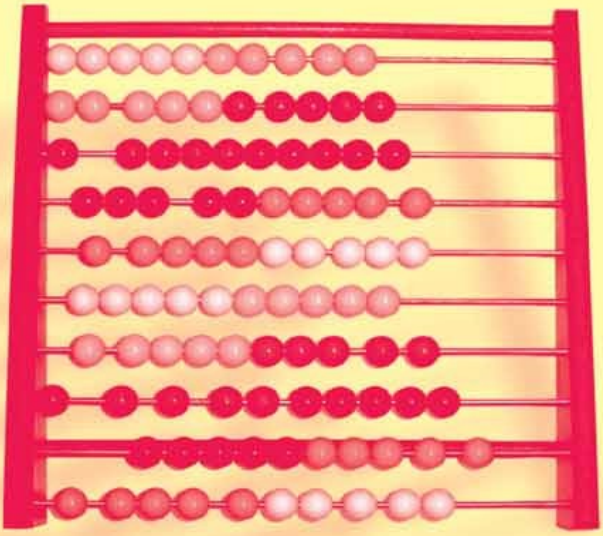
CHRISTINE
DULLER

Einführung in die Statistik mit EXCEL und SPSS

Ein anwendungsorientiertes Lehr- und Arbeitsbuch



Physica-Verlag
Ein Unternehmen
von Springer

Physica-Lehrbuch

Physica-Lehrbuch

Bannier, Christina E.
Vertragstheorie
Eine Einführung
mit finanzökonomischen Beispielen
und Anwendungen
2005, XVI, 218 S.

Basler, Herbert
Grundbegriffe der Wahrscheinlichkeitsrechnung und Statistischen Methodenlehre
11. Aufl. 1994, X, 292 S.

Bossert, Rainer · Manz, Ulrich L.
Externe Unternehmensrechnung
Grundlagen der Einzelrechnungslegung, Konzernrechnungslegung und internationalen Rechnungslegung.
1997, XVIII, 407 S.

Duller, Christine
Einführung in die Statistik mit EXCEL und SPSS
Ein anwendungsorientiertes Lehr- und Arbeitsbuch
2006, XII, 279 S.

Endres, Alfred
Ökonomische Grundlagen des Haftungsrechts
1991, XIX, 216 S.

Farmer, Karl · Wendner, Ronald
Wachstum und Außenhandel
Eine Einführung
in die Gleichgewichtstheorie
der Wachstums-
und Außenhandelsdynamik
2. Aufl. 1999, XVIII, 423 S.

Ferschl, Franz
Deskriptive Statistik
3. Aufl. 1985, 308 S.

Fink, Andreas
Schneidereit, Gabriele · Voß, Stefan
Grundlagen der Wirtschaftsinformatik
2. Aufl. 2005, XVIII, 316 S.

Gaube, Thomas u. a.
Arbeitsbuch Finanzwissenschaft
1996, X, 282 S.

Göcke, Matthias · Köhler, Thomas
Außenwirtschaft
Ein Lern- und Übungsbuch
2002, XIII, 359 S.

Graf, Gerhard
Grundlagen der Volkswirtschaftslehre
2. Aufl. 2002, XIV, 335 S.

Graf, Gerhard
Grundlagen der Finanzwissenschaft
2. Aufl. 2005, XII, 334 S.

Hax, Herbert
Investitionstheorie
5. Aufl., korrigierter Nachdruck
1993, 208 S.

Heiduk, Günter S.
Außenwirtschaft
Theorie, Empirie und Politik der interdependenten Weltwirtschaft
2005, XII, 429 S.

Heno, Rudolf
Jahresabschluss nach Handelsrecht, Steuerrecht und internationalen Standards (IAS/IFRS)
4. Aufl. 2004, XIX, 535 S.

Hofmann, Ulrich
Netzwerk-Ökonomie
2001, X, 242 S.

Huch, Burkhard u. a.
Rechnungswesen-orientiertes Controlling
Ein Leitfaden für Studium
und Praxis
4. Aufl. 2004, XX, 510 S.

Kistner, Klaus-Peter
Produktions- und Kostentheorie
2. Aufl. 1993, XII, 293 S.

Kistner, Klaus-Peter
Optimierungsmethoden
Einführung
in die Unternehmensforschung
für Wirtschaftswissenschaftler
3. Aufl. 2003, XII, 293 S.

Kistner, Klaus-Peter
Steven, Marion
Produktionsplanung
3. Aufl. 2001, XIII, 372 S.

Kistner, Klaus-Peter
Steven, Marion
Betriebswirtschaftslehre im Grundstudium
Band 1: Produktion, Absatz, Finanzierung
4. Aufl. 2002, XIV, 510 S.
Band 2: Buchführung, Kostenrechnung, Bilanzen
1997, XVI, 451 S.

König, Rolf
Wosnitza, Michael
Betriebswirtschaftliche Steuerplanungs- und Steuerwirkungslehre
2004, XIV, 288 S.

Kortmann, Walter
Mikroökonomik
Anwendungsbezogene Grundlagen
3. Aufl. 2002, XVIII, 674 S.

Kraft, Manfred · Landes, Thomas
Statistische Methoden
3. Aufl. 1996, X, 236 S.

Marti, Kurt · Gröger, Detlef
Einführung in die lineare und nichtlineare Optimierung
2000, VII, 206 S.

Marti, Kurt · Gröger, Detlef
Grundkurs Mathematik für Ingenieure, Natur- und Wirtschaftswissenschaftler
2. Aufl. 2003, X, 267 S.

Michaelis, Peter
Ökonomische Instrumente in der Umweltpolitik
Eine anwendungsorientierte Einführung
1996, XII, 190 S.

Nissen, Hans-Peter
Einführung in die makroökonomische Theorie
1999, XVI, 341 S.

Nissen, Hans-Peter
Das Europäische System Volkswirtschaftlicher Gesamtrechnungen
5. Aufl. 2004, XVI, 362 S.

Risse, Joachim
Buchführung und Bilanz für Einsteiger
2. Aufl. 2004, VIII, 296 S.

Schäfer, Henry
Unternehmensfinanzen
Grundzüge in Theorie
und Management
2. Aufl. 2002, XVIII, 522 S.

Schäfer, Henry
Unternehmensinvestitionen
Grundzüge in Theorie
und Management
2. Aufl. 2005, XVI, 439 S.

Sesselmeier, Werner
Blauermel, Gregor
Arbeitsmarkttheorien
2. Aufl. 1998, XIV, 308 S.

Steven, Marion
Hierarchische Produktionsplanung
2. Aufl. 1994, X, 262 S.

Steven, Marion
Kistner, Klaus-Peter
Übungsbuch zur Betriebswirtschaftslehre im Grundstudium
2000, XVIII, 423 S.

Swoboda, Peter
Betriebliche Finanzierung
3. Aufl. 1994, 305 S.

Tomann, Horst
Volkswirtschaftslehre
Eine Einführung
in das ökonomische Denken
2005, XII, 186 S.

Weise, Peter u. a.
Neue Mikroökonomie
5. Aufl. 2005, XI, 645 S.

Zweifel, Peter
Heller, Robert H.
Internationaler Handel
Theorie und Empirie
3. Aufl. 1997, XXII, 418 S.

Christine Duller

Einführung in die Statistik mit EXCEL und SPSS

Ein anwendungsorientiertes Lehr- und Arbeitsbuch

Mit 70 Abbildungen
und 25 Tabellen

Physica-Verlag
Ein Unternehmen
von Springer

Ass. Professorin DI Dr. Christine Duller

Johannes Kepler Universität Linz
IFAS – Institut für Angewandte Statistik
Altenberger Straße 69
4040 Linz
Österreich
E-mail: christine.duller@jku.at

ISBN 3-7908-1641-8 Physica-Verlag Heidelberg

Bibliografische Information Der Deutschen Bibliothek

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

Physica-Verlag ist ein Unternehmen von Springer Science+Business Media
springer.de

Umschlaggestaltung: Erich Kirchner
Herstellung: Helmut Petri
Druck: Strauss Offsetdruck

SPIN 11540601 154/3153 – 5 4 3 2 1 0 – Gedruckt auf säurefreiem Papier

Vorwort

Dieses Buch soll auf allgemein verständlichem Niveau die Grundlagen der Statistik vermitteln. Die LeserInnen sollen die Fähigkeit erwerben, die vorgestellten statistischen Verfahren korrekt anwenden zu können und die daraus resultierenden Ergebnisse richtig und verständlich zu interpretieren. Voraussetzungen sind Mathematik auf Maturaniveau und Grundkenntnisse im Umgang mit dem Computer, insbesondere mit EXCEL. Um das Verständnis zu erleichtern werden zahlreiche Beispiele mit Lösungen angeführt, wobei viele Beispiele auch mit den gängigen Programmen EXCEL und SPSS gelöst werden.

Bücher bieten aber nicht nur Inhalte, sondern transportieren immer auch Überzeugungen der Autorinnen und Autoren. Dieses Lehrbuch soll den Studierenden ein nützliches Hilfsmittel sein und auch einen Beitrag zur bewußten Integration von Studentinnen leisten. Die sonst üblichen ausschließlich männlichen Bezeichnungen werden durch eine Ausdrucksweise ersetzt, die alle Studierende ansprechen soll und daher zwischen männlicher, weiblicher und gemischter Ausdrucksweise abwechselt.

Im ersten Teil des Buches werden grundsätzliche Überlegungen zum Thema Statistik, sowie Kurzeinführungen in EXCEL und SPSS geboten. Der zweite Teil setzt sich intensiv mit den Verfahren der deskriptiven Statistik auseinander und legt so den Grundstein für erste statistische Analysen. Die Wahrscheinlichkeitsrechnung, die im dritten Teil beleuchtet wird, dient als Grundlage für die induktive Statistik, die im vierten und letzten Teil behandelt wird.

Um auch den Zweck als Nachschlagewerk zu erfüllen, ist im Anhang neben den wichtigsten Verteilungstabellen auch eine Übersicht mit Notationen und Formeln angeführt. Unter http://www.ifas.jku.at/personal/duller/duller.htm wird ein Link zu diesem Buch angeboten, wo man Ergänzungen und ausführlichere Lösungen zu den Beispielen findet.

Mein Dank gilt Norbert Faschinger, der mir vor allem beim Erstellen der zahlreichen Grafiken eine große Hilfe war. Für Anregungen und Korrekturvorschläge bedanke ich mich bei meinen Testleserinnen, insbeondere bei Frau Eva Obermayr, Frau Martha Pfleger und Frau Dr. Helga Wagner.

Dem Physica-Verlag aus dem Hause Springer möchte ich danken für die Erstellung dieses Lehrbuches und die gute und problemlose Zusammenarbeit, insbesondere gilt mein Dank Frau Dipl.-Math. Lilith Braun, Frau Gabriele Keidel M.A. und Herrn Dr. Werner Müller, die durch ihre Unterstützung mein Debüt als Lehrbuchautorin erst ermöglicht haben.

Über Anregungen meiner Leserinnen und Leser würde ich mich sehr freuen (christine.duller@jku.at) und ich wünsche allen viel Spaß mit der Statistik.

Linz, August 2005 *Christine Duller*

Inhaltsverzeichnis

Teil I Einführung

1 Was ist Statistik? . . . 3
1.1 Der Begriff Statistik . . . 3
1.2 Wozu Statistik? . . . 4
1.3 Grundbegriffe . . . 6
1.4 Teilbereiche der Statistik . . . 9
Übungsaufgaben . . . 10

2 Ablauf einer statistischen Analyse . . . 11
2.1 Planung . . . 11
2.2 Merkmale und Merkmalstypen . . . 12
2.2.1 Skalenniveaus von Merkmalen . . . 12
2.2.2 Stetige und diskrete Merkmale . . . 14
2.3 Methode der Datengewinnung . . . 15
2.4 Datenerfassung und -aufbereitung . . . 18
2.5 Abschlussbericht . . . 20
2.6 Problemfelder in der Praxis . . . 22
2.6.1 Datenschutz, Anonymität . . . 22
2.6.2 Unzureichendes Studiendesign . . . 23
2.6.3 Sekundärstatistiken . . . 23
2.6.4 Fehlende Daten . . . 24
Übungsaufgaben . . . 24

3 Anmerkungen zum Umgang mit dem Computer 27
3.1 Grundlagen 27
3.2 Nützliche Tasten und Tastenkombinationen 28
3.3 Drag and Drop 29
3.4 Konventionen zur Beschreibung 29

4 Das Tabellenkalkulationsprogramm EXCEL 31
4.1 Grundelemente in EXCEL 31
4.2 Formatierung in EXCEL 33
4.3 Dateneingabe 36
4.4 Statistische Analysen 37

5 Das Statistikpaket SPSS 39
5.1 Erste Schritte in SPSS 39
5.2 Der Dateneditor 41
5.3 Datenquellen 47
5.4 Der Viewer 49
5.5 Datenaufbereitung 50
5.5.1 Fehlende Werte 50
5.5.2 Umkodieren von Variablen 50
5.5.3 Transformieren von Variablen 53
5.5.4 Fälle gewichten 54
5.6 Tipps im Umgang mit SPSS 57

Teil II Deskriptive Statistik

6 Eindimensionale Häufigkeitsverteilungen 61
6.1 Diskrete Merkmale 61
6.1.1 Häufigkeitsverteilung in EXCEL 63
6.1.2 Häufigkeitsverteilungen in SPSS 66
6.2 Stetige Merkmale 68
6.2.1 Stetige Häufigkeitsverteilung in EXCEL 70
6.2.2 Stetige Häufigkeitsverteilung in SPSS 70
6.3 Grafische Darstellung von Verteilungen 71

6.3.1 Kreis- oder Tortendiagramm ... 71
6.3.2 Balken-, Säulen- oder Stabdiagramm ... 72
6.3.3 Histogramm ... 73
6.3.4 Qualitätskriterien für Grafiken ... 76
6.3.5 Auswahl der passenden Darstellungsform ... 80
6.3.6 Grafiken in EXCEL ... 81
6.3.7 Erstellen von Histogrammen in EXCEL ... 83
6.3.8 Grafiken in SPSS ... 84
6.4 Die empirische Verteilungsfunktion ... 85
6.4.1 Abbild der empirischen Verteilungsfunktion ... 86
6.4.2 Rechnen mit der empirischen Verteilungsfunktion ... 88
6.4.3 Die empirische Verteilungsfunktion in EXCEL ... 91
6.4.4 Die empirische Verteilungsfunktion in SPSS ... 91
Übungsaufgaben ... 92

7 Maßzahlen für eindimensionale Verteilungen ... 95
7.1 Lagemaße ... 95
7.1.1 Arithmetisches Mittel ... 95
7.1.2 Median ... 98
7.1.3 Modus ... 101
7.1.4 Geometrisches Mittel ... 102
7.1.5 Quantile ... 104
7.1.6 Lagekennzahlen in EXCEL ... 105
7.1.7 Lagekennzahlen in SPSS ... 107
7.2 Streuungsmaße ... 108
7.3 Eigenschaften von Lage- und Streuungsmaßen ... 111
7.3.1 Maßeinheiten ... 111
7.3.2 Minimaleigenschaften ... 112
7.3.3 Robustheit ... 113
7.4 Auswahl geeigneter Lagemaßzahlen ... 113
7.5 Maßzahlen der Schiefe und Wölbung ... 113
7.6 Streuung, Schiefe und Wölbung in EXCEL ... 117
7.7 Streuung, Schiefe und Wölbung in SPSS ... 118

Übungsaufgaben 118

8 Multivariate deskriptive Statistik 121
8.1 Zweidimensionale Häufigkeitsverteilungen 121
8.2 Randverteilungen 123
8.3 Bedingte Verteilung 123
8.4 Maße für den Zusammenhang zweier Merkmale 125
8.4.1 Zusammenhang zweier nominaler Merkmale 125
8.4.2 Zusammenhang zweier ordinaler Merkmale 128
8.4.3 Zusammenhang zweier metrischer Merkmale 130
8.5 Grafische Darstellung zweidimensionaler metrischer Merkmale . 134
8.6 Korrelation und Kausalität 135
8.7 Zweidimensionale Merkmale in EXCEL 136
8.8 Zweidimensionale Merkmale in SPSS 137
8.9 Tipps und Tricks 142
Übungsaufgaben 142

9 Die Regressionsanalyse 145
9.1 Die lineare Einfachregression 145
9.2 Regressionsanalyse in EXCEL 151
9.3 Regressionsanalyse in SPSS 151
Übungsaufgaben 155

Teil III Wahrscheinlichkeitsrechnung

10 Wahrscheinlichkeitsrechnung 159
10.1 Exkurs: Mengenlehre 159
10.2 Grundbegriffe der Wahrscheinlichkeitsrechnung 160
10.3 Denkmodelle für den Wahrscheinlichkeitsbegriff 162
10.3.1 Wahrscheinlichkeit als Anteil 162
10.3.2 Wahrscheinlichkeit als relative Häufigkeit 163
10.4 Rechnen mit Wahrscheinlichkeiten 163
10.4.1 Axiome von Kolmogorov 164
10.4.2 Bedingte Wahrscheinlichkeiten 165

10.4.3 Stochastisch unabhängige Ereignisse 166
10.4.4 Das Theorem von Bayes 167
Übungsaufgaben 169

11 Diskrete Wahrscheinlichkeitsverteilungen 173
11.1 Dichte und Verteilungsfunktion 173
11.2 Lage- und Streuungsparameter 175
11.3 Spezielle diskrete Verteilungen 177
11.3.1 Alternativverteilung 177
11.3.2 Diskrete Gleichverteilung 178
11.3.3 Binomialverteilung 179
11.3.4 Hypergeometrische Verteilung 181
11.3.5 Poissonverteilung 183
11.4 Rechnen mit diskreten Verteilungen 185
Übungsaufgaben 187

12 Stetige Wahrscheinlichkeitsverteilungen 191
12.1 Dichte und Verteilungsfunktion 191
12.2 Unabhängigkeit zweier stetiger Zufallsvariablen 195
12.3 Lage- und Streuungsparameter 196
12.4 Die stetige Gleichverteilung 198
12.5 Die Normalverteilung 199
12.6 Approximationen durch die Normalverteilung 206
12.6.1 Gesetz der großen Zahlen und Grenzwertsätze 206
12.6.2 Approximationen durch die Normalverteilung 208
Übungsaufgaben 210

Teil IV Schließende Statistik

13 Die Gedankenwelt der schließenden Statistik 215
13.1 Stichprobenverteilung 215
13.2 Parameterschätzung 217
13.3 Schätzen von Anteilen 219
13.4 Schätzen von Mittelwerten 221

13.5 Konfidenzintervalle in EXCEL 224
13.6 Konfidenzintervalle in SPSS 224
Übungsaufgaben 225

14 Statistisches Testen 227
14.1 Grundbegriffe der Testtheorie 227
14.2 Testen von Hypothesen über Anteile 231
14.2.1 Testen von zweiseitigen Hypothesen 231
14.2.2 Testen von einseitigen Hypothesen 234
14.3 Testen von Hypothesen über einen Mittelwert 236
14.3.1 Testen von zweiseitigen Hypothesen 237
14.3.2 Testen von einseitigen Hypothesen 238
14.4 Testen von Hypothesen in EXCEL und SPSS 241
14.5 Der Chi-Quadrat-Test auf Unabhängigkeit 242
Übungsaufgaben 245

Tabellen 247

Lösungen zu den Übungsaufgaben 253

Symbolverzeichnis 267

Literaturverzeichnis 271

Sachverzeichnis 273

Teil I

Einführung

1

Was ist Statistik?

1.1 Der Begriff Statistik

Die Statistik ist heute ein selbstständiger Teilbereich der Mathematik und hat sich zur Jahrhundertwende vom neunzehnten zum zwanzigsten Jahrhundert als eigene Disziplin herauskristallisiert.

Der Ursprung der Statistik liegt im Staatswesen, insbesondere in der Erhebung von Daten über die Bevölkerung und den Handel. Der Name wurde unter anderem von Gottfried Achenwall (1719-1772) durch seine Vorlesung über Staatskunde unter dem Titel „Notitia politica vulgo statistica" geprägt. Bis zu diesem Zeitpunkt wurde nur statista - Staatskundiger oder eine adjektivische Form davon verwendet.

Achenwall definiert Statistik folgendermaßen:

„Wenn ich einen einzelnen Staat ansehe, so erblicke ich eine unendliche Menge von Sachen, so darinnen wirklich angetroffen werden. Unter diesen sind einige, welche seine Wohlfahrt in einem merklichen Grade angehen, entweder dass sie solche hindern oder befördern. Man kann selbige Staatsmerkwürdigkeiten nennen. Der Inbegriff der wirklichen Staatsmerkwürdigkeiten eines Reiches oder einer Republik macht ihre Staatsverfassung im weiteren Sinne aus, und die Lehre von der Staatsverfassung eines oder mehrerer einzelner Staaten ist die Statistik."

Erst gegen Ende des neunzehnten Jahrhunderts wurde die Statistik als allgemeine Theorie zur Analyse zufallsabhängiger Probleme anerkannt.

Heute kann der Begriff Statistik folgendermaßen definiert werden:

> Unter Statistik versteht man die Erfassung, Zusammenfassung, Analyse und Darstellung von Massendaten, sowie die Methoden zum vernünftigen Entscheiden bei Unsicherheit.

Aus dieser Definition leitet sich auch der Inhalt dieses Buches ab. Der erste Teil ist der beschreibenden (deskriptiven) Statistik mit der Erfassung, Zusammenfassung, Analyse und Darstellung von Daten gewidmet. Im Bereich der induktiven Statistik werden die Methoden der schließenden Statistik bereitgestellt, die sich unter anderem mit dem vernünftigen Entscheiden bei Unsicherheit auseinandersetzen. Der Hauptaugenmerk wird auf ein Verständnis der Grundlagen und auf eine korrekte Anwendung der Methoden gelegt, dabei wird insbesondere auf mögliche Fallen und Fehlinterpretationen eingegangen.

Die allgemeine Meinung über Statistiken und deren Zuverlässigkeit ist nicht sehr hoch, wie einige bekannte Zitate zeigen:

- „Ich traue einer Statistik nie - es sei denn ich habe sie selbst gefälscht." Winston Churchill (1874 - 1965), Joseph Goebbels (1897 - 1945).
- „Es gibt drei Arten von Lügen: Lügen, verdammte Lügen und Statistiken." Benjamin Disraeli, Earl of Beaconsfield (1804 - 1881), britischer konservativer Staatsmann und Schriftsteller.
- „Ich stehe Statistiken etwas skeptisch gegenüber, denn laut Statistik haben ein Millionär und ein Habenichts je eine halbe Million." Franklin Delano Roosevelt (1882 - 1945), 32. Präsident der Vereinigten Staaten von Amerika.

Statistikerinnen und Statistiker haben verständlicherweise eine andere Sicht auf die Rolle der Statistik, die sich etwa mit folgendem Zitat beschreiben lässt:

> „Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."
> H.G. Wells

1.2 Wozu Statistik?

Die pragmatische Antwort für einen Teil der LeserInnen (genauer gesagt für meine StudentInnen) lautet, dass sie gar keine andere Wahl haben, weil Statistik ein Pflichtfach im Rahmen der meisten Studienrichtungen ist.

Dieses Buch soll aber auch zeigen, dass Kenntnisse aus diesem Bereich im Alltag nützlich sind, dass die Grundlagen der Statistik keineswegs so kompliziert sind, wie allgemein befürchtet wird, und dass - last but not least - Statistik auch Spaß machen kann.

Ohne es bewusst wahrzunehmen, kommen wir sehr oft mit Statistik in Berührung. Einige konkrete Beispiele sollen die vielfältigen Anwendungsgebiete illustrieren.

Die Volkszählung
Wie in den meisten Staaten der Welt wird auch in Österreich alle 10 Jahre eine Volkszählung (= Zensus) durchgeführt, die letzte fand am 15. Mai 2001 statt. Dabei werden mittels Fragebogen wichtige demographische und wirtschaftliche Daten aller BewohnerInnen erhoben. Die Ergebnisse einer Volkszählung zeigen die Struktur der Bevölkerung und dienen als Grundlage für zahlreiche Maßnahmen der öffentlichen Verwaltung. Beispielsweise basiert auf diesen Zahlen die Aufteilung von Steuermittel auf Bundesländer und Gemeinden. Man kann aus den Daten den Bedarf an Verkehrseinrichtungen für PendlerInnen ablesen oder Bevölkerungszahl und -struktur prognostizieren.

Der Mikrozensus
Bei der Volkszählung werden die Daten aller BewohnerInnen erhoben, dem entsprechend aufwändig ist die Erhebung. Um auch in den Jahren zwischen den Volkszählungen über aktuelle Daten zu verfügen, wird viermal jährlich ein Mikrozensus erhoben, also eine „kleine Volkszählung" durchgeführt. Hier werden die Daten nicht von allen BewohnerInnen erhoben, sondern nur von einem Teil. In Österreich umfasst der Mikrozensus etwa 35.000 Haushalte.

Meinungsumfragen
Die Meisten sind bereits mit Meinungsumfragen in Berührung gekommen, sei es, dass sie selbst den einen oder anderen Fragebogen ausgefüllt haben, oder sei es, dass man in den Medien über Ergebnisse solcher Meinungsumfragen informiert wird. Besonders häufig werden Ergebnisse von Meinungsumfragen vor Wahlen präsentiert, in Österreich hat sich mittlerweile der Begriff „Sonntagsfrage" etabliert, hinter dem die Frage steht „Wenn kommenden Sonntag Wahlen wären, wen würden Sie wählen?".

Informationen aller Art
Selbst auf den ersten Blick unverdächtig erscheinende Themenbereiche wie Sport- oder Wetterberichte sind durchsetzt mit Statistik. Statistiken beim Tennis informieren beispielsweise über die Anzahl der Doppelfehler, die Anzahl der selbstverschuldeten Fehler oder auch über das Verhältnis der Siege bei gleicher Konstellation von SpielerInnen.
Der Wetterbericht informiert uns über das morgen wahrscheinliche Wetter und prognostiziert den weiteren Verlauf für die nächsten fünf Tage.

1.3 Grundbegriffe

Jede statistische Erhebung beginnt mit der Bestimmung der Grundgesamtheit. Das ist die Menge aller Objekte (Personen, Betriebe oder ähnliches), über die man Informationen gewinnen will. Dabei ist eine exakte räumliche, zeitliche und sachliche Abgrenzung notwendig. Die Notwendigkeit einer exakten Abgrenzung ist plausibel, in der Praxis allerdings nicht immer leicht durchzuführen.

Beispiel 1.1. MitarbeiterInnenzufriedenheit
In einem Unternehmen soll die Zufriedenheit von MitarbeiterInnen untersucht werden. Die Grundgesamtheit umfasst somit alle MitarbeiterInnen des Unternehmens. Der Begriff „MitarbeiterIn im Unternehmen" muss nun so weit konkretisiert werden, dass für jede einzelne Person unzweifelhaft festgestellt werden kann, ob diese zur Grundgesamtheit gehört oder nicht. Eine räumliche Abgrenzung legt beispielsweise fest, ob alle Standorte des Unternehmens untersucht werden sollen oder nur ausgewählte. Die zeitliche Abgrenzung könnte dazu führen, dass alle Personen befragt werden, die an einem bestimmten Stichtag im Unternehmen beschäftigt waren. Die sachliche Abgrenzung beschäftigt sich mit der Frage, wer MitarbeiterIn ist, ob beispielsweise auch Leasingpersonal oder Aushilfskräfte befragt werden sollen oder nicht. Diese Abgrenzungen sind in Hinblick auf das verfolgte Untersuchungsziel vor der Datenerhebung festzulegen.

Die einzelnen in der Grundgesamtheit zusammengefassten Elemente nennt man Erhebungseinheiten oder Merkmalsträger. Die Anzahl der Erhebungseinheiten ist auch der Umfang der Grundgesamtheit und wird üblicherweise mit N bezeichnet.

Beispiel 1.2. MitarbeiterInnenzufriedenheit
(Fortsetzung von Beispiel 1.1) Erhebungseinheit ist in diesem Beispiel eine einzelne Mitarbeiterin, welche die Abgrenzungskriterien in räumlicher, zeitlicher und sachlicher Hinsicht erfüllt.

Nachdem feststeht, von welchen Personen die Daten erhoben werden, ist nun festzulegen, was genau erhoben werden soll. Die interessierenden Eigenschaften der Erhebungseinheiten nennt man Merkmale, wobei jedes Merkmal verschiedene Ausprägungen besitzt. Alle prinzipiell möglichen Ausprägungen zusammen nennt man den Wertebereich des Merkmals.

Beispiel 1.3. MitarbeiterInnenzufriedenheit
(Fortsetzung von Beispiel 1.1) Neben verschiedenen Fragen über die Zufriedenheit wird man auch das Geschlecht erheben, da es ja sein könnte, dass

es geschlechtsspezifische Unterschiede in der Zufriedenheit gibt. Somit ist von jeder Person das Merkmal Geschlecht mit den möglichen Ausprägungen weiblich und männlich zu erheben. Anders formuliert besteht der Wertebereich des Merkmals Geschlecht aus den Ausprägungen weiblich und männlich.

Oft wird nicht die Grundgesamtheit untersucht, sondern nur eine Teilmenge davon, die sogenannte Stichprobe. Dafür können folgende Gründe sprechen:

- Kostenersparnis
- Zeitersparnis
- Die Teilgesamtheit kann gründlicher untersucht werden, als es bei der Grundgesamtheit der Fall wäre.
- Eine Untersuchung der Grundgesamtheit ist nicht möglich, weil beispielsweise das Objekt bei der Untersuchung zerstört wird (z.B. bei Materialtests).

Für die Methoden der induktiven Statistik (vgl. Kapitel 1.4) darf diese Teilmenge nicht willkürlich aus der Grundgesamtheit ausgewählt werden, sondern die Stichprobe muss gewissen Ansprüchen genügen.

Eine Stichprobe muss repräsentativ sein, das bedeutet, dass die Stichprobe ein möglichst genaues Abbild der Grundgesamtheit darstellt. Das Gegenteil einer repräsentativen Stichprobe ist eine verzerrte Stichprobe, die nur dann für Aussagen über die Grundgesamtheit verwendet werden kann, wenn sich die Verzerrung rechnerisch beheben lässt, was aber sehr selten der Fall ist. Ob eine Stichprobe repräsentativ ist oder nicht lässt sich leider nicht berechnen, da hilft nur ein kritischer Blick auf das Studiendesign.

Das Auswahlverfahren legt fest, welche konkreten Elemente der Grundgesamtheit in die Stichprobe aufgenommen werden. Für die meisten Verfahren der angewandten Statistik wird von einer einfachen Zufallsauswahl ausgegangen, dies bedeutet, dass jedes Element der Grundgesamtheit prinzipiell die gleiche Chance haben muss in die Stichprobe zu gelangen. Eine einfache Zufallsstichprobe ist meistens auch repräsentativ. Es gibt auch Verfahren der bewussten Auswahl, wie z.B. das sogenannte Quotenverfahren, welche in diesem Buch aber nicht näher beleuchtet werden.

Beispiel 1.4. MitarbeiterInnenzufriedenheit
(Fortsetzung von Beispiel 1.1) Man möchte nicht die Grundgesamtheit untersuchen, sondern lediglich eine Stichprobe. Es ist offensichtlich, dass man eine

verzerrte (und damit unbrauchbare) Stichprobe erhält, wenn man ausschließlich Frauen oder nur die leitenden Angestellten befragt. Weniger offensichtlich und daher häufig begangener Fehler ist die Auswahl von MitarbeiterInnen in der Kantine. Man befragt einfach die Personen, die man mittags in der Kantine antrifft. Dies ist zwar naheliegend und bequem, führt aber ebenfalls zu einer verzerrten Stichprobe, denn vielleicht sind die anderen Personen aufgrund ihrer Unzufriedenheit nicht in der Kantine. Auch Teilzeitbeschäftigte werden die Kantine nur selten aufsuchen. Folgende Möglichkeit bietet sich zur Erstellung einer Zufallsstichprobe an: Alle Personen aus der Grundgesamtheit werden willkürlich durchnummeriert, im nächsten Schritt werden mittels Computer zufällig Zahlen ausgewählt. Befragt werden jene Personen, die diesen Zahlen zugeordnet sind.

Die folgende Übersicht enthält die wichtigsten Begriffe, die in diesem Kapitel eingeführt wurden.

Grundgesamtheit
Die Menge aller Objekte, über die man Informationen gewinnen will. Eine exakte räumliche, zeitliche und sachliche Abgrenzung ist notwendig.

Erhebungseinheit
Ein einzelnes Element der Grundgesamtheit. Die Anzahl der Erhebungseinheiten bildet den Umfang der Grundgesamtheit (= N).

Merkmal
Die interessierende Eigenschaft der Erhebungseinheiten. Jedes Merkmal besitzt verschiedene **Ausprägungen**.

Wertebereich
Alle Ausprägungen eines Merkmals bilden den Wertebereich.

Stichprobe
Eine Teilmenge der Grundgesamtheit.

Repräsentative Stichprobe
Die Stichprobe zeichnet ein möglichst genaues Abbild der Grundgesamtheit.

Einfache Zufallsstichprobe
Jedes Element der Grundgesamtheit hat die gleiche Chance in die Stichprobe zu gelangen.

1.4 Teilbereiche der Statistik

Man unterscheidet in der Statistik drei inhaltliche Teilbereiche:

Deskriptive Statistik
Die deskriptive (= beschreibende) Statistik ist Ausgangspunkt jeder Datenanalyse, hier erfolgt die Beschreibung und Darstellung der Daten. Dazu gehört die Aufbereitung der Daten in Form von Tabellen und Grafiken und die Berechnung einfacher statistischer Kennzahlen. Mit den Methoden der deskriptiven Statistik verschafft man sich einen ersten Überblick über die Datensituation, manchmal gibt man sich damit auch schon zufrieden.

Induktive Statistik
Ist der erhobene Datensatz eine repräsentative Stichprobe einer interessierenden Grundgesamtheit, so erlauben die Methoden der induktiven (= schließenden) Statistik Rückschlüsse von der Stichprobe auf die Grundgesamtheit. Obwohl man also nur einen Auszug aus der Grundgesamtheit kennt, ist es trotzdem möglich, Aussagen über diese unbekannte Grundgesamtheit zu treffen. Diese Aussagen sind zwar mit Unsicherheit behaftet, die sich jedoch abschätzen lässt.

Explorative Datenanalyse
Mittlerweile hat sich die explorative Datenanalyse als eigener Bereich im Übergang zwischen deskriptiver und induktiver Statistik etabliert. Die explorative Datenanalyse dient dem Suchen nach Strukturen, nach möglichen Fragestellungen und Hypothesen (=Behauptungen). Diese Hypothesen werden anschließend mit den Methoden der induktiven Statistik überprüft.

Eine andere Möglichkeit die Statistik in Teilgebiete aufzusplittern erhält man, wenn man die Anzahl der gleichzeitig betrachteten Merkmale als Unterscheidungskriterium heranzieht.

Univariate, bivariate und multivariate Statistik
Üblicherweise werden an den Erhebungseinheiten mehrere Merkmale erhoben. Greift man zur Analyse nur ein einziges Merkmal heraus, so spricht man von univariater Statistik. Dem entsprechend betrachtet man bei bivariaten Verfahren zwei Merkmale, die multivariate Statistik analysiert mehrere Variablen gleichzeitig.

Übungsaufgaben

1.1. Notenverteilung

Man möchte die Notenverteilung einer Statistik-Klausur untersuchen.

a) Was ist die Grundgesamtheit?
b) Was ist eine Erhebungseinheit?
c) Welche Merkmale könnten von Interesse sein?
d) Welche Wertebereiche haben diese Merkmale? Nennen Sie jeweils einige Ausprägungen.

1.2. Medizinische Studie

Man möchte eine medizinische Studie zum Vergleich zweier Medikamente gegen Bluthochdruck durchführen.

a) Was ist die Grundgesamtheit?
b) Was ist eine Erhebungseinheit?
c) Welche Merkmale könnten von Interesse sein?
d) Welche Wertebereiche haben diese Merkmale? Nennen Sie jeweils einige Ausprägungen.

2

Ablauf einer statistischen Analyse

Dieses Kapitel skizziert die Schritte, die vor bzw. nach der eigentlichen statistischen Auswertung notwendig sind. Newcomer in der Statistik kennen zwar die Methoden zur Datenauswertung, in der Praxis liegen aber die Probleme oft in den Phasen vor bzw. nach der statistischen Auswertung.

2.1 Planung

Die Planung ist ein sehr wichtiger Teil einer statistischen Erhebung. Erst eine gute und durchdachte Planung ermöglicht den Erfolg.
In einem ersten Schritt sollte das meist mehr oder weniger diffus vorliegende Forschungsziel konkretisiert werden. Dazu gehört die möglichst exakte Zielformulierung und die Abgrenzung der Grundgesamtheit (vgl. Kapitel 1.3), weiters sind die zu erhebenden Merkmale festzulegen. Der nächste Schritt liegt in der Wahl der Methode zur Datengewinnung und die Datengewinnung selbst. Nach einer Probeerhebung (Pretest) werden notwendige Verbesserungen beim Erhebungsinstrument und bei den Merkmalen durchgeführt, dann kann mit der eigentlichen Datenerhebung begonnen werden. Danach erfolgt die Auswertung der Daten mit den Methoden der Statistik. Den Abschluss bildet ein schriftlicher Bericht. Die einzelnen Schritte sollten im Vorfeld terminiert werden und alle Arbeiten sind ausreichend zu dokumentieren.

Praxistipp:
Bei der Erstellung des Zeitplanes sollte die veranschlagte Zeit immer großzügig bemessen sein. Dokumentieren ist eine lästige Arbeit, aber in der Praxis unumgänglich. Eine unzureichende Planung und Vorbereitung kann später nicht mehr korrigiert werden.

Ablauf einer statistischen Analyse

1. Vorbereitung
   - Zielformulierung
   - Zeitplan erstellen
   - Festlegen von Grundgesamtheit und Merkmalen
   - Methode der Datengewinnung festlegen
   - Auswertungsmethode festlegen
   - Probeerhebung und Durchführung von Korrekturen
2. Datengewinnung
3. Datenerfassung und -aufbereitung
4. Statistische Auswertung
5. Bericht

Eine ausreichende Dokumentation ist unumgänglich.

2.2 Merkmale und Merkmalstypen

Bereits in der Planungsphase muss man sich über Eigenschaften von Merkmalen im Klaren sein. Merkmale lassen sich im Wesentlichen nach zwei Prinzipien einteilen: Einerseits gibt es verschiedene Skalenniveaus, andererseits unterscheidet man diskrete und stetige Merkmale.

2.2.1 Skalenniveaus von Merkmalen

Hinsichtlich des Skalenniveaus gibt es metrische, ordinale und nominale Merkmale.

Ein Merkmal heißt **metrisch** (= quantitativ, kardinalskaliert), wenn seine Ausprägungen Vielfache einer Einheit sind (z.B. Länge, Einkommen). Die Ausprägungen sind voneinander verschieden, haben eine eindeutige Anordnung und einen eindeutig definierten Abstand. Bei metrischen Merkmalen kann man zwischen intervallskalierten und verhältnisskalierten Merkmalen unterscheiden.
Bei **verhältnisskalierten** Merkmalen gibt es einen natürlichen Nullpunkt (z.B. Preis) und das Verhältnis zweier Ausprägungen lässt sich sinnvoll interpretieren (Produkt A ist doppelt so teuer wie Produkt B).
Intervallskalierte Merkmale haben keinen natürlichen Nullpunkt, daher können auch Verhältnisse nicht sinnvoll interpretiert werden (z.B. Temperatur in Grad Celsius).

Ein Merkmal heißt **ordinal**, wenn die Ausprägungen nur in einer Ordnungsbeziehung wie größer, kleiner, besser oder schlechter zueinander stehen (z.B. Schulnoten, Güteklassen). Die Ausprägungen sind voneinander verschieden und haben eine eindeutige Anordnung, der Abstand zweier Merkmalsausprägungen ist hingegen nicht klar definiert und daher auch nicht interpretierbar. Es ist unumstritten, dass im österreichischen Schulnotensystem die Note 1 besser ist als die Note 2, der Abstand zwischen diesen beiden Noten lässt sich aber weder interpretieren noch mit dem Abstand zwischen den Noten 4 und 5 gleichsetzen.

Ein Merkmal heißt **nominal**, wenn seine Ausprägungen nicht in eindeutiger Weise geordnet werden können, sondern nur durch ihre Bezeichnungen unterschieden sind (z.B. Geschlecht, Familienstand, Beruf). Die Ausprägungen sind von einander verschieden, es gibt keine eindeutige Anordnung, der Abstand zweier Merkmalsausprägungen ist nicht definiert. Diese Merkmale werden auch als qualitative oder kategoriale Merkmale bezeichnet.

Diese Skalenniveaus sind hierarchisch aufgebaut, an unterster Stelle stehen die nominalen Merkmale, dann folgen die ordinalen Merkmale und an der Spitze stehen die metrischen Merkmale.

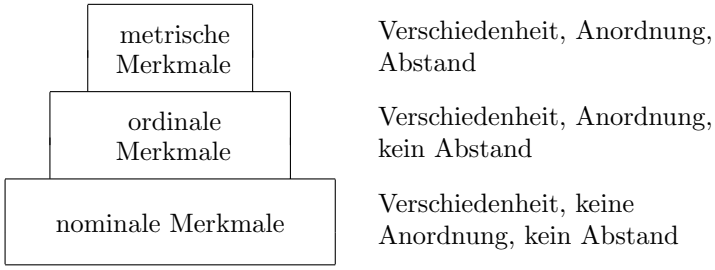


Abb. 2.1. Hierarchie der Skalenniveaus

Daraus ergeben sich zwei Konsequenzen:

- Jedes Merkmal aus einer höheren Hierarchiestufe kann durch Zusammenfassen und Umbenennen von Merkmalsausprägungen in ein Merkmal der niedrigeren Stufe umgewandelt werden, allerdings entsteht dadurch ein Informationsverlust. Das Merkmal Körpergröße gemessen in cm kann auch angegeben werden mit den Ausprägungen klein - mittel - groß, die Berechnung von Größenunterschieden ist aber nach dieser Transformation nicht mehr möglich.

- Alle Verfahren, die für ein Merkmal aus einer bestimmten Stufe zulässig sind, sind auch zulässig für Merkmale aus darüber liegenden Stufen.

Das Skalenniveau eines Merkmals bestimmt welche Verfahren und Berechnungen im Umgang mit dem Merkmal gestattet sind. Tabelle 2.1 gibt einen ersten Überblick über sinnvolle Berechnungen auf den verschiedenen Skalenniveaus.

Tabelle 2.1. Zulässige Verfahren für Skalenniveaus

Skalenniveau	Auszählen	Ordnen	Summen Differenzen	Quotienten
verhältnisskaliert	ja	ja	ja	ja
intervallskaliert	ja	ja	ja	nein
ordinal	ja	ja	nein	nein
nominal	ja	nein	nein	nein

2.2.2 Stetige und diskrete Merkmale

Ein Merkmal heißt **stetig**, wenn seine Ausprägungen beliebige Zahlenwerte aus einem Intervall annehmen können (z.B. Länge, Gewicht).

Ein Merkmal heißt **diskret**, wenn seine Ausprägungen bei geeigneter Skalierung (bzw. Kodierung) nur ganzzahlige Werte annehmen können (z.B. Fehlerzahlen, Schulnoten, Geschlecht). Diskrete Merkmale haben abzählbar viele Ausprägungen.

Dichotome Merkmale sind eine Sonderform von diskreten Merkmalen und besitzen nur zwei Ausprägungen.

Von **quasistetigen** Merkmalen spricht man bei Merkmalen, die aufgrund der Definition diskret sind, gleichzeitig aber über eine so feine Abstufung verfügen, dass man sie als stetige Merkmale behandeln kann. Insbesondere zählen hierzu alle monetären Merkmale (Preis, Kredithöhe, Miete, ...).

Die Bezeichnung **diskretisierte** Merkmale wird verwendet, wenn stetige Merkmale nur in diskreter Form erfasst werden, beispielsweise die Frage nach dem Alter in ganzen Jahren. Die Zusammenfassung von Ausprägungen eines Merkmals in Gruppen wird als **Gruppieren** bezeichnet.

Merkmale

metrisch:	Abstand, Anordnung, Verschiedenheit der Ausprägungen intervall- oder verhältnisskaliert intervallskaliert: kein absoluter Nullpunkt verhältnisskaliert: absoluter Nullpunkt
ordinal:	kein Abstand, aber Anordnung und Verschiedenheit
nominal:	Ausprägungen sind Namen kein Abstand, keine Anordnung, nur Verschiedenheit
stetig:	Ausprägungen sind beliebige Werte aus einem Intervall
diskret:	abzählbar viele Ausprägungen
dichotom:	zwei Ausprägungen

2.3 Methode der Datengewinnung

Die Datengewinnung ist ein wesentlicher Teil der statistischen Analyse, der oft in den Überlegungen etwas zu kurz kommt. Dies ist aber ein fataler Fehler, denn wenn die Daten schlecht oder unzureichend erhoben werden, so kann diese Unzulänglichkeit auch mit den Methoden der Statistik nicht behoben werden. Die Daten sind die Basis der Analyse und dessen sollte man sich immer bewusst sein.

Werden die Daten speziell für einen bestimmten Zweck erhoben, spricht man von primärstatistischen Erhebungen. Sekundärstatistische Erhebungen greifen hingegen auf bereits vorhandenes Zahlenmaterial zurück. Im technisch-naturwissenschaftlichen Bereich werden Daten meist im Rahmen von Experimenten gewonnen, in den Sozialwissenschaften stammen die Daten im Normalfall aus einer Befragung oder einer Beobachtung. Die Befragung wird mit einem Fragbogen durchgeführt, daher werden die wesentlichen Aspekte der Fragebogengestaltung kurz dargestellt.

Generell ist bei der Erstellung eines Fragebogens darauf zu achten, dass die Komplexität und die Wortwahl dem Zielpublikum angepasst sind. Der Fragebogen beginnt mit einfachen und harmlosen Fragen, z.B. nach dem Geschlecht oder dem Familienstand. Nach dieser Aufwärmphase kommen die Fragen zum Thema, die einen sinnvollen Aufbau haben sollten. Besonders heikle Fragen (z.B. nach dem Einkommen) sollten möglichst am Ende des Fragebogens platziert werden, um einen vorzeitigen Abbruch der Befragung zu vermeiden. Ein Fragebogen sollte so lang wie notwendig und so kurz wie möglich sein, ein ansprechendes Layout erhöht die Motivation der Befragten. Jeder Fragebogen endet höflicherweise mit einem Dank für die Mitarbeit.

Bei **geschlossenen Fragen** werden die möglichen Antwortalternativen im Fragebogen angeführt. Dabei ist darauf zu achten, dass sich einerseits die Alternativen nicht überschneiden (ausschließend), andererseits aber alle möglichen Alternativen angeführt werden (erschöpfend). Im Zweifel kann die Vollständigkeit immer mit der Kategorie *Sonstiges* erreicht werden. Ein freies Textfeld neben der Kategorie *Sonstiges* ermöglicht den Befragten eine nähere Spezifizierung.

Geschlossene Fragen werden manchmal dazu verwendet, besonders heikle Fragen zu entschärfen. Als Beispiel sei die Frage nach dem Einkommen erwähnt, die eher beantwortet wird, wenn die Befragten nur eine Einkommensklasse ankreuzen müssen anstatt das exakte Einkommen anzugeben. Bei der Erstellung der Antwortalternativen ist Fingerspitzengefühl notwendig. Die Einkommensklassen sollen einerseits möglichst klein sein, weil man an detaillierten Informationen interessiert ist, andererseits führen zu kleine Klassen wieder zu vermehrter Verweigerung. Auf jeden Fall sollten die Klassen möglichst so angelegt werden, dass niemand die niedrigste oder die höchste Kategorie auswählen muss, weil dies für die Befragten unangenehm wäre.

Eine besondere Stellung nehmen die **Bewertungsfragen** ein, bei denen die Befragten beispielsweise die Qualität von Produkten bewerten müssen. Hier ist es meist sinnvoll ein Bewertungsschema zu verwenden, das dem Schulnotensystem angepasst ist (1 = Sehr Gut bis 5 = Nicht Genügend). Das Bewertungsschema muss jedenfalls so aufgebaut sein, dass die Befragten in ihrer Meinungsäußerung nicht beeinflusst oder eingeschränkt werden.

Beispiel 2.1. Bewertungsfrage Mensa
Die Mensa an der Universität Linz stellte im Rahmen einer Befragung im Jahr 2003 folgende Frage: „Sagt Ihnen unser Suppenangebot zu?" Als mögliche Antwort waren lediglich die Alternativen sehr gut, gut und befriedigend vorgesehen. Die Beeinflussung der Befragten durch diese Art der Antwortvorgabe ist offensichtlich ....

Eine fünfteilige Bewertungsskala ist in den meisten Fällen ein gutes Mittelmaß, zudem ist die Assoziation mit den Schulnoten für viele Befragte eine Erleichterung. Die Anzahl der Kategorien rechts und links von der Mittelkategorie muss jedenfalls gleich sein und auch die verbale Formulierung sollte möglichst symmetrisch sein. Bei einer ungeraden Anzahl von Kategorien kann die mittlere Kategorie von manchen Befragten als Fluchtkategorie verwendet werden, manchmal spiegelt aber die Mittelkategorie auch eine echte Einstellung wieder.

Bei **offenen Fragen** werden keine Antwortalternativen vorgegeben. Diese Fragestruktur sollte nur sehr eingeschränkt verwendet werden, da einerseits die Auswertung von offenen Fragen sehr aufwändig ist und andererseits ge-

schlossene Fragen den Befragten weniger Mühe bereiten. Offene Fragestellungen sind zu bevorzugen, wenn man über den Forschungsgegenstand nicht gut genug Bescheid weiß, um alle Antwortalternativen einer geschlossenen Frage angeben zu können. In manchen Fällen wird bewusst eine offene Frage gestellt, weil deren Beantwortung andere Fähigkeiten von Befragten verlangt als die einer geschlossenen Frage. Fragt man nach Fakten, so müssen sich Befragte bei offenen Fragen erinnern, bei geschlossenen Fragen reicht das Wiedererkennen der richtigen Antwort. Offene Fragen sind auch dazu geeignet, um bei Mehrfachantworten die spontane Reihenfolge zu erheben, z.B. bei der Frage „Welche PolitikerInnen der Gegenwart sind Ihnen am sympathischsten?". Auch für sehr einfache Fragen mit vielen Antwortmöglichkeiten kann eine offene Frage besser sein als eine geschlossene, z.B. die Frage nach dem Alter.

Sind **Mehrfachantworten** möglich, so sollte dies auch eindeutig aus der Fragestellung hervorgehen bzw. direkt nach der Frage vermerkt werden.

Die Erstellung eines guten Fragebogens erfordert Erfahrung und spezielles Wissen über die Fragebogengestaltung. Viele unterliegen dem Trugschluss, dass sie ohne Hilfe einen Fragebogen erstellen können, weil Fragen zu stellen eine einfache Sache ist. Gerade bei der erstmaligen Durchführung einer Befragung sollte man jedoch unbedingt auf professionelle Hilfe zurückgreifen, weil ansonsten die Gefahr zu groß ist, dass die erhobenen Daten unbrauchbar sind.

Datengewinnung	
Primärstatistik	Neue Daten werden z.B. mittels Befragung oder Experiment erhoben
Sekundärstatistik	Zurückgreifen auf bereits veröffentlichtes Datenmaterial
Fragetypen	
geschlossene Fragen	Antwortalternativen sind vorgegeben erschöpfende und ausschließende Alternativen
Bewertungsfragen	am besten dem Schulnotensystem entsprechend symmetrische Kategorien
offene Fragen	keine Antwortalternativen vorgegeben
Mehrfachantworten	Hinweis im Fragebogen

2.4 Datenerfassung und -aufbereitung

Die Informationen müssen aus den ausgefüllten Fragebögen entnommen und aufbereitet werden. Im Normalfall wird die Auswertung mit Hilfe eines Computers erfolgen, im einfachsten Fall unter Zuhilfenahme des Tabellenkalkulationsprogramms EXCEL. Wir wollen daher die Daten in einem ersten Schritt für EXCEL aufbereiten, die Grundprinzipien der Aufbereitung sind ohnehin für alle gängigen Programmpakete gleich.

Der Fragebogen in Abbildung 2.2 dient als Beispiel zur Datenerfassung und -aufbereitung.

Vor der Dateneingabe werden die Fragebögen durchnummeriert, danach wird der Kodeplan erstellt. Der **Kodeplan** ist eine Anleitung für die Übersetzung von Informationen aus dem Fragebogen in Zahlen. Die Informationen des Fragebogens werden in Tabellenform erfasst. Jede Zeile entspricht einem Fragebogen und damit einer Erhebungseinheit und jede Spalte der Tabelle enthält ein Merkmal, was aber nicht immer gleichzusetzen ist mit einer Frage.

Fragen, bei denen nur eine Antwort möglich ist lassen sich durch eine Variable (= ein Merkmal) darstellen. Metrische Merkmale werden als Zahlenwerte übertragen, wobei Maße immer in der gleichen Maßeinheit und mit der gleichen Anzahl an Dezimalstellen erfasst werden sollten. Ordinale Merkmale werden durch natürliche Zahlen kodiert, die Reihenfolge der Ordnung muss dabei erhalten werden. Üblich und auch sinnvoll ist die Verwendung von aufeinander folgenden natürlichen Zahlen, mit 1 beginnend. Nominale Merkmale werden ebenfalls mit natürlichen Zahlen kodiert. Welche Ausprägung welche Kodierung erhält ist unwesentlich, aber auch hier werden üblicherweise von 1 ansteigende ganze Zahlen verwendet.

Bei Fragen mit Mehrfachantworten muss jede Antwortalternative als eigenes dichotomes Merkmal mit den Ausprägungen „ausgewählt" und „nicht ausgewählt" kodiert werden (vgl. Tabelle 2.2 auf Seite 20).

Datenerfassung

- Informationen werden nach Möglichkeit als Zahlen erfasst
- Erste Zeile der Datentabelle enthält die Variablennamen
- Erste Spalte enthält eine laufende Nummer
- Jede Zeile entspricht einer Erhebungseinheit
- Jede Spalte entspricht einer Variable (Fragen mit Mehrfachantworten benötigen für jede Antwortalternative eine Variable)

Fragebogen
Einführung in die Methodenlehre

Geschlecht: □ weiblich □ männlich

Alter: _____ Jahre

Höchste abgeschlossene Schulbildung:
- □ Pflichtschule
- □ Höhere Schule ohne Matura
- □ Höhere Schule mit Matura
- □ Universität, Fachhochschule

Art der Studienzulassung:
- □ AHS □ HAK □ Studienberechtigung
- □ HBLA □ HTL □ Sonstige:_____

Studienrichtung: (Mehrfachnennungen möglich)
- □ Wirtschaftspädagogik □ Statistik
- □ Wirtschaftswissenschaften □ Soziologie
- □ Sozialwirtschaft □ Sonstiges: _____

Maturanote (bzw. Abschlussnote) in Mathematik _____

Maturanote (bzw. Abschlussnote) in Deutsch _____

Körpergröße: _____ cm Körpergewicht: _____ kg

Vielen Dank für Ihre Mitarbeit!

Abb. 2.2. Fragebogen

Tabelle 2.2. Auszug aus dem Kodeplan

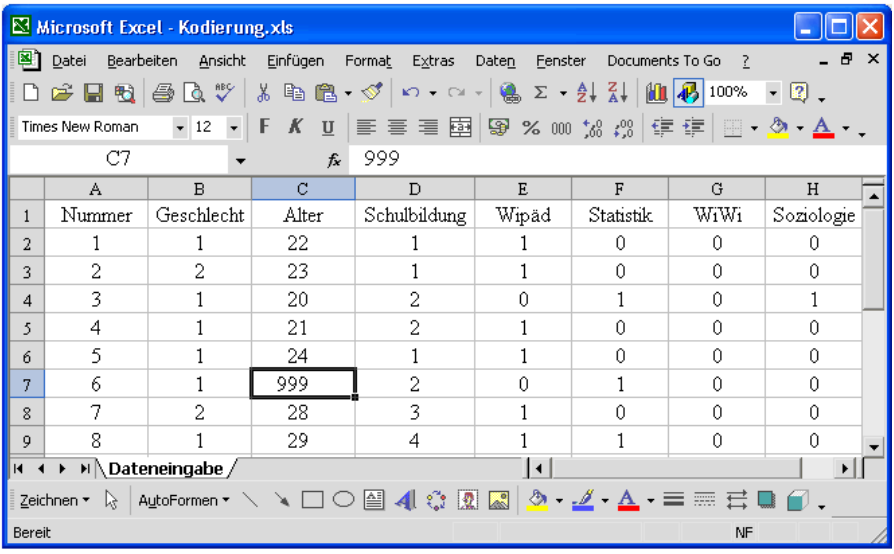
Merkmal	Merkmalstyp	Ausprägung	Kodierung
Geschlecht	nominal, diskret	weiblich	1
		männlich	2
Alter	metrisch, stetig verhältnisskaliert	Messwerte	nicht notwendig
Schulbildung	ordinal, diskret	Pflichtschule	1
		Höhere Schule ohne Matura	2
		Höhere Schule mit Matura	3
		Universität, Fachhochschule	4
Studien-zulassung	nominal, diskret	AHS	1
		HBLA	2
		HAK	3
		HTL	4
		Studienberechtigung	5
		Sonstige	6
Wirtschafts-pädagogik	nominal, dichotom	nicht ausgewählt	0
		ausgewählt	1

Nun kann mit der Dateneingabe begonnen werden, fehlende Antworten werden im einfachsten Fall mit leeren Zellen (kein Leerzeichen, sondern eine Zelle ohne Inhalt) kodiert. Empfehlenswerter ist es, auch fehlende Werte mit einer Zahl zu kodieren. Die Kodierung für einen fehlenden Wert darf natürlich nicht als sinnvolle Ausprägung möglich sein (fehlende Werte für das Merkmal Alter würde man beispielsweise mit 999 kodieren). Bei Fragen mit Mehrfachnennungen werden nicht ausgewählte Alternativen mit 0 und ausgewählte Alternativen mit 1 kodiert. In der ersten Zeile der Tabelle werden die Variablennamen vermerkt. Das Ergebnis der Datenerfassung ist eine Datenmatrix wie in Abbildung 2.3 ersichtlich.

Die nun folgende Datenauswertung wird in diesem Buch ausführlich beschrieben. Daher wird bei der Beschreibung des Ablaufes auf diesen Teil verzichtet und gleich der Abschlußbericht als letzter Teil näher beleuchtet.

2.5 Abschlussbericht

Beim Verfassen des Abschlussberichtes ist darauf zu achten, dass nicht alle LeserInnen ausgebildete StatistikerInnen sind. In einer Einleitung wird kurz das Problem geschildert, die Grundgesamtheit und die Untersuchungsmethode werden dargestellt. Der Auswertungsteil geht von der einfachen zur kompli-



	A	B	C	D	E	F	G	H
1	Nummer	Geschlecht	Alter	Schulbildung	Wipäd	Statistik	WiWi	Soziologie
2	1	1	22	1	1	0	0	0
3	2	2	23	1	1	0	0	0
4	3	1	20	2	0	1	0	1
5	4	1	21	2	1	0	0	0
6	5	1	24	1	1	0	0	0
7	6	1	999	2	0	1	0	0
8	7	2	28	3	1	0	0	0
9	8	1	29	4	1	1	0	0

Abb. 2.3. Datenmatrix

zierten Auswertung, die Ergebnisse sind neben einer verbalen Beschreibung auch in Tabellenform oder als Grafik dargestellt. Ein Resümee bildet den Abschluss des Berichtes. Ergänzt wird dieser Bericht durch ein ansprechendes Deckblatt und einem Inhaltsverzeichnis. Sind im Bericht viele Tabellen bzw. Grafiken vorhanden, so werden LeserInnen ein Tabellen- bzw. ein Abbildungsverzeichnis zu schätzen wissen. Wurde die Datenerhebung mittels Fragebogen durchgeführt, so ist im Anhang der verwendete Fragebogen beizulegen. Generell sollte man auf ein ansprechendes Layout Wert legen, die Verwendung von Kopf- und Fußzeilen kann durchaus hilfreich sein.

Bevor man die Arbeit aus den Händen gibt können folgende Checklisten wertvolle Dienste leisten:

Checkliste Grafiken

- Übersichtlichkeit
- Dreidimensionale Darstellungen vermeiden
- Informative Überschriften und Legenden
- Achsenbezeichnungen, Achsenbeschriftungen
- Unverzerrte Skalierungen
- Achsen beginnen beim Nullpunkt (Ausnahme: geschultes Publikum)
- Quellenangabe bei Verwendung bereits vorhandenen Datenmaterials
- Färbige Grafiken nur bei Farbdruck
- Beachtung bestehender Konventionen (vgl. Kapitel 6.3.4)

Checkliste Tabellen

- Ansprechendes Layout
- Informative Überschriften
- Angabe von Maßeinheiten
- Quellenangabe bei Verwendung bereits vorhandenen Datenmaterials

Checkliste Gesamtbericht

- Deckblatt besonders gründlich auf Tippfehler untersuchen
- Verzeichnisse überprüfen: Inhalt, Tabellen, Grafiken
- Korrekte Nummerierung: Seiten, Kapitel, Tabellen, Grafiken
- Querverweise überprüfen
- Fragebogen im Anhang
- Verwendete Quellen anführen
- Rechtschreibung kontrollieren

2.6 Problemfelder in der Praxis

Die Problemfelder in der Praxis sind vielfältig, hier sollen nur einige davon kurz angeschnitten werden.

2.6.1 Datenschutz, Anonymität

In den meisten Ländern gibt es umfassende gesetzliche Bestimmungen zum Datenschutz, die man kennen und selbstverständlich auch respektieren sollte. So wird im österreichischen Bundesgesetz über den Schutz personenbezogener Daten (kurz: Datenschutzgesetz 2000) genau geregelt, wer unter welchen Voraussetzungen welche Daten erheben und verwenden darf. Insbesondere die Verwendung von personenbezogenen Daten unterliegt besonderen Schutzbestimmungen, so dass man im Normalfall auf nicht personenbezogene Daten zurückgreifen wird. Bei der Erhebung der Daten ist die Anonymität der Personen unbedingt zu wahren, auch das Kennzeichnen von Fragebögen oder Rückkuverts ist rechtswidrig. Freiwillige Angaben von personenbezogenen Daten sind zwar zulässig, man sollte aber berücksichtigen, dass man ehrlichere Antworten erhält, wenn die Anonymität der Befragten garantiert werden kann.

2.6.2 Unzureichendes Studiendesign

Schlechte Planung kann dazu führen, dass man während der Auswertung feststellt, dass interessierende Merkmale nicht oder zu ungenau erhoben wurden. Andere mögliche Fehler liegen in der Stichprobenbildung, wenn festgestellt werden muss, dass die Stichprobe nicht repräsentativ und somit unbrauchbar ist. In der Praxis werden StatistikerInnen immer wieder gefragt, wie viele Personen zu befragen sind um eine repräsentative Stichprobe zu erhalten. Repräsentativität kann nicht über die Anzahl der befragten Personen erreicht werden, sondern nur durch die (zufällige) Auswahl der Personen. Will man die Meinung der Österreicherinnen und Österreicher erheben, so wird man kein repräsentatives Ergebnis erhalten, wenn ausschließlich Frauen befragt werden (auch dann nicht, wenn man alle 4 Millionen Frauen befragt ...). Der Stichprobenumfang darf trotzdem nicht zu klein gewählt werden, denn aus der Meinung von 20 zufällig ausgewählten ÖsterreicherInnen kann auch kein allgemeines Meinungsbild abgelesen werden. Zudem soll die Stichprobe nicht nur hinsichtlich des Merkmals Geschlecht ein repräsentatives Abbild der Grundgesamtheit sein, sondern auch hinsichtlich anderer Merkmale, wie z.B. Alter oder Bildungsniveau.

Auch bei Telefonbefragungen und Internetumfragen muss man sich aber über die Repräsentativität der Stichprobe Gedanken machen. Es ist unmittelbar einsichtig, dass Menschen im Ruhestand leichter telefonisch am Festnetz zu erreichen sind als junge, sehr aktive Menschen. Bei den Mobiltelefonen gibt es zudem viele BenutzerInnen, die ein unregistriertes Handy benutzen und deren Telefonnummer daher nicht bekannt ist. Auch wenn man zufällig Telefonnummern heraussucht, wird man also nicht unbedingt auch eine repräsentative Stichprobe erhalten. Bevor man eine Telefon- oder Internetumfrage initiiert, sollte man sich mit diesen Risiken auseinandersetzen und sie möglichst schon im Vorfeld reduzieren.

2.6.3 Sekundärstatistiken

In manchen Analysen kann man auf Sekundärstatistiken zurückgreifen, das heißt man verwendet Zahlen, die bereits veröffentlicht wurden. In Österreich werden eine Reihe von Statistiken von Statistik Austria erstellt. Bei der Verwendung solcher Daten ist aber größte Vorsicht geboten, weil man sehr genau recherchieren muss, was tatsächlich erhoben wurde.

Statistik Austria veröffentlicht beispielsweise die Anzahl der Erwerbstätigen, wobei diese Zahlen nach dem durch EUROSTAT vorgeschriebenen Labour-Force-Konzept definiert und berechnet werden. Bei den Erwerbstätigen handelt es sich nach dieser Definition um die Summe aus den Selbstständigen,

den bei ihnen ohne Bezahlung mitarbeitenden Familienangehörigen und den Unselbstständigen. Einbezogen sind auch die geringfügig Erwerbstätigen, d.h. alle Personen, die in der Referenzwoche mindestens eine Stunde gearbeitet haben und alle, die in der Referenzwoche nicht gearbeitet haben (z.B. wegen Krankheit), sonst aber erwerbstätig sind.

In den Statistiken des Hauptverbandes der Sozialversicherungen werden geringfügig Erwerbstätige als nicht erwerbstätig eingestuft. Je nach Datenquelle ist damit die Anzahl der Erwerbstätigen unterschiedlich. Dies gilt natürlich in weit größerem Ausmaß für den Vergleich von internationalen Daten. Will man also die Arbeitslosenrate in Österreich mit der in Japan vergleichen, so sollte man zuerst die unterschiedlichen Definitionen recherchieren, damit man nur Werte vergleicht, die auch miteinander vergleichbar sind.

2.6.4 Fehlende Daten

Ein besonderer Augenmerk ist im Zusammenhang mit fehlenden Daten auf den Pretest zu legen. Gibt es im Pretest Fragen, die eine auffällig hohe Ausfallsquote haben, sollte man diese auf jeden Fall überarbeiten. Möglicherweise liegt die hohe Verweigerungsrate an einer missverständlich formulierten Frage oder eine für die Befragten heikle Frage wurde nicht mit genügender Sensibilität gestellt.

Übungsaufgaben

2.1. Skalenniveaus von Merkmalen
Geben Sie für folgende Merkmale die Skalenniveaus an. Sind diese Merkmale stetig oder diskret?
a) Augenfarbe von Personen
b) Produktionsdauer
c) Alter von Personen
d) Preis einer Ware in Euro
e) Parteipräferenz
f) Einwohnerzahl
g) Körpergröße in cm
h) Platzierung in einem Schönheitswettbewerb
i) Gewicht von Gegenständen in kg
j) Schwierigkeitsgrad einer Klettertour
k) Intensität von Luftströmungen in m/s

2.2. Volkszählung
Es wird in der österreichischen Volkszählung z.B. erhoben, in welchem Bundesland die Befragten wohnen, wann sie geboren wurden oder welchem Familienstand sie angehören. Welches Skalenniveau besitzen diese Merkmale, sind sie stetig oder diskret?

3

Anmerkungen zum Umgang mit dem Computer

Dieses Kapitel soll wesentliche Begriffe im Umgang mit dem Computer kurz erläutern. Daneben werden auch einige Aspekte angesprochen, die sonst meistens unerwähnt bleiben und erst nach einigen leidvollen Erfahrungen von BenutzerInnen selbst entdeckt werden.

3.1 Grundlagen

Ordnerstruktur und Namen
Eines der grundlegenden Dinge im Umgang mit Computern ist ein sinnvolles Ablagesystem für einzelne Dateien. Am besten lässt sich dieses Ablagesystem mit einer Ordnerstruktur verwirklichen. Zusammengehörende Files werden in eigene Ordner zusammengefasst und nach Möglichkeit sinnvoll in Unterordner aufgeteilt. Dies erleichtert die Übersicht vor allem dann, wenn viele einzelne Dateien vorhanden sind. Es sollte sich von selbst verstehen, dass die Ordner und Dateien informative Namen aufweisen, damit man möglichst ohne eine Datei zu öffnen weiß, was sich darin befindet. Im Umgang mit Daten empfiehlt es sich, vor dem Beginn der Analysen einen eigenen Sicherungsordner anzulegen, in dem die Originaldaten abgelegt werden und der für die weitere Analyse tabu ist. In diesen Sicherungsordner gehören dann auch z.B. verwendeter Fragebogen und Kodierplan.

Sichern
Aus eigener Erfahrung kann ich konsequentes Sichern der Dateien am besten auch auf externe Speichermedien nur empfehlen. Mit den heutigen technischen Hilfsmittel bedeutet regelmäßiges Sichern (beispielsweise auf einen USB-Stick) nur noch wenig Aufwand. Im Falle einer defekten Festplatte sind Sicherungen aber von nahezu unschätzbaren Wert.

Hilfe
Alle Programme bieten Hilfefunktionen an, deren Qualitäten allerdings sehr unterschiedlich sind. Weitere Hilfe kann man im Internet finden, für EXCEL und SPSS gibt es zudem ausreichend Literatur in Buchform.

3.2 Nützliche Tasten und Tastenkombinationen

Für den effizienten Umgang mit dem Computer sind einige besondere Tasten von großer Bedeutung und sollen daher kurz beschrieben werden.

Strg-Taste
Als verbale Formulierung der Strg-Taste wird meistens Steuerung verwendet, daneben gibt es auch noch die Bezeichnungen Ctrl-Taste (Control). Zu finden ist die Taste meistens im Buchstabenteil der Tastatur in der rechten und linken unteren Ecke. Diese Taste verändert die Funktion der Tastatur- und Maustasten. Mit der Tastenkombination *Strg + C* kann man ausgewählte Bereiche in eine sogenannte Zwischenablage kopieren, mit der Tastenkombination *Strg + V* den Inhalt der Zwischenablage an gewünschter Stelle wieder einfügen. Diese Tastenkombinationen funktionieren nicht nur in EXCEL, sondern in nahezu allen Programmen.

Umschalt-Taste
Direkt über der Strg-Taste befindet sich die Umschalt-Taste (Taste mit Pfeil nach oben), die auch zur Großschreibung verwendet wird, diese Taste wird auch als Shift oder Shifttaste bezeichnet.

Enter-Taste
Zuletzt sei noch die Eingabe-Taste oder Enter-Taste erwähnt. Diese dient zur Bestätigung von Schaltflächen (Buttons). Sie wird auch als Return-Taste bezeichnet. Man erkennt sie meist an der Aufschrift Enter oder einem kleinen Pfeil nach links, der am Anfang einen kleinen Strich nach oben hat. Auf PC's beinhaltet diese Taste zusätzlich die Absatz-Funktion, d.h. in Textprogrammen kann man damit einen Zeilenumbruch erreichen.

Konvention zur Beschreibung von Tastenkombinationen
Im vorliegenden Buch werden Tastenkombinationen durch die Verknüpfung + dargestellt. Beispielsweise meint man mit Tastenkombination *Strg + C* folgende Vorgehensweise: zuerst wird die Strg-Taste gedrückt, man lässt diese auch gedrückt und drückt dann zusätzlich noch die C-Taste. Es ist unmittelbar einsichtig, dass Tastenkombinationen meist aus maximal drei gleichzeitig zu drückenden Tasten bestehen ...

3.3 Drag and Drop

Unter Drag und Drop versteht man das Verschieben von Objekten mit Hilfe der Maus. Dazu wird ein Objekt mit der linken Maustaste ausgewählt und gedrückt gehalten. Durch Verändern der Mausposition wird auch das Objekt verschoben. Das Objekt wird an jener Stelle platziert, an der man die Maustaste auslässt.

3.4 Konventionen zur Beschreibung

Viele Anweisungen werden über die Menüleiste angesteuert. Das Standardmenü ist im Normalfall stets sichtbar, klickt man mit der linken Maustaste auf einen Menüpunkt, so erscheint ein passendes Untermenü, aus dem man wieder per Mausklick auswählen kann. Beispielsweise enthält das EXCEL-Standardmenü einen Menüpunkt Format, im Untermenü findet man als ersten Unterpunkt „Zelle ...". Zur einfacheren Beschreibung werden in diesem Buch Punkte und andere in Anweisungen angeführte Zeichen ignoriert, und eine Verwendung dieser Anweisung würde folgendermaßen aussehen: „Unter dem Menüpunkt *Format* → *Zelle* erscheint ein Dialogfenster, in dem das Registerblatt *Zahlen* aktiviert ist."

Es gibt immer mehrere Möglichkeiten zum Ziel zu gelangen, beispielsweise hätte man statt anklicken der Menüleiste auch mit der Tastenkombination *Strg + 1* das Formatierungs-Dialogfeld erhalten können. In diesem Buch wird meistens nur eine Vorgehensweise beschrieben, was aber nicht gleichbedeutend damit ist, dass es nur eine gibt.

Umgang mit dem Computer

- Sicherungsordner anlegen
- Informative Namen verwenden
- Regelmäßiges Sichern auch auf externe Datenträger
- *Strg + C* kopiert Daten in Zwischenablage
- *Strg + V* fügt Daten aus der Zwischenablage ein

4

Das Tabellenkalkulationsprogramm EXCEL

EXCEL ist das Tabellenkalkulationsprogramm aus dem Office-Paket von Microsoft und damit wohl auf den meisten PCs zu finden. Mittlerweile bietet EXCEL auch viele interessante Möglichkeiten im Bereich der Statistik. Die Ausführungen in diesem Buch beziehen sich auf EXCEL 2002 unter Windows.

4.1 Grundelemente in EXCEL

Das Tabellenkalkulationsprogramm EXCEL verwendet standardmäßig Dateien des Types *.xls, sogenannte Arbeitsmappen. Diese bestehen aus mehreren Arbeitsblättern, die Namen der Arbeitsblätter erscheinen auf einem Register am unteren Rand der Arbeitsmappe (vgl. Abbildung 4.1). Die in Kapitel 3 gestellten Forderungen nach sprechenden Namen und Ordnung gilt auch hier. Für die Verwendung von EXCEL-Files wird empfohlen, die einzelnen Arbeitsblätter mit sinnvollen Namen zu versehen und es nicht bei Tabelle1 bis Tabelle3 zu belassen.

Am oberen Rand der Arbeitsmappe findet man die **Symbolleisten**, die ein effizientes Arbeiten erleichtern und auch den persönlichen Anforderungen angepasst werden können. Unter dem Menüpunkt *Ansicht* → *Symbolleisten* können gewünschte Symbolleisten an- oder abgewählt werden oder bestehende Symbolleisten an eigene Anforderungen angepasst werden. Standardmäßig werden die Symbolleisten Standard und Format angezeigt.

Unter den Symbolleisten befindet sich die **Bearbeitungsleiste**, in der Adresse und Inhalt der aktiven Zelle angezeigt werden, wobei die Adresse üblicherweise aus einer Spaltenbezeichnung in Form von Buchstaben und einer Zeilenbezeichnung im Zahlenformat besteht.

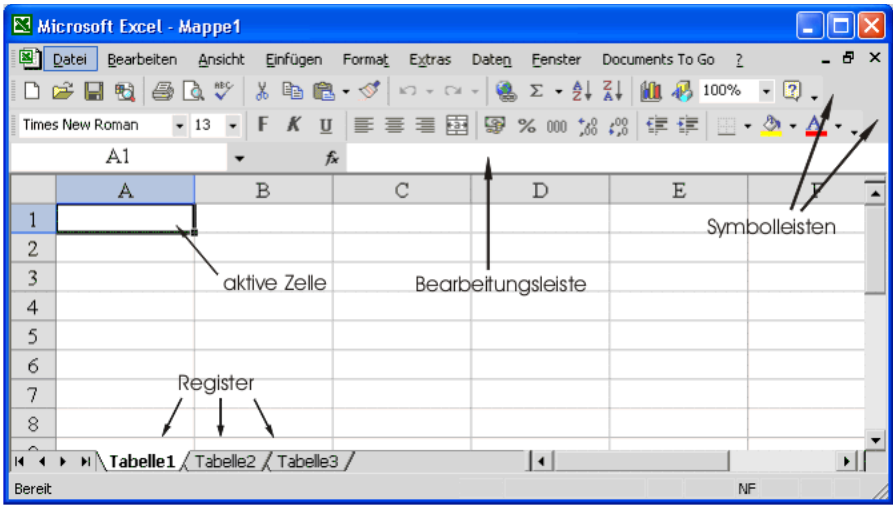


Abb. 4.1. EXCEL-Arbeitsmappe

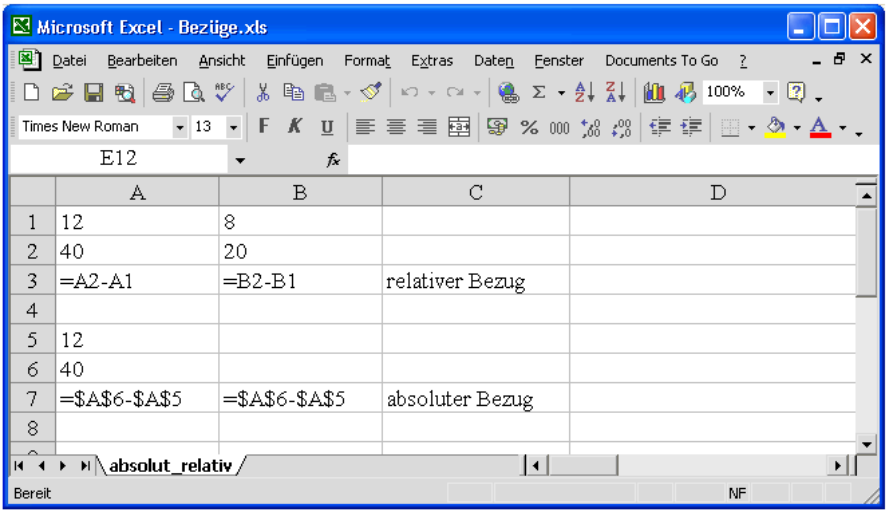
Für die Verarbeitung von Daten werden in EXCEL Formeln verwendet, die in einer Zelle eingegeben werden können. Eine Formel beginnt immer mit einem Gleichheitszeichen und kann als Bestandteile unter anderem Konstanten (Zahlen), mathematische Operatoren (+, -, ...), integrierte Funktionen (z.B. Summe) und Zellbezüge enthalten.

Man unterscheidet zwei Arten von Zellbezügen

- Relativer Zellbezug:
  Kopiert man einen Zellbezug in eine andere Zelle, so verändert sich der Bezug der Zelle in derselben Relation, wie sich die Position verändert hat. Relativ zur Ergebniszelle hat sich also der Bezug nicht verändert, wohl aber die konkrete Adresse der Zelle.

- Absoluter Zellbezug:
  Kopiert man einen absoluten Zellenbezug in eine andere Zelle, so bleibt der Bezugspunkt unverändert. Die konkrete Adresse des Bezuges bleibt also unverändert. Einen absoluten Zellenbezug erhält man wenn man vor die Spalten- bzw. Zeilenbezeichnung ein $ setzt. Daneben kann auch nur ein Teil der Zelladresse absolut gesetzt werden, indem beispielsweise nur vor die Spaltenbezeichnung ein $ gesetzt wird.

In Abbildung 4.2 wurde der Zelleninhalt von A3 nach B3 kopiert. Durch die relativen Bezüge in A3 ändern sich nun die konkreten Bezüge, lediglich die relative Position zur Ergebniszelle bleibt erhalten. Beim Kopieren des Zellen-



	A	B	C	D
1	12	8		
2	40	20		
3	=A2-A1	=B2-B1	relativer Bezug	
4				
5	12			
6	40			
7	=$A$6-$A$5	=$A$6-$A$5	absoluter Bezug	
8				

Abb. 4.2. EXCEL: Absolute und relative Bezüge

inhaltes von A7 nach B7 bleiben die konkreten Bezüge erhalten, weil in diesem Fall absolute Bezüge verwendet wurden.

Zellbezüge

Relativer Bezug	konkreter Bezug ändert sich beim Kopieren oder Verschieben
Absoluter Bezug	Zellbezug bleibt unverändert $ vor Spalten- bzw. Zeilenbezeichnung

Praxistipp

Durch das Kopieren und Verschieben von Formeln können sehr leicht Fehler entstehen, weil man zu wenig Aufmerksamkeit auf die Unterscheidung zwischen absolute und relative Bezüge gelegt hat. Daher sollte man die Formeln nach dem Kopieren bzw. nach dem Verschieben noch einmal kontrollieren.

4.2 Formatierung in EXCEL

Standardmäßig sind Zellen in EXCEL im sogenannten Standardformat formatiert und verfügen damit über kein bestimmtes Zahlenformat. Für ein ansprechendes Tabellenlayout empfiehlt sich eine benutzerdefinierte Formatierung.

Unter dem Menüpunkt *Format* → *Zelle* erscheint ein Dialogfenster, in dem das Registerblatt *Zahlen* als aktives Registerblatt geöffnet wird. Als letzte Option der angegebenen Kategorien im linken Fensterbereich findet man *Benutzerdefiniert*, ein Anklicken dieser Option öffnet im rechten Teil des Fensters eine weitere Auswahlleiste mit der Bezeichnung *Typ* (vgl. Abbildung 4.3).

Für viele Anwendungen ist hier die Auswahl #.##0,00 empfehlenswert, die bewirkt, dass Tausendertrennzeichen - sofern benötigt - eingefügt werden und zwei Dezimalstellen angegeben werden. Die zwei Dezimalstellen werden jedenfalls angeführt, die Stellen, die mit einer Raute definiert sind werden nur dann angeführt, wenn sie notwendig sind.

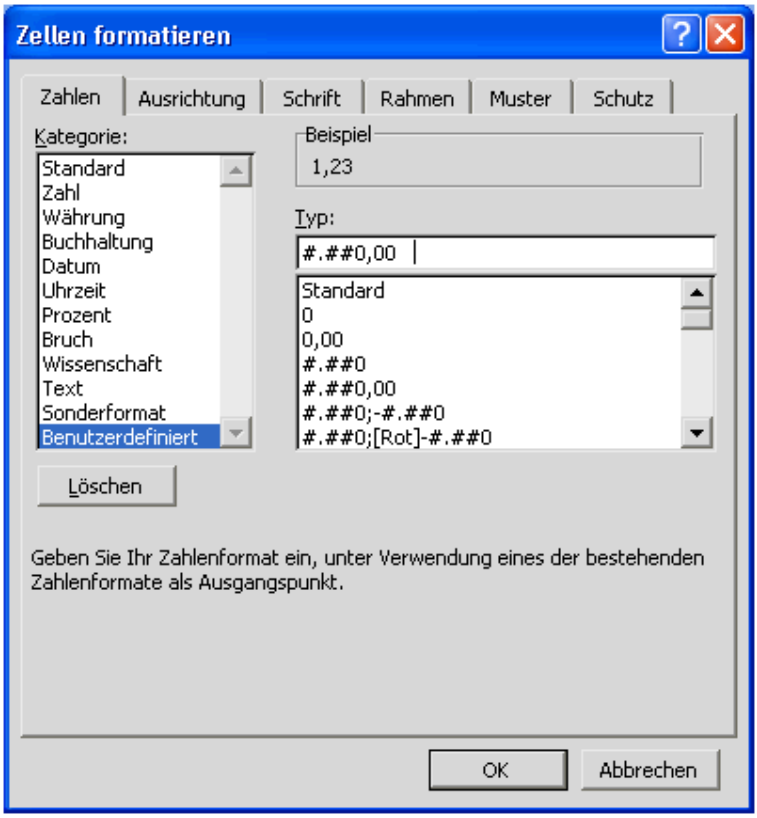


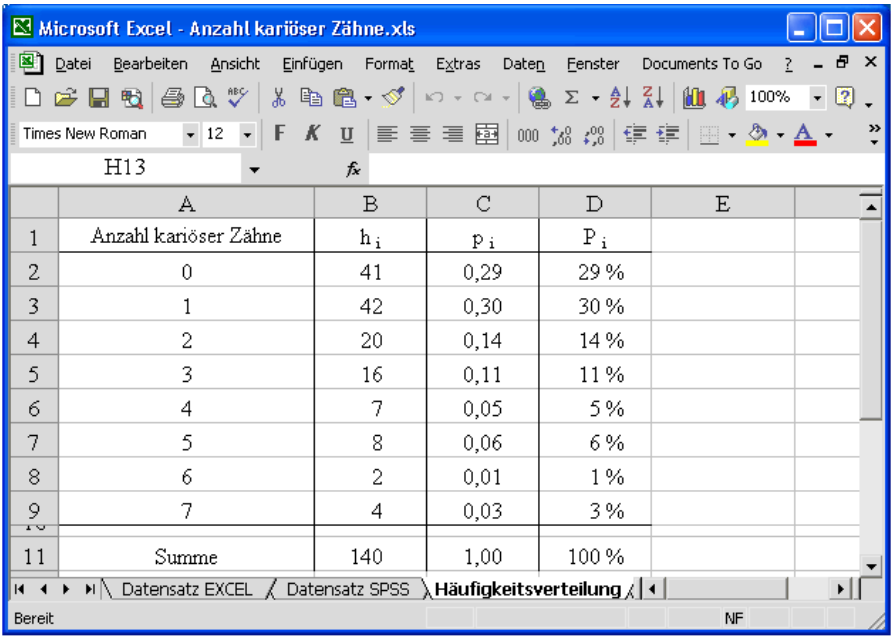
Abb. 4.3. EXCEL: Zellen formatieren

Tabelle 4.1 listet einige Beispiele auf, aus denen entnommen werden kann, wie sich unterschiedliche Formatierungen auswirken.

Tabelle 4.1. Auswirkung ausgewählter Formatierungen

Eingabe	Formattyp	Darstellung
1,23	0,0	1,2
1,23	0,00	1,23
1,23	0,000	1,230
1,23	0,00#	1,23
1,23	0,00 €	1,23 €
1,23	00,00	01,23
1,23	#0,00	1,23
1,23	0.000,00	0.001,23
1,23	#.##0,00	1,23

Unabhängig von der konkreten Auswahl der Formatierung empfiehlt sich das Einfügen von Leerzeichen nach der ausgewählten Formatierung. Zusammen mit einer rechtsbündigen Ausrichtung ergibt sich nun eine Formatierung, in der die Zahlen vom Zellenrand etwas abgerückt sind und trotzdem rechtsbündig angeordnet sind, wie in Abbildung 4.4 ersichtlich.



	A	B	C	D	E
1	Anzahl kariöser Zähne	h_i	p_i	P_i	
2	0	41	0,29	29 %	
3	1	42	0,30	30 %	
4	2	20	0,14	14 %	
5	3	16	0,11	11 %	
6	4	7	0,05	5 %	
7	5	8	0,06	6 %	
8	6	2	0,01	1 %	
9	7	4	0,03	3 %	
11	Summe	140	1,00	100 %	

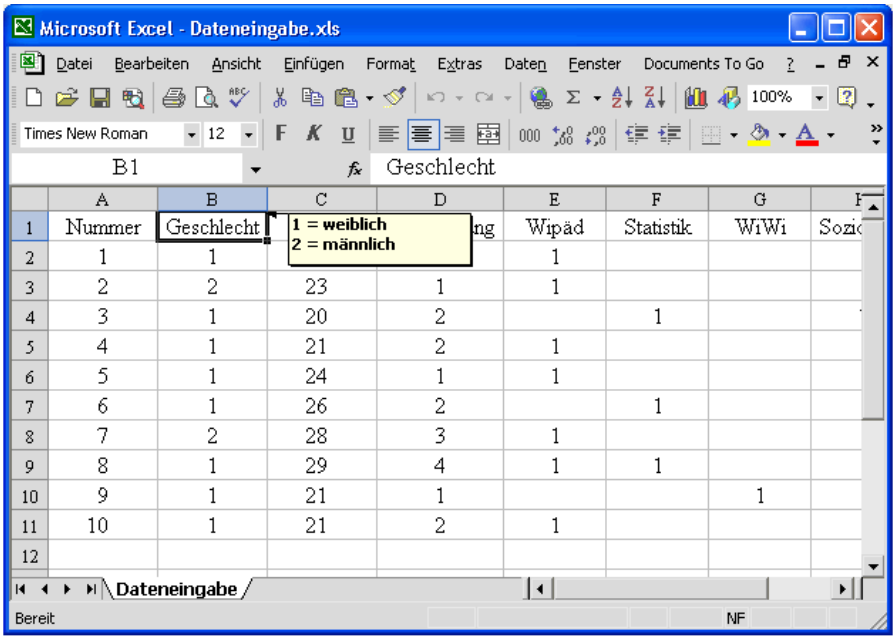
Abb. 4.4. EXCEL: formatierte Darstellung

4.3 Dateneingabe

Die Dateneingabe in EXCEL erfolgt in einem Arbeitsblatt in Tabellenform, wobei jede Zeile einer Erhebungseinheit (z.B. einer Person) und jede Spalte einer Variablen (z.B. Geschlecht) entspricht. Diese Variable beinhaltet im Normalfall ein einzelnes Merkmal, im Falle von Datensätzen, die aus Fragen mit Mehrfachantworten entstanden sind, enthält eine Variable eine einzelne Merkmalsausprägung. Jede Zelle enthält somit die Ausprägung eines Merkmals einer bestimmten Erhebungseinheit.

In der Praxis ist es sinnvoll in der ersten Spalte eine Nummerierung der Erhebungseinheiten mitzuführen, in der ersten Zeile die Bezeichnungen der Merkmale anzuführen und Ausprägungen nur in kodierter Form (vgl. Kapitel 2.4) zu verwenden.

Die Kodierung kann unter dem Menüpunkt *Einfügen* → *Kommentar* als Kommentar zum Namen des Merkmals hinzugefügt werden (Abbildung 4.5).



	A	B	C	D	E	F	G	H
1	Nummer	Geschlecht		ng	Wipäd	Statistik	WiWi	Sozi
2	1	1			1			
3	2	2	23	1	1			
4	3	1	20	2		1		
5	4	1	21	2	1			
6	5	1	24	1	1			
7	6	1	26	2		1		
8	7	2	28	3	1			
9	8	1	29	4	1	1		
10	9	1	21	1			1	
11	10	1	21	2	1			
12								

Abb. 4.5. Dateneingabe in einem EXCEL-Arbeitsblatt

4.4 Statistische Analysen

In EXCEL gibt es mehrere Möglichkeiten der statistischen Analyse:

- Statistik im Rahmen der Tabellenkalkulation:
  Hier werden die zugrundeliegenden Formeln selbst eingegeben.

- Statistik mit Hilfe des Funktionsassistenten:
  Unter dem Menüpunkt *Einfügen* → *Funktion* erscheint der Funktionsassistent, der neben statistischen Funktionen auch andere wichtige Funktionen beinhaltet.

  Der Funktionsassistent lässt sich auch mit der Schaltfläche  aufrufen, die standardmäßig in der Menüleiste angeführt ist. Ist dies nicht der Fall, kann die Schaltfläche auf folgende Art erstellt werden: Unter *Ansicht* → *Symbolleisten* → *Anpassen* erscheint eine Dialogbox, dort findet man im Blatt *Befehle* in der Kategorie *Einfügen* den Befehl *Funktion einfügen*.

  Durch Drag and Drop kann der zugehörige Button in die Symbolleiste gezogen werden und steht ab diesem Zeitpunkt dort zur Verfügung. Auch in der Bearbeitungsleiste kann dieser Button angeführt sein.

- Statistik mit den Analysefunktionen:
  Die Analysefunktionen müssen zuerst über den Add-Inn-Manager aktiviert werden (Menüpunkt *Extras* → *Add-Inn-Manager*), dadurch werden unter dem Menüpunkt *Extras* die Analysefunktionen angeboten.

Praxistipps im Umgang mit EXCEL

- Bezüge kontrollieren
- Informative Namen verwenden
- Ausreichend dokumentieren
- Dateneingabe
  - eine Zeile = eine Erhebungseinheit
  - eine Spalte = ein Merkmal bzw. eine Merkmalsausprägung
  - Fälle durchnummerieren in der ersten Spalte
  - Bezeichnungen der Merkmale in erster Zeile
  - Kodierungen als Kommentare anführen
- EXCEL eignet sich gut zum Layoutieren von Tabellen und Grafiken

5

Das Statistikpaket SPSS

SPSS (Statistical Package for the Social Sciences) ist ein Softwarepaket, welches speziell für statistische Auswertungen konzipiert ist. Die Ausführungen in diesem Buch beziehen sich auf die Version 12.0 für Windows.

5.1 Erste Schritte in SPSS

Nach dem Aufrufen des Programms (entweder mittels Doppelklick auf die Verknüpfung am Desktop oder mittels *Start → Programme → SPSS für Windows → SPSS 12.0 für Windows*) erscheint das Dialogfeld aus Abbildung 5.1.

Dieses Dialogfeld enthält ein Kontrollkästchen, in dem man auswählen kann, dass dieses Dialogfeld nicht mehr angezeigt werden soll. Man sollte sich jedoch diese Entscheidung gut überlegen, denn die einzige Möglichkeit dies wieder zu ändern ist eine vollständige Neuinstallation von SPSS.

Das Lernprogramm starten bietet die Möglichkeit sich mit der neuen Software mit Hilfe einer geführten Tour vertraut zu machen. Dieses Lernprogramm ist vor allem für BenutzerInnen geeignet, die wenig Erfahrung im Umgang mit Software-Paketen aller Art haben, BenutzerInnen mit mehr Computererfahrung finden sich auch ohne Lernprogramm sehr schnell zurecht.

Die beiden Optionen *Eine vorhandene Abfrage ausführen* und *Neue Abfrage mit Datenbank-Assistent anlegen* sind ausschließlich für fortgeschrittene BenutzerInnen interessant, auf eine Beschreibung dieser Optionen wird daher verzichtet und auf die einschlägige Literatur verwiesen.

Die Option *Vorhandene Datenquelle öffnen* ermöglicht den Zugriff auf bereits vorhandene SPSS-Datendateien. Viele solcher Datendateien werden standardmäßig bei der Installation von SPSS zur Verfügung gestellt und sind im

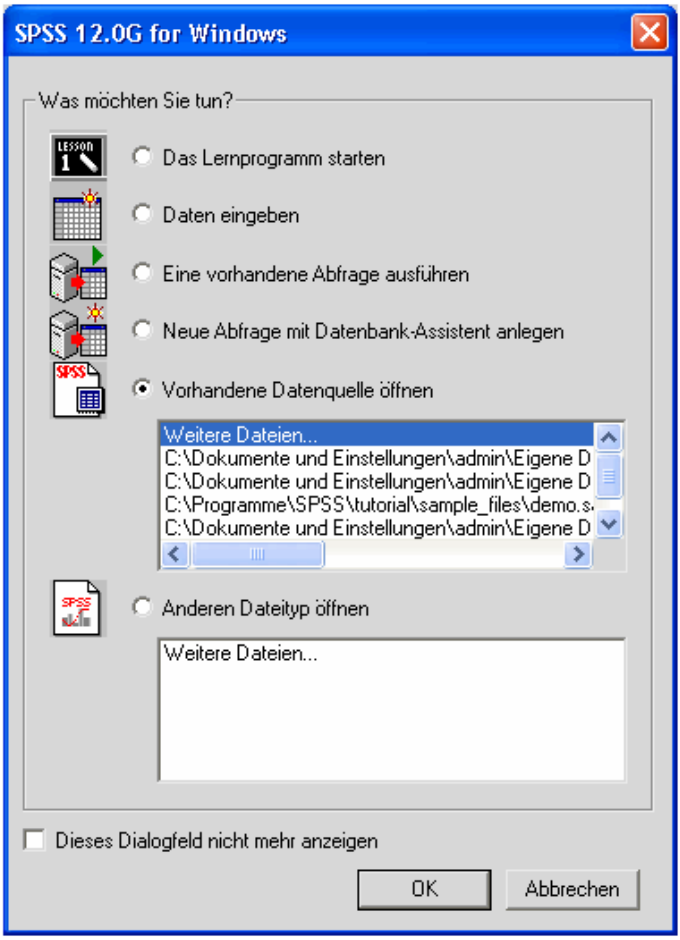


Abb. 5.1. Eröffnungsdialogfeld in SPSS

SPSS-Ordner bzw. in verschiedenen Unterordnern abgelegt. Die Standardeinstellung *Weitere Dateien* führt automatisch in den SPSS-Ordner, sofern von BenutzerInnen nichts anderes festgelegt wurde.

Die Option *Anderen Dateityp öffnen* ermöglicht das unkomplizierte Einlesen anderer Datenformate, wie z.B. auch von *.xls-Dateien, wie in Kapitel 5.3 beschrieben wird.

Gleichzeitig mit dem Erscheinen dieses Dialogfeldes wird im Hintergrund der Dateneditor von SPSS geöffnet. Mit der Option *Daten eingeben* oder mit der Schaltfläche *Abbrechen* schließt das Dialogfeld und der Dateneditor wird aktiviert.

5.2 Der Dateneditor

Mit dem Dateneditor können Dateien vom Typ $*.sav$ bearbeitet werden, die Daten in tabellarischer Form enthalten.
Der Dateneditor bietet zwei verschiedene Ansichten, die Datenansicht (Abbildung 5.2) und die Variablenansicht (Abbildung 5.3 und 5.5). Der Wechsel zwischen den Ansichten erfolgt durch Anklicken der entsprechenden Registerblätter im linken unteren Fensterbereich.

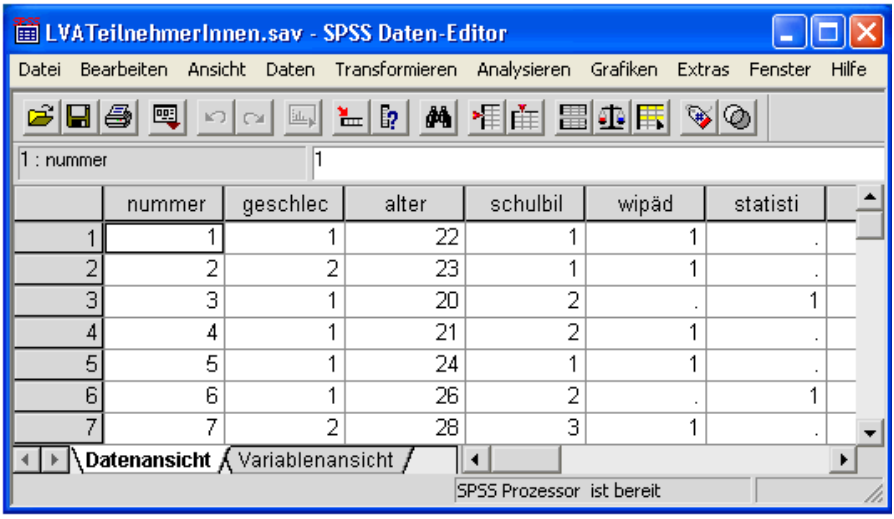


Abb. 5.2. Dateneditor in SPSS - Datenansicht

Datenansicht
In der Datenansicht kann man Daten wie in einer EXCEL-Arbeitsmappe eingeben. Auch hier gilt wieder das Prinzip jede Zeile eine Erhebungseinheit, jede Spalte ein Merkmal bzw. eine Merkmalsausprägung (vgl. Kapitel 2.4) und jede Zelle eine kodierte Ausprägung. Die Namen der Merkmale werden im Unterschied zu EXCEL nicht in der ersten Zeile vermerkt, sondern sind als Spaltenbezeichnung sichtbar. Die Zeilen sind durchnummeriert, analog zu EXCEL sollte als erste Variable stets die laufende Nummer des Datensatzes angelegt werden.

Variablenansicht
In der Variablenansicht entspricht jede Zeile einem Merkmal und damit einer Spalte in der Datenansicht. Für jedes Merkmal können verschieden Informationen gespeichert werden, die in den einzelnen Spalten der Variablenansicht angezeigt werden.

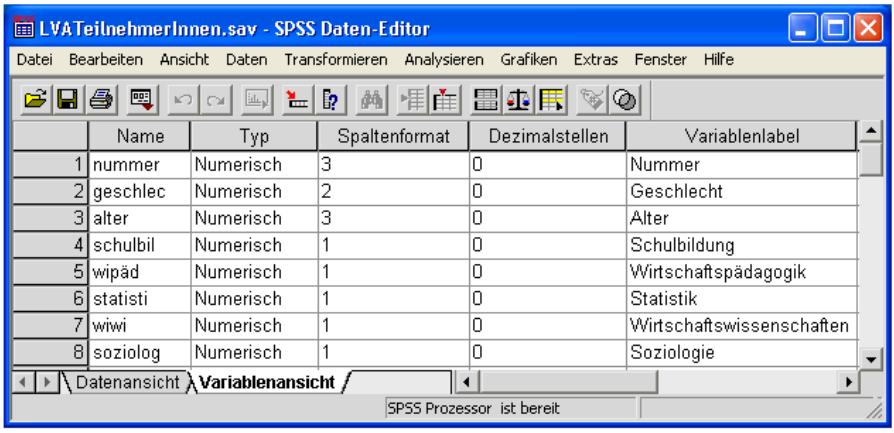


Abb. 5.3. Dateneditor in SPSS - Variablenansicht, linker Bereich

Name
Die wichtigste Information ist der Variablenname, für den in SPSS gewisse Beschränkungen gelten.

Kriterien für Namen in SPSS

- Der Name darf maximal 64 Zeichen lang sein.
- Zulässige Zeichen sind Buchstaben, Ziffern, manche Zeichen und die Sonderzeichen _ (Unterstrich), . (Punkt), @ (Klammeraffe) und # (Raute).
- Der Name muss mit einem Buchstaben beginnen.
- Der Name darf nicht mit einem Punkt enden.
- Groß- und Kleinschreibung wird nicht unterschieden.

Diese Beschränkungen sind teilweise von der verwendeten Version abhängig, z.B. darf bis zur Version 11 der Name lediglich 8 Zeichen umfassen. Die beschriebenen Konventionen sind nur die wesentlichsten, es gibt noch weitere Konventionen, auf deren detaillierte Beschreibung hier verzichtet wird. Wählt man einen Namen, der nicht den Konventionen von SPSS entspricht, so erhält man umgehend eine informative Fehlermeldung und kann den Namen entsprechend anpassen.

Typ
In der Spalte Typ kann man aus verschiedenen Variablentypen wählen, als Standardeinstellung ist der Variablentyp Numerisch festgelegt.

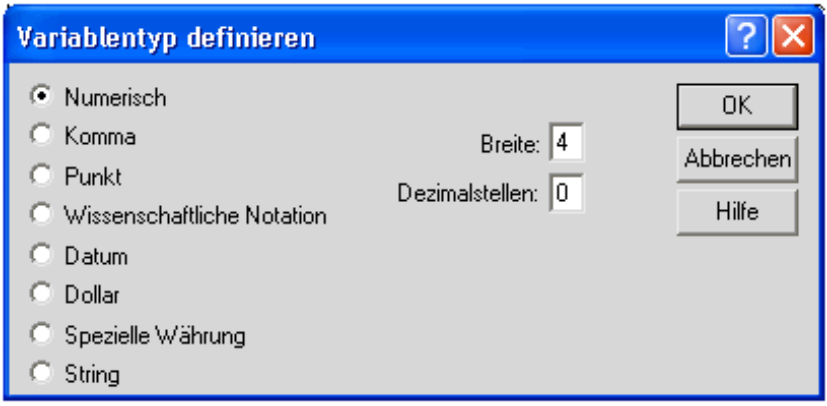


Abb. 5.4. SPSS Variablentypen

Eine Kurzbeschreibung der Variablentypen kann aus folgender Übersicht entnommen werden:

Numerisch	Enthält Ziffern, ein Dezimaltrennzeichen und falls notwendig ein vorangestelltes Minuszeichen. Im Feld Breite wird die maximale Stellenanzahl eingetragen, einschließlich einer Stelle für das Dezimaltrennzeichen. Im Feld Dezimalstellen werden die gewünschten Dezimalstellen angegeben, maximal ist Breite abzüglich 1 möglich.
Komma	Enthält Ziffern, einen Punkt als Dezimaltrennzeichen, und falls notwendig ein vorangestelltes Minuszeichen und Kommas als Tausendertrennzeichen. Im Feld Breite wird die maximale Stellenanzahl eingetragen, einschließlich der Stellen für das Dezimaltrennzeichen und die Kommas. Im Feld Dezimalstellen werden die gewünschten Dezimalstellen angegeben.
Punkt	Enthält Ziffern, ein Komma als Dezimaltrennzeichen, und falls notwendig ein vorangestelltes Minuszeichen und Punkte als Tausendertrennzeichen. Im Feld Breite wird die maximale Stellenanzahl eingetragen, einschließlich der Stellen für das Dezimaltrennzeichen und die Tausendertrennzeichen. Im Feld Dezimalstellen werden die gewünschten Dezimalstellen angegeben.

Wissenschaftliche Notation	Numerische Werte werden in Exponentenschreibweise angezeigt. Dem Exponenten kann entweder ein E oder ein D (mit oder ohne Vorzeichen) oder ein Vorzeichen allein vorangestellt werden. Demnach besitzen folgende Darstellungen den gleichen Zahlenwert: 123; 1,23E2; 1,23D2; 1,23E+2 und 1,23+2.
Datum	Geeignet für Datums- und/oder Zeitangaben, es kann aus verschiedenen Möglichkeiten der Formatierung aus einem Pull-Down-Menü ausgewählt werden.
Dollar	Enthält neben den Ziffern ein Dollarzeichen und je nach Bedarf und Einstellung ein Dezimaltrennzeichen und Tausendertrennzeichen.
Spezielle Währung	Numerische Werte in einem Währungsformat, das unter *Bearbeiten → Optionen* im Registerblatt *Währung* definiert werden kann.
String	Enthält Zeichenketten mit maximal 255 Zeichen. Da Stringvariablen nicht in Berechnungen verwendet werden können sollte man diesen Variablentyp weitgehend vermeiden.

Bei numerischen Formaten ist die Art des Dezimaltrennzeichens (Punkt oder Komma) von der Spezifikation in den Regions- und Sprachoptionen in der Windows-Systemsteuerung abhängig. Intern wird der vollständige Wert gespeichert und dieser wird auch für die Berechnungen verwendet, angezeigt wird nur die gewünschte Zahl an Dezimalstellen.

Spaltenformat
Das Spaltenformat ergibt sich aus der Breite, die beim Variablentyp festgelegt wurde. Änderungen, die beim Spaltenformat durchgeführt werden, werden automatisch im Variablentyp angepasst.

Dezimalstellen
Die Anzahl der Dezimalstellen wird vom Variablentyp übernommen. Änderungen bei den Dezimalstellen werden auch im Variablentyp angepasst. Für bestimmte Variablentypen, z.B. für String oder Datum, können sinnvollerweise keine Dezimalstellen angegeben werden, daher ist die Standardeinstellung 0 Dezimalstellen.

Variablenlabel
Als Variablenlabel kann man eine lange Version des Namens oder auch die gesamte zugehörige Frage angeben. Standardmäßig werden bei der Ausgabe nicht die Namen sondern die Labels verwendet, dies lässt sich unter dem Menüpunkt *Bearbeiten → Optionen* im Blatt *Beschriftung der Ausgabe* ändern.

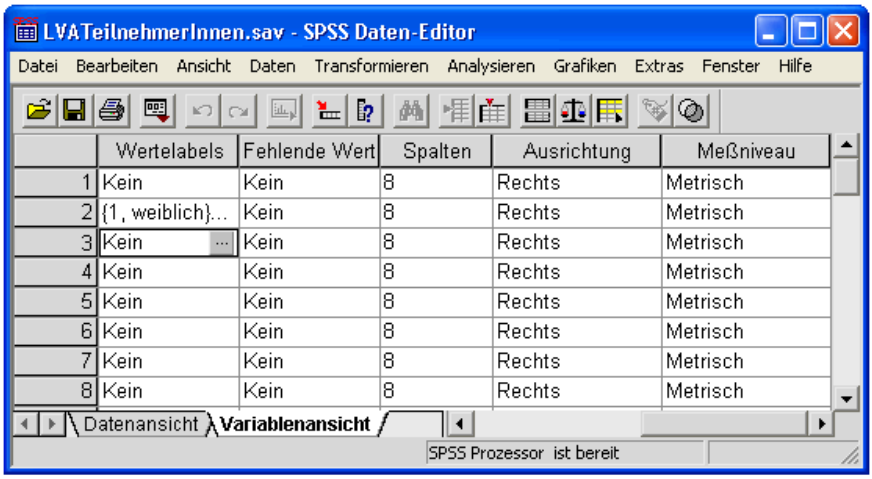


Abb. 5.5. Dateneditor in SPSS - Variablenansicht, rechter Bereich

Wertelabels
In der Spalte Wertelabels können die Kodierungen der Ausprägungen mit verbalen Beschreibungen versehen werden. Durch Anklicken der betreffenden Zelle erscheint an der rechten Seite der Zelle ein graues Kästchen mit drei Punkten, durch Anklicken des Kästchens erscheint die Dialogbox für die Wertelabels (vgl. Abbildung 5.5). Jede Zuweisung muss gesondert durch die Schaltfläche *Hinzufügen* oder die Enter-Taste bestätigt werden. Das Ende aller Zuweisungen wird mit dem OK-Button bestätigt. In der Ausgabe werden auch bei den Ausprägungen standardmäßig die Labels verwendet, auch diese Option kann unter dem Menüpunkt *Bearbeiten* → *Optionen* im Blatt *Beschriftung der Ausgabe* geändert werden.

Fehlende Werte
Fehlende Werte können mehrere Ursachen haben, beispielsweise können Befragte die Beantwortung einer Frage verweigern, oder die Beantwortung einer Frage ist den Befragten nicht möglich, weil sie die dafür notwendigen Voraussetzungen nicht erfüllen. Fragt man beispielsweise zuerst nach der höchsten abgeschlossenen Schulbildung und dann nach der absolvierten Studienrichtung, falls ein Universitätsstudium abgeschlossen wurde, so ist unmittelbar einsichtig, dass die Frage nach der Studienrichtung nur von jenen Befragten beantwortet werden kann, die ein Universitätsstudium als höchste Schulausbildung angegeben haben.

Die Behandlung von fehlenden Werten in der Datenmatrix kann auf zwei verschiedene Wege erfolgen:

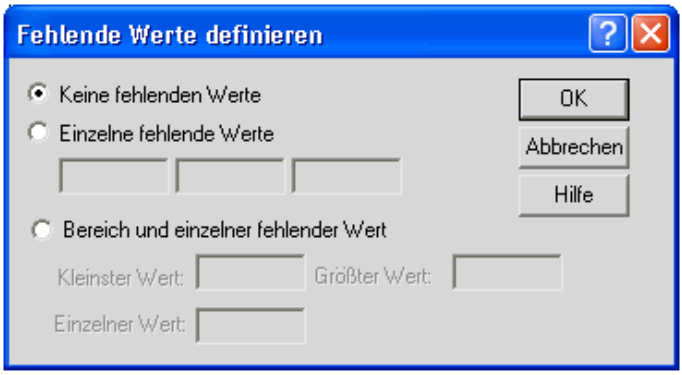


Abb. 5.6. SPSS Fehlende Werte

- Systemdefinierte fehlende Werte:
  Werden in einer Datenmatrix numerische Daten nicht ausgefüllt, so weist SPSS ihnen systemdefinierte fehlende Werte zu, in der Datenansicht sind diese als Punkt bzw. Komma sichtbar. Die Voreinstellung *Keine fehlenden Werte* entspricht dieser Vorgehensweise.

- Benutzerdefinierte fehlende Werte:
  Hier können BenutzerInnen für die verschiedenen Gründe eines fehlenden Wertes Kodierungen einführen und diese Kodierungen dann als fehlende Werte definieren. Die Kodierung für einen fehlenden Wert darf natürlich nicht als sinnvolle Ausprägung möglich sein (fehlende Werte für das Merkmal Alter würde man beispielsweise mit 999 kodieren). Auch die Kodierungen für fehlende Werte können mit Wertelabels versehen werden.

Spalten
Die hier eingegebene Anzahl von Spalten bezieht sich auf die Darstellung der Daten in der Datenansicht. Sinnvollerweise sollte die hier angegebene Anzahl mit der Breite übereinstimmen oder diese übersteigen. Ist die gewählte Anzahl zu niedrig, so erscheinen in der Datenansicht als Zelleninhalt Sterne.

Ausrichtung
Hier kann die Ausrichtung des Zellinhaltes der Datenansicht innerhalb einer Zelle eingestellt werden. Zur Verfügung stehen *Rechts* für rechtsbündige, *Mitte* für zentrierte und *Links* für linksbündige Darstellungen. Die Voreinstellung für numerische Werte ist rechtsbündig, für Stringvariablen linksbündig.

Messniveau
In der letzten Spalte kann das Skalenniveau der Merkmale (nominal, ordinal oder metrisch) festgelegt werden. Allerdings überprüft SPSS nicht, ob ein

Verfahren für ein bestimmtes Messniveau zulässig ist. Damit haben die BenutzerInnen dafür zu sorgen, dass nur zulässige Verfahren verwendet werden.

5.3 Datenquellen

In SPSS lassen sich über das Dialogfeld, das beim Starten von SPSS geöffnet wird, auch Datenquellen mit anderen Dateiformaten problemlos verwenden. Als Beispiel sei hier das Öffnen und Bearbeiten einer EXCEL-Datei angeführt. Beim Start von SPSS wird im Eröffnungsdialogfeld standardmäßig die Option *Vorhandene Datenquelle öffnen* mit Auswahl *Weitere Dateien* verwendet. Bestätigt man diese Einstellung, so öffnet sich ein Dialogfenster zum Datei öffnen. Dieses Dialogfenster erhält man auch unter dem Menüpunkt *Datei* → *Öffnen* → *Daten*. In diesem kann nun im unteren Bereich der Dateityp auf EXCEL geändert werden und man kann wie gewohnt die gewünschte Datei auswählen (Abbildung 5.7).

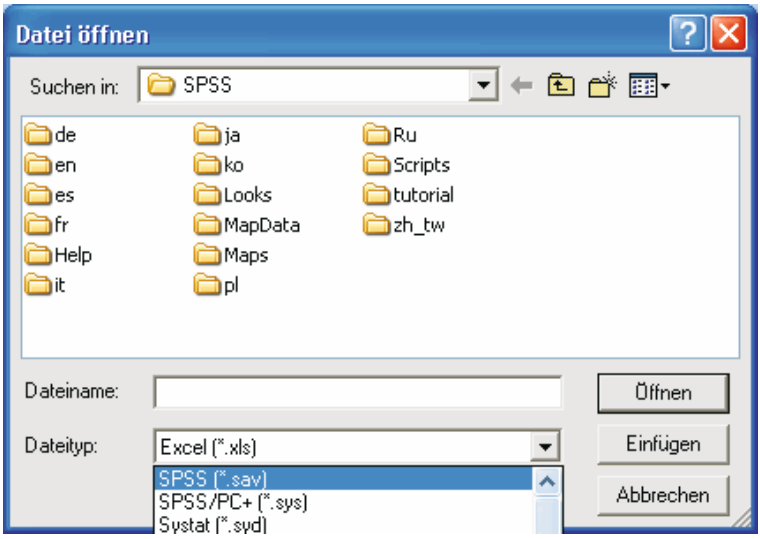


Abb. 5.7. Dialogfeld Datei öffnen

Nach der Auswahl der Datei kann man in einem weiteren Dialogfenster auswählen, welches Arbeitsblatt und welcher Bereich im Arbeitsblatt verwendet werden sollen. Des weiteren ist hier auch festzulegen, ob die Variablennamen aus der ersten Zeile der EXCEL-Datei übernommen werden können (Abbildung 5.8).

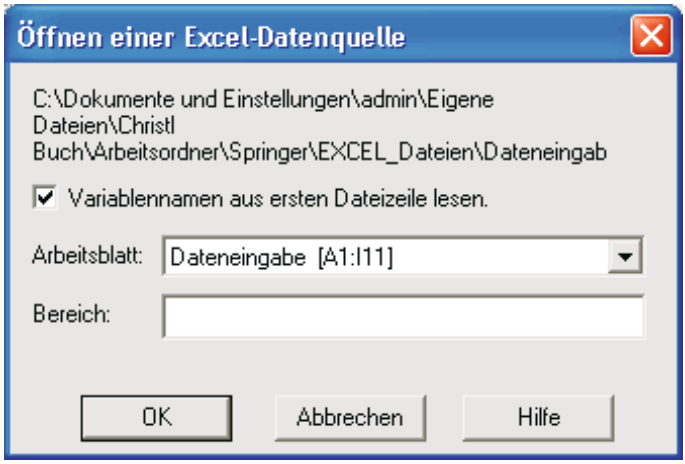


Abb. 5.8. Dialogfeld EXCEL-Datei öffnen

Die Inhalte der EXCEL-Datei stehen nun in einer neuen, noch unbenannten $*.sav$-Datei zur Verfügung. In der Variablenansicht sind die Einstellungen ersichtlich, die SPSS beim Importieren ausgewählt hat, diese sind gegebenenfalls zu ändern.

Beim Einlesen einer EXCEL-Datei können manchmal Fehlermeldungen oder Warnungen auftreten, die aber selbsterklärend sind. Mögliche Fehlerquellen dabei sind die Namensgleichheit von Variablen und gänzlich leere Spalten oder Zeilen.

Umgang mit SPSS

- *.sav sind Datendateien, die mit dem Dateneditor bearbeitet werden können
- *.spo sind Ausgabedateien, die im Viewer bearbeitet werden können
- Datendateien im Dateneditor haben eine Datenansicht und eine Variablenansicht
- Namen unterliegen Beschränkungen, Variablenlabels dienen als Langform für Namen
- Wertelabels enthalten die Bezeichnungen der Kodierungen
- EXCEL-Dateien öffnen mit *Datei* → *Öffnen* → *Daten*, Dateityp auf EXCEL(*.xls) ändern und auswählen

5.4 Der Viewer

Im Viewer können die Ergebnisdateien (Dateityp *.spo) betrachtet und auch bearbeitet werden. Hier werden nicht nur Ergebnisse von Analysen ausgegeben, sondern auch Warnungen und Fehlermeldungen. Als Beispiel für die Funktionsweise des Viewers wollen wir uns Informationen über die geöffnete Datendatei geben lassen. Dazu wählen wir im Dateneditor den Menüpunkt *Datei → Datendatei-Informationen anzeigen → Arbeitsdatei* (für die Version 11 *Extras → Datei-Info*) aus. Das Programm startet den Viewer und zeigt die in Abbildung 5.9 sichtbare Ausgabedatei.

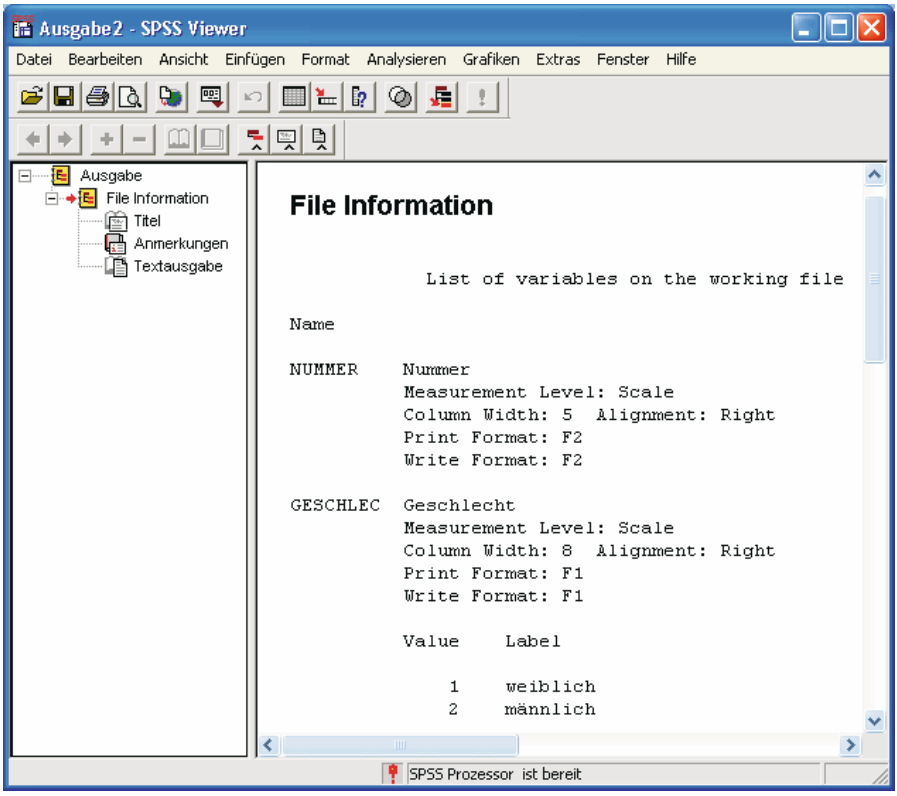


Abb. 5.9. Viewer-Ansicht

Die Menüleiste im Viewer entspricht derjenigen des Dateneditors. Damit können Analysen und Befehle auch vom Viewer aus gestartet werden können, sofern die benötigte Datendatei geöffnet ist.

5.5 Datenaufbereitung

Bevor Daten analysiert werden können, ist es fast immer notwendig die Daten zuerst in geeigneter Form aufzubereiten. Die vorgenommenen Änderungen oder Erweiterungen im Datensatz sollten genau dokumentiert werden, oberste Priorität in der Datenanalyse ist Transparenz und Nachvollziehbarkeit. Die drei wesentlichsten Teilbereiche der Datenaufbereitung werden im Folgenden näher betrachtet.

5.5.1 Fehlende Werte

Wie fehlende Werte entstehen können wurde bereits in Kapitel 5.2 aufgezeigt. Dort wurde auch angeführt, wie man fehlende Werte in der Datenmatrix behandelt. Offen ist die Frage, wie sich fehlende Werte in der weiteren Analyse auswirken. Die einfachste Methode, mit fehlenden Werten in der Analyse umzugehen, ist das Ausschließen aller Datensätze, die in einer der gerade verwendeten Variablen einen fehlenden Wert aufweisen. Das ist zwar sicher die einfachste Methode mit fehlenden Werten umzugehen, aber in der Praxis möglicherweise nicht die beste. Einerseits wird der Datensatz dadurch reduziert, was bei vielen fehlenden Werten zum Problem werden kann, andererseits sind manchmal auch diese fehlenden Werte Informationsträger (Befragte verweigern nicht zufällig eine Antwort, sondern sie tun dies aus ganz bestimmten Gründen). Damit würde ein Ausschluss dieser Datensätze zu einer verzerrten Stichprobe führen, was den Erklärungsgehalt der Ergebnisse zunichte macht.

Im Forschungsbereich Missing Data werden Verfahren entwickelt, mit denen man versucht diese fehlenden Werte in den Griff zu bekommen. Auch in SPSS gibt es dafür bereits ein eigenes Tool mit dem Namen Missing Value Analysis, welches aber in der Standardversion nicht enthalten ist. Da diese Verfahren zudem für ErstbenutzerInnen ungeeignet sind, sei an dieser Stelle auf die weiterführende Literatur verwiesen.

5.5.2 Umkodieren von Variablen

Bei stetigen Variablen bzw. Variablen mit sehr vielen verschiedenen Ausprägungen kann eine Häufigkeitstabelle sehr schnell unübersichtlich und uninformativ werden. Man kann die Anzahl der Ausprägungen reduzieren, in dem man geeignete Ausprägungen zusammenfasst. Bei stetigen Variablen bedeutet dies, dass man statt der einzelnen Ausprägungen Intervalle festlegt und die Häufigkeiten in den einzelnen Intervallen betrachtet (vgl. Kapitel 6).

Als Beispiel dient uns die Variable *age* der Datei *demo.sav*, diese Datei wurde bei der Installation von SPSS mitinstalliert und befindet sich im Ordner *SPSS\tutorial\sample_files*. Wie viele Intervalle verwendet werden und wie die Intervallbreiten festgelegt werden sollen, entscheiden die BenutzerInnen. Es gibt hier keine einheitliche Regeln, die man anwenden könnte, dazu benötigt man etwas Erfahrung. Für unerfahrene BenutzerInnen sind in einem ersten Schritt Intervalle gleicher Breite empfehlenswert, die Anzahl sollte etwa zwischen 5 und 10 liegen, so dass zweckmäßige Intervallgrenzen entstehen. In diesem Beispiel sollen die Intervallgrenzen folgendermaßen festgelegt werden: bis 20 Jahre, mehr als 20 bis 40 Jahre, mehr als 40 bis 60 Jahre und älter als 60 Jahre. Dazu muss die Variable Alter umkodiert werden, man wählt daher aus dem Menü den Bereich *Transformieren → Umkodieren → in andere Variablen*. Es ist empfehlenswert in eine andere Variable zu transformieren, weil so die Originalvariable erhalten bleibt. Durch das Zusammenfassen in Intervalle entsteht ein gewisser Informationsverlust, der beispielsweise für die Berechnung von Mittelwerten oder anderen Maßzahlen vermieden werden sollte. Für weitere Analysen ist also die ungruppierte Originalvariable meist besser geeignet als die intervallskalierte Hilfsvariable.

Abb. 5.10. Dialogfenster Umkodieren

In dem Dialogfenster Umkodieren (vgl. Abbildung 5.10) wird zuerst aus der Variablenliste die Variable ausgewählt, die umkodiert werden soll. Dann wird für die neu anzulegende Ausgabevariable ein Name (beispielsweise *nalter*) und eine Beschriftung (= Variablenlabel, beispielsweise *Alter in Intervallskalierung*) festgelegt und mit der Schaltfläche *Ändern* bestätigt.
In der Option *Falls* kann man festlegen, ob diese Umkodierung für alle Fälle durchgeführt werden soll oder nur unter gewissen Bedingungen, die ebenfalls

hier definiert werden können. In diesem Beispiel werden wir sinnvollerweise die Transformation für alle Fälle durchführen, also verwenden wir die Standardeinstellung dieser Option. Die Schaltfläche *Alte und neue Werte* erlaubt nun die Festlegung der Umkodierungen (vgl. Abbildung 5.11).

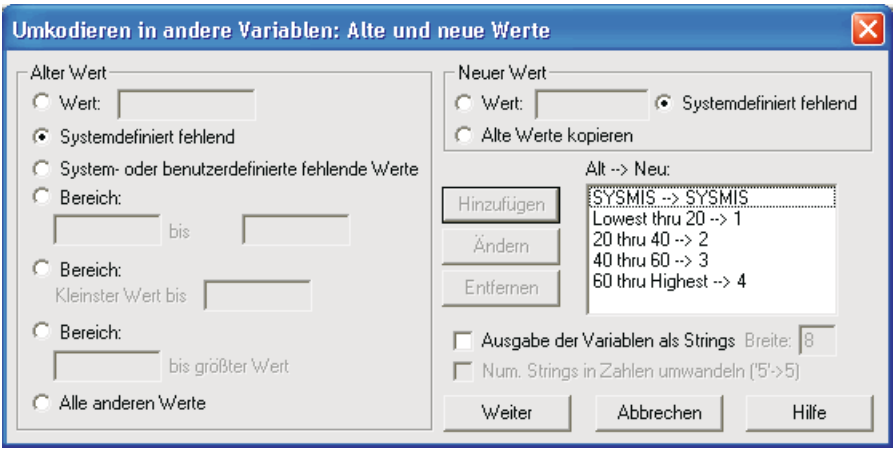


Abb. 5.11. Umkodieren: Alte und Neue Werte

Im linken Bereich können die alten Werte eingegeben werden, im rechten oberen Bereich werden die neuen Werte festgelegt und durch die Schaltfläche *Hinzufügen* werden die Umkodierungen als Liste im Fenster sichtbar. Wir beginnen mit dem Intervall „bis 20", dazu wählt man die Option *Bereich: Kleinster Wert bis* ___ und fügt 20 ein, der neue Wert wird mit 1 belegt als Kodierung für das erste Intervall. Für die folgenden Intervalle wählt man jeweils die Option *Bereich:* ___ *bis* ___ und fügt die jeweilige Unter- und Obergrenzen ein und vergibt als neue Werte die laufenden Intervallnummern. Für das letzte Intervall wählt man die Option *Bereich:* ___ *bis größter Wert* und verwendet 60 als untere Intervallgrenze. Der Vollständigkeit halber werden auch systemdefiniert fehlende Werte als systemdefiniert fehlend definiert (vgl. Kapitel 5.2).

Durch Bestätigung mit *Weiter* und *OK* wird nun die Transformation durchgeführt. Die neue Variable *nalter* wird als letzte Spalte dem Datensatz hinzugefügt und auch in der Variablenansicht ist die neue Variable in der letzten Zeile sichtbar. In der Variablenansicht lassen sich nun auch Name und Label verändern, falls dies erwünscht ist. Auf jeden Fall sollte man aber Wertelabels anlegen, weil die Kodierung 1 bis 4 für eine Variable Alter nicht aussagekräftig ist.

Umkodieren von Variablen

- *Transformieren* → *Umkodieren* → *in andere Variablen*
- Variable auswählen
- Name und Beschriftung der neuen Variable festlegen, mit *Ändern* bestätigen
- Alte und neue Werte festlegen
- Falls notwendig unter der Option *Falls* Einschränkung der Umkodierung auf bestimmte Fälle vornehmen
- Nach Umkodierung Wertelabels für Kodierungen festlegen

5.5.3 Transformieren von Variablen

Liegen metrische Variablen in unterschiedlichen Maßeinheiten (z.B. Längenangaben einmal in Meter und einmal in Zentimeter) vor, so kann es hilfreich sein, die Variablen vor der Analyse in einheitliche Maßeinheiten zu transformieren.

Als Beispiel dient uns die Variable *Sales* der Datei *car_sales.sav* aus dem Ordner *SPSS\tutorial\sample_files*. Diese Variable beinhaltet für jede Erhebungseinheit die Verkaufszahlen in Tausend. Zu Übungszwecken soll eine neue Variable Verkauf angelegt werden, welche die absoluten Verkaufszahlen enthält. Man wählt aus dem Menü den Bereich *Transformieren* → *Berechnen* und erhält das Dialogfenster aus Abbildung 5.12.

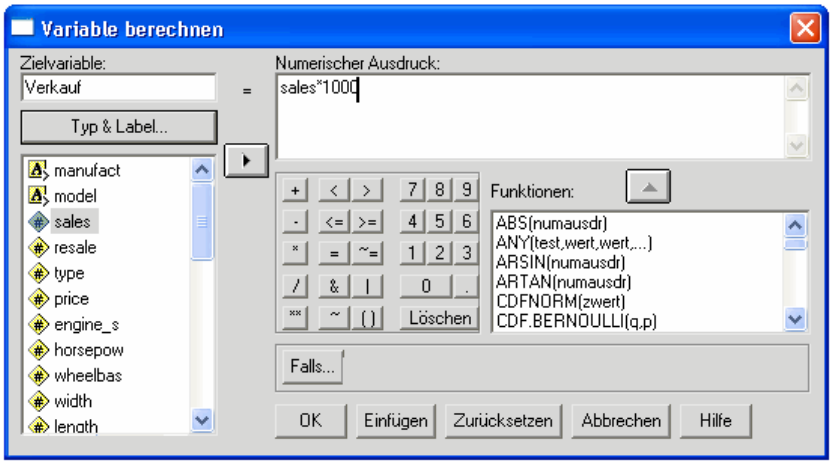


Abb. 5.12. Transformieren von Variablen

Man legt den Namen der Zielvariable fest, bei Bedarf auch deren Typ und Label und kann die Berechnungsvorschrift im rechten Fensterteil definieren. Neben den Grundrechnungsarten stehen für diese Berechnungen auch zahlreiche andere mathematische Funktionen zur Verfügung.

Bestätigen der Eingabe fügt die neue Variable am Ende des Datensatzes ein.

Transformieren von Variablen

- *Transformieren → Berechnen*
- Zielvariable festlegen (Name, Typ, Label)
- Berechnungsvorschrift eingeben
- Falls notwendig Einschränkung der Berechnung auf bestimmte Fälle vornehmen unter der Option *Falls*
- Nach Transformation Wertelabels festlegen, falls notwendig

5.5.4 Fälle gewichten

In der Praxis kann es passieren, dass Daten nicht als Datensatz zur Verfügung stehen, sondern dass bereits vorhandene Tabellen zu analysieren sind. Üblicherweise geht SPSS davon aus, dass eine Zeile einem Datensatz, also einer Erhebungseinheit entspricht. Man kann dies aber auch ändern. Im einfachsten Fall wollen wir eine Tabelle verarbeiten, die Informationen über ein einziges Merkmal enthält.

Beispiel 5.1. Geschlecht von LVA-TeilnehmerInnen
An einer Lehrveranstaltung nehmen 400 Personen teil, davon sind 180 Personen weiblich.

Gemäß der bisherigen Beschreibung müssten wir einen Datensatz anlegen mit 400 Zeilen entsprechend den Erhebungseinheiten und einer Spalte für das Merkmal Geschlecht mit den Ausprägungen 1 = männlich und 2 = weiblich. Diese Vorgehensweise wäre zwar korrekt, aber sehr zeitaufwändig.

Wesentlich effizienter ist die Dateneingabe in Tabellenform. Dazu geben wir die Daten in folgender modifizierter Form ein: Die erste Spalte beinhaltet das Merkmal und die zweite Spalte die zugehörigen Häufigkeiten, die Zeilen werden durch die verschiedenen Ausprägungen festgelegt (vgl. Abbildung 5.13).

Nun muss auch dem System mitgeteilt werden, dass hier kein Datensatz vorliegt, sondern bereits eine fertige Tabelle, dazu wählt man den Menüpunkt *Daten → Fälle gewichten* (vgl. Abbildung 5.14).

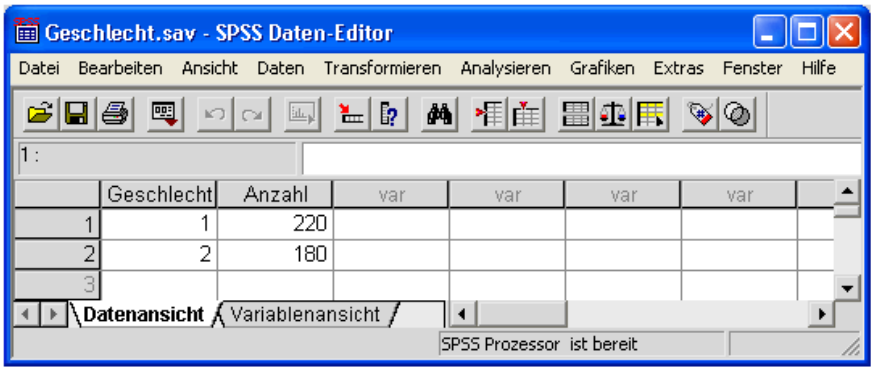


Abb. 5.13. Tabellen eingeben

Abb. 5.14. SPSS Fälle gewichten

Nun wählt man die Option *Fälle gewichten mit* und als Häufigkeitsvariable wählt man die Variable *Anzahl*. Durch *OK* wird diese Auswahl übernommen und für alle weiteren Berechnungen verwendet. Diese Einstellung bleibt auch beim Abspeichern der Datei erhalten, wenn man also das nächste Mal diese Datei aufruft, ist die Gewichtung bereits berücksichtigt.

Will man mehrdimensionale Tabellen verarbeiten ist für jede mögliche Ausprägungskombination eine Zeile anzulegen.

Beispiel 5.2. Einfluss von Strategietraining
In einer Studie wird der Einfluss von Strategietraining bei N = 235 zufällig ausgewählten Führungskräften auf den Unternehmenserfolg untersucht. Das Ergebnis der Untersuchung kann aus nachstehender Tabelle entnommen werden.

	kein Erfolg	Erfolg	Summe
kein Training	40	75	115
mit Training	30	90	120
Summe	70	165	235

Wir kodieren beispielsweise *kein Training* mit 1 und *Training* mit 2, die Variable *Erfolg* wird analog kodiert. Dem entsprechend muss die Datendatei aussehen wie in Abbildung 5.15. Nun gewichtet man wieder unter dem Menüpunkt *Daten* → *Fälle gewichten* und kann danach beliebige Analysen durchführen.

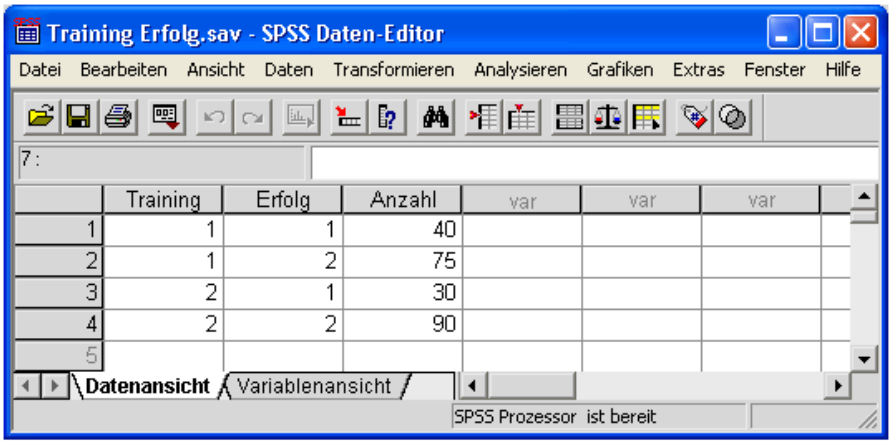


Abb. 5.15. Eingabe mehrdimensionaler Tabellen in SPSS

Fälle gewichten, Tabelleneingabe in SPSS

- Daten eingeben, für jede Ausprägung bzw. Ausprägungskombination ist eine Zeile anzulegen. Neben den Spalten für die Merkmale eine zusätzliche Spalte mit den Häufigkeiten versehen.
- Namen, Variablenlabels und Wertelabels vergeben
- *Daten* → *Fälle gewichten* als Gewichtungsvariable die Häufigkeiten auswählen
- Gewichtung bleibt beim Speichern der Datei erhalten

5.6 Tipps im Umgang mit SPSS

Für die Erstellung eines Ergebnisberichtes sollten die Ergebnisse in ansprechender Form aufbereitet sein. Für das Layout empfiehlt es sich die Ergebnisse aus SPSS zu exportieren, beispielsweise als EXCEL-File oder als HTML-Dokument, falls die Ergebnisse für das Internet aufbereitet werden sollen. Der Export erfolgt unter dem Menüpunkt *Datei* → *Exportieren*, der ein Dialogfenster öffnet, in dem die Einzelheiten für den Export festgelegt werden können. Ein Export in eine EXCEL-Datei erscheint vor allem dann sinnvoll, wenn auch Grafiken erstellt werden sollen, weil das Layoutieren von Grafiken in EXCEL benutzerfreundlicher ist als in SPSS.

Unter dem Menüpunkt *Bearbeiten* → *Optionen* erscheint ein Dialogfenster mit sehr vielen Möglichkeiten zur Anpassung von Bearbeitungseinstellungen. Beispielsweise lässt sich im Registerblatt *Allgemein* festlegen, ob bei der Anzeige von Variablenlisten die Namen der Variablen oder die Labels verwendet werden sollen. Eine vollständige Beschreibung aller Optionen würde hier zu weit führen, die meisten Optionen sind selbsterklärend, ansonsten muss auf spezielle SPSS-Literatur zurückgegriffen werden.

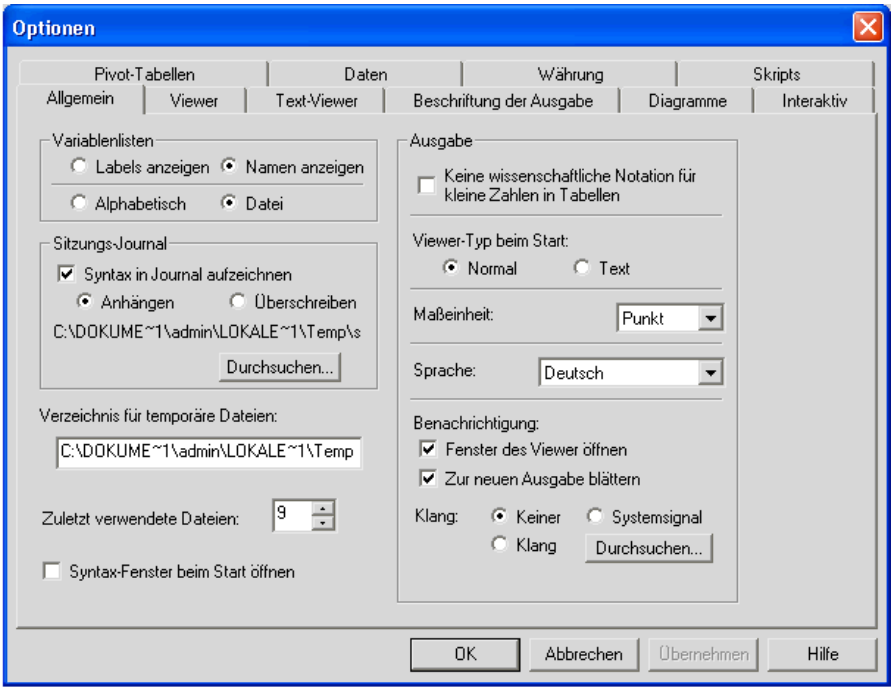


Abb. 5.16. Optionen für die Bearbeitung in SPSS

Teil II

Deskriptive Statistik

6

Eindimensionale Häufigkeitsverteilungen

Häufigkeitsverteilungen geben einen guten Überblick über wesentliche Informationen aus den Urdaten. Dazu werden die Daten nach Ausprägungen zusammengefasst und in Tabellen oder Grafiken dargestellt. Je nach Skalenniveau und Anzahl der Ausprägungen gibt es unterschiedliche Darstellungsformen, die in diesem Kapitel beleuchtet werden.

6.1 Diskrete Merkmale

Ausgangspunkt der Betrachtung sei die Erhebung eines diskreten Merkmales X (z.B. die Anzahl kariöser Zähne bei Kindern) mit verschiedenen Ausprägungen. Mit x_i bezeichne man die i-te Ausprägung, beispielsweise $x_i = 1$ für einen kariösen Zahn.

Beispiel 6.1. Kariöse Zähne von Kindern
Bei $N = 140$ Kindern einer Jahrgangsstufe werden die Zähne auf Karies untersucht. Bei jedem Kind wird das Merkmal Anzahl der kariösen Zähne erhoben, wobei sich folgende Ergebnisse einstellen:

4	0	0	3	1	5	1	2	2	0	5	0	5	2	1	0	1	0	0	4
0	1	1	3	0	1	1	1	3	1	0	1	4	2	0	3	1	1	7	2
0	2	1	3	0	0	0	0	6	1	1	2	1	0	1	0	3	0	1	3
0	5	2	3	0	2	4	0	1	1	3	0	6	2	1	5	1	1	2	2
0	3	0	1	0	1	0	0	0	5	0	4	1	2	2	7	1	3	1	5
1	0	1	0	0	4	0	3	1	1	7	2	1	0	3	0	1	3	2	2
2	7	1	3	1	5	1	0	0	0	2	1	0	3	1	4	0	2	1	1

Diese Urliste ist unübersichtlich, daher werden die Daten in einem ersten Schritt in übersichtlicher Form dargestellt. Dazu wird eine Tabelle erstellt, in der für jede Merkmalsausprägung angeführt ist, wie oft diese in der Untersuchungsgesamtheit aufscheint. Diese Anzahl wird als absolute Häufigkeit bezeichnet. Summiert man die absoluten Häufigkeiten aller möglichen Merkmalsausprägungen, so erhält man wiederum den Umfang der Untersuchungsgesamtheit. Die relativen Häufigkeiten geben die Anteile der jeweiligen Merkmalsausprägung an der Untersuchungsgesamtheit in Dezimalschreibweise oder als Prozentwerte an.

Bezeichnungen

N	Untersuchungsumfang
r	Anzahl an verschiedenen Ausprägungen
x_i	Ausprägung, $i = 1, \ldots, r$
h_i	absolute Häufigkeit der Ausprägung x_i
$p_i = h_i/N$	relative Häufigkeit der Ausprägung x_i
$P_i = 100 \cdot p_i$	relative Häufigkeit der Ausprägung x_i in Prozent

Damit kann die Urliste in Tabellenform zusammengefasst werden.

Tabelle 6.1. Häufigkeitsverteilung zu Beispiel 6.1

	Ausprägung	absolute Häufigkeiten	relative Häufigkeiten	relative Häufigkeiten in Prozent
i	x_i	h_i	p_i	P_i
1	0	41	0,29	29%
2	1	42	0,30	30%
3	2	20	0,14	14%
4	3	16	0,11	11%
5	4	7	0,05	5%
6	5	8	0,06	6%
7	6	2	0,01	1%
8	7	4	0,03	3%
	Summe	140	≈ 1	$\approx 100\%$

Diese Häufigkeitsverteilung kann nun folgendermaßen interpretiert werden (Zeile i = 3): 20 Schulkinder haben je 2 kariöse Zähne, anders ausgedrückt haben 14% der untersuchten Kinder je zwei kariöse Zähne.

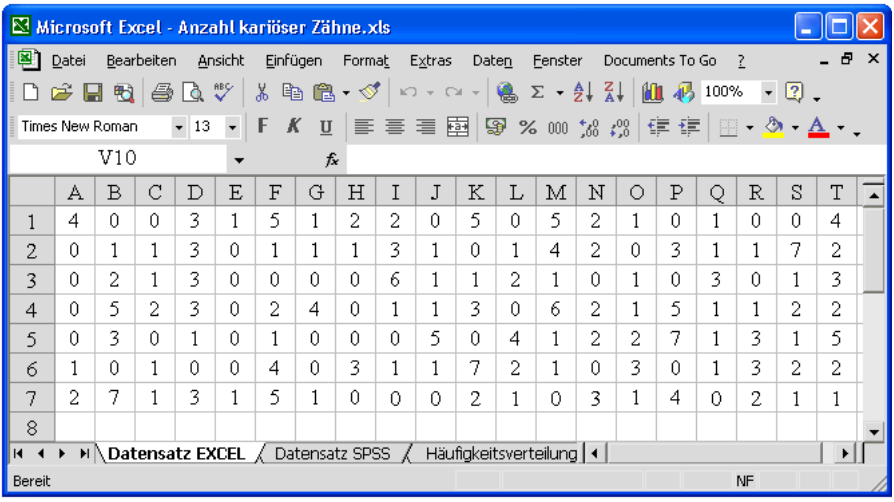
Summiert man die relativen Häufigkeiten über alle Ausprägungen, so erhält man als Summe 1, die Summe der absoluten Häufigkeiten ergibt immer N. Die relativen Häufigkeiten wurden auf zwei Nachkommastellen gerundet, daher ergibt die Summe nicht exakt, sondern nur annähernd 1 bzw. 100%.

Summen von Häufigkeiten

$$\sum_{i=1}^{r} h_i = N \qquad \sum_{i=1}^{r} p_i = 1 \qquad \sum_{i=1}^{r} P_i = 100\%$$

6.1.1 Häufigkeitsverteilung in EXCEL

Im Beispiel 6.1 wurde lediglich ein Merkmal erhoben. Den Ausführungen in Kapitel 4 folgend müsste man für jedes Merkmal eine Spalte reservieren und in jede Zeile einen Datensatz eintragen. Die Datenmatrix in unserem Beispiel würde aus einer Spalte und 140 Zeilen bestehen. In diesem speziellen Fall, in dem nur ein Merkmal bearbeitet wird, kann man in EXCEL die Daten aber auch als Datenfeld beliebiger Größe belassen, beispielsweise bestehend aus 20 Spalten und 7 Zeilen.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	4	0	0	3	1	5	1	2	2	0	5	0	5	2	1	0	1	0	0	4
2	0	1	1	3	0	1	1	1	3	1	0	1	4	2	0	3	1	1	7	2
3	0	2	1	3	0	0	0	0	6	1	1	2	1	0	1	0	3	0	1	3
4	0	5	2	3	0	2	4	0	1	1	3	0	6	2	1	5	1	1	2	2
5	0	3	0	1	0	1	0	0	0	5	0	4	1	2	2	7	1	3	1	5
6	1	0	1	0	0	4	0	3	1	1	7	2	1	0	3	0	1	3	2	2
7	2	7	1	3	1	5	1	0	0	0	2	1	0	3	1	4	0	2	1	1
8																				

Abb. 6.1. EXCEL: Datensatz zu Beispiel 6.1

Zur Erhebung der Häufigkeiten mit EXCEL sollte zuerst ein Raster für die Ergebnisse angelegt werden (vgl. Abbildung 6.2).

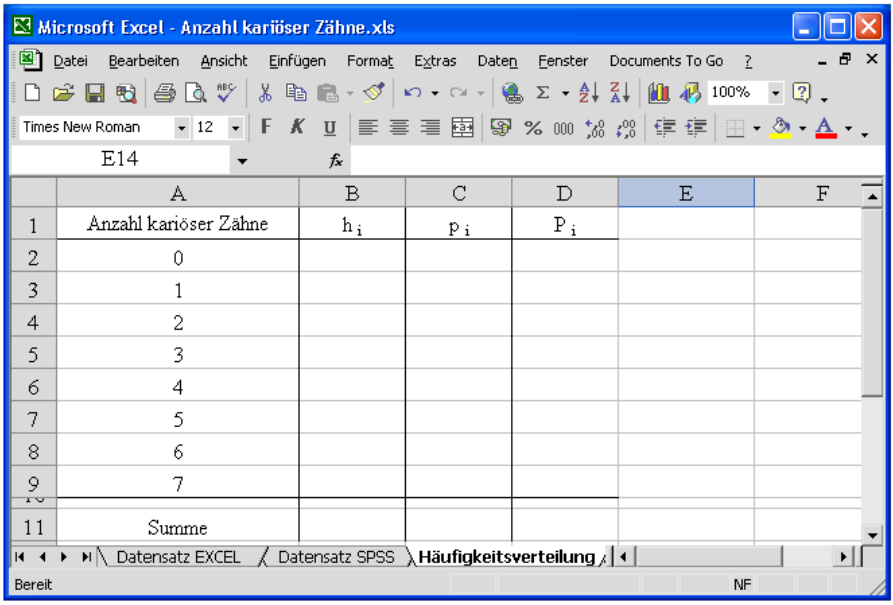


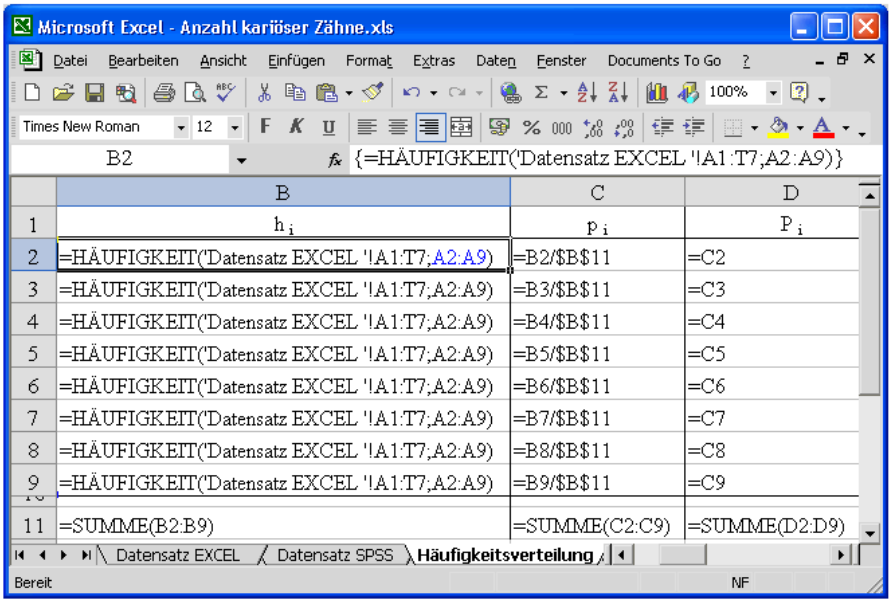
Abb. 6.2. EXCEL: Vorbereitung für die Funktion Häufigkeiten

EXCEL bietet für das Erstellen einer Häufigkeitsverteilung die Funktion *Häufigkeit* an. Für die Verwendung dieser Funktion sind folgende Einzelschritte notwendig.

Ermitteln von Häufigkeiten in EXCEL

- Ergebnistabelle vorbereiten (Abbildung 6.2)
- Markieren des Ergebnisbereiches von oben nach unten (hier *B2:B9*)
- Aufrufen der Funktion *Häufigkeit*
- Bezug für Daten eingeben bzw. markieren (hier *Datensatz EXCEL!A1:T7*, vgl. Abbildung 6.1)
- Bezug für Klassen (=Ausprägungen) eingeben oder markieren (hier *Häufigkeitsverteilung!A2:A9*)
- Bestätigen mit *OK*
- In der ersten Zeile des Ergebnisbereiches (*B2*) steht nun die Häufigkeit für die erste Ausprägung. Nun muss der Cursor an die letzte Stelle in der Bearbeitungsleiste positioniert werden.
- Durch die Tastenkombination *Strg + Umschalt + Enter* erhält man alle gewünschten Häufigkeiten.

Für die Summenbildung stellt EXCEL die Funktion *Summe* zur Verfügung, hier muss nur der Bezug angepasst werden. Die relativen Häufigkeiten werden mit Hilfe der Formel $p_i = h_i/N$ ermittelt. Die relativen Häufigkeiten in Prozent werden auf die gleiche Weise errechnet, nur die Formatierung der Zelle wird auf Prozent geändert. Die verwendeten Formeln können aus Abbildung 6.3 entnommen werden.

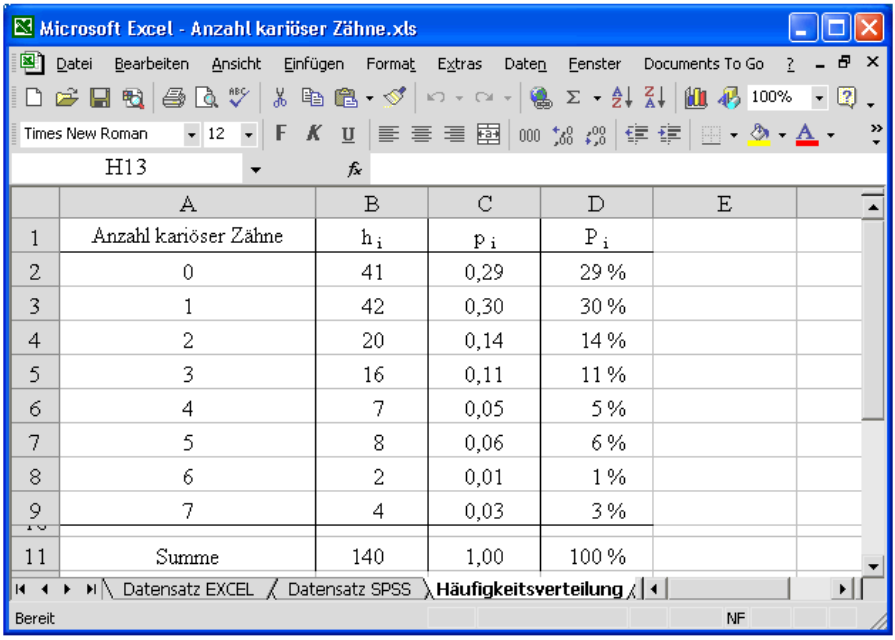


	B	C	D
1	h_i	p_i	P_i
2	=HÄUFIGKEIT('Datensatz EXCEL '!A1:T7;A2:A9)	=B2/$B$11	=C2
3	=HÄUFIGKEIT('Datensatz EXCEL '!A1:T7;A2:A9)	=B3/$B$11	=C3
4	=HÄUFIGKEIT('Datensatz EXCEL '!A1:T7;A2:A9)	=B4/$B$11	=C4
5	=HÄUFIGKEIT('Datensatz EXCEL '!A1:T7;A2:A9)	=B5/$B$11	=C5
6	=HÄUFIGKEIT('Datensatz EXCEL '!A1:T7;A2:A9)	=B6/$B$11	=C6
7	=HÄUFIGKEIT('Datensatz EXCEL '!A1:T7;A2:A9)	=B7/$B$11	=C7
8	=HÄUFIGKEIT('Datensatz EXCEL '!A1:T7;A2:A9)	=B8/$B$11	=C8
9	=HÄUFIGKEIT('Datensatz EXCEL '!A1:T7;A2:A9)	=B9/$B$11	=C9
11	=SUMME(B2:B9)	=SUMME(C2:C9)	=SUMME(D2:D9)

Abb. 6.3. EXCEL: Formeln für die Erstellung der Häufigkeitsverteilung

Die Funktion *Häufigkeit* gibt als Ergebnis eine einspaltige Matrix zurück, daher muss die Formel als Matrixformel eingegeben werden, was durch die vorher beschriebene Vorgehensweise automatisch erreicht wird. Das Ergebnis lässt sich aus Abbildung 6.4 entnehmen.

Matrixformeln sind in der Bearbeitungsleiste an einer Einbettung in geschwungene Klammern erkennbar. Änderungen in Matrixformeln sind eher schwierig, besser ist es, die gesamte Ergebnismatrix zu entfernen und die geänderte Formel neu einzugeben.



	A	B	C	D	E
1	Anzahl kariöser Zähne	h_i	p_i	P_i	
2	0	41	0,29	29 %	
3	1	42	0,30	30 %	
4	2	20	0,14	14 %	
5	3	16	0,11	11 %	
6	4	7	0,05	5 %	
7	5	8	0,06	6 %	
8	6	2	0,01	1 %	
9	7	4	0,03	3 %	
11	Summe	140	1,00	100 %	

Abb. 6.4. EXCEL: Häufigkeitsverteilung Endergebnis

6.1.2 Häufigkeitsverteilungen in SPSS

Zur Ermittlung der Häufigkeitsverteilung mit Hilfe von SPSS benötigt man den Datenfile in der in Kapitel 5.2 beschriebenen Form. Unter dem Menüpunkt *Analysieren* → *Deskriptive Statistiken* → *Häufigkeiten* erscheint das Dialogfeld für Häufigkeiten (Abbildung 6.5).

Im linken Bereich des Fensters werden alle Variablen angezeigt, die im Datenfile vorhanden sind. Die Variablen sind standardmäßig alphabetisch geordnet und es sind lediglich die kurzen Variablennamen, nicht jedoch die längeren Variablenlabels sichtbar. Unter dem Menüpunkt *Bearbeiten* → *Optionen* kann man diese Einstellungen verändern. In unserem Beispiel ist in der Datei lediglich eine Variable vorhanden.

Durch Anklicken mit der Maus wird eine Variable markiert. Betätigt man dann mit der Maus den Button mit dem Pfeil, wird diese Variable ausgewählt, und erscheint nun im rechten Fenster. Ein Doppelklick auf die Variable im linken Fenster hat den gleichen Effekt. Durch Markieren im rechten Fenster und neuerliches Anklicken des Pfeil-Buttons (die Richtung des Pfeils auf dem Button hat sich übrigens jetzt umgedreht) kann eine Variable jederzeit wieder

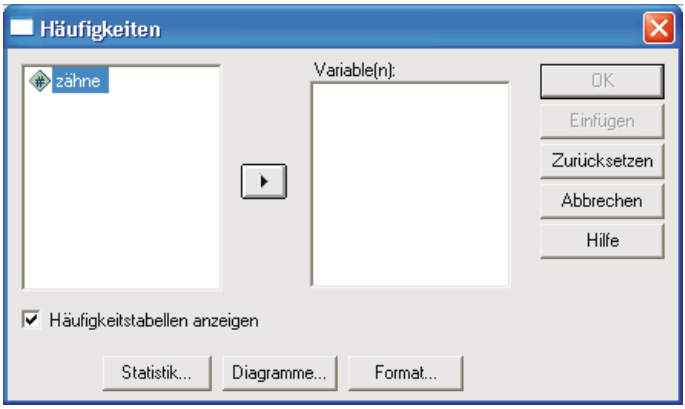


Abb. 6.5. SPSS: Häufigkeiten Eingabe

abgewählt werden. Auch hier kann als Alternative ein Doppelklick auf die Variable verwendet werden.

Das Kontrollkästchen links unten für die Anzeige von Häufigkeitstabellen sollte mit einem Häkchen versehen sein, durch Anklicken kann diese Einstellung verändert werden (Abbildung 6.5).

Im unteren Bereich des Fensters sind drei Schaltflächen zu sehen, die weitere Einstellungen zulassen. Wir belassen es vorerst bei den Standardeinstellungen und beschäftigen uns erst später mit diesen Schaltflächen. Ein Bestätigen der Eingabe öffnet im Viewer das Ergebnisfile (Abbildung 6.6).

Folgende Informationen können abgelesen werden: Insgesamt wurden 140 Kinder untersucht, von allen liegt ein gültiges Ergebnis vor, es gibt keine fehlenden Werte. Die Spalte „Prozent" gibt die relative Häufigkeit in Prozent an (bezogen auf die Gesamtheit inklusive derjenigen, von denen keine gültige Information vorliegt), die Spalte gültige Prozent bezieht den Anteil auf die Erhebungseinheiten, von denen eine gültige Ausprägung erhoben wurde. In unserem Fall sind beide gleich, weil von allen untersuchten Kindern ein gültiges Ergebnis vorliegt. Betrachtet man nun für die Zeile der Ausprägung 2 die kumulierten Prozente, so kann man ablesen, dass 73,6% der untersuchten Kinder höchstens zwei kariöse Zähne haben.

Häufigkeitstabellen erstellen mit SPSS

- *Analysieren* → *Deskriptive Statistiken* → *Häufigkeiten*
- Variablen auswählen
- Häkchen im Kontrollkästchen für die Anzeige von Häufigkeitstabellen

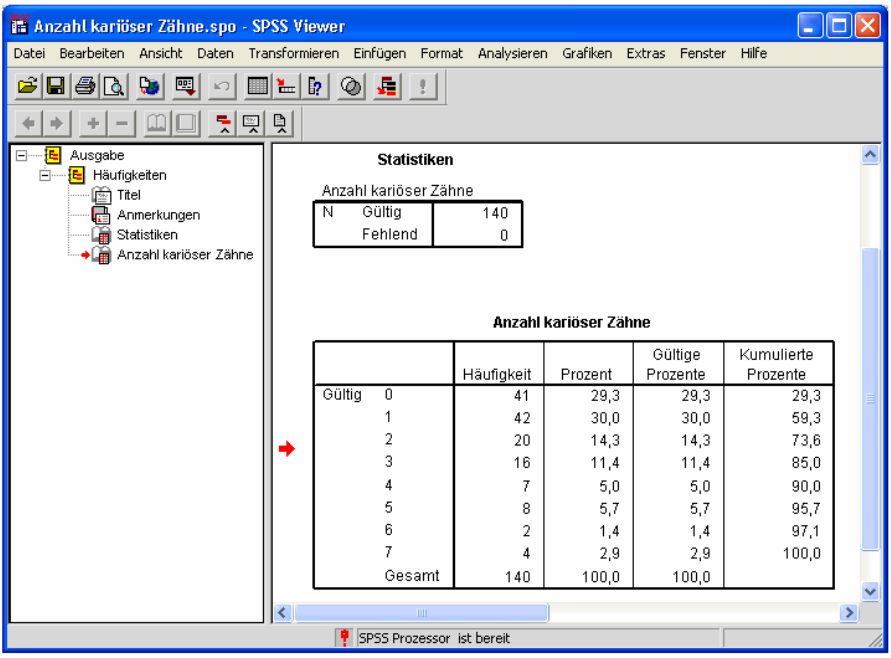


Abb. 6.6. SPSS: Häufigkeiten Ergebnis

6.2 Stetige Merkmale

Bei stetigen Merkmalen ist es für die Erstellung einer Häufigkeitstabelle zielführend, den gesamten Wertebereich in Intervalle zu gliedern und die absoluten, relativen Häufigkeiten und relativen Häufigkeiten in Prozent für diese Intervalle zu bestimmen.

Bezeichnungen

N	Untersuchungsumfang
r	Anzahl der Intervalle
e_{i-1}	Untergrenze des i-ten Intervalls, $i = 1, \ldots, r$
e_i	Obergrenze des i-ten Intervalls
h_i	$= h(e_{i-1} < x \leq e_i)$, absolute Häufigkeit des Intervalls $I_i =]e_{i-1}, e_i]$
$p_i = h_i/N$	relative Häufigkeit des Intervalls i
$P_i = 100 \cdot p_i$	relative Häufigkeit des Intervalls i in Prozent

Beispiel 6.2. Körpergröße von Studierenden
Bei $N = 65$ Studierenden einer Jahrgangsstufe wurde die Körpergröße in cm erhoben. Diese Erhebung führte zu folgendem Ergebnis:

178	182	166	162	181	168	170	164	171	170	165	164	176
169	165	165	175	180	164	188	170	194	171	185	168	164
180	174	183	193	172	178	165	162	174	174	162	163	170
172	176	168	160	170	170	171	166	165	160	175	183	182
189	162	168	160	178	175	168	158	172	163	172	183	164

Um für die Körpergröße eine sinnvolle Häufigkeitstabelle erstellen zu können, sind die Ausprägungen in Intervalle zu gruppieren. Die Intervalle müssen so angelegt werden, dass jede Ausprägung in genau einem Intervall liegt. Vor der Festlegung der Intervalle muss die gewünschte Anzahl an Intervallen fixiert werden. Dabei ist zu bedenken, dass durch die Gruppierung einerseits Information verloren geht, andererseits aber Übersichtlichkeit gewonnen wird. Man sollte daher die Anzahl groß genug wählen, damit die wesentlichen Informationen erhalten bleiben, und klein genug, um Übersicht zu gewinnen. Wir wollen in einem ersten Schritt unsere Daten in etwa 10 Intervalle gruppieren. Zuerst werden der kleinste Wert (158) und der größte Wert (194) aus dem Datensatz gesucht. Die abzudeckende Breite ergibt sich aus der Differenz (36). Nachdem wir etwa 10 Intervalle anstreben, würde dies einer Intervallbreite von 3,6 entsprechen. Um einigermaßen zweckmäßige Intervallgrenzen zu erhalten, begnügen wir uns mit 8 Intervallen und wählen als Intervallbreite 5. Als Untergrenze des ersten Intervalls wählen wir 155, die restlichen Intervallgrenzen ergeben sich automatisch.

Damit kann die Urliste in Tabellenform dargestellt werden (Tabelle 6.2).

Tabelle 6.2. Häufigkeitsverteilung Körpergröße

Intervall	Größe	Häufigkeiten		
i	$e_{i-1} < x \leq e_i$	h_i	p_i	P_i
1	$155 < x \leq 160$	4	0,06	6%
2	$160 < x \leq 165$	16	0,25	25%
3	$165 < x \leq 170$	14	0,22	22%
4	$170 < x \leq 175$	13	0,20	20%
5	$175 < x \leq 180$	7	0,11	11%
6	$180 < x \leq 185$	7	0,11	11%
7	$185 < x \leq 190$	2	0,03	3%
8	$190 < x \leq 195$	2	0,03	3%
	Summe	65	≈ 1	$\approx 100\%$

Aus dieser Tabelle lassen sich folgende Informationen ablesen (i = 4): 13 Personen (das entspricht 20% der untersuchten Personen) sind größer als 170 cm aber höchstens 175 cm groß.

Intervallskalierung

- Anzahl der Intervalle festlegen
- Minimum, Maximum, Differenz zwischen Maximum und Minimum der Ausprägungen bestimmen
- Intervallbreiten berechnen (Differenz/Anzahl)
- Intervallgrenzen fixieren
- Intervalle müssen alle Ausprägungen abdecken und dürfen sich nicht überschneiden

6.2.1 Stetige Häufigkeitsverteilung in EXCEL

Die Erstellung der Häufigkeitstabelle mit EXCEL verläuft analog zur Beschreibung in Kapitel 6.1.1, als Klassen werden die Obergrenzen der Intervalle verwendet.

6.2.2 Stetige Häufigkeitsverteilung in SPSS

In SPSS muss in einem ersten Schritt die stetige Variable umkodiert werden. Mit der Anweisung *Variable kategorisieren* unter dem Menüpunkt *Transformieren* legt SPSS die Intervallgrenzen so fest, dass die Häufigkeiten in jedem Intervall in etwa gleich groß sind. Dies ist gleichbedeutend damit, dass die Intervalle sehr unterschiedliche Breite haben können. Zudem sind die genauen Intervallgrenzen, die SPSS verwendet, nicht ersichtlich. Es wird daher das manuelle Umkodieren (vgl. Kapitel 5.5.2) in eine andere Variable empfohlen. Für die Erstellung einer Häufigkeitstabelle sollte die in Intervalle umkodierte Variable verwendet werden.

Häufigkeitsverteilungen für stetige Merkmale
(bzw. für Merkmale mit vielen Ausprägungen)

- Stetige Merkmale in Intervalle gruppieren
- EXCEL: Als Klassen sind die Obergrenzen der Intervalle zu verwenden
- SPSS: Eine in Intervalle umkodierte Variable verwenden, die Originalvariable sollte beim Umkodieren erhalten bleiben.

6.3 Grafische Darstellung von Verteilungen

Es gibt viele verschiedene grafische Darstellungsformen für Häufigkeitsverteilungen, deren Verwendungsmöglichkeiten hier kurz beschrieben werden.

6.3.1 Kreis- oder Tortendiagramm

Bei einem Kreis- oder Tortendiagramm werden die Häufigkeiten als Kreissektoren dargestellt. Der **Zentriwinkel**, das ist der Winkel in der Mitte des Kreises, ist proportional zu den relativen Häufigkeiten. Zur Berechnung der Zentriwinkel werden die relativen Häufigkeiten mit 360 ° multipliziert. Geeignet ist diese Darstellungsform für nominale Merkmale mit wenig Ausprägungen.

Beispiel 6.3. Geschlecht von LVA-TeilnehmerInnen
An einer Lehrveranstaltung nehmen 25 Personen teil, davon sind 10 Personen weiblich. Das Merkmal Geschlecht soll mittels Kreisdiagramm dargestellt werden.

Geschlecht	Anzahl h_i	relative Häufigkeiten p_i	Zentriwinkel α_i
weiblich	10	0,4	144 °
männlich	15	0,6	216 °
Summe	25	1,0	360 °

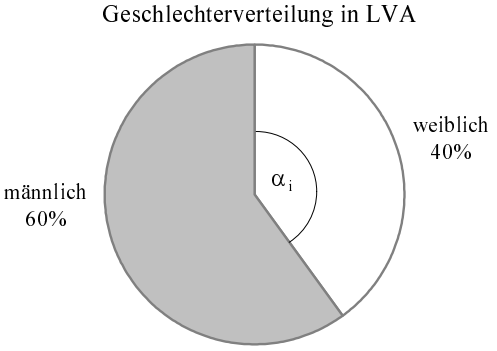


Abb. 6.7. Kreisdiagramm zu Beispiel 6.3

6.3.2 Balken-, Säulen- oder Stabdiagramm

Bei einem Säulen- oder Stabdiagramm werden auf der Abszisse (x-Achse) die Merkmalsausprägungen aufgetragen und auf der Ordinate (y-Achse) die jeweiligen Häufigkeiten. Dazu können die absoluten Häufigkeiten, die relativen Häufigkeiten oder die relativen Häufigkeiten in Prozent verwendet werden. Die Breite des Stabes hat keine Aussagekraft, entscheidend ist lediglich die Höhe. Geeignet ist diese Darstellungsform für ordinale Merkmale, für diskrete metrische Merkmale mit wenigen Ausprägungen oder für nominale Merkmale mit vielen Ausprägungen.

Die Begriffe Säulendiagramm und Stabdiagramm werden synonym verwendet. Als Balkendiagramm bezeichnet man ein um 90 ° gedrehtes Säulendiagramm, in dem die sonst vertikalen Säulen als horizontale Balken zu sehen sind.

Beispiel 6.4. Kariöse Zähne von Schulkindern
(Fortsetzung von Beispiel 6.1, Seite 61)
Das Merkmal Anzahl kariöser Zähne soll in einem Stabdiagramm und in einem Balkendiagramm dargestellt werden.

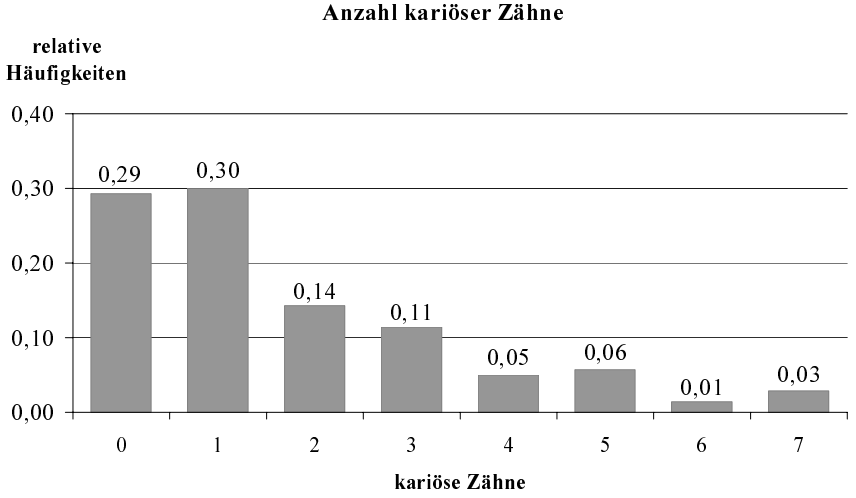


Abb. 6.8. Stabdiagramm zu Beispiel 6.1

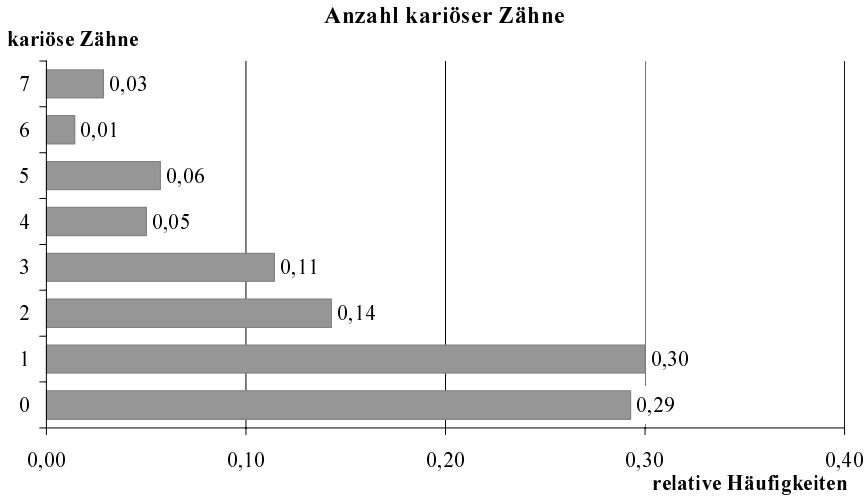


Abb. 6.9. Balkendiagramm zu Beispiel 6.1

6.3.3 Histogramm

Ein Histogramm ist für metrische stetige Merkmale geeignet, deren Ausprägungen in Intervalle zusammengefasst wurden. Bei einem Histogramm werden auf der x-Achse die Ausprägungen und auf der y-Achse die Dichten aufgetragen, wobei die Dichte der Quotient aus Häufigkeit und Intervallbreite ist. In dieser Darstellungsform sind die relativen Häufigkeiten als Flächen sichtbar.

Beispiel 6.5. Körpergröße von Studierenden
(Fortsetzung von Beispiel 6.2, Seite 69)
Zu den gemessenen Körpergrößen ist ein Histogramm anzufertigen.

Tabelle 6.3. Dichte zu Beispiel 6.2

Intervall i	Größe $e_{i-1} < x \leq e_i$	rel. Häufigkeit p_i	Intervallbreite d_i	Dichte f_i
1	$155 < x \leq 160$	0,062	5	0,012
2	$160 < x \leq 165$	0,246	5	0,049
3	$165 < x \leq 170$	0,215	5	0,043
4	$170 < x \leq 175$	0,200	5	0,040
5	$175 < x \leq 180$	0,108	5	0,022
6	$180 < x \leq 185$	0,108	5	0,022
7	$185 < x \leq 190$	0,031	5	0,006
8	$190 < x \leq 195$	0,031	5	0,006

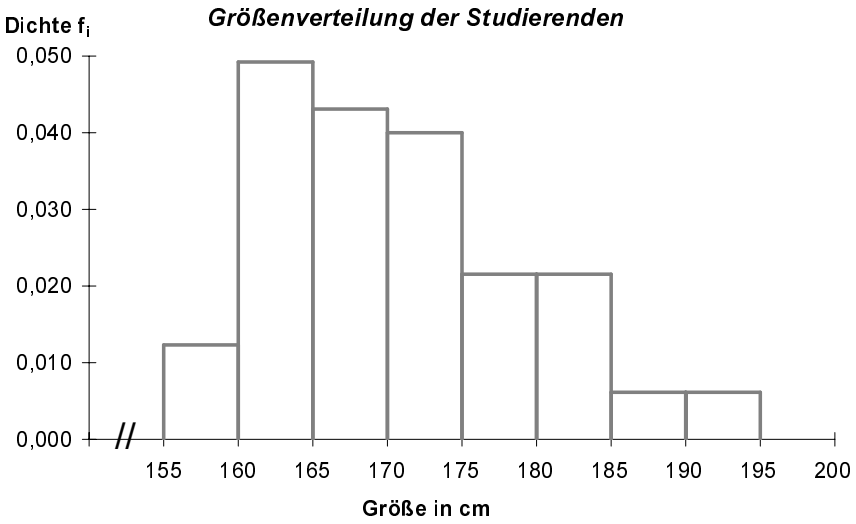


Abb. 6.10. Histogramm zu Beispiel 6.2

Ausschließlich bei gleich breiten Intervallen ist es zulässig, statt der Dichten die Häufigkeiten aufzutragen. Dies liegt daran, dass man bei der Betrachtung eines Histogramms automatisch die Verhältnisse der Flächen wahrnimmt und nicht die Relationen der Rechteckshöhen. Eine kleine Veränderung im Beispiel 6.2 soll die Notwendigkeit der Verwendung von Dichten bei unterschiedlichen Intervallbreiten verdeutlichen.

Beispiel 6.6. Körpergröße von Studierenden
(Fortsetzung von Beispiel 6.2, Seite 69)
Die gemessenen Körpergrößen wurden in unterschiedlich breite Intervalle aufgeteilt. Dafür ist ein Histogramm anzufertigen.

Tabelle 6.4. Dichte bei unterschiedlichen Intervallbreiten

Intervall i	Größe $e_{i-1} < x \leq e_i$	rel. Häufigkeit p_i	Intervallbreite d_i	Dichte f_i
1	$155 < x \leq 160$	0,062	5	0,012
2	$160 < x \leq 165$	0,246	5	0,049
3	$165 < x \leq 170$	0,215	5	0,043
4	$170 < x \leq 175$	0,200	5	0,040
5	$175 < x \leq 185$	0,215	10	0,022
6	$185 < x \leq 195$	0,062	10	0,006

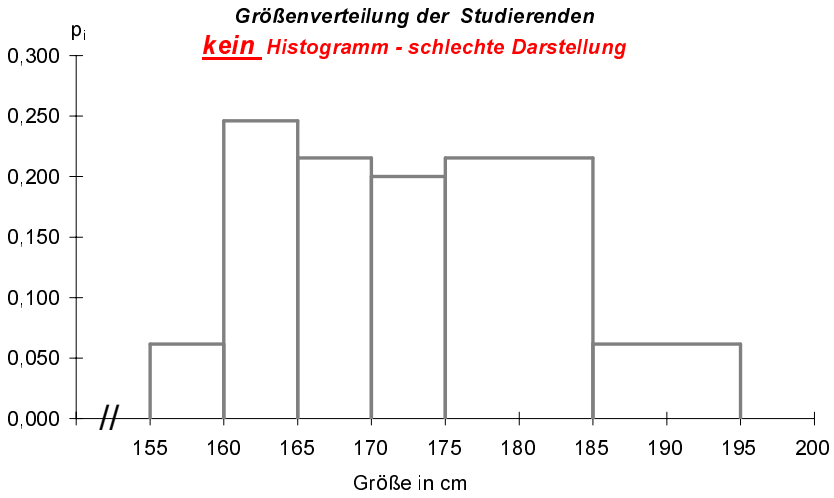


Abb. 6.11. Kein Histogramm

Abbildung 6.11 vermittelt den Eindruck, dass das Intervall zwischen 175 und 185 etwa doppelt so viele Personen umfasst, wie das Intervall zwischen 165 und 170. Ein Blick auf die Tabelle zeigt hingegen, dass die relativen Häufigkeiten in den beiden Intervallen gleich groß sind. Daher würde diese Form der Darstellung optisch einen falschen Eindruck vermitteln.

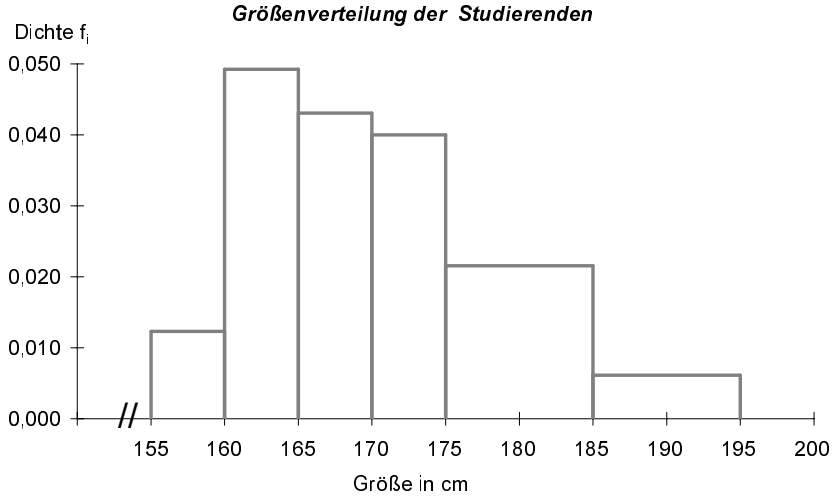


Abb. 6.12. Histogramm zu Beispiel 6.6

Ein Histogramm hingegen vermittelt die Proportionen der Häufigkeiten in den einzelnen Intervallen auf einen Blick.

In den drei Abbildungen 6.10 bis 6.12 sind auf der x-Achse jeweils zwei Schrägstriche erkennbar. Diese sollen darauf hinweisen, dass die x-Achse nicht bei Null beginnt, sondern bei einem anderen Wert.

6.3.4 Qualitätskriterien für Grafiken

Gute Grafiken sollen den BetrachterInnen möglichst auf einen Blick Ergebnisse statistischer Erhebungen vermitteln. Einerseits ist es mit den modernen Hilfsmitteln leicht geworden gute Grafiken zu erstellen, andererseits besteht aber mehr denn je die Gefahr unsinniger Grafiken. In vielen Bereichen dienen Grafiken nur als Auflockerung oder Farbtupfer im eintönigen Zahlendschungel, in anderen Bereichen werden Grafiken gezielt eingesetzt um Halbwahrheiten zu vermitteln.

Eine gute Grafik muss gewisse Anforderungen erfüllen. Jede Grafik benötigt einen Titel, der den wesentlichen Inhalt beschreiben sollte. Sind Koordinatenachsen vorhanden, so sind diese auch zu bezeichnen. Der Nullpunkt der Achsen sollte erkennbar sein, ein Abschneiden der Achsen (wie z.B. in Abbildung 6.12 die x-Achse) ist nur zulässig bei geschultem Publikum oder wenn dies zusätzlich kenntlich gemacht wird. Die Skalierungen auf den Achsen müssen maßstabsgetreu sein und generell sind bestehende Konventionen einzuhalten. Eine solche Konvention ist beispielsweise, dass auf der x-Achse die Werte nach rechts ansteigend sortiert sind. Sind keine Koordinatenachsen vorhanden, so wird die Grafik mit Hilfe einer Legende näher beschrieben.

Seitdem Grafiken nicht mehr händisch angefertigt werden müssen, haben manche Unarten überhand genommen. Erwähnt seien beispielsweise dreidimensionale Darstellungen und Grafiken, die so viel Information beinhalten, dass man das Gefühl hat, vor lauter Wald die Bäume nicht mehr zu sehen.

Die beiden folgenden Beispiele sind willkürlich ausgewählt und zeigen, wie eine Grafik nicht aussehen sollte.

Im ersten Beispiel wurden Daten aus dem Internet verwendet und die Grafik wurde in Analogie zu der gebotenen Grafik erstellt.
(Quelle, zuletzt aufgerufen am 23.5.2005:
http://www.landesbibliothek.at/content/ueber_uns/statistik/Entl98-04.gif)

Beispiel 6.7. Landesbibliothek
In einer Landesbibliothek liegen folgende Entlehnungszahlen vor:

Tabelle 6.5. Entlehnungen Landesbibliothek 1998 - 2004

	Jan	Feb	Mär	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
1998	3.973	3.871	3.524	3.656	2.988	2.625	2.755	2.556	3.000	3.503	3.794	3.721
1999	3.584	3.726	4.551	3.402	2.649	2.386	2.870	2.515	2.742	3.735	3.532	3.502
2000	3.722	4.067	4.051	3.138	3.057	2.520	2.425	2.433	2.908	3.262	2.968	2.989
2001	3.723	3.660	3.771	3.193	2.769	2.330	2.612	2.599	2.835	4.410	3.404	3.937
2002	4.167	4.098	4.401	4.000	3.746	3.162	3.691	3.490	4.196	5.085	4.492	4.599
2003	5.393	5.234	5.161	5.459	4.530	3.798	4.874	4.035	5.244	5.558	5.151	5.621
2004	5.838	5.866	6.183	5.925	5.235	5.139	5.186	5.238	5.077	5.982	6.657	6.832

Folgende Grafik wurde zur Darstellung verwendet:

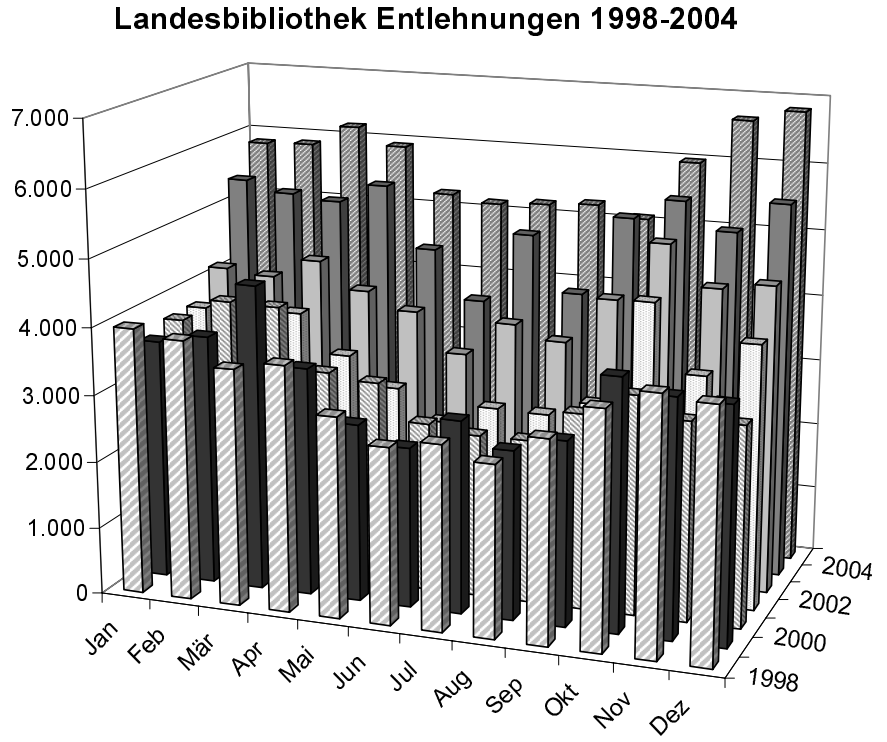


Abb. 6.13. Entlehnungen Landesbibliothek

Man sieht, dass man nichts sieht. Dies liegt einerseits an der dreidimensionalen Darstellung, andererseits an der Fülle von Informationen, die in dieser Grafik dargestellt wurde. Eine bessere Möglichkeit der Darstellung zeigt Abbildung 6.14.

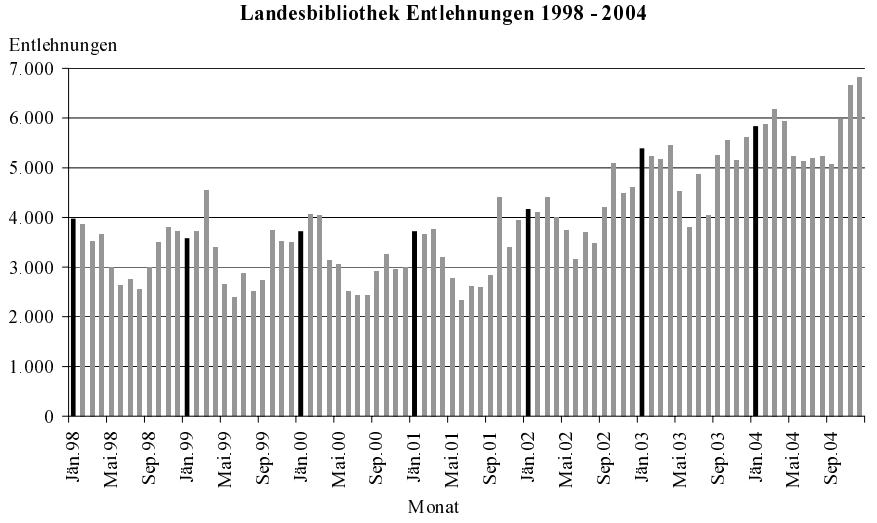


Abb. 6.14. Entlehnungen Landesbibliothek, verbesserte Darstellung

Diese Grafik stellt die gleiche Information dar wie die vorhergehende dreidimensionale Abbildung. Es wurde ein Stabdiagramm angefertigt, wobei der Jänner jeden Jahres mit einem dunkleren Stab dargestellt wurde, um die Orientierung zu erleichtern. Jetzt lässt sich problemlos ablesen, dass insgesamt die Entlehnungen ansteigen und jedes Jahr saisonale Schwankungen aufweist. In den Sommermonaten gibt es immer einen Rückgang der Entlehnungen, daneben sind in den einzelnen Jahren einige Spitzenmonate (z.B. März 1999, Oktober 2001 und 2002) erkennbar. Mit dem Hintergrundwissen, dass diese Bibliothek in erster Linie von Studierenden genutzt wird, lassen sich Schwankungen und Spitzenwerte leicht erklären, denn an der Universität sind von Juli bis September Ferien, im März und Oktober ist jeweils Semesterbeginn.

Mit Grafiken kann auch - bewusst oder unbewusst - manipuliert werden. Manchmal ist erst auf den zweiten Blick die tatsächlich verpackte Information erkennbar, wie folgendes Beispiel zeigt:

Beispiel 6.8. Höhenflug der Presse
Auf der Titelseite der Zeitung „Die Presse" vom 25.5.2000 findet sich eine Grafik über den Höhenflug der Presse (vgl. Abbildung 6.15).

Abb. 6.15. Höhenflug der Presse

Aus Sicht einer Statistikerin würde sich ein etwas anderer Höhenflug der Presse ergeben (vgl. Abbildung 6.16).

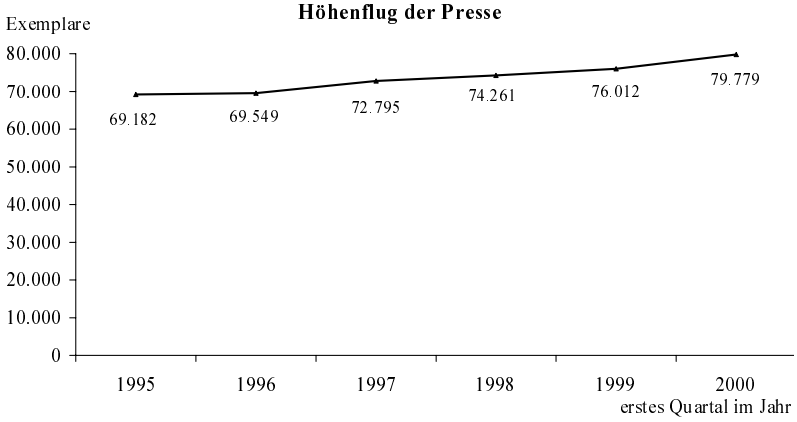


Abb. 6.16. Tatsächlicher Höhenflug der Presse

Abbildung 6.15 suggeriert steil ansteigende Verkaufzahlen der Presse. Dieser Effekt wurde unter anderem erreicht durch ein „Abschneiden" der y-Achse (beginnt bei 70.000 statt bei 0), weil dadurch Unterschiede immer besonders betont werden. Ein kleiner Pfeil nach oben im Jahr 2000 deutet an, dass der Höhenflug (natürlich) weitergeht.

Anforderungen an Grafiken

- Informativer Titel
- Beschriftung von Koordinatenachsen, Legende einfügen
- Nullpunkte von Achsen kennzeichnen
- Unverzerrte Skalierungen
- Einhalten von Konventionen
- Vermeidung dreidimensionaler Darstellungen
- Nicht zu viele Informationen in einer Grafik

Die einfachste Grafik ist zumeist auch die Beste!

6.3.5 Auswahl der passenden Darstellungsform

Die passende Darstellungsform richtet sich nach dem Skalenniveau des Merkmals und nach der Anzahl der Ausprägungen.

Kreis- oder Tortendiagramm

- relative Häufigkeiten als Kreissektoren
- Zentriwinkel $\alpha_i = 360 \cdot p_i$
- nominale Merkmale mit wenigen Ausprägungen

Säulen- oder Stabdiagramm

- x-Achse: Merkmalsausprägungen
- y-Achse: Häufigkeiten (absolute, relative oder in Prozent)
- nominale Merkmale mit vielen Ausprägungen
- ordinale Merkmale
- diskrete metrische Merkmale mit wenigen Ausprägungen
- Balkendiagramm = ein um 90 ° gedrehtes Stabdiagramm

Histogramm

- x-Achse: Merkmalsausprägungen
- y-Achse: Dichten = relative Häufigkeiten / Intervallbreiten
- metrische Merkmale mit Intervalleinteilung
- relative Häufigkeiten sind als Flächen sichtbar

6.3.6 Grafiken in EXCEL

Ausgangspunkt für die Erstellung einer Grafik mit EXCEL ist eine vorliegende Häufigkeitsverteilung in Tabellenform. Unter dem Menüpunkt *Einfügen* → *Diagramm* öffnet sich der Diagramm-Assistent. Der Diagramm-Assistent kann auch durch Anklicken des Symbols geöffnet werden. Der erste Schritt ist die Auswahl des Diagrammtyps. Entsprechend den vorangegangenen Ausführungen können wir Säulen-, Balken- oder Kreisdiagramme auswählen. Die Erstellung eines Histogramm erfolgt mit dem Diagrammtyp *Punkt (XY)*.

Als Beispiel soll ein Säulendiagramms zu der Anzahl kariöser Zähne erstellt werden (vgl. Beispiel 6.4, Seite 72). Zu jedem Diagrammtyp gibt es verschiedene Diagrammuntertypen die im rechten Fensterbereich angezeigt werden. Für ein erstes Diagramm wählen wir den vorgeschlagenen Untertyp *Gruppierte Säulen*. Mit der Betätigung des *Weiter*-Buttons gelangen wir zum zweiten Schritt, der Datenquelleneingabe (vgl. Abbildung 6.17).

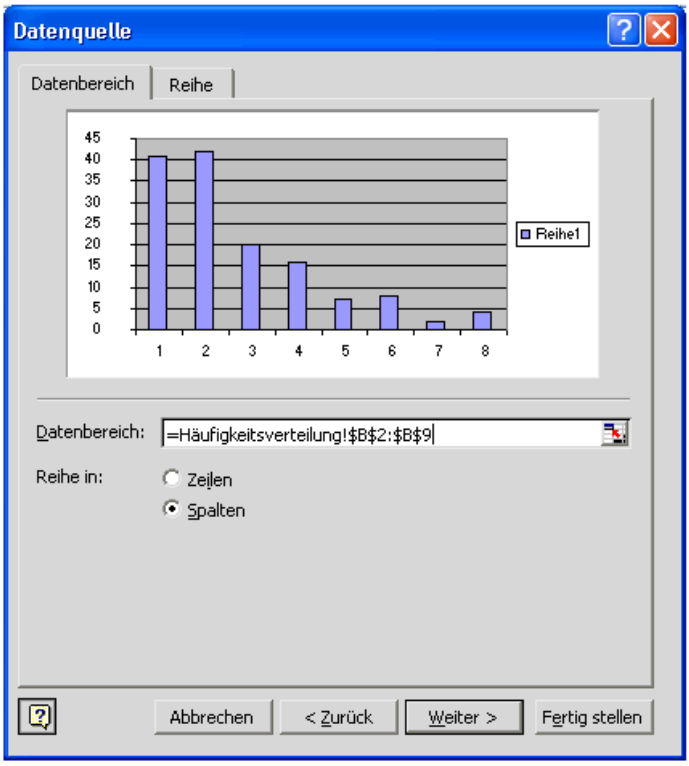


Abb. 6.17. EXCEL: Grafiken Datenbereich

Zuerst ist der Datenbereich einzugeben, am besten markiert man den Datenbereich mit der Maus (in unserem Fall *Häufigkeitsverteilung!$B$2:$B$9*). EXCEL schlägt dann für *Reihe in:* die Auswahl *Spalten* vor. Diese Auswahl ist für unser Beispiel richtig, weil eine Spalte einem Merkmal entspricht.

Im Blatt *Reihe* kann der Name der Datenreihe und die Beschriftung der x-Achse als Bezug eingegeben werden. Damit sind die Eingaben der Datenquelle abgeschlossen und werden mit *Weiter* bestätigt (vgl. Abbildung 6.18).

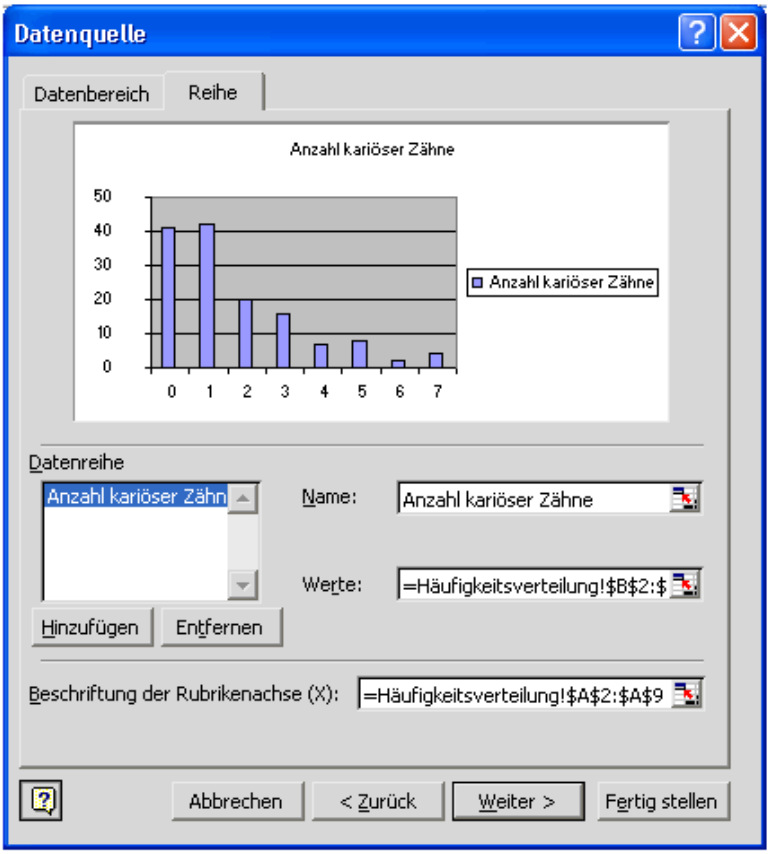


Abb. 6.18. EXCEL: Grafiken Reihe

Im dritten Schritt können weitere Optionen, wie z.B. der Diagrammtitel geändert werden (vgl. Abbildung 6.19). Im vierten und letzten Schritt wird die Art der Implementierung festgelegt: Das fertige Diagramm wird demnach entweder als eigenes Blatt in die Datei, oder als Objekt in das bestehende

Arbeitsblatt eingefügt. Alle Eingaben, die während des Erstellens mit dem Diagrammassistenten gemacht werden, können später geändert werden.

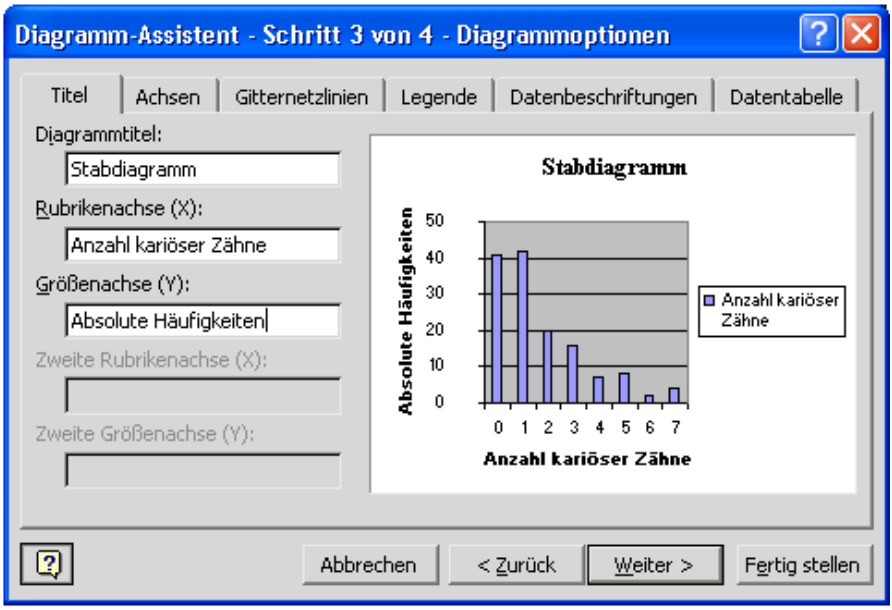


Abb. 6.19. EXCEL: Grafiken Optionen

Nach der Erstellung des Diagramms sollte das Layout noch überarbeitet werden, damit eine ansprechende Grafik entsteht. Die Vorteile von EXCEL-Grafiken liegen einerseits in der einfachen Handhabung, andererseits beim problemlosen Einfügen von Grafiken in Berichte.

6.3.7 Erstellen von Histogrammen in EXCEL

Histogramme sind leider nicht standardmäßig als EXCEL-Grafik vorgesehen, daher müssen die vorhandenen Möglichkeiten etwas adaptiert werden, um wirklich ein Histogramm erstellen zu können. Für die Erstellung eines Histogramms wird der Diagrammtyp *Punkt (XY)*, Diagrammuntertyp *Punkte mit Linien ohne Datenpunkte* verwendet. Die Datenreihe wird aber in veränderter Form eingegeben. Jeder einzelne Eckpunkt bzw. Anstoßpunkt einer Linie muss getrennt eingegeben werden, für das Histogramm aus Abbildung 6.10 (Seite 74) wäre das die Datenreihe aus Tabelle 6.6. Als *Datenbereich* wählt man den Bereich der Dichten aus, im Blatt *Reihe* wird dieser Bereich nun automatisch als *Y-Werte* angeführt. Als *X-Werte* dient der Bereich mit den Körpergrößen.

Tabelle 6.6. Dateneingabe Histogramm

Körpergröße	Dichte
155	0,0000
155	0,0123
160	0,0123
160	0,0000
160	0,0492
165	0,0492
165	0,0000
165	0,0431
⋮	⋮
190	0,0062
190	0,0000
190	0,0062
195	0,0062
195	0,0000

Grafiken erstellen in EXCEL

- Diagramm-Assistent aufrufen 
- Diagrammtyp auswählen
  (für Histogramm *Punkt (XY)*)
- Datenbezug eingeben
  (beim Historgramm jeden einzelnen Eckpunkt bzw. Anstoßpunkt)
- Layout an Anforderungen anpassen

6.3.8 Grafiken in SPSS

SPSS stellt unter dem Menüpunkt *Grafiken* eine Vielzahl von Grafiken bereit, darunter auch Kreisdiagramme und Balkendiagramme. Auch eine Grafik mit der Bezeichnung Histogramm kann angefordert werden, allerdings entspricht diese nicht den Anforderungen an ein Histogramm und sollte daher unbedingt vermieden werden. Das Layoutieren in SPSS ist weniger benutzerfreundlich als in EXCEL, daher wird empfohlen Ergebnisse aus SPSS in eine EXCEL-Datei zu exportieren und allfällige Grafiken in EXCEL zu erstellen. Für das Exportieren wird im Viewer die gewünschte Tabelle markiert, die rechte Maustaste öffnet dann ein Menü, in dem *Exportieren* angeboten wird. Das nun erscheinende Dialogfenster erlaubt nähere Spezifikationen, wie z.B. Speicherort oder gewünschtes Dateiformat.

6.4 Die empirische Verteilungsfunktion

Die empirische Verteilungsfunktion weist jeder Ausprägung x_i den Anteil von Erhebungseinheiten zu, deren Ausprägung höchstens x_i ist. Sie ist nur für ordinale oder metrische Merkmale sinnvoll interpretierbar. Die Werte der empirischen Verteilungsfunktion werden auch als kumulierte Häufigkeiten bezeichnet.

Die Verteilungsfunktion wird mit $F(x_i)$ oder $p(x \leq x_i)$ bezeichnet.

Die empirische Verteilungsfunktion
Bezeichnungen und Eigenschaften

- $F(x_i) = p(x \leq x_i) = p_1 + \ldots + p_i$
  der Anteil an Objekten, die höchstens die Ausprägung x_i aufweisen
  = empirische Verteilungsfunktion, kumulierte Häufigkeiten

- $0 \leq p(x \leq x_i) \leq 1$
  Die Werte der Verteilungsfunktion liegen im Intervall $[0, 1]$

- $p(x \leq x_i) \leq p(x \leq x_{i+1})$
  Die Verteilungsfunktion ist monoton steigend.

Beispiel 6.9. Kariöse Zähne von Kindern
Fortsetzung von Beispiel 6.1, Seite 61
Der Wert der empirischen Verteilungsfunktion (kumulierte Häufigkeit) einer Ausprägung ergibt sich aus der Addition der relativen Häufigkeiten aller Ausprägungen, die kleiner oder gleich dieser Ausprägung sind.

Tabelle 6.7. Empirische Verteilungsfunktion zu Beispiel 6.1

i	x_i	h_i	p_i	$p(x \leq x_i)$
1	0	41	0,29	0,29
2	1	42	0,30	0,59
3	2	20	0,14	0,74
4	3	16	0,11	0,85
5	4	7	0,05	0,90
6	5	8	0,06	0,96
7	6	2	0,01	0,97
8	7	4	0,03	1,00
Summe		140	≈ 1	nicht sinnvoll

Aus der empirischen Verteilungsfunktion lässt sich in diesem Beispiel folgende Information ablesen (Zeile i = 4): 85 Prozent der Kinder haben höchstens 3 kariöse Zähne.

Beispiel 6.10. Körpergrößen von Studierenden
Fortsetzung von Beispiel 6.2, Seite 69
Der Wert der empirischen Verteilungsfunktion (kumulierte Häufigkeit) an der Intervallobergrenze ergibt sich aus der Addition der relativen Häufigkeiten aller Intervalle, deren Obergrenze kleiner oder gleich dieser Intervallobergrenze sind.

Tabelle 6.8. Empirische Verteilungsfunktion zu Beispiel 6.2

Intervall i	Größe $e_{i-1} < x \leq e_i$	rel. Häufigkeit p_i	kum. Häufigkeit $p(x \leq e_i)$
1	$155 < x \leq 160$	0,062	0,062
2	$160 < x \leq 165$	0,246	0,308
3	$165 < x \leq 170$	0,215	0,523
4	$170 < x \leq 175$	0,200	0,723
5	$175 < x \leq 180$	0,108	0,831
6	$180 < x \leq 185$	0,108	0,938
7	$185 < x \leq 190$	0,031	0,969
8	$190 < x \leq 195$	0,031	1,000
	Summe	≈ 1	nicht sinnvoll

Aus der empirischen Verteilungsfunktion lässt sich z.B. ablesen (Zeile i = 4), dass 72,3 Prozent der untersuchten Personen höchstens 175 cm groß sind.

6.4.1 Abbild der empirischen Verteilungsfunktion

Das Aussehen des Abbildes der empirischen Verteilungsfunktion ist vom Merkmalstyp abhängig. Für diskrete Merkmale bildet die Verteilungsfunktion eine Treppenfunktion, für intervallskalierte Merkmale ist das Abbild eine Kurve mit geraden Teilstücken. Für beide Formen werden die Ausprägungen auf der x-Achse aufgetragen und die kumulierten Häufigkeiten auf der y-Achse, der Unterschied liegt lediglich in der Art und Weise, wie diese Punkte verbunden werden.

Abbild der Verteilungsfunktion

- Diskrete Merkmale: Treppenfunktion
- Intervallskalierte Merkmale: Kurve mit geraden Teilstücken

Beispiel 6.11. Kariöse Zähne von Schulkindern
Fortsetzung von Beispiel 6.1
Nachdem das Merkmal Anzahl kariöser Zähne diskret ist, muss das Abbild einer Treppenfunktion entsprechen. Zuerst werden die Datenpunkte $(x_i, F(x_i))$ in ein Koordinatensystem eingetragen, für die Verbindung der Punkte beginnt man beim Ursprung. Falls die erste Ausprägung ungleich 0 ist, beginnt man mit einer horizontalen Linie bis zur ersten Ausprägung. Danach wird die erste Ausprägung bzw. der Ursprung durch eine vertikale Linie mit dem ersten Datenpunkt verbunden. Vom Datenpunkt zieht man eine horizontale Linie bis zur nächsten Ausprägung und von dieser wiederum eine vertikale Linie bis zum nächsten Datenpunkt. In dieser Weise werden alle Datenpunkte verbunden. Nach dem letzten Datenpunkt kann die Linie horizontal weitergeführt werden. Zu beachten ist, dass an den Sprungstellen (also bei den jeweiligen Ausprägungen) das obere Ende der Verbindungslinie der zugehörige Funktionswert ist.

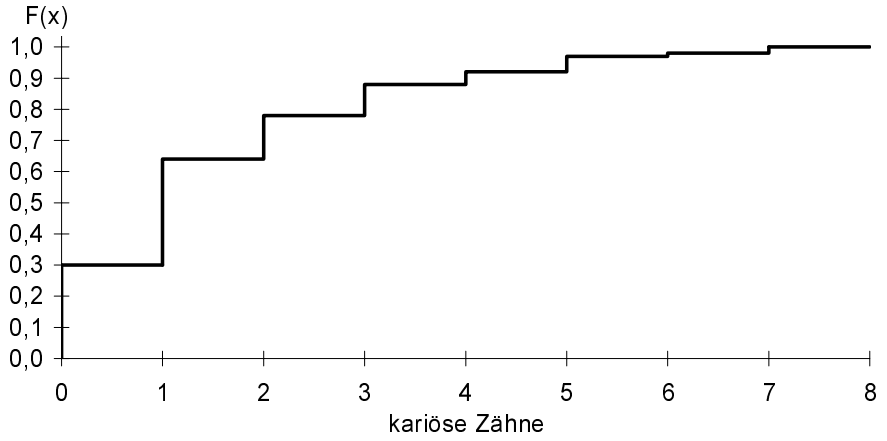


Abb. 6.20. Verteilungsfunktion zu Beispiel 6.1

Beispiel 6.12. Körpergröße
Fortsetzung von Beispiel 6.2
Das Abbild des intervallskalierten Merkmals Körpergröße ist eine Kurve mit geraden Teilstücken. Auch hier werden zuerst die Datenpunkte $(e_i, F(e_i))$ in ein Koordinatensystem eingetragen. Die Punkte werden anschließend durch gerade Linien verbunden.

In diesem Beispiel ist es sinnvoll, die x-Achse erst bei 155 beginnen zu lassen. Auf diesen Umstand muss aber auch in der Grafik gesondert hingewiesen werden, etwa durch die Kennzeichnung mit einem doppelten Schrägstrich (Abbildung 6.21).

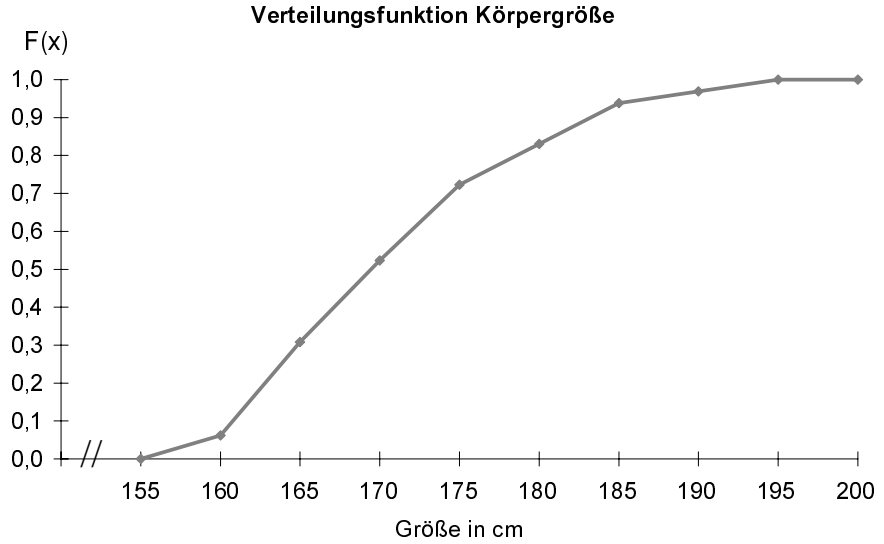


Abb. 6.21. Verteilungsfunktion zu Beispiel 6.2

Die Verteilungsfunktion lässt sich vor dem ersten und nach dem letzten Datenpunkt beliebig horizontal fortsetzen. Auch Interpretationen sind für diese Erweiterungen kein Problem. Im Beispiel der Körpergrößen wäre z.B. der Ausprägung 130 der Funktionswert 0 zugeordnet, was bedeutet, dass 0% der Personen höchstens 130 cm groß sind. Diese Aussage ist sicherlich richtig, wenn auch nicht besonders informativ.

6.4.2 Rechnen mit der empirischen Verteilungsfunktion

Für die Berechnung der empirischen Verteilungsfunktion mit EXCEL ist die **rekursive Darstellung**

$$p(x \leq x_i) = p(x \leq x_{i-1}) + p_i$$

hilfreich.

Die **Zuverlässigkeitsfunktion** weist jeder Ausprägung i den Anteil von Objekten zu, deren Ausprägung größer als i ist, in formaler Schreibweise

$$p(x > x_i) = 1 - p(x \leq x_i)$$

Beispiel 6.13. Körpergröße
Fortsetzung von Beispiel 6.2
Die Werte der Zuverlässigkeitsfunktion an den Intervallobergrenzen sind zu berechnen und zu interpretieren.

Tabelle 6.9. Zuverlässigkeitsfunktion zu Beispiel 6.2

Intervall i	Größe $e_{i-1} < x \leq e_i$	Verteilungsfunktion $p(x \leq e_i)$	Zuverlässigkeitsfunktion $p(x > e_i)$
1	$155 < x \leq 160$	0,062	0,938
2	$160 < x \leq 165$	0,308	0,692
3	$165 < x \leq 170$	0,523	0,477
4	$170 < x \leq 175$	0,723	0,277
5	$175 < x \leq 180$	0,831	0,169
6	$180 < x \leq 185$	0,938	0,062
7	$185 < x \leq 190$	0,969	0,031
8	$190 < x \leq 195$	1,000	0,000

27,7 Prozent der untersuchten Personen sind größer als 175 cm (Zeile i = 4).

Aus der Differenz zweier Verteilungsfunktionswerte kann man den Anteil an Objekten in einem Intervall berechnen

$$p(a < x \leq b) = p(x \leq b) - p(x \leq a)$$

Beispiel 6.14. Körpergröße
Fortsetzung von Beispiel 6.2
FlugbegleiterInnen müssen größer als 170 cm sein, andererseits darf aber eine Körpergröße von 185 cm nicht überschritten werden. Welcher Anteil an Personen erfüllt diese Kriterien?

$$p(170 < x \leq 185) = p(x \leq 185) - p(x \leq 170) = 0,938 - 0,523 = 0,415$$

41,5% der untersuchten Personen erfüllen die Größenanforderungen für FlugbegleiterInnen.

Für intervallskalierte Merkmale lassen sich die Werte der Verteilungsfunktion für Ausprägungen innerhalb der Intervalle mit folgender Formel berechnen:

Für die Ausprägung j $\in (e_{i-1}; e_i]$ gilt

$$p(x \leq j) = p(x \leq e_{i-1}) + p_i \cdot \frac{j - e_{i-1}}{e_i - e_{i-1}} \tag{6.1}$$

Diese Formel lässt sich mit Hilfe von Abbildung 6.22 leicht herleiten.

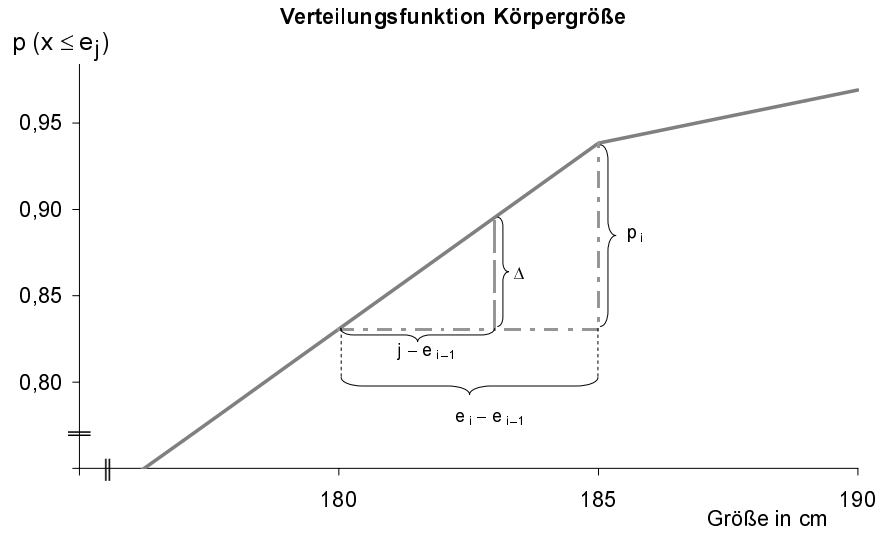


Abb. 6.22. Detail einer Verteilungsfunktion

Bereits bekannt sind die Intervallgrenzen e_{i-1} und e_i, sowie die Intervallbreite $e_i - e_{i-1}$. Auch die relative Häufigkeit dieses Intervalls p_i ist bekannt. Liegt die Ausprägung j im Intervall $(e_{i-1}; e_i]$ so kann aus der Ähnlichkeit der Dreiecke in Abbildung 6.22 der Zusammenhang

$$(j - e_{i-1}) : (e_i - e_{i-1}) = \Delta : p_i$$

abgelesen werden. Die Auflösung der Gleichung nach Δ ergibt

$$\Delta = p_i \cdot \frac{j - e_{i-1}}{e_i - e_{i-1}}$$

Addiert man nun diesen Wert zum Wert der Verteilungsfunktion an der Untergrenze des Intervalls $p(x \leq e_{i-1})$ so erhält man (6.1) und damit den Wert der Verteilungsfunktion an der Stelle j.

Beispiel 6.15. Körpergröße
Fortsetzung von Beispiel 6.2
Welcher Anteil an Personen ist höchstens 183 cm groß?

Um den Wert der Verteilungsfunktion für die Ausprägