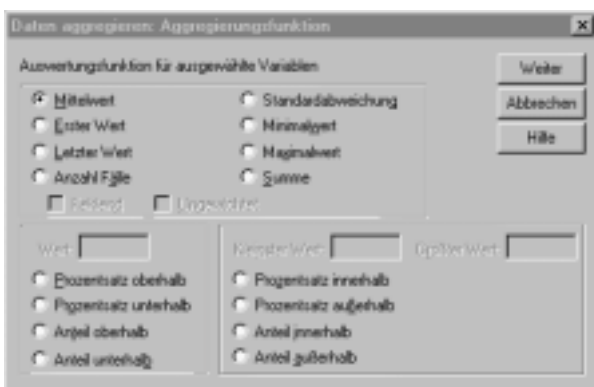
Da der zeitliche Verlauf der Angstzustände nicht weiter interessieren soll, wird als neue Variable nur noch der Mittelwert der Angstzustände jeder Person in der Arbeitsdatei auftauchen. Verwenden Sie als **Break-Variable** **person** und als Aggregierungsmethode den Mittelwert über die Variable **angst**. Wählen Sie **Data-> Aggregate**.



Wählen Sie als **Break-Variable** diejenige Variable aus, für die Daten akkumuliert werden sollen, hier die Variable **person**.

Wählen Sie als Aggregierung den Mittelwert der Variablen **angst** über alle Beobachtungen der selben Person.

Erzeugen Sie eine neue Arbeitsdatei **angr1.sav**.



Es ist über die Schaltfläche „*Funktion*“ möglich, andere Aggregierungsfunktionen zu wählen, z.B. den Minimal- oder Maximal-Wert oder die Summe.

Anmerkung:

Durch das Aggregieren geht im vorliegenden Fall die Information über die zeitliche Entwicklung von Angst verloren; Aggregieren bedeutet in jedem Fall einen Verlust von Information.

Aggregieren (Zusammenfassen) von Daten in eine neue SPSS Datei

Laden der aggregierten Daten

Überprüfen Sie nun die Aggregation, indem Sie die neu erstellte Arbeitsdatei öffnen über **File -> Open** `aggr1.sav`

person	angst_1	n_break
1,00	2,00	3
2,00	4,00	3
3,00	5,00	3

SPSS erzeugt eine neue Datei `aggr1.sav` und dort eine neue Variable `angst_1`, die für jeden Probanden (Break Variable: Person) den Mittelwert von Angst über alle Tage bildet.

Die Variable `tag_nr` taucht entsprechend in der neuen Arbeitsdatei nicht mehr auf, weil sie beim Aggregieren „herausgefallen“ ist. Die neue Variable `n_break` ist die Anzahl der Beobachtungen, die bei der Mittelwertbildung berücksichtigt wurden.

Aufgaben

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 dunkel, daß ich in Veranstaltungen zur Statistik etwas über Median, Mittelwert und Standardabweichung gehört habe ...

Aufgaben der deskriptiven Statistik

Die **deskriptive Statistik** (beschreibende Statistik) befaßt sich mit der tabellarischen und grafischen Darstellung von Daten und 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.

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

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

In der Regel werden von einem Objekt oder einer Person mehrere Variablen gleichzeitig beobachtet, so daß eine Beobachtung aus mehreren Variablen besteht. Z.B. könnten bei einer Person gleichzeitig die zufälligen Variablen Größe, Gewicht, Alter und Geschlecht beobachtet werden; d.h. eine Beobachtung \mathbf{x} einer Person setzt sich aus vier Datenwerten zusammen:

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

Die Beobachtung \mathbf{x} wird dann als **vektoriell** oder **multivariat** bezeichnet. Im Gegensatz hierzu wird eine Beobachtung, die aus nur einer Variablen besteht, bzw. eine Analyse, die nur eine Variable berücksichtigt, als **univariat** bezeichnet. Ein häufiger Spezialfall sind **bivariate** Auswertungen, die sich auf 2 Variablen beziehen. Zusammenhänge zwischen 3 und mehr Variablen lassen sich nur mit Mühe tabellarisch und grafisch darstellen.

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).

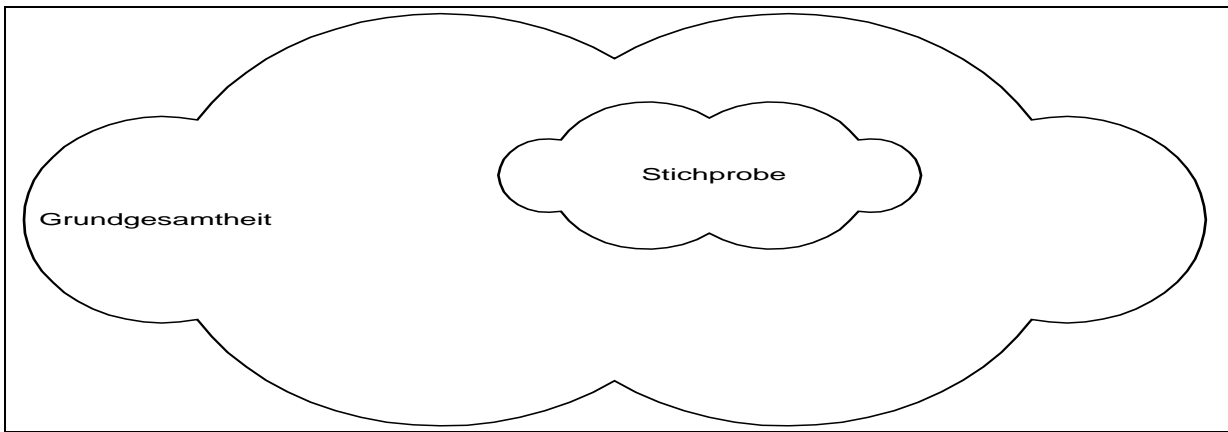


Abb./Tab. 17: Grundgesamtheit und Stichprobe

Eine **Stichprobe** S wird u. a. aus folgenden **Gründen** durchgeführt:

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

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

Beispiele für Stichproben sind:

- repräsentative Umfragen vor Bundestagswahlen
- Mikrozensus
- zufälliges Entnehmen von Wasserproben
- Testen einzelner Blitzlichter (Hierbei wird das Testobjekt zerstört!)
- Markieren einzelner Vögel und Beobachten ihrer Lebensweise
- Messen der Nutzung von Fernsehkanälen in ausgesuchten Haushalten

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

Messung von Variablen

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

Eine Variable ist **nominal-skaliert** (klassifizierend, kategorisierend, gruppenbildend), wenn sie die Stichprobe in disjunkte "Kategorien", "Teilmengen", "Klassen" oder "Gruppen" einteilt, wobei die möglichen Werte (Ausprägungen) keine Ordnung besitzen. Beispiele sind die Variablen **sex** (Geschlecht), die die Stichprobe in die Gruppen "Männer" und "Frauen" zerlegt, oder **nation** (Nationalität), die eine Stichprobe nach Staatsangehörigkeit untergliedert. Ein Spezialfall sind **dichotom-skalierte** Variablen mit genau 2 möglichen Werten.

Überblick über die deskriptive Statistik

Bei einer **nominalen Variablen** werden die möglichen Werte zur Abkürzung willkürlich auf Zahlen abgebildet. Die Zahl hat keine andere Bedeutung als daß sie stellvertretend für einen Wert steht. Z.B. können die Bundesländer mit Zahlen von 1 bis 16 durchnummeriert werden, wobei die Zahl 1 für das Bundesland Berlin steht, 2=Brandenburg usw. Es ist deshalb z.B. unsinnig, für nominal gemessene Variablen einen Mittelwert zu bilden oder ein Histogramm zu erzeugen.

Bei einer **dichotomen Messung** wird nur zwischen zwei möglichen Werten einer Variablen unterschieden. Werte für Variablen, die dichotom gemessen werden, werden häufig zur Abkürzung mit den Zahlen 0 und 1 dargestellt (kodiert), wobei die Zahlen selbst keine inhärente Bedeutung haben, sondern einfach nur einen von zwei möglichen Werten repräsentieren. Beispiele sind {1=Wahr, 0=Falsch} oder {1=Krank, 0=Gesund}. Die Unterscheidung kann dabei auch willkürlich gewählt werden, z.B. {0=Älter als 18, 1=Jünger als 18}.

Eine Variable ist **ordinal-skaliert** (ordnend, quantitativ), wenn ihr Wertebereich aus Zahlen besteht, zwischen denen eine natürliche (Rang-) Ordnung existiert. Beispiele sind Schulnoten oder Sympathie-Werte für Politiker. Ordinal-skalierte Variablen sind auch zur Klassifikation einsetzbar. Eine feinere Unterscheidung ist möglich.

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. Z.B. 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 Rang-Ordnung ist allerdings zu vage, um behaupten zu können, daß die Abstände zwischen den möglichen Ausprägungen gleichbedeutend sind.

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

Bei der **Messung** von Variablen sind also folgende Wertebereiche (Skalen) und Bedeutungen der möglichen Werte zu unterscheiden, zusätzlich ist zwischen diskreten und kontinuierlichen Wertebereichen zu differenzieren:

Typ der Messung	Wertebereich
dichotome Messung	{0,1} oder ähnlich, keine Rang-Ordnung, nur zur Klassifikation verwendbar
nominale Messung	{0,1,2,...,n} oder ähnlich, keine Ordnung, nur zur Klassifikation verwendbar
ordinale Messung	{0,1,2,...,n}, [a,b] oder ähnlich, nur Ordnung
Metrische Messung	[a,b], Ordnung, Differenzen aussagekräftig

Abb./Tab. 18: Levels of Measurement

Einige Verfahren der deskriptiven oder mathematischen Statistik können nur angewendet werden, wenn gewisse Voraussetzungen bzgl. der Skalierung (*levels of measurement*) vorliegen. Es ist z.B. sinnlos, den Mittelwert von nominal-skalierten Variablen zu berechnen oder die Korrelation zwischen zwei dichotom-skalierten Variablen. Ordinal-skalierte Variablen sollten durch eine Rang-Transformation auf Ränge abgebildet werden.

Kenngrößen von Stichproben

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

Überblick über die deskriptive Statistik

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

Wichtige **Kenngößen** der Stichprobe sind für eine numerische Variable im folgenden aufgelistet. Bekannt sind auch die Bezeichnungen **Lagemaße** für Mittelwert und emp. Median und **Streuemaße** für emp. Varianz und emp. Standardabweichung der Stichprobe. Der Zusatz (emp.) für empirisch soll jeweils verdeutlichen, daß es sich um eine Kenngröße der Stichprobe handelt, die sich von der Kenngröße der Grundgesamtheit unterscheidet.

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

$$\bar{x} = (x_1 + \dots + x_n) / n$$

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:

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

Die **empirische Standardabweichung** (*emp. standard deviation, stdev*) s der Stichprobe S ist definiert als Wurzel aus der empirischen Varianz:

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

Die **geordnete Stichprobe** (*ordered sample*) enthält die nach aufsteigender Reihenfolge geordneten Werte x_1, \dots, x_n der Stichprobe. 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 Maximum und Minimum 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 n gerade) und ist definiert als:

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

Die **relative Häufigkeit** (*relative frequency*) beschreibt die Anzahl der Beobachtungen x_i , die gleich einem vorgegebenen Wert x sind, in Relation zur Gesamtanzahl n aller Beobachtungen und ist definiert als:

$$h(x) = (\text{Anzahl beobachtete Werte } x_i = x) / n \\ = \#(x_i = x) / n$$

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

$$F^-(x) = (\text{Anzahl beobachtete Werte } x_i \leq x) / n \\ = \#(x_i \leq x) / n$$

$$0 = F^-(x_{(1)}) < \dots < F^-(x_{(n)}) = 1$$

Aufgaben

1. Welche Skalierung (Wertebereich, Maßeinheit, diskret oder kontinuierlich, Meßgenauigkeit) und Kodierung (Werte und Werte-Etiketten; d.h. Bedeutung der möglichen Werte) würden Sie für folgende Variablen wählen:

Sympathiewerte für Politiker
Schulnoten
Europäische Staaten
Zeiten für 50km-Skilanglauf
Zeiten für 100-m-Lauf

2. (*) Berechnen Sie wichtige Kenngrößen für die Stichprobe $S=(1,4,3,5,2,3,3,1,6)$, die die Ergebnisse von 9 Würfelwürfen darstellen soll.

Erstellen von einfachen Tabellen und Berechnen von Kennzahlen

In diesem Kapitel werden einige Methoden zum tabellarischen Darstellen von Variablen vorgestellt.

Ich will mir zunächst einen tabellarischen Überblick über meine Daten verschaffen, da ich noch keine Idee habe, welche Zusammenhänge überhaupt existieren könnten ...

Überblick - Darstellen des Datenmaterials in tabellarischer Form und Berechnen von Kennzahlen

Sie können sich in SPSS u.a. über folgende Menüpunkte nähere (beschreibende) Informationen über das Datenmaterial in der aktuellen SPSS Arbeitsdatei verschaffen, indem Sie Tabellen erstellen und zusätzlich charakterisierende Kennzahlen berechnen.

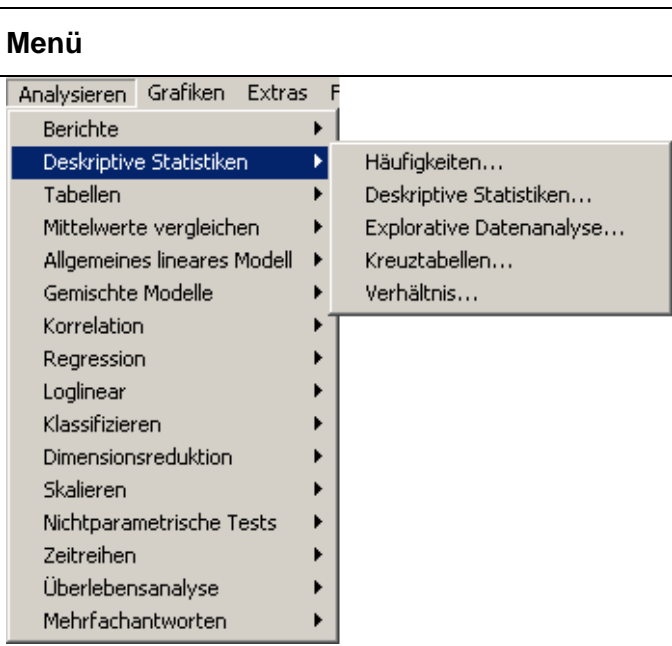
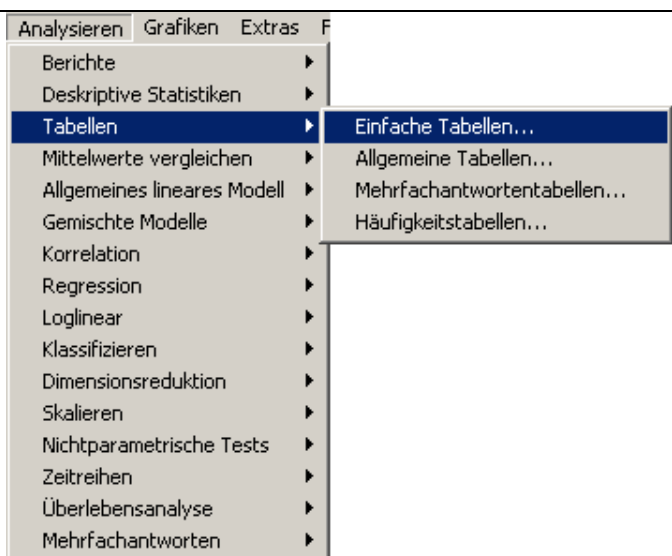
Menü	Funktion
	<p>... berechnet für die ausgewählten Variablen u.a. Häufigkeiten, Kreuztabellen, Kennzahlen (Streu- und Lagemaße wie Mittelwert, Median, Varianz, Standardabweichung).</p>
	<p>... erzeugt tabellarische Berichte über das Datenmaterial, ggf. gruppiert und sortiert und angereichert mit vielfältigen Statistiken.</p>

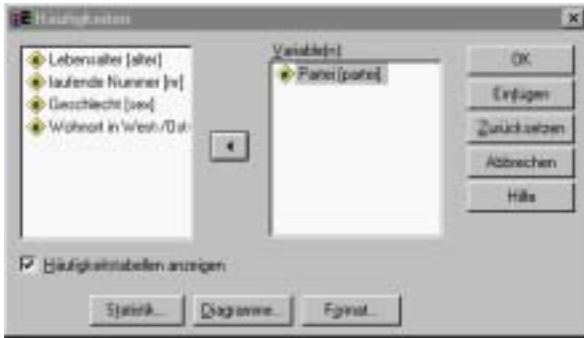
Abb./Tab. 19: Menüpunkte für Tabellen und Kennzahlen

Vorgehensweise - Berechnen von Häufigkeiten

Im folgenden Beispiel führen Sie für die Variable `partei` aus `wahl.sav` eine Häufigkeitsauszählung durch:

Analysieren > Deskriptive Statistiken > Häufigkeiten

Erstellen von einfachen Tabellen und Berechnen von Kennzahlen



Wählen Sie zunächst aus der Liste der Variablen die zu bearbeitenden Variablen aus, hier **partei** (Partei).

Aufgrund der gewählten Einstellungen erhalten Sie im Ausgabe-Fenster (*SPSS Viewer*) folgendes Ergebnis der Auszählung angezeigt: Wechseln Sie mit **Fenster -> Ausgabe** in das Ausgabefenster.

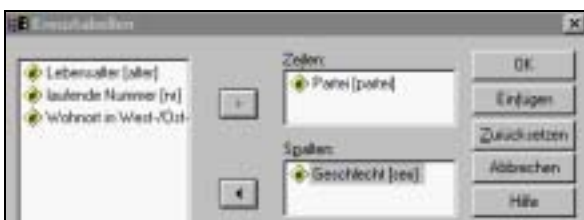
Partei					
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	CDU/CSU	10	33,3	34,5	34,5
	FDP	2	6,7	6,9	41,4
	SPD	9	30,0	31,0	72,4
	Grüne/Bündnis 90	3	10,0	10,3	82,8
	PDS	1	3,3	3,4	86,2
	Republikaner	2	6,7	6,9	93,1
	Sonstige	2	6,7	6,9	100,0
	Gesamt	29	96,7	100,0	
Fehlend	keine Angabe	1	3,3		
Gesamt		30	100,0		

Abb./Tab. 20: Häufigkeitsauszählung für die Variable **partei** aus *wahl.sav*

Vorgehensweise - Erstellen einer Kreuztabelle

Im folgenden Beispiel führen Sie für die Variablen **partei** und **sex** (Geschlecht) aus *wahl.sav* eine Häufigkeitsauszählung durch:

Analysieren > Deskriptive Statistiken > Kreuztabellen



Wählen Sie zunächst aus der Liste der Variablen die zu bearbeitenden Variablen aus, hier **partei** (Partei) und **sex** (Geschlecht).

Nehmen Sie ggf. in den Dialogboxen zu den Aktionsschaltflächen *Statistiken*, *Zellen* und *Format* weitere Einstellungen vor.

Fenster -> Ausgabe

Erstellen von einfachen Tabellen und Berechnen von Kennzahlen

Partei * Geschlecht Kreuztabelle

Anzahl

		Geschlecht		Gesamt
		weiblich	männlich	
Partei	CDU/CSU	5	5	10
	FDP	1	1	2
	SPD	4	5	9
	Grüne/Bündnis 90	2	1	3
	PDS	1		1
	Republikaner		2	2
	Sonstige	1	1	2
Gesamt		14	15	29

Abb. 21: Kreuztabelle für die Variablen `partei` und `sex` aus `wahl.sav`

Vorgehensweise - Berechnen von charakterisierenden Kennzahlen

Im folgenden Beispiel berechnen Sie für die numerische Variable `alter` aus `wahl.sav` einige für die Stichprobe charakteristische Kennzahlen wie z.B. den Mittelwert und die Standardabweichung:

Analysieren > Deskriptive Statistiken > Deskriptive Statistiken



Wählen Sie zunächst aus der Liste der Variablen die zu bearbeitenden Variablen aus, hier `alter` (Alter).



Kreuzen Sie in der Dialogbox zu *Optionen* die gewünschten Kennzahlen an.

Fenster -> Ausgabe

Erstellen von einfachen Tabellen und Berechnen von Kennzahlen

Deskriptive Statistik					
	N	Spannweite	Minimum	Maximum	Mittelwert
Lebensalter	30	79	0	79	42,80
Gültige Werte (Listenweise)	30				

Abb. 22: Kennzahlen für die Variable `alter` aus `wahl.sav` (Ausschnitt)

SPSS berechnet als Kennzahlen u.a. den Mittelwert (*mean*), die emp. Standardabweichung (*Std Dev*) und Maximum und Minimum sowie die Anzahl fehlender Beobachtungen. Die Bedeutung der anderen Kennzahlen ist über das Hilfesystem abrufbar.

Aufgaben

1. Erzeugen Sie für die SPSS Arbeitsdatei `schueler.sav` Häufigkeitsauszählungen für die Variable `deutsch` und berechnen Sie für `physik` Mittelwert, Varianz, Median und 25%- und 75%- Quantile.
2. Erzeugen Sie für `wahl.sav` über *Kreuztabellen* eine Kreuztabelle mit `alter`/`partei` als Zeilen/Spalten.
Führen Sie danach eine zusätzliche Gruppierung nach `sex` durch.
3. (*) Welche Bedeutung haben die weiteren berechneten Kennzahlen wie z.B. Quantil (Perzentil), Schiefe und Kurtosis?
Kann eine Stichprobe mit "extrem schiefer" empirischer Verteilung approximativ normal-verteilt sein?
Hinweis: Konsultieren Sie die Online Hilfe oder ein gutes Statistik-Lehrbuch.

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.

Ich will mir zunächst visuellen Überblick über meine Daten verschaffen, da ich noch keine Idee habe, welche Zusammenhänge überhaupt existieren könnten ...

Überblick - Visualisieren von Daten

Sie können u.a. über folgende Menüpunkte grafische Darstellungen des Datenmaterial anfordern:

Menü	Funktion
	<p>.. erstellt hochauflösende Grafiken mit unterschiedlichen statistischen Diagramm-Typen wie z.B. Histogramm, Balkendiagramm, Flächendiagramm, Boxplot, Streudiagramm, Fehlerbalken.</p> <p>Es folgen einige Beispiele aus der Galerie:</p> <div style="display: flex; flex-wrap: wrap; justify-content: space-around;"> <div style="text-align: center; margin: 5px;">  <p>Balkendiagramme</p> </div> <div style="text-align: center; margin: 5px;">  <p>Streudiagramme</p> </div> <div style="text-align: center; margin: 5px;">  <p>Linendiagramme</p> </div> <div style="text-align: center; margin: 5px;">  <p>Histogramme</p> </div> <div style="text-align: center; margin: 5px;">  <p>Flächendiagramme</p> </div> <div style="text-align: center; margin: 5px;">  <p>Kreisdiagramme</p> </div> <div style="text-align: center; margin: 5px;">  <p>Hoch-Tief-Diagramme</p> </div> </div>

Erstellen von Diagrammen

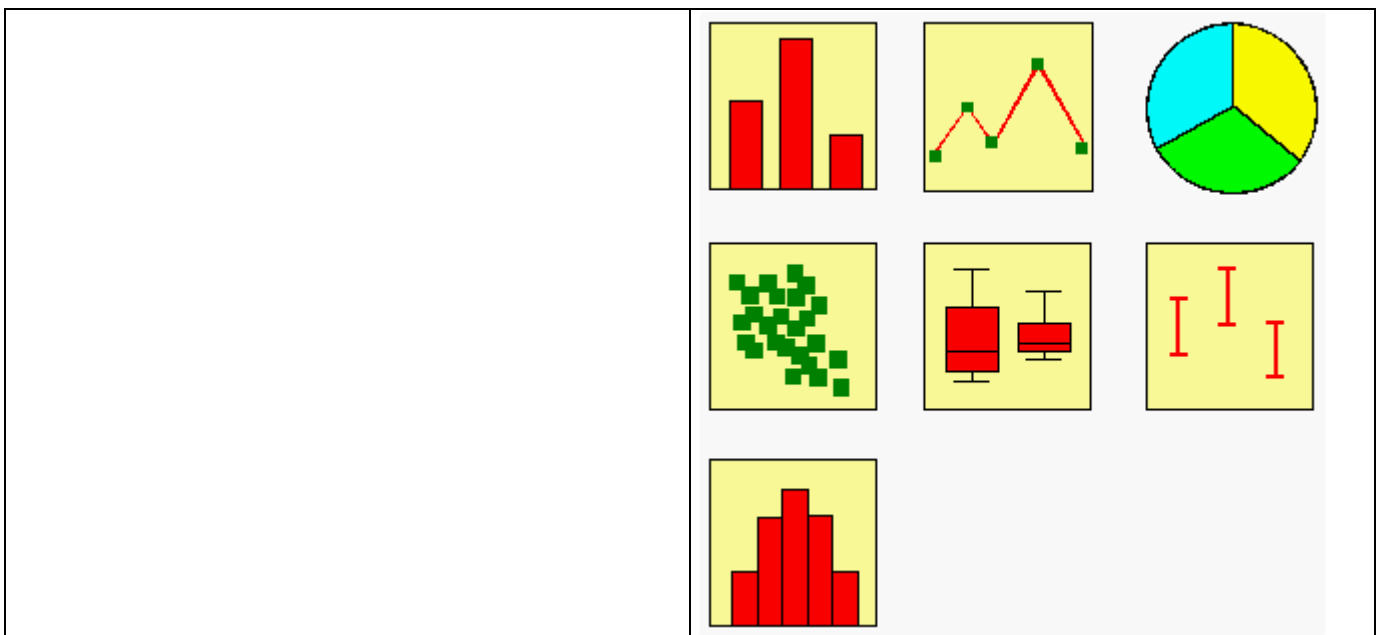


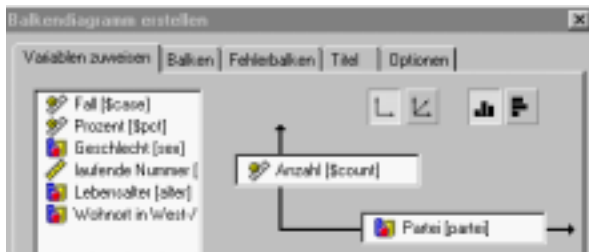
Abb./Tab. 23: Menüpunkte für Grafik

Im folgenden werden Diagramme aus dem Menü **Grafiken > Interaktiv** betrachtet, die sehr komfortabel interaktiv nachbearbeitet werden können.

Vorgehensweise - Erstellen eines einfachen Balkendiagramms

Stellen Sie die Häufigkeit der Kategorien (die unterschiedlichen beobachteten Werte) von der Variablen **partei** aus der Arbeitsdatei **wahl.sav** als Balkendiagramm dar:

Grafiken -> Interaktiv > Balken



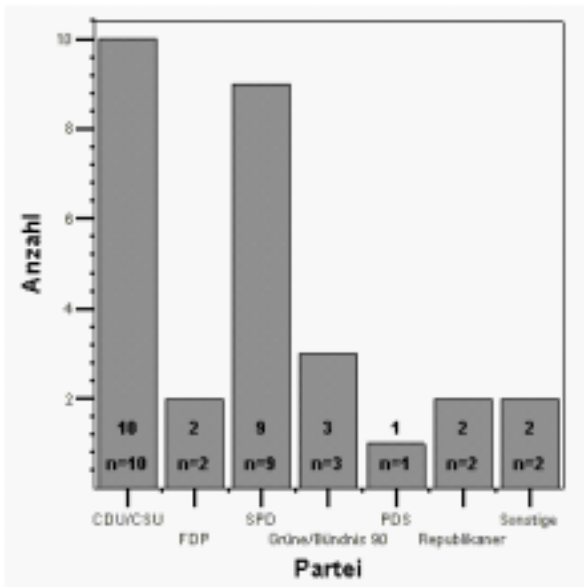
Wählen Sie ein einfaches (x-y) Balkendiagramm aus.

Die x-Achse repräsentiert die unterschiedlichen Ausprägungen (Kategorien, unterschiedliche beobachtete Werte) der Variablen (hier: **partei**).

Wählen Sie die Anzahl der Fälle als darzustellende Größe in y-Richtung aus; d.h. die Höhe eines Balkens repräsentiert die Anzahl der Beobachtungen mit dem an der x-Achse markierten Wert der Variablen.

Fenster -> Ansicht

Erstellen von Diagrammen



Das Balkendiagramm zeigt die Anzahl der Beobachtungen für jede Kategorie der Variablen `partei` an. Im vorliegenden Fall existieren 8 Kategorien vor (7 unterschiedliche Werte und sämtliche fehlenden Werte zusammengefaßt als 8. Kategorie)

Es gibt z.B. insgesamt 2 Personen, die FDP wählen würden.

Vorgehensweise - Erstellen eines gruppierten Balkendiagramms

Stellen Sie die Häufigkeit der Kategorien von der Variablen `partei` als Balkendiagramm dar, wobei Sie eine Untergliederung nach West-/Ostdeutschland) vornehmen:

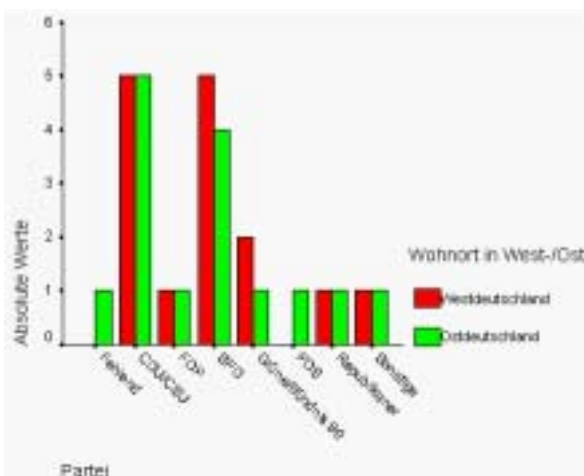
Grafiken -> Balkendiagramme



Wählen Sie nun ein gruppiertes Balkendiagramm aus.

Die x-Achse repräsentiert die unterschiedlichen Ausprägungen (Kategorien, unterschiedliche beobachtete Werte) der Variablen .

Fenster -> Ansicht



Es gibt z.B. 5 Personen aus Westdeutschland, die SPD wählen würden und 4 Personen aus Ostdeutschland.

Vorgehensweise - Erstellen eines gestapelten Flächendiagramms

Die folgenden Diagramme beruhen auf Datenmaterial über die Entwicklung von Studentenzahlen im Fach Informatik in den Jahren von 1975 bis 1993. Die SPSS Arbeitsdatei `inform.sav` enthält folgende Variablen:

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

Die Arbeitsdatei enthält u.a. folgende Beobachtungen:

JAHR	STUD_GES	STUD_W	STUD_M	ERST_GES	ERST_W	ERST_M
1975	5003	682	4321	1439	209	1230
1976	5820	832	4988	1491	247	1244
1977	6374	970	5404	1525	263	1262
1978	7558	1418	6140	2156	449	1707
...						
1992	30889	2837	28052	5005	381	4624
1993	31005	2958	28047	4345	320	4025

Erzeugen Sie ein gestapeltes Flächendiagramm, um die zeitliche Entwicklung der Studentenzahlen zu verdeutlichen:

Grafiken -> Flächendiagramme

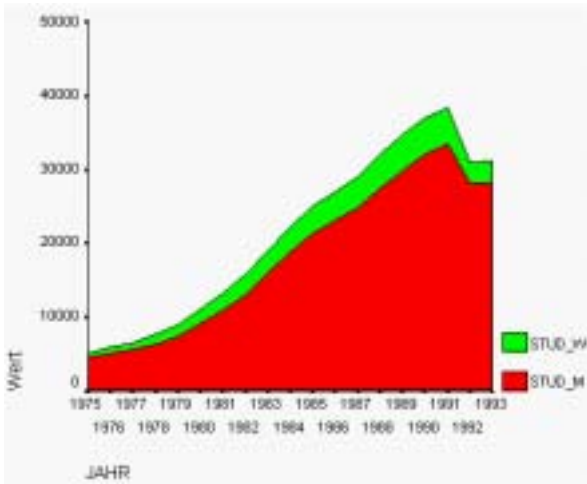


Wählen Sie hier "Werte einzelner Fälle" aus, da die x-Achse die einzelnen **Fälle** (Beobachtungen) repräsentiert.

(Vgl. Balkendiagramm: Hierbei repräsentiert die x-Achse unterschiedliche **Kategorien** einer Variablen).

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

Fenster -> Ansicht



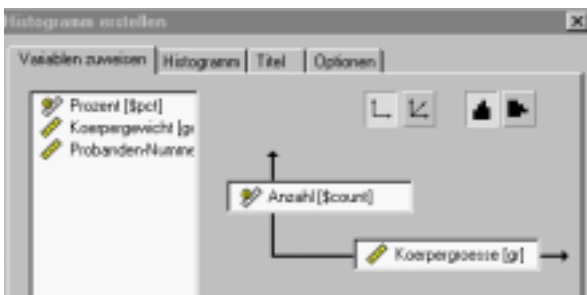
Das Flächendiagramm zeigt die Werte einzelner Beobachtungen. Statt der laufenden Nummer der Beobachtung wurde die Variable **jahr** zur Beschriftung der x-Achse ausgewählt.

Es gibt z.B. im Jahr 1973 682 weibliche und 4321 männliche Studenten im Fach Informatik und insgesamt 5003 Studenten.

Vorgehensweise - Erstellen eines Histogramms (empirische Dichte)

Im folgenden Beispiel erstellen Sie für die intervall-skalierte Variable **gr** aus **broca.sav** ein Histogramm mit überlagerter Normalverteilungskurve:

Grafiken -> Interaktiv > Histogramm



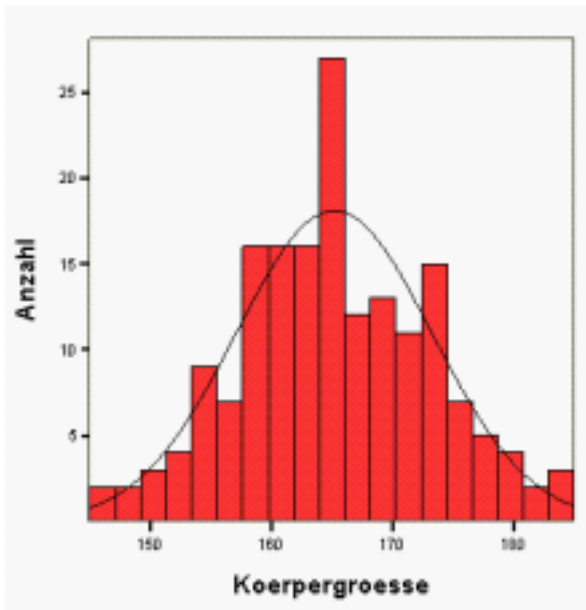
Wählen Sie mindestens ordinal-skalierte Variablen aus, für die Histogramme erstellt werden sollen, hier: **alter**.



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

Fenster -> Ansicht

Erstellen von Diagrammen



Das Histogramm zeigt, wieviele Beobachtungen in die vorgegebenen Intervalle fallen. Ein Histogramm enthält also aggregierte Daten.

Die überlagerte Normalverteilungskurve mit Schätzwerten für Erwartungswert μ und Varianz σ^2 gibt Anlaß zur Vermutung, daß die Stichprobe repräsentativ für eine normal-verteilte Grundgesamtheit ist.

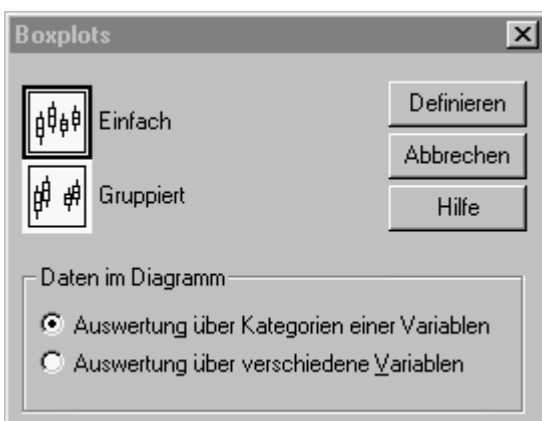
Vorgehensweise - Vergleichen von empirischen Verteilungen mit Hilfe von Boxplots

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)
	Verbindungslinie	
+-----+	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
	Verbindungslinie	
-----	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

Vergleichen Sie für `wahl1.sav` die Verteilung des Alters der befragten Frauen mit der der befragten Männer mit Hilfe eines **Box- und Whisker-Plots**:

Grafik -> Boxplot



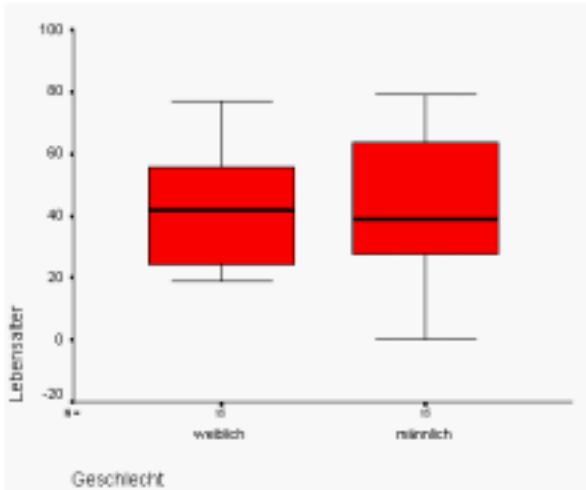
Wählen Sie einen einfachen Boxplot, wobei die x-Achse die unterschiedlichen Kategorien einer Variablen enthält.

Erstellen von Diagrammen



Wählen Sie nun als darzustellende Variable das Alter **alter** und als gruppenbildende Variable das Geschlecht **sex**.

Fenster -> Ansicht



Der Boxplot zeigt die Verteilung der Werte der Variablen **alter**, gruppiert nach der Kategorie **sex** (Geschlecht)

Die Boxplots sind leicht gegeneinander verschoben, d.h. die befragten Männer waren im Mittel älter die befragten Frauen. Es gibt hier keine Extremwerte und keine Ausreißer.

Vorgehensweise - Bearbeiten von Diagrammen

Sie können Grafiken, die Sie im Navigator betrachten, individuell nachbearbeiten.

Klicken Sie hierzu auf das Diagramm. Die aktuell angezeigte Grafik 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 der Grafik, 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

Aufgaben

1. Erzeugen Sie über *Grafik* folgende Diagramme für Variablen der SPSS Arbeitsdatei `wahl.sav`:
 - a) Tortendiagramm für **partei**
 - b) Histogramm für **alter** mit überlagerter Normalverteilungskurve
 - c) Boxplot (auch Box-and-Whisker-Plot genannt) für **alter**
 - d) Streudiagramm mit **alter** als x-Achse und **partei** als y-Achse
 - e) gestapeltes Balkendiagramm für **partei** mit Gruppierung nach **sex**
2. Fügen Sie das Tortendiagramm aus 1a) über die Windows Zwischenablage in ein Textverarbeitungsprogramm wie z.B. **Word für Windows** ein.
Hinweis: Kopieren Sie das Diagramm durch Klicken der rechten Maustaste, dann *Kopieren*, in die Windows Zwischenablage. Wechseln Sie nun über den **Windows Task Manager** in das vorher bereits gestartete Textverarbeitungsprogramm. Wählen Sie in **Word für Windows** unter *Bearbeiten* den Menüpunkt *Inhalte einfügen* und fügen Sie das Diagramm als *Grafik* ohne Verknüpfung ein.

Erstellen von Diagrammen

3. (**) Verschaffen Sie sich einen Überblick über die verfügbaren Diagrammtypen. Experimentieren Sie z.B. mit Hoch-Tief-Diagrammen (Aktien) oder Fehlerbalken-Diagramm (Meßfehler).

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 nie verstanden, was „Wahrscheinlichkeit“ eigentlich im mathematischen Sinne bedeutet.

Zufallsexperiment und Wahrscheinlichkeit

Ein Vorgang oder Versuch, dessen Durchführung "zufällig" zu genau einem von mehreren möglichen Ergebnissen führt, wird als **Zufallsexperiment** oder **Zufallsvorgang** bezeichnet. Derartige Vorgänge werden auch als nicht-deterministisch 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 Wahrscheinlichkeitsrechnung bezeichnet als "Auswahl einer Stichprobe aus einer Grundgesamtheit, bei der m Objekte "zufällig" ohne Zurücklegen aus der Grundgesamtheit mit n Objekten ausgewählt werden ($m \leq n$)".

Sei ein Zufallsexperiment mit der Ergebnismenge Ω (Menge der möglichen Ergebnisse) gegeben. Die **Wahrscheinlichkeit P** ist eine Abbildung von Ω in das Intervall $[0,1]$, die jedem **Ergebnis** des Zufallsexperimentes eine positive Zahl p (die Wahrscheinlichkeit für das Eintreten des Ergebnisses) zuordnet.

Die Wahrscheinlichkeit P für die Ergebnisse oder Ereignisse eines Zufallsexperimentes wird nicht "bewiesen", sondern ihre Existenz wird als plausible Annahme (Axiom) vorausgesetzt. Intuitiv ist $P(w_i)$ die "stabilisierte" relative Häufigkeit für das Ergebnis w_i für eine große Anzahl von Versuchen. Die Abbildung P , die Ergebnissen "Wahrscheinlichkeiten" zuweist, wird aufgrund von plausiblen Annahmen, Erfahrungswerten oder Schätzungen aufgestellt.

Zufallsvariablen und ihre Verteilung

Bei einem Zufallsexperiment seien die Ergebnisse $\Omega = \{w_1, \dots, w_n\}$ möglich. Jedem Ergebnis werde durch die 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 Ergebnis w abgeleitete (Zufalls-) Variable $X(w)$. Bei einem Angelwettbewerb interessieren z.B. in der Regel nicht die einzelnen gefangenen Fische, sondern nur die Anzahl oder das Gesamtgewicht aller gefangenen Fische oder der schwerste gefangene Fisch.

Die (Wahrscheinlichkeits-) **Verteilung** P^X einer Zufallsvariablen X wird über die Wahrscheinlichkeit der Urbilder in der Ergebnismenge Ω definiert; d.h. die Summe aller Wahrscheinlichkeiten für Ergebnisse w_i , 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.

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 im folgenden immer mit Großbuchstaben (gebräuchlich sind X, Y, Z) bezeichnet, während die tatsächlich beobachteten Werte für eine (Zufalls-) Variable mit Kleinbuchstaben (x_i : Wert von X für die i .te Beobachtung) bezeichnet werden:

Zufallsexperimente, Zufallsvariablen und Wahrscheinlichkeit

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

Es 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).

Aufgaben

1. Kennen Sie Spiele, die auf Zufallsexperimenten beruhen?
(Spielen Sie lieber Schach oder BlackJack?)
2. (*) Wie lautet die Verteilung der Augenzahl beim Zufallsexperiment "Werfen von 2 echten Würfeln"? Zeichnen Sie die Verteilungsfunktion und berechnen Sie die Kenngrößen der Verteilung (Grundgesamtheit). Führen Sie das Zufallsexperiment 10-mal 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.
3. (*) Was sagt Ihnen der Begriff **Gauß'sche Glockenkurve**? Unter welchem Namen ist die zugehörige Verteilung bekannt? Worin besteht die besondere Bedeutung dieser Verteilung? (Wie lautet der zentrale Grenzwertsatz? Wie ist er zu interpretieren?)
Hinweis: Der "alte" 10-DM-Schein enthält sowohl eine Grafik der Dichtefunktion wie auch die recht komplizierte explizite Formel der Dichtefunktion.
4. (*) Welche weiteren diskreten und kontinuierlichen Verteilungen sind Ihnen bekannt und wie lauten jeweils die Verteilungsfunktionen?

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.

Haben Hypothesen etwas mit ... zu tun ?

Ziehen von Rückschlüssen aus einer Stichprobe

In der mathematischen oder analytischen Statistik werden **Verfahren** entwickelt und angewendet, um anhand einer Stichprobe (d.h. Auswählen einer Teilmenge der Grundgesamtheit) **Rückschlüsse** oder **Folgerungen** (*statistical inference*) für die Grundgesamtheit ziehen zu können.

Die mathematische Statistik hat sich aus der "Politischen Arithmetik" entwickelt, die sich hauptsächlich mit Tauf-, Heirats- und Sterberegistern befaßte, um Geschlechtsverhältnis, Altersaufbau und Sterblichkeit der Bevölkerung abzuschätzen. Es wurde bereits sehr früh der Versuch unternommen, aus repräsentativen Beobachtungsdaten Gesetzmäßigkeiten abzuleiten, die über den Beobachtungszeitraum und -ort hinaus gültig waren.

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 ³ von bestimmten Werten $h(k)=\#(x_i=k)/n$
Verteilungsfunktion von X $F(k)=P(X\leq k)$	Empirische Verteilungsfunktion von S $F^*(k)=\#(x_i\leq k)/n$
Erwartungswert $\mu=E[X]=k_1 P(X=k_1)+\dots$	Mittelwert von S $\bar{x}=(k_1 + \dots)/n$
Varianz $\sigma^2=\sigma_{XX}=E[(X-\mu)^2]=(k_1-\mu)^2 P(X=k_1)+\dots$	Empirische Varianz von S $s^2=s_{XX}=(k_1-\bar{x})^2 + \dots$
Standardabweichung $\sigma=\sqrt{\sigma^2}$	Empirische Standardabweichung von S $s=\sqrt{s^2}$
Median $m=F^{-1}(0.5)$	Empirischer Median von S $m^*=x_{\lceil n/2 \rceil}$
Kovarianz $\rho_{XY}=\text{Cov}(X,Y)=E[(X-E[X])(Y-E[Y])]$	Empirische Kovarianz $s_{XY}=\frac{((k_1-\bar{x})(l_1-\bar{y}) + \dots)}{(s_{XX} s_{YY})}$

Die beobachteten **Kennzahlen der Stichprobe** stimmen in der Regel nicht mit den entsprechenden **Kennzahlen der Grundgesamtheit** überein. 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 Kennzahlen der Grundgesamtheit durchzuführen.

Durchführen von Schätzungen und Hypothesentests

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

³ Das Zeichen # dient zur Abkürzung für Anzahl, z.B. #(Augenzahl=5) steht für: Anzahl der Würfe, bei denen die gewürfelte Augenzahl 5 beträgt.

Welche Aussage über eine **unbekannte** Kennzahl (wahrer Parameter) der Grundgesamtheit kann aufgrund der Beobachtung der korrespondierenden **realisierten** (empirischen, beobachteten) 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 Beobachtungsdaten, **mit denen sich "vernünftige" Entscheidungen bei ungewisser Sachlage treffen lassen**⁴.

Vernünftig heißt in diesem Zusammenhang, daß die Sicherheit, mit der ein statistisches Verfahren zu einer richtigen Entscheidung führt, vertrauenswürdig 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.

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 die Getränke für den Rest des Jahres reichen ...

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, Varianz) unterscheiden. Die einfachere Aufgabe besteht in diesem Fall nun darin, mit Hilfe eines statistischen Verfahrens gesicherte Aussagen über die unbekanntes 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 Verteilung der Grundgesamtheit (wenigstens approximativ) eine **Normalverteilung** mit unbekanntes Parametern μ und σ 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.

⁴ Da ungewisse Sachlagen eigentlich den Normalfall im Leben darstellen, treffen wir mehr oder weniger bewußt sehr häufig statistisch motivierte Entscheidungen.

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 μ der gewürfelten Augenzahl. Bei einem "echten" Würfel berechnet sich der Erwartungswert μ aus Symmetriegründen zu 3.5, aber vielleicht ist der Würfel manipuliert!

1. Welcher **Schätzwert** $T(x_1, \dots, x_n)$ für den Parameter (Erwartungswert) μ kann aus der Stichprobe $S=(x_1, \dots, x_n)$ abgeleitet werden?
(*Punkt-Schätzung*)
2. Welcher **Schätzwert für ein Intervall** $[a, b] = [CI_L(x_1, \dots, x_n), CI_R(x_1, \dots, x_n)]$, das den unbekannt wahren Parameter (Erwartungswert) μ 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 α (Annahme der Hypothese, obwohl sie falsch ist) und 2. Art β (Ablehnung der Hypothese, obwohl sie wahr ist)?
(*Hypothesen-Test*)

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 keine Aussage über den Wahrheitsgehalt der Null-Hypothese H treffen, weil Sie den Wert von β nicht kennen (siehe unten). Sie können nur eine Aussage darüber treffen, mit welcher Wahrscheinlichkeit Sie die Hypothese irrtümlicherweise verworfen haben⁵.

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.

Abb. 24 : Entscheidungsregel für Hypothesentests

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**:

<i>Wahrheit</i>	<i>Hypothese ist wahr.</i>	<i>Hypothese ist falsch.</i>
<i>Ihre Entscheidung</i>		

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

Überblick über die mathematische Statistik

Hypothese wird angenommen.	Richtige Entscheidung	Falsche Entscheidung Fehler 2. Art β
Hypothese wird abgelehnt.	Falsche Entscheidung Fehler 1. Art α	Richtige Entscheidung

Abb. 25: Dilemma beim Hypothesen-Test

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 α und entsprechend ein **Konfidenz-Niveau** $(1-\alpha)$ zugeordnet.

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

Abb./Tab. 26: 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 Annahmebereich für die Null-Hypothese H wird, desto größer wird das Konfidenz-Niveau $(1-\alpha)$ und desto kleiner wird die Irrtumswahrscheinlichkeit α (Wahrscheinlichkeit, die Hypothese fälschlicherweise abzulehnen)⁶. Umgekehrt gilt: Je größer α gewählt wird, desto größer wird der Ablehnungsbereich und desto größer wird die Irrtumswahrscheinlichkeit dafür, die Hypothese fälschlicherweise abzulehnen.

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

Sie brauchen nun nur noch die von Ihnen gewünschte oder von anderen Personen vorgegebene Irrtumswahrscheinlichkeit α (z.B. 0.01) mit der von SPSS berechneten Irrtumswahrscheinlichkeit **p** zu vergleichen und entscheiden nun folgendermaßen:

1. Ist **p** (von SPSS berechnet) $< \alpha$ (von Ihnen gewünscht), sollten Sie die Null-Hypothese ablehnen.
2. Ist **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 α 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 p ist zu interpretieren als die **minimale** Irrtumswahrscheinlichkeit, bei der die Null-Hypothese H noch abgelehnt werden kann. SPSS kann nicht wissen, welche Irrtumswahrscheinlichkeit α Sie ansetzen möchten und berechnet deshalb die **minimal** zulässige Irrtumswahrscheinlichkeit, die zur Ablehnung der Hypothese führen kann.

⁶ Allerdings wird die Wahrscheinlichkeit für einen Fehler 2. Art immer größer, nämlich die Hypothese H anzunehmen, obwohl sie falsch ist.

Aufgaben

1. Was halten Sie davon, den Erwartungswert im obigen Beispiel des Würfelwurfes durch folgende Punktschätzer $T(X_1, \dots, X_n)$ zu schätzen:
 - a) T_1 : Schätzwert ist Ergebnis des 1. Würfelwurfes
 - b) T_2 : Schätzwert ist Mittelwert von 1. und letztem Würfelwurf
 - c) T_3 : Schätzwert ist Median aller Würfelwürfe
 - d) T_4 : Schätzwert ist 3.5, unabhängig davon, was gewürfelt wurde

Welchen Punktschätzer T würden Sie verwenden?

Hinweis: 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 α die Antwort (Annahme/Ablehnung) berechnen lassen.

2. (*) Kriterien für einen "guten" Punktschätzer $T(X_1, \dots, X_n)$ sind z.B. a) Erwartungstreue sowie b) Konsistenz. Interpretieren Sie diese Kriterien.
Hinweis:
 - a) $E(T) = \mu$
 - b) $\text{Var}(T) \rightarrow 0$ bei wachsendem Stichprobenumfang
3. Nennen Sie "Alltagssituationen", in denen Sie oder andere Punkt- oder Bereichsschätzungen vornehmen und entsprechend Entscheidungen treffen.
Hinweise:
Verbrauch an Lebensmitteln am Wochenende, Dimensionierung von Parkhäusern, Planung von Zugkapazitäten, ...
4. Erläutern Sie mit eigenen Worten, welche Probleme sich bei einem Hypothesentest ergeben.
5. (*) Wie würden Sie die Irrtumswahrscheinlichkeit α festlegen
 - a) für einen genetischen Test ("genetischer Fingerabdruck"), der in einem Vergewaltigungs- und Mordprozeß zur Urteilsfindung herangezogen werden soll,
 - b) für eine Marketing-Untersuchung,
 - c) für den Nachweis der Wirksamkeit eines Medikamentes als Befürworter/Gegner des Medikamentes?
6. (*) Interpretieren Sie folgende statistische Grundweisheit für Konfidenz-Intervalle:
"Sichere Aussagen sind unscharf, scharfe Aussagen sind unsicher."
Hinweis:
Welcher Zusammenhang besteht zwischen Irrtums-Wahrscheinlichkeit und Länge von des Konfidenz-Intervalls?

Exploratives Analysieren von Daten und Überprüfen von Voraussetzungen

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.

Motivation - Exploratives Analysieren von Daten und Überprüfen von Voraussetzungen für statistische Tests

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**:

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

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

Hierzu dienen u.a. die Teststatistik von Kolmogorov-Smirnov (**Test auf Normalverteilung**) und die Teststatistik von Levene (**Test 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

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!).

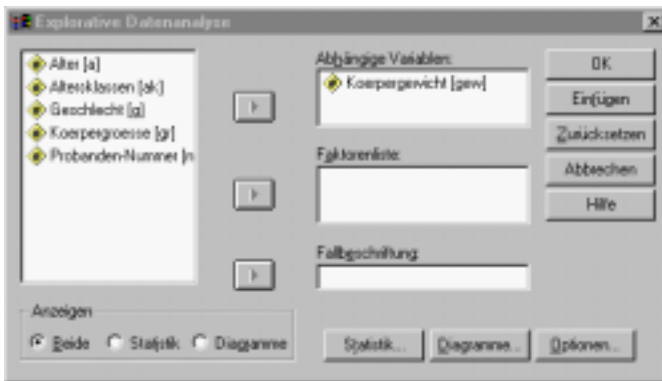
Die Ergebnisse der genannten Tests beeinflussen die Auswahl weiterer statistischer Verfahren, da diese häufig genau diese Voraussetzungen an die Stichprobe stellen. Falls die Voraussetzungen nicht erfüllt sind, können Sie z.B. auf nicht-parametrische Verfahren zurückgreifen.

Vorgehensweise - Testen auf Normalverteilung

Im folgenden Beispiel untersuchen Sie die Variable **gew** aus der Arbeitsdatei **broca.sav**. Testen Sie die Verteilung der Variablen **gew** (Gewicht) auf Normalverteilung.

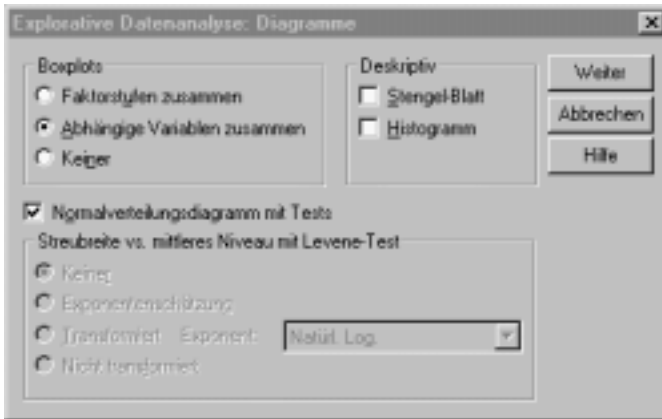
Analysieren > Deskriptive Statistiken > Explorative Datenanalyse

Exploratives Analysieren von Daten und Überprüfen von Voraussetzungen



Wählen Sie zunächst die Variable **gew** als abhängige Variable

Aktivieren Sie nun **Diagramme** und nehmen Sie dort weitere Einstellungen vor.



Fordern Sie **Boxplots** an

Kreuzen Sie Normalverteilungsplots mit Tests an.

Fenster -> Ausgabe

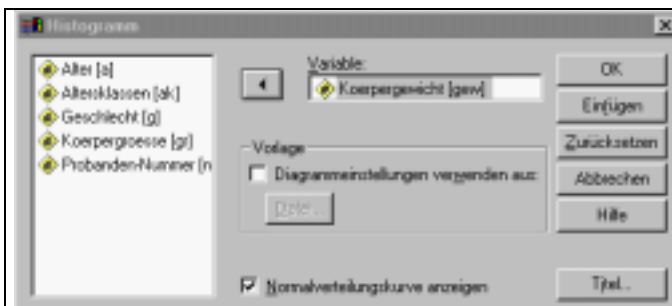
Tests auf Normalverteilung			
	Kolmogorov-Smirnov ^a		
	Statistik	df	Signifikanz
Körpergewicht	,082	174	,006

a. Signifikanzkorrektur nach Lilliefors

Abb./Tab. 27: K-S-Test

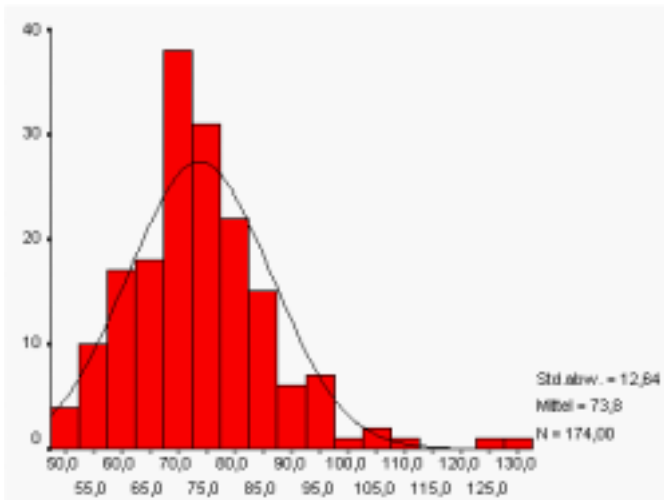
Die explorative Datenanalyse liefert Ihnen einen Wert $t=0.082$ für die Kolmogorov-Smirnov-Testgröße K-S (Normalverteilung). Die zugehörige Irrtumswahrscheinlichkeit beträgt $p=0.006$. Die Null-Hypothese sollte deswegen abgelehnt werden; d.h. es handelt sich nicht um eine Normalverteilung..

Grafik > Histogramm



Fordern Sie ein Histogramm mit überlageter Normalverteilungskurve an.

Exploratives Analysieren von Daten und Überprüfen von Voraussetzungen

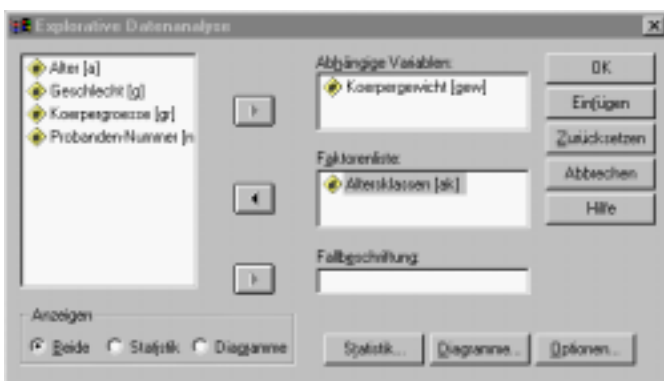


Die Ablehnung der Null-Hypothese wird auch visuell durch ein Histogramm mit eingezeichneter Normalverteilungskurve für die Variable **gew** unterstützt (Histogramm als visuelles Hilfsmittel zur Überprüfung der Normalverteilungsannahme).

Vorgehensweise - Testen auf Varianzhomogenität

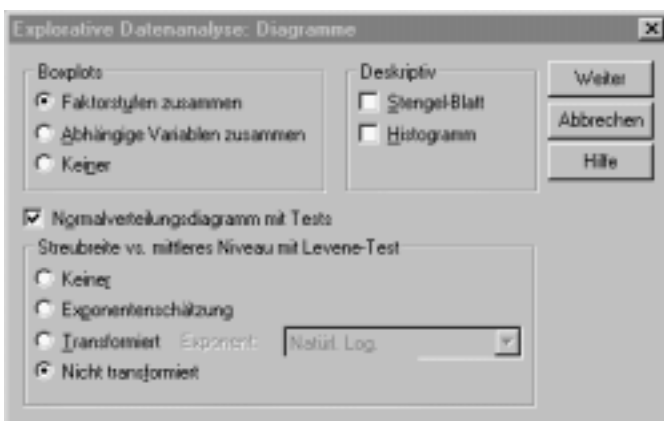
Im folgenden Beispiel untersuchen Sie die Variable **gew** aus der Arbeitsdatei **broca.sav** bzgl. **ak** (Altersklasse) auf Varianzhomogenität:

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.

Fenster -> Ausgabe

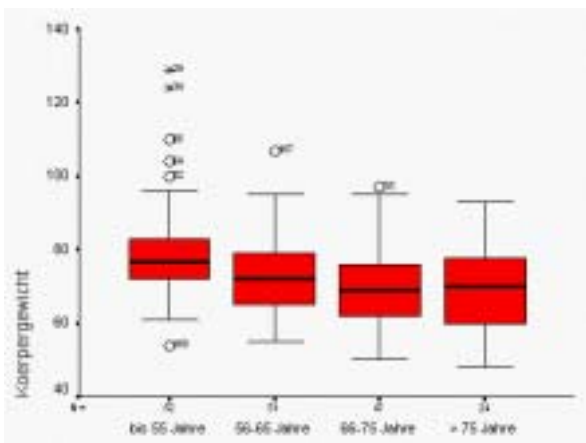
Exploratives Analysieren von Daten und Überprüfen von Voraussetzungen

Test auf Homogenität der Varianz					
		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

Abb./Tab. 28: Levene Test

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.

Aufgaben

1. Erzeugen Sie für `broca.sav` Boxplots für die Variable `gr` nach `g` (Geschlecht) gruppiert und führen Sie 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.
1. Überprüfen Sie, ob die Variable `gr` (annähernd) normalverteilt ist.
Testen Sie nochmals getrennt für Männer und Frauen. Formulieren Sie Ihre Interpretation ähnlich wie im Skript.
3. (*) Nennen Sie für Ihnen bekannte statistische Verfahren die zu überprüfenden Voraussetzungen.
Wie könnten diese Voraussetzungen mit SPSS überprüft werden?

Berechnen eines Vertrauensbereiches (Konfidenz-Intervalls)

In diesem Kapitel wird für eine numerische Variable ein Vertrauensbereich für den 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?

Motivation - Interpretieren von Vertrauensbereichen

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

Ein **Vertrauensbereich (Konfidenz-Intervall)** enthält einen unbekannt Parameter μ , hier den Erwartungswert, einer Verteilung mit einer **Sicherheit (Konfidenz-Niveau)** von z.B. 95% und entsprechen einer **Irrtumswahrscheinlichkeit** α von 5%.

Das **Konfidenz-Niveau** ist folgendermaßen zu interpretieren:

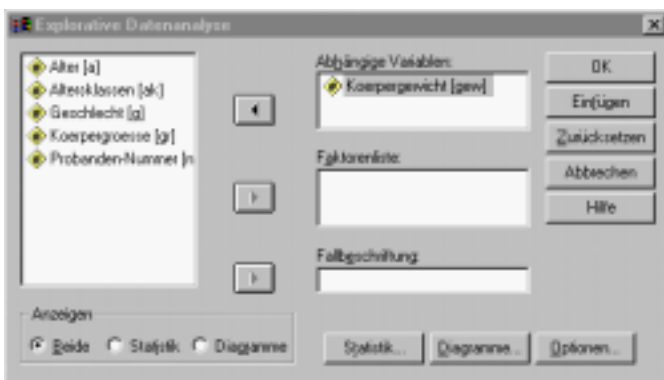
Falls Sie das ausgewählte Verfahren 100-mal durchführen würden (was Sie aber nicht tun!), erhalten Sie im Mittel 95-mal einen **Vertrauensbereich**, der den unbekannt Parameter tatsächlich enthält, allerdings auch 5-mal einen Vertrauensbereich, der ihn nicht enthält. Da Sie i.d.R. nur eine und nicht 100 Untersuchungen durchführen, kann Ihre aktuelle Untersuchung also zu den 5 von 100 Untersuchungen gehören, bei denen das Verfahren einen "falschen" Vertrauensbereich liefert, der den wahren Parameter μ **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 also falsch zu behaupten, daß das berechnete Konfidenz-Intervall den wahren Parameter enthält. Das berechnete Konfidenz-Intervall enthält den wahren Parameter mit großer Wahrscheinlichkeit ...

Vorgehensweise - Berechnen eines Vertrauensbereichs

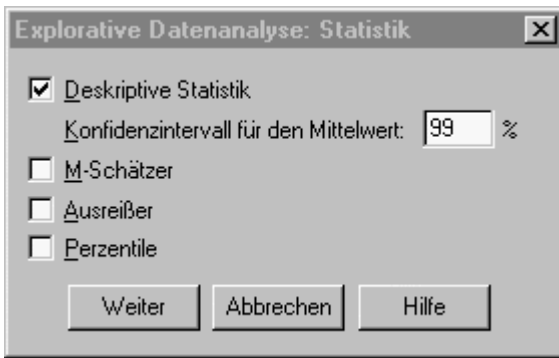
Im folgenden Beispiel berechnen Sie für die Variable **gew** (Gewicht) aus **broca.sav** ein 99%-Vertrauensbereich (**Konfidenz-Intervall**, *confidence interval*, CI) :

Analysieren > Deskriptive Statistiken > Explorative Datenanalyse



Wählen Sie **gew** als abhängige Variable.

Berechnen eines Vertrauensbereiches (Konfidenz-Intervalls)



Geben Sie die gewünschte Sicherheit (Konfidenz-Niveau) S in Prozent ein, hier S=99%.

Die Irrtumswahrscheinlichkeit ist somit:
 $\alpha=(1-S)$, hier: $(1-S)=1\%$ oder 0.01

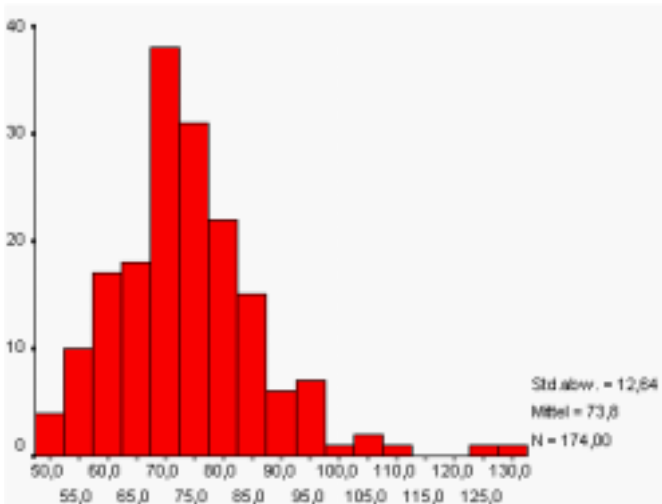
Fenster -> Ausgabe

Univariate Statistiken			
		Statistik	Standardfehler
Körpergewicht	Mittelwert	73,79	,96
	99%		
	Untergrenze	71,29	
	Konfidenzintervall		
	Obergrenze	76,28	

Abb. 29: Ergebnis von SPSS für das 99%- Konfidenz-Intervall

Der 99%-Vertrauensbereich lautet demnach: CI= [71.29, 76.28]; d.h. $71,29 < \mu < 76,28$. Dieses Vertrauensbereich enthält den unbekanntem Erwartungswert μ mit einer Irrtumswahrscheinlichkeit von 1%.

Fenster -> Karussell



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

Statistischer Hintergrund - Ableiten eines Vertrauensbereichs

Es sei \bar{X} der arithmetische Mittelwert der n Beobachtungen in der Stichprobe. Dann gilt unter den Voraussetzung, daß alle X_i identisch und normalverteilt⁷ sind, in Analogie zum zentralen Grenzwertsatz für beliebige Werte c:

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

⁷ Hier macht sich eine angenehme Eigenschaft von unabhängigen normal-verteilten Zufallsvariablen bemerkbar, daß deren Summe selbst wieder normal-verteilt ist.

Berechnen eines Vertrauensbereiches (Konfidenz-Intervalls)

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

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

Die Formel (2) läßt sich nach dem Erwartungswert μ umstellen:

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

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

$$(4) \alpha = 0.05$$

Damit der Erwartungswert μ 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:

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

Diese kritischen Werte c mußten Sie früher in Tabellen nachschlagen ...

Die "zufällige" Länge L des Konfidenz-Intervalls beträgt:

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

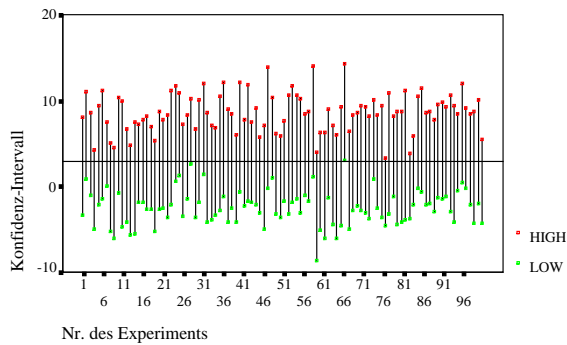
Hierbei ist c (durch Vorgabe von α) vor Beginn der Untersuchung vorgegeben, n ist der Stichprobenumfang und $s = \sigma^{\wedge}$ (emp. Standardabweichung) ist ein "zufälliger" Wert, der erst nach Durchführung der Untersuchung bekannt ist. Insgesamt handelt es sich also bei L um eine **vor** Durchführung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

Beachten Sie, daß Sie selbst das Konfidenz-Niveau $(1-\alpha)$ und damit den kritischen Wert c festlegen. Sie "erkaufen" sich eine größere Sicherheit Ihrer Schätzung (kleines α) durch ein längeres Konfidenzintervall L und umgekehrt erhalten Sie bei kleinerer Sicherheit ein kürzeres Konfidenzintervall L . Sie können das Konfidenz-Intervall natürlich auch verkleinern, indem Sie die Stichprobe vergrößern. $n^{1/2}$ im Nenner verkleinert die Länge des Konfidenzintervalls L mit wachsendem n , im Grenzfall wird aus dem Intervall ein Punkt.

Die folgende Abbildung zeigt für 100 Durchführungen⁸ eines Experiments die jeweils ermittelten 99%-Konfidenz-Intervalle (Experiment: Ziehen einer Stichprobe mit 25 Beobachtungen aus einer Grundgesamtheit mit Normalverteilung $\mu=3$ und $\sigma=10$).

⁸ Die 100 Ziehungen sind über den SPSS Zufallszahlengenerator mit RV.NORMAL(3.0,10.0) realisiert, siehe: normal.sav und normal1.sav.

Berechnen eines Vertrauensbereiches (Konfidenz-Intervalls)



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 einer Durchführung nicht⁹.

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

(Hoch-Tief-Diagramm)

Aufgaben

1. Berechnen Sie einen 95%-Vertrauensbereich für den Erwartungswert von **physik** (Schulnote für Physik) aus **schueler.sav**.
(*) Halten Sie einen Rückschluß auf die Gesamtbevölkerung für sinnvoll?
2. (*) Führen Sie nun die Berechnung aus 1) 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.
3. (*) Überprüfen Sie, ob für die Variable **physik** (Note im Fach Physik) aus **schueler.sav** die Normalverteilungsannahme gerechtfertigt ist. Sehen Sie prinzipiell Probleme aufgrund der Skalierung der Variablen?
Hinweis:
Normalverteilte Zufallsvariablen können (zumindest theoretisch) alle Werte zwischen $-\infty$ bis $+\infty$ annehmen.

⁹ Zugegebenermaßen habe ich einige Male probiert, bis es so genau geklappt hat.

Testen der Unabhängigkeit von 2 Variablen

In diesem Kapitel wird der **Chi-Quadrat-Test** zum Überprüfen der Unabhängigkeit von kategorialskalierten 2 Variablen X und Y in einer r x s **Kontingenztafel** (r Kategorien von X und s Kategorien von Y) behandelt.

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

Motivation - Ableiten der 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.

$$(1) H: p_{ij} = P(X=i, Y=j) = v_i u_j = P(X=i) P(Y=j) \\ \text{für alle möglichen Kombinationen von } i \text{ aus } W_1 \text{ und } j \text{ aus } W_2$$

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 :

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

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:

$$(3) T = \sum_{ij} (h_b - h_e)^2 / h_e$$

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

Vorgehensweise - Berechnen der Chi-Quadrat-Testgröße

Im folgenden untersuchen Sie aggregiertes Datenmaterial über die Religionszugehörigkeit von Braut und Bräutigam bei Eheschließungen (hier: Köln im Jahr 1970) aus der Arbeitsdatei `heirat.sav`.

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

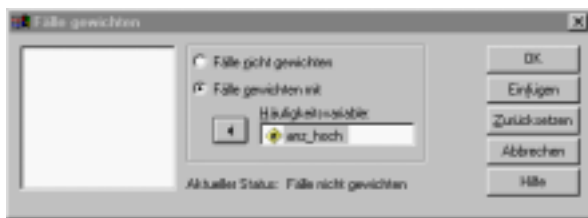
H: X=**braut** und Y=**braeut** sind unabhängig.

Gewichten Sie zunächst die Beobachtungen mit der Variablen `anz_hoch` (Anzahl der Hochzeiten pro Kombination der Religionszugehörigkeit von Braut und Bräutigam)¹⁰:

Daten -> Fälle gewichten ...

¹⁰ Die Beobachtungen liegen bereits gewichtet (aggregiert) vor.

Testen der Unabhängigkeit von 2 Variablen



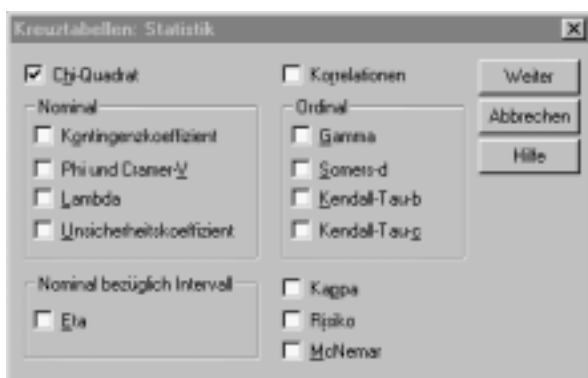
Wählen Sie die Variable `anz_hoch` zur Gewichtung.

Analysieren > Deskriptive Statistiken > Kreuztabellen



Wählen Sie die Variablen für die Kreuztabulation aus, hier: **braut** und **braeut**.

Wählen Sie über *Statistiken* die nachgeschaltete Dialogbox Zellen aus.



Fordern Sie einen Chi-Quadrat Test an.



Fordern Sie auch die erwarteten Häufigkeiten an.

Fenster -> Ausgabe

Testen der Unabhängigkeit von 2 Variablen

BRAEUTIG * BRAUT Kreuztabelle							
			BRAUT				Gesamt
			ev	ohneB	rk	sonstB	
BRAEUTIG	ev	Anzahl	784	47	1193	14	2038
		Erwartete Anzahl	608,4	58,0	1314,9	56,8	2038,0
ohneB	Anzahl	Erwartete Anzahl	122	78	152	6	358
		Anzahl	106,9	10,2	231,0	10,0	358,0
rk	Anzahl	Erwartete Anzahl	1100	56	2987	25	4168
		Anzahl	1244,2	118,6	2689,1	116,1	4168,0
sonstB	Anzahl	Erwartete Anzahl	40	14	90	146	290
		Anzahl	86,6	8,3	187,1	8,1	290,0
Gesamt	Anzahl	Erwartete Anzahl	2046	195	4422	191	6854
		Anzahl	2046,0	195,0	4422,0	191,0	6854,0

Abb./Tab. 30: Relative und erwartete Häufigkeiten in einer 4x4 Kontingenztafel

Chi-Quadrat-Tests			
	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	3166,034 ^a	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.

Abb./Tab. 31: Ergebnis für den Chi-Quadrat Test

Der Chi-Quadrat Test liefert einen verschwindend kleinen und deshalb auf Null abgerundeten Wert für p. Sie sollten die Null-Hypothese H₀ deshalb verwerfen (aufgrund des extrem hohen Wertes der Teststatistik und der damit verbundenen extrem kleinen Irrtumswahrscheinlichkeit $\alpha < 0.00001$ für eine irrtümliche Ablehnung). Die Religionszugehörigkeit spielt sehr wohl eine Rolle bei der Auswahl des Ehepartners.

Aufgaben

1. Untersuchen Sie für das Datenmaterial aus der Arbeitsdatei **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 **hautf** (Hautfarbe des Verurteilten) auf Unabhängigkeit. Die Gewichtung (*Daten -> Fälle gewichten ...*) erfolgt über die Variable **anzahl1**. Messen Sie dieser Untersuchung politische Bedeutung zu?
2. Überlegen, wie Sie obige Ergebnis für **heirat.sav** begründen können. Hinweis: 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?

Berechnen von Korrelationskoeffizienten

In diesem Kapitel wird der (Pearson-) **Korrelationskoeffizient** behandelt, der ein Maß für die lineare Abhängigkeit zwischen zwei (intervall-skalierten) 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?

Motivation - Festlegen eines Maßes für den linearen Zusammenhang

Die Korrelation zweier intervall-skalierten Zufallsvariablen X und Y berechnet sich über Erwartungswert und Varianz¹¹:

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

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

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 ...

Vorgehensweise - Ermitteln des Korrelationskoeffizientens

Im folgenden Beispiel berechnen Sie für Abiturnoten aus `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:

Grafiken -> Streudiagramm > Einfach

¹¹ Für unabhängige Variablen ist die Korrelation gleich Null; d.h. aus Unabhängigkeit folgt Unkorreliertheit. Die Umkehrung gilt nicht.

Berechnen von Korrelationskoeffizienten



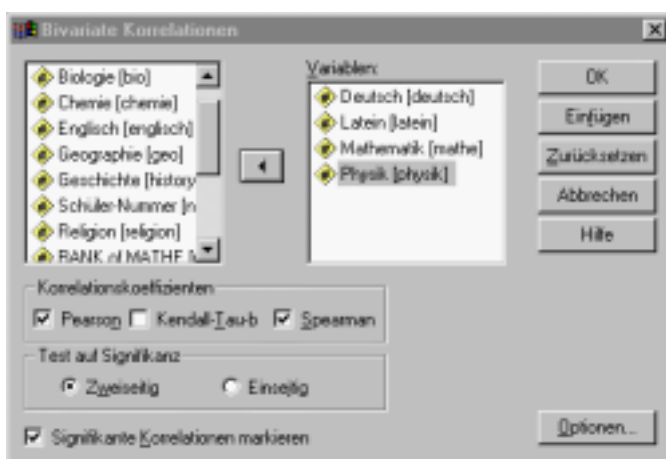
Wählen Sie alle Variablen aus, für die paarweise Streudiagramme erzeugt werden sollen, hier die Fächer Deutsch, Mathematik, Physik und Latein.

fenster -> Grafik-Karussell



Die Streudiagramme legen nahe, daß nur für Mathematik und Physik eine hohe emp. Korrelation besteht.

Analysieren -> Korrelation -> Bivariat



Wählen Sie alle Variablen aus, für die der emp. Korrelationskoeffizient berechnet werden soll, hier die Fächer Deutsch, Mathematik, Physik und Latein.

Fenster -> Ausgabe

Berechnen von Korrelationskoeffizienten

Korrelationen			Deutsch	Latein	Mathematik	Physik
Spearman-Rho	Deutsch	Korrelationskoeffizient	1,000	,419*	,171	,202
		Sig. (2-seitig)	,	,042	,424	,344
		N	24	24	24	24
	Latein	Korrelationskoeffizient	,419*	1,000	,234	,154
		Sig. (2-seitig)	,042	,	,270	,471
		N	24	24	24	24
	Mathematik	Korrelationskoeffizient	,171	,234	1,000	,737**
		Sig. (2-seitig)	,424	,270	,	,000
		N	24	24	24	24
	Physik	Korrelationskoeffizient	,202	,154	,737**	1,000
		Sig. (2-seitig)	,344	,471	,000	,
		N	24	24	24	24

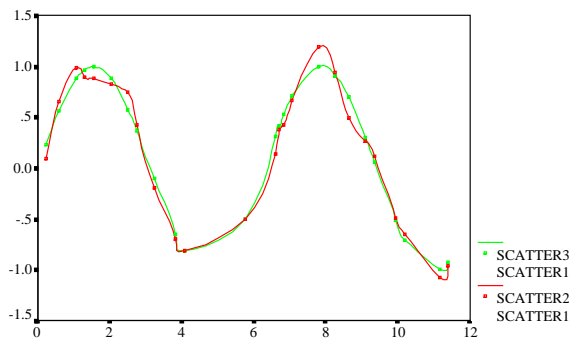
*. Korrelation ist auf dem Niveau von 0,05 signifikant (2-seitig).
 **. Korrelation ist auf dem Niveau von 0,01 signifikant (2-seitig).

Abb./Tab. 32: Emp. Korrelationskoeffizienten

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

Exkurs – nicht-lineare Zusammenhänge

Das folgende Streudiagramm für einen sinusförmigen Zusammenhang zwischen $x = \text{scatter1}$ und $y = \text{scatter2}$ verdeutlicht, daß ein kleiner emp. Korrelationskoeffizient r nur eine Aussage über einen **linearen** Zusammenhang ermöglicht, nicht aber über andere (nicht-lineare Zusammenhänge) wie z.B. einen sinusförmigen Zusammenhang zwischen X und Y wie in $Y = \sin(X)$, entsprechend auch: logarithmisch $Y = \ln(X)$, exponentiell $Y = \exp(X)$ oder polynomial $Y = a_n X^n + \dots + a_0$.



Die überlagerte Sinus-Kurve für scatter1 und $\text{scatter3} = \sin(\text{scatter1})$ macht deutlich, daß ein signifikanter nicht-linearer Zusammenhang, hier: sinusförmig, zwischen scatter1 und scatter2 existiert.

Der emp. Korrelationskoeffizient r , der nur den **linearen** Zusammenhang erfaßt, beträgt -0.44.

Aufgaben

- Nehmen Sie in die Untersuchung der Korrelation für `schueler.sav` zusätzlich die Schulfächer Biologie und Chemie auf.
- Der sozio-ökonomische Status (*socioeconomic status*, **SES**) einer Person werde auf einer Skala von 11 (niedrig) bis 77 (hoch) gemessen. SES ist dabei ein 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-

Berechnen von Korrelationskoeffizienten

Achse) und **vater** (x-Achse).

(*) Wie würden Sie SES definieren?

Hinweis: Unterscheiden Sie zwischen Familien mit niedrigem, mittlerem und hohem SES.

Beachten Sie, daß SES nach oben und unten beschränkt ist.

3. (*) Für ordinal-skalierte Variablen wie SES empfiehlt sich anstelle des **Pearson** der **Spearman Rang-Korrelationskoeffizient**, da er sich auf die rang-transformierten 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 .

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, sowie alternative Modelle zur Approximation von x-y-Beobachtungspunkten.

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?

Motivation - 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.

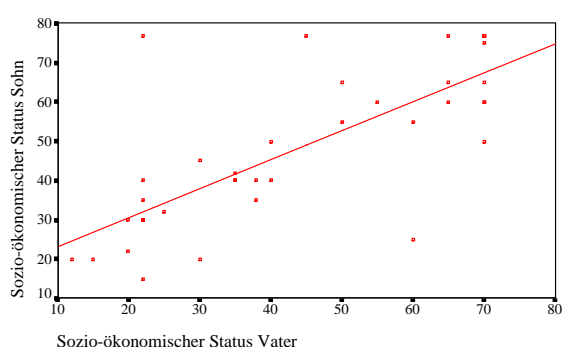
Bei einer Regression postulieren Sie allgemein ein Modell, bei dem eine abhängige Variable **Y funktional** von einer unabhängigen Variablen X abgeleitet wird, im Fall der linearen Regression wird einschränkend ein lineares Modell vermutet.

Sie untersuchen mit der linearen Regression die Null-Hypothese H^{12} , 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.

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

So könnte z.B. die Vermutung bestehen, daß die Anzahl der verkauften Bücher Y für jedes Jahr X steigend ist.

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 $g(x)$ kleinstmöglich ist (**Gauß'sche Methode der kleinsten Quadrate**). Der für X erwartete Wert von Y auf der Regressions-Geraden wird im folgenden als Y^{\wedge} (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. $y_i^{\wedge} = g(x_i)$.

Bei der linearen Regression wird zusammenfassend 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)

¹² Die Alternative umfaßt alle nicht-linearen oder keinerlei Zusammenhänge.

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

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)

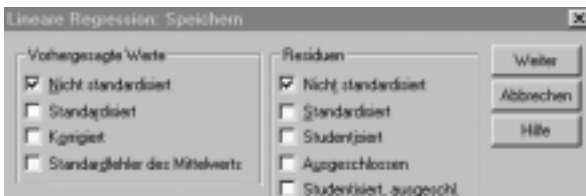
Vorgehensweise - Durchführen einer linearen Regression

In diesem Beispiel untersuchen Sie die Variablen **anz** (Anzahl produzierter Bücher) und **jahr** aus `bu-echer.sav` in Hinblick auf einen linearen Zusammenhang:

Analysieren -> Regression -> Linear



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



Berechnen Sie zusätzliche Variablen für erwarteten Wert \hat{y}_i und Konfidenzintervalle für \hat{y}_i (siehe unten). Fordern Sie u.a. obere und untere Grenze der Konfidenzintervalle als neue Variablen der Arbeitsdatei an.

Fenster -> Ausgabe (hier modifiziert!)

Variable	B					
JAHR	1583.6 = m					
(Constant)	-3081413.9 = b	-> $g(x) = 1593.6 x + (-3081413.9)$				
Koeffizienten ^a						
Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	-3081414	218574,437		-14,098	,000
	Jahr	1583,687	110,586	,938	14,321	,000

a. Abhängige Variable: Anzahl

Abb./Tab. 33: Ergebnis der linearen Regression

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

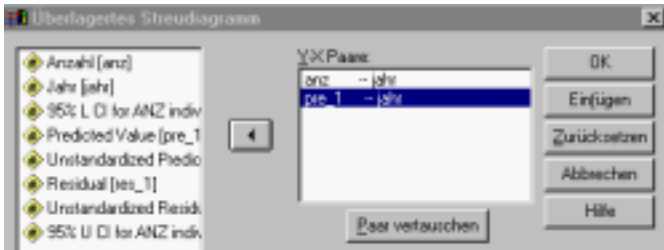
Die Gleichung der Regressions-Geraden $g(x)$ lautet:

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

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

Erzeugen Sie nun auf Grundlage der neuen Variablen **pre_1** (erwarteter Wert) überlagerte Streudiagramme mit den Beobachtungspunkten und der Regressionsgeraden:

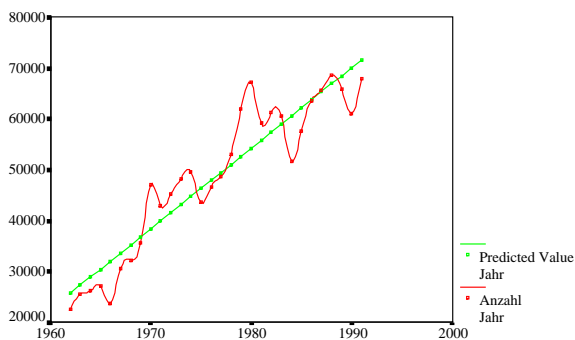
Grafiken -> Streudiagramm -> Überlagert



Geben Sie die y-x Paare ein, die im Streudiagramm überlagert dargestellt werden sollen.

Verwenden Sie als x-Variable jeweils **jahr** und als y-Variable **anz** (Anzahl) und **pre_1** ($g(x)$).

Fenster -> Karussell (hier nachbearbeitet!)



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

Statistischer Hintergrund - 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 $n > 3$ i.d.R. 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**):

$$(1) \quad \mathbf{SSE} = [y_1 - (m \cdot x_1 + b)]^2 + [(y_2 - (m \cdot x_2 + b))]^2 + \dots + [y_n - (m \cdot x_n + b)]^2$$

$$\leq [y_1 - (m \cdot x_1 + b)]^2 + [(y_2 - (m \cdot x_2 + b))]^2 + \dots + [y_n - (m \cdot x_n + b)]^2$$

für beliebige b und m

(SSE: *Sum of Squares Errors or Residual*)

Die Schätzwerte b (Achsenabschnitt, *intercept*) und 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.

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

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:

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

Das Verhältnis¹³ 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:

$$(4) \quad F = (SSM/1) / (SSE/(n-2)) \quad \text{ist verteilt nach } \mathbf{F(x;1,n-2)}$$

$\mathbf{F(x;1,n-2)}$ 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 SSE¹⁴ 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 Varianz beiträgt (vgl. (3)). Sie können die Testgröße F für einen formalen Hypothesentest verwenden, da die Verteilung von F bekannt ist.

Exkurs: Vorgehensweise -Approximieren durch andere Kurven

Falls die Approximation durch das lineare Modells für Ihre Zwecke nicht ausreichend ist; d.h. der Zusammenhang zwischen Y und X nicht ausreichend erklärt wird, sollten Sie ein anderes Modell verwenden, z.B. durch Hinzunahme weiterer erklärender Variablen oder durch Verwenden eines anderen funktionalen Zusammenhangs wie z.B in den Modellgleichungen $Y=aX^2+bX$ oder $Y=\ln(X)$, $Y=\exp(X)$, $Y=X^{1/2}$

Aufgaben

1. Führen Sie für das Datenmaterial aus `allbus90.sav` eine lineare Regression für die Abhängigkeit zwischen Alter (unabhängige Variable) und Einkommen (abhängige Variable) durch.
2. Führen Sie für das Datenmaterial aus `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`.
3. Welche Prognosen können Sie aus den linearen Modellen aus Aufgabe 1 für das Jahr 2000 ablesen (*forecasting*) und inwieweit können Sie den Prognosen vertrauen? Hinweis: $g(x) = mx + b$, $x = 2000$

¹³ 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.

¹⁴ Optimal wäre SSE=0. In diesem Fall besteht ein genauer linearer Zusammenhang. Je größer SSE wird, desto größer sind die störenden Einflüsse.

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

4. (*) Führen Sie für das Datenmaterial aus `inform.sav` (Informatikstudenten) eine Kurvenanpassung der Variablen `stud_m`, `stud_w` und `stud_ges` nach `jahr` (Jahr) durch. Testen Sie ggf. auch andere Modelle (linear, quadratisch, ...).

Vergleichen von 2 Gruppenmittelwerten (t-Test)

In diesem Kapitel werden Verfahren vorgestellt, um die gemessenen arithmetischen Mittelwerte zweier Gruppen miteinander zu vergleichen und zu entscheiden, ob ein Unterschied zwischen den Gruppen zufällig zu erklären ist oder nicht zufällig (**signifikant**) ist.

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

Motivation - 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.

Als Beispiel dient die Variable Behandlungserfolg (= "Senkung des Blutdruckes bei Bluthochdruck-Patienten"), 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 .

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

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.

Vorgehensweise - Testen auf gleiche Erwartungswerte

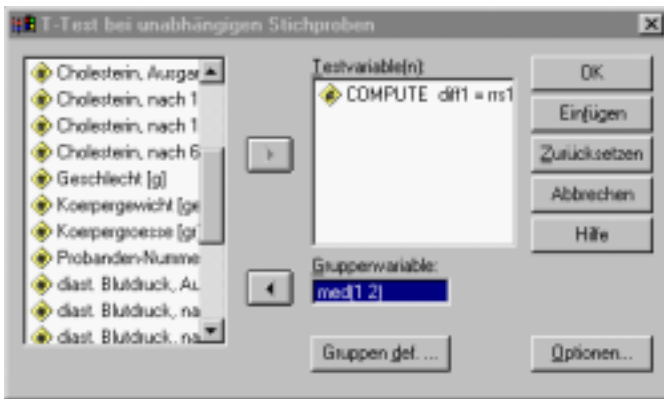
Im folgenden Beispiel werden für `hyper.sav` die beiden fiktiven Medikamente **alphasan** (`med=1`) und **betasan** (`med=2`) hinsichtlich der Senkung des Blutdrucks während einer 1-monatigen Behandlung (`diff=rrs1-rrs0`) untersucht.

Definieren Sie zunächst eine neue Variable `diff1=rrs1-rrs0` (Behandlungserfolg durch Absenkung des Blutdrucks). Die Null-Hypothese H lautet, daß die Erwartungswerte μ_1 und μ_2 von `diff1` für die Gruppen, die durch Medikament 1 bzw. 2 festgelegt werden, übereinstimmen.

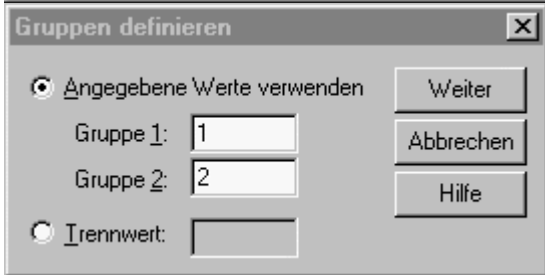
Führen Sie zur statistischen Absicherung Ihrer Vermutung, daß die Mittelwerte signifikant unterschiedlich sind, einen statistischen Test durch, hier den **t-Test**. Ihre Null-Hypothese H – die Sie verwerfen wollen - lautet hierbei, daß der Erfolg der Medikamente in Hinblick auf Blutdrucksenkung gleich ist.

Analysieren -> Mittelwertvergleiche -> t-Test für unabhängige Stichproben

Vergleichen von 2 Gruppenmittelwerten (t-Test)



Wählen Sie die zu analysierende Variable aus, hier **diff1**, und die Variable, nach der gruppiert werden soll, hier **med**.



Zusätzlich müssen die beiden Werte eingegeben werden, nach denen die Gruppen unterschieden werden, hier: med=1 und med=2..

Fenster -> Ausgabe

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

Abb./Tab. 34: t-Test

Der t-Test liefert Ihnen den Wert der Teststatistik (*t-value*) und die zugehörige Irrtumswahrscheinlichkeit p . Die Irrtumswahrscheinlichkeit α , die Null-Hypothese $H_0 (\mu_1 = \mu_2)$ 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 Mittelwerten ist zu signifikant, um allein auf zufällige Schwankungen zurückgeführt werden zu können.

Aufgaben

- Führen Sie einen t-Test durch für die Variable Einkommen aus der Arbeitsdatei `allbus90.sav`, wobei Sie nach Geschlecht unterscheiden
- Führen Sie einen t-Test durch für die Variable `physik` (Abiturnote einer Klasse in Physik) aus `schueler.sav`, wobei Sie nach `sex` (Geschlecht) unterscheiden.
- (*) Vergleichen Sie mit einem nicht-parametrischen Test wie z.B. den **Mann-Whitney U-Test**, der nicht die Mittelwerte, sondern die Ränge, miteinander vergleicht.
Hinweis:
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 (Robustheit)?

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 ...

Motivation - 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, $i=1, \dots, m$. (Beim t-Test gilt $m=2$.)

Im folgenden Beispiel soll die Merkfähigkeit in Abhängigkeit vom Alter untersucht werden.

Sie wollen nun i.d.R. die Null-Hypothese H überprüfen, daß alle Gruppen den selben Erwartungswert μ besitzen (zusammengesetzte Hypothese, da nicht ein Vergleich, sondern viele Vergleiche durchgeführt werden müssen, bei $n=4$ z.B. 6 Vergleiche):

$$H: \quad \begin{aligned} \mu_1 &= \mu_2 = \dots = \mu_m \\ \mu_1 &= E(X|G_1) \text{ usw.} \end{aligned}$$

Bei der **Varianz-Analyse** zerlegen Sie nun 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 entscheiden, ob die Gruppen hinsichtlich ihrer Erwartungswerte als "gleich" (SSM klein im Vergleich zu SSE) oder als "unterschiedlich" (SSE klein im Vergleich zu SSM) zu betrachten sind¹⁵.

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 Gruppenmittelwerte auf signifikante Unterschiede:

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

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

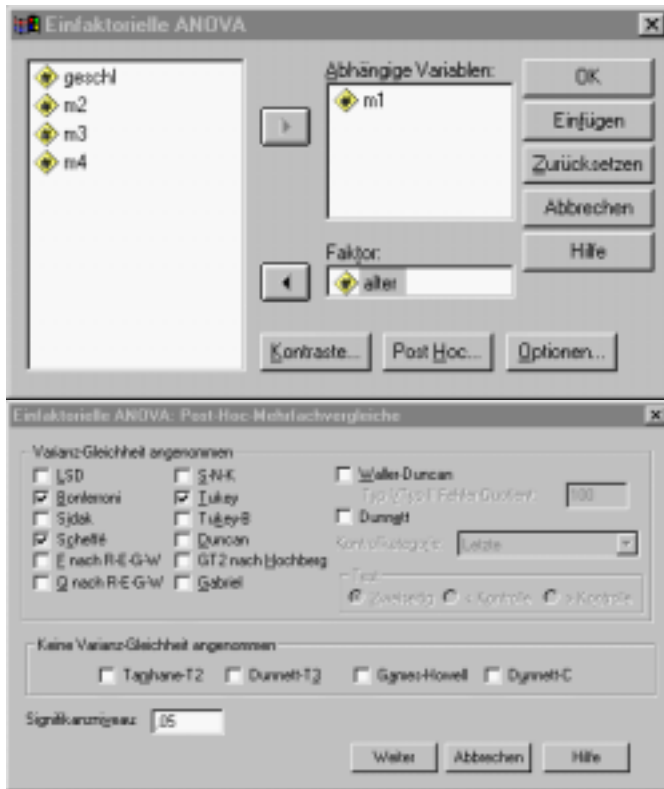
Vorgehensweise - Vergleichen von mehreren unabhängigen Stichproben

Im folgenden Beispiel verwenden Sie einen Merkfähigkeitstest in der Arbeitsdatei `varana.sav`, bei dem die Merkfähigkeit `m1` untersucht werden soll. Als Gruppenvariable dient die Variable `alter` (Altersklasse) mit 3 unterschiedlichen Ausprägungen.

Analysieren-> Mittelwertvergleiche -> Ein-faktorielle ANOVA

¹⁵ 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.

Vergleichen mehrerer Gruppenmittelwerte (Varianz-Analyse)



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

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

Fenster -> Ausgabe

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			

Abb./Tabelle 35: Ergebnis der Varianz-Analyse (I)

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.

Vergleichen mehrerer Gruppenmittelwerte (Varianz-Analyse)

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.

Abb./Tabelle 36: Ergebnis der Varianz-Analyse (II)

Exkurs: Motivation - 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.

Vorgehensweise - Durchführen einer 2-faktoriellen Varianzanalyse

Im folgenden Beispiel untersuchen Sie erneut den Merkfähigkeitstest. Als Faktoren werden nun die Variablen **gesch1** (Geschlecht) mit 2 unterschiedlichen Ausprägungen (Faktor-Stufen) und **alter** (Altersklasse) mit 3 unterschiedlichen Ausprägungen verwendet. Es gibt mithin insgesamt 6 Zellen im Design:

$$(\text{Anzahl Faktor-Stufen gesch1}) \times (\text{Anzahl Faktor-Stufen alter}) = 2 \times 3 = 6$$

Faktor 1 / Faktor 2	alter=1	alter=2	Alter=3
sex=1	Zelle 1/1	Zelle 1/2	Zelle 1/3
sex=2	Zelle 2/1	Zelle 2/2	Zelle 2/3

Die Auswertung erfolgt dann über *Statistik* -> *ANOVA Modelle* -> *Einfach mehrfaktoriell*

Exkurs: Statistischer Hintergrund - 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.

Aufgaben

- 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),

Vergleichen mehrerer Gruppenmittelwerte (Varianz-Analyse)

s (Schulabschluß), **pu** (Putzhäufigkeit), **zb** (Wechsel der Zahnbürste) und **beruf** als Faktoren durch. Wie lauten Ihre Hypothesen? Sollten die Hypothesen verworfen oder nicht verworfen werden? Welche Mittelwerte sind jeweils signifikant unterschiedlich (**post-hoc** Tests)?
(*) Führen Sie eine mehrfach-faktorielle Varianzanalyse durch. Untersuchen Sie auch die Interaktionsterme.

2. (*) Welche Voraussetzungen sollten **vor** einer Varianz-Analyse überprüft werden und welche Möglichkeiten gibt es hierzu? (Stichworte: Normalverteilung, Varianzhomogenität)
3. (*) Besteht für das Einkommen ein signifikanter Unterschied bzgl. der Schulbildung? Verwenden Sie die Variablen Einkommen und Schulbildung aus `alibus90.sav`, um diese Fragestellung zu untersuchen. Welche anderen Faktoren könnten einen Einfluß haben? (Geschlecht?)

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, ...

Motivation - Ermitteln von gemeinsamen Faktoren

Ausgangspunkt einer Faktor-Analyse ist eine Vielzahl von Variablen, bei denen nicht a-priori bekannt ist, ob in und welcher Weise sie miteinander etwas zu tun haben. Gesucht werden „**Hintergrund-Variablen**“ wie z.B. „Kreativität“, „Qualifikation“, „Allgemeine oder sprachliche Intelligenz“, die im Rahmen der Faktor-Analyse als **Faktoren** bezeichnet werden. Besonders bei einer großer ("unüberschaubaren") Anzahl von Variablen wird bei Ihnen der Wunsch bestehen, diese auf einige grundlegende, allerdings von Ihnen nicht direkt beobachtbare oder quantifizierbare Hintergrund-Variablen oder Faktoren zurückzuführen.

Die "Lebensgefühl in einem Wohngebiet" stellt z.B. 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.

Ein anderer, nicht direkt meßbarer Faktor ist die "Liebe" zwischen zwei Personen, der sich in Antworten auf Fragen wie "Schickt mir Liebesbriefe", "Hört mir immer bewundernd zu", "Liest mir Wünsche von den Augen ab" usw. manifestiert.

Sie verwenden die **Faktor-Analyse** als ein mathematisches (!) 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 Mengen¹⁶ 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 in Schritt 4, 7 und 8 Entscheidungen treffen bzw. eine Interpretation vornehmen müssen, die restlichen, sehr technischen Schritte werden automatisch von SPSS durchgeführt:

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

¹⁶ 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.

Reduzieren der Variablenanzahl (Faktor-Analyse)

Vorgehensweise - Durchführen einer Faktoren-Analyse

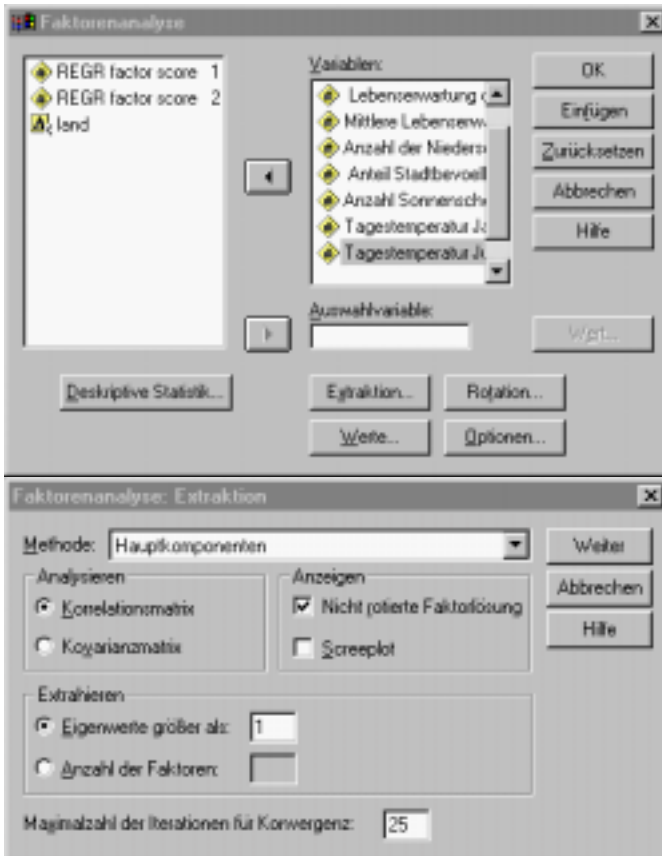
Im folgenden führen Sie eine Faktor-Analyse für die demoskopischen und klimarelevanten Variablen **sb** (Anteil der Stadtbevölkerung), **lem** (Lebenserwartung der Männer), **lew** (Lebenserwartung der Frauen), **ks** (Kindersterblichkeit), **so** (Anzahl Sonnenscheintage pro Jahr), **nt** (Anzahl Regentage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

Die Korrelationsmatrix sieht folgendermaßen aus:

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

Abb./Tab. 37: Korrelationsmatrix der an der Faktor-Analyse beteiligten Variablen

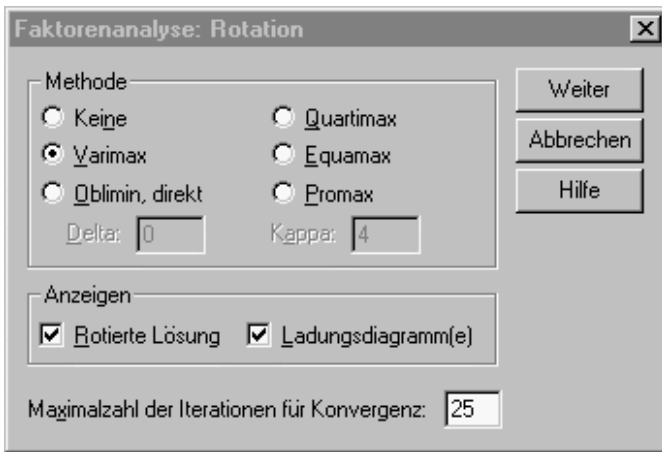
Analysieren -> Dimensionsreduktion -> Faktor-Analyse



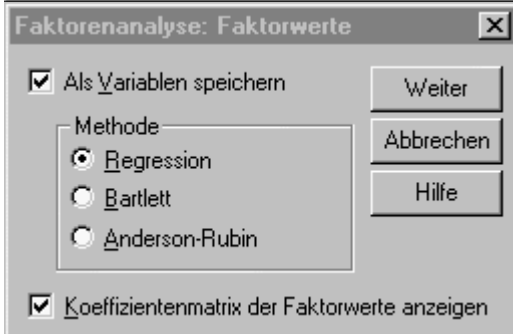
Wählen Sie die Variablen aus, die in eine mit Standardeinstellungen durchgeführte Faktor-Analyse einbezogen werden sollen, hier:) **sb**, **lem**, **lew**, **ks**, **so**, **nt**, **tjan** und **tjul**

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

Reduzieren der Variablenanzahl (Faktor-Analyse)



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).

Fenster -> Ausgabe

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

Abb./Tab. 38: Eigenwerte und Anteil an Gesamtvarianz

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

Reduzieren der Variablenanzahl (Faktor-Analyse)

Rotierte Komponentenmatrix ^a		
	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.

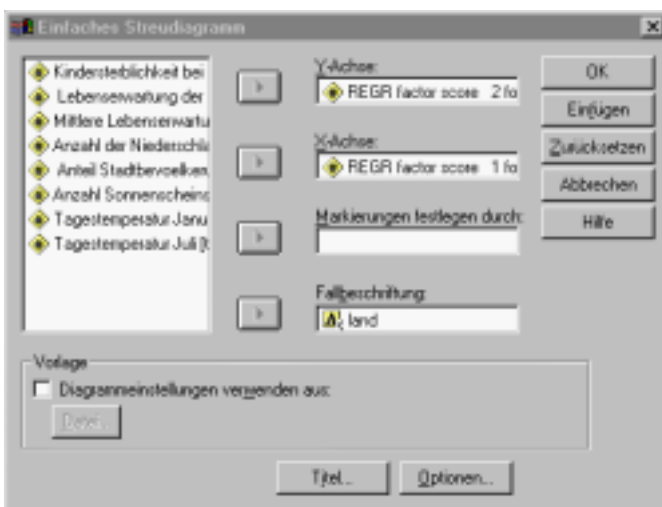
a. Die Rotation ist in 3 Iterationen konvergiert.

Abb./Tab. 39: Faktorladungen (factor score coefficients)

Die von SPSS vorgenommene „Rotation“ verfolgt das Ziel, jede Variable auf nur genau einen Faktor "hochzuladen". Sie können 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 ¹⁷.

Erstellen Sie nun das Streudiagramm für die beiden Faktoren:

Grafiken -> Streudiagramm -> Einfach



Tragen Sie die beiden Faktoren gegeneinander auf und wählen Sie die Variable **land** als Fallbeschriftung.

Fenster -> Grafik-Karussell

¹⁷ Eine Interpretation der Faktoren ist einfacher, wenn alle einem Faktor zugeordneten Variablen zum Faktoren 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. ks_neu=1-ks.

Reduzieren der Variablenanzahl (Faktor-Analyse)



Der 1. Faktor ist ein Maß für "Lebensdauer", der 2. Faktor 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 oben zu finden.

Stat. Hintergrund - Reduzieren der Variablenanzahl

Zunächst nehmen Sie an, daß sich p von Ihnen ausgewählte Variablen $X_1 - X_p$ als Linearkombinationen von k (zunächst unbekannt) Faktoren $F_1 - F_k$ ($k < p$) darstellen lassen¹⁸:

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

Hierbei sind $F_1 - F_k$ 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 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.

Aufgaben

1. Führen Sie für eine Untersuchung über die Einstellung zu Ausländern in `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"):

`a01`: Die Integration der Ausländer muß verbessert werden.

`a02`: Das Flüchtlingselend muß gemindert werden.

¹⁸ Es sollte $k < p$ gelten, denn sonst führt dieser Ansatz nicht zu einer Reduktion der Variablenanzahl.

Reduzieren der Variablenanzahl (Faktor-Analyse)

- 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.

Hinweis

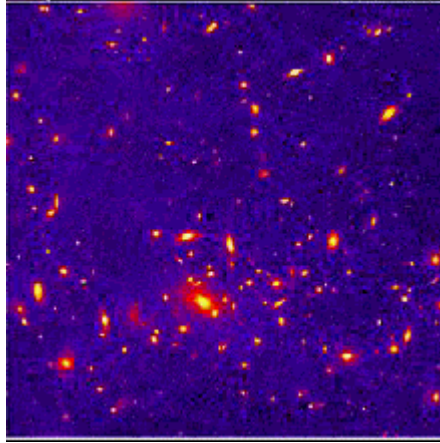
Wählen Sie 3 Faktoren aus. Eine mögliche Aufteilung in 3 Gruppen ist z.B. (1,12,13,15,3,4,7), (5,7), (6,14).

2. Transformieren Sie die Variablen aus der Arbeitsdatei `europa.sav` so in neue Variablen, so daß für diese neuen Variablen nur positive Korrelationen zu "hochgeladenen" Faktoren bestehen (siehe Fußnote). Interpretieren Sie die neuen Faktoren hinsichtlich ihrer "realen" Bedeutung. Hinweis: Die Variablen `ks` und `nt` mit „negativer“ Bedeutung sollten in Variablen mit „positiver“ Bedeutung transformiert werden. Vorschlag: `ks-neu=1000-ks`, `nt-neu=365-nt`. Verwenden Sie erneut 2 Faktoren.
3. (*) In einer Studie zu Frühgeburten (`fruehgeb.sav`) werden die Faktoren *allgemeine Intelligenz* (AI) und *sprachliche Intelligenz* (SI) vermutet. Versuchen Sie, hierfür eine Faktoren-Analyse durchzuführen. (siehe Brosius/Brosius, 1995, S817ff)

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** zusammengefaßt werden.

Siehst Du 1000 Sternlein stehen ... (<http://aitzu3.ait.physik.uni-tuebingen.de/~stuhli/html/astro.html>)



Motivation - Zusammenfassen von Beobachtungen

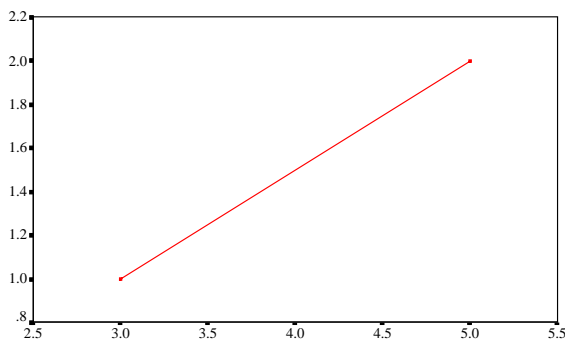
Aufgrund der grafischen Darstellung von intervall-skalierten Beobachtungen in Streudiagrammen erkennen Sie häufig visuell auffällige Ansammlungen oder Punktwolken von „dicht zusammenliegenden Beobachtungen“.

Die **Cluster-Analyse** liefert Ihnen Verfahren, um derartige Cluster (Gruppen, 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 nur 2 Variablen läßt sich der Abstand geometrisch 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:

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

Zusammenfassen von Beobachtungen in Clustern (Cluster-Analyse)

Der Abstand zwischen zwei Clustern C_1 und C_2 lässt sich z.B. über folgende 2 Abstandsmaße definieren (**mittlerer Abstand** bzw. **maximaler Abstand** zwischen Beobachtungspunkten aus unterschiedlichen Clustern):

1. **mittlerer Abstand:**

$$\text{Abstand_Zwischen_Cluster}(C_1, C_2) = \sum_{ij} \text{Abstand}(P_i, P_j) / (n \cdot m)$$

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:**

$$\text{Abstand_Zwischen_Cluster}(C_1, C_2) = \text{Maximum} \{ \text{Abstand}(P_i, P_j), i \in C_1, j \in C_2 \}$$

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 "aussagekräftige" Interpretation möglich ist.

Vorgehensweise - Ermitteln von hierarchisch geordneten Clustern

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

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*.

Zusammenfassen von Beobachtungen in Clustern (Cluster-Analyse)

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.

Fenster -> Ausgabe

Die oben angeforderte Cluster-Analyse liefert eine sehr umfangreiche Ausgabe, u.a. eine **Abstandsmatrix** für die Beobachtungen, einen **Ablauf der Verschmelzung** (Fusionierungsschema, *Agglomeration Schedule*), d.h. die Reihenfolge, in der Beobachtungen bzw. bereits vorhandene Cluster zu neuen Clustern verschmolzen werden und ein **Dendrogramm**.

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	

Abb 40: Euklidische Abstandsmatrix zwischen den Beobachtungen (Ausschnitt)

Zusammenfassen von Beobachtungen in Clustern (Cluster-Analyse)

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

Abb 41: 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.

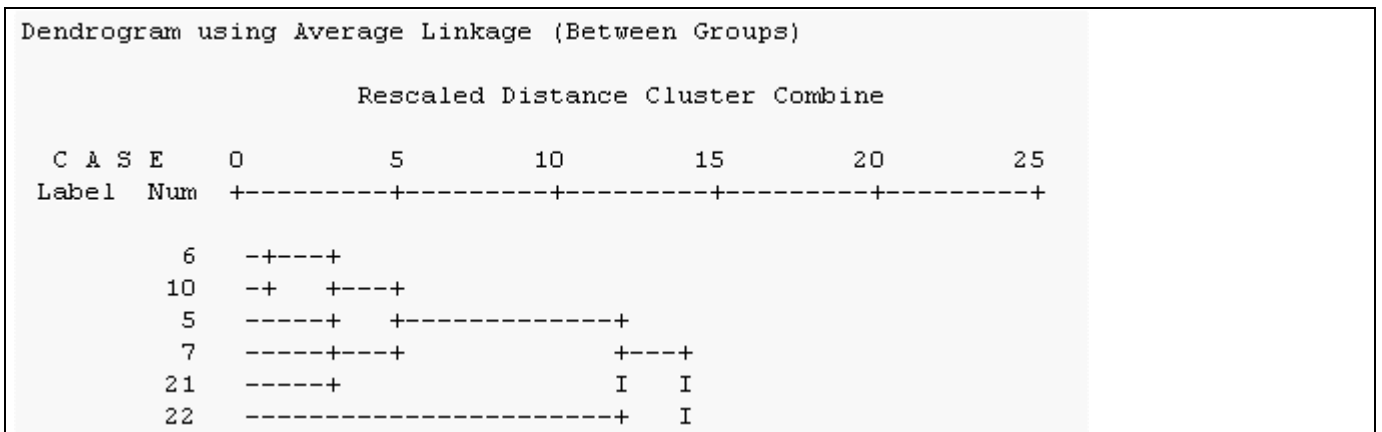


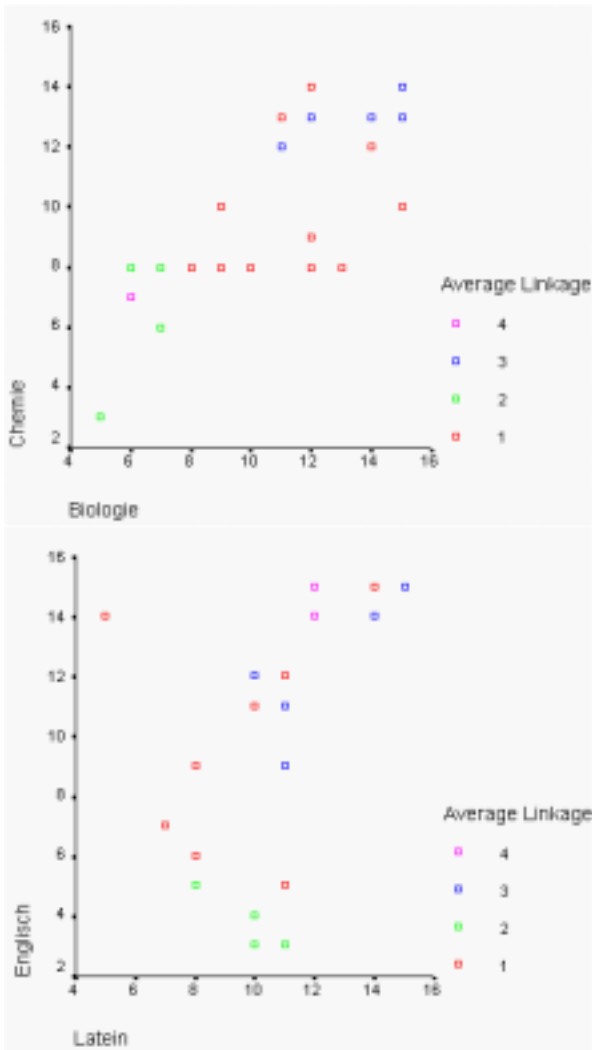
Abb 42: Dendrogram (Ausschnitt)

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

Zusammenfassen von Beobachtungen in Clustern (Cluster-Analyse)

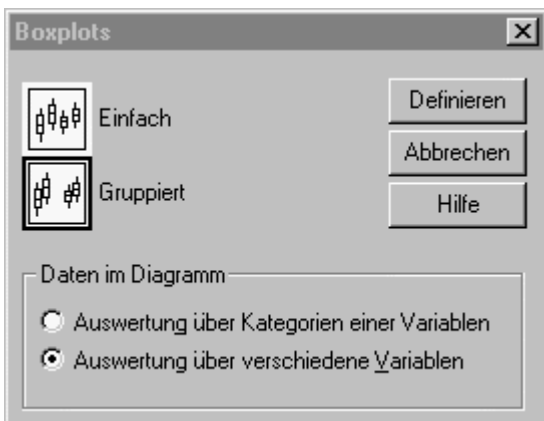


Das Streudiagramm zeigt die räumliche Verteilung der 4 Cluster bezogen auf die beiden Fächer Biologie und Chemie (leider nur in der PDF Version mit Farbe unterscheidbar, sorry!).

Das Streudiagramm zeigt die räumliche Verteilung der 4 Cluster bezogen auf die beiden Fächer Englisch und Latein (leider nur in der PDF Version mit Farbe unterscheidbar, sorry!).

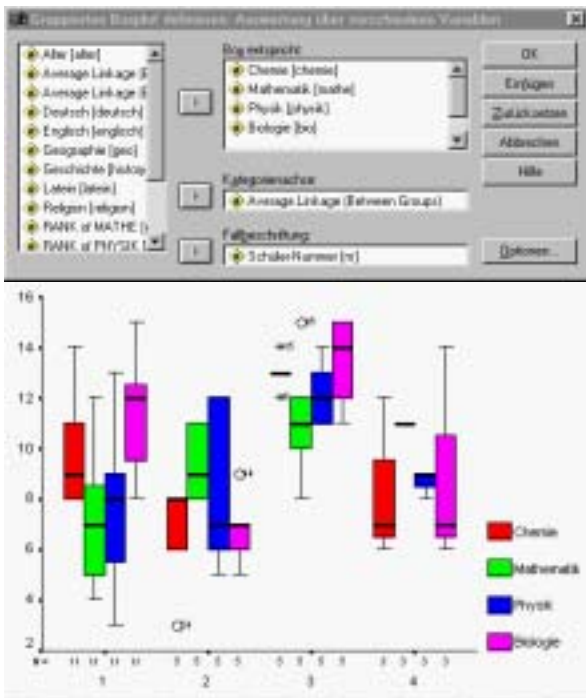
Stellen Sie die 4 Cluster zusätzlich in einem gruppierten Boxplot bzgl. der naturwissenschaftlichen Fächer gegenüber:

Grafiken -> Boxplot



Fordern Sie einen gruppierten Boxplot für Variablen an.

Zusammenfassen von Beobachtungen in Clustern (Cluster-Analyse)



Wählen Sie die naturwissenschaftlichen Fächer aus.

Die Trennung zwischen Cluster 3 und den anderen Clustern gelingt gut.

Cluster 3 enthält die „Spitzengruppe“ in den naturwissenschaftlichen Fächern – auch in den anderen?

Cluster 1 zeichnet sich durch bessere Noten in Chemie und Biologie und schlechteren Noten in Mathematik und Physik im Vergleich zu Cluster 2 mit besseren Noten in Mathematik und Physik und schlechteren in Chemie und Biologie..

Aufgaben

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. (*) Welche Auswirkungen haben unterschiedliche Wertebereiche bei den in der Cluster-Analyse beteiligten Variablen? Wie könnten alle Variablen "gleichermaßen" berücksichtigt werden?
Hinweis: Normierung auf einen einheitlichen Wertebereich [0,1] durch Z-Transformation.
4. (*) 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.
5. (*) Können Sie aus der Cluster-Analyse aus `schueler.sav` allgemeine Aussagen über Kombinationen von Begabungen ableiten?

Index

(mehrstufige Zufallsauswahl 22

.sav Format 11

Ablehnungsbereich 43

Abstand 77

Abstandsmaß 77

Abstandsmatrix 79

Annahmebereich 43

Anwendungsfenster 5

Ausreißer 35

Bedingung 12

benutzerdefinierte fehlende Werte 9

Beobachtung 21

Beobachtungen 12

Between-Groups-Variance 67

bivariate 21

Box- und Whisker-Plot 35

Chi-Quadrat-Test 53

Cluster 77

Cluster-Analyse 77

Datenanalyse 4

Dateneditor 4, 9

Datenformat 9

Datenmaterial 11

Datentyp 9

Datums- oder Zeitangaben 16

dBase 11

den **Mann-Whitney U-Test** 66

Dendrogramm 79

deskriptive Statistik 21

Dialogboxen 4

Dilemma 43

Eigenwerte 73

Eintrag 10

empirische Median 24

empirische Mittelwert 24

empirische Standardabweichung 24

empirische Varianz 24

empirische Verteilungsfunktion 25

Entscheidungsregeln 42

Erfolgswahrscheinlichkeit, 41

Erhebungsverfahren 22

Erhebungszeitraum 22

Erwartungswert 49

Etikett für den Variablenname 9

Etiketten für einzelne Werte 9

euklidischer Abstand. 77

Excel für Windows 11

Extremwerte 35

Faktor-Analyse 71

Faktor-Ladungen 73

Faktor-Stufen 69

Fehler 1. Art 43

Fehler 2. Art 43

Fragestellung 40

Funktionen 17

Fusionierungsschema 79

Gauß'sche Methode der kleinsten Quadrate
60

geordnete Stichprobe 24

Gesamterhebung 22, 49

grafische Benutzeroberfläche 4

Grenzwertsatz 42

Grundgesamtheit 21, 40

Güte der Modellgleichung 63

Hypothese 42

Hypothesentest 42

Interaktionsterme 69

Interpretation 1

Irrtumswahrscheinlichkeit 41, 51

Kenngößen 24

kleinste Informationseinheit 10

Konfidenz-Intervall 49

Konfidenz-Niveau 41

Kontingenztafel 53

Korrelationskoeffizient 56

Korrelationsmatrix 71

kritischer Wert 43

lineare Abhängigkeit 56

lineare Regression 60

lineares Modell 60, 69

mathematische Statistik 40, 41

Maximum 24

Menüsystem 4

Meßfehler 60

Messung 23

minimale Irrtumswahrscheinlichkeit 44

Minimum 24

Name 9

nominal-skaliert 22

Normalverteilung 41, 45

Normalverteilungskurve 34

Null-Hypothese 43

ordinal-skaliert 23

Parameter 41
Parametrische Statistik 41
Programmiersprache 4

Regressions-Gerade 60
relative Häufigkeit 24

Satz des Pythagoras 77
Schätzung 49
Schätzwert 42
Sicherheit 41
Sitzung 5
Spannweite (24
SPSS für Windows 1
Stichprobe 21, 22, 23, 40, 49
Stichprobenumfang 51
Syntax-Fenster 4

Test auf Normalverteilung 45
Test auf Varianzhomogenität 45
Testgröße 43
Teststatistik von Kolgomorov-Smirnov 45
Teststatistik von Levene 45
t-Test 65
t-Verteilung 50
Typen von falschen Entscheidungen 43

Übungen 1

Variablen 12, 21
Varianz des Modells 63
Varianzanalyse 67
Varianzhomogenität 45
vektorwertig 21
Verfahren 40
Verteilung 38
Verteilungsannahme 41
Vertrauensbereich 49
visuelle Darstellung 30

Wahrscheinlichkeit 38
Wechselwirkungen 69
Windows 1
Within-Group-Variance 67

zufälliger Fehler 60
Zufallsexperiment 38
Zufallsvariable 38
Zufallsvariablen 21
Zufallsvorgang 38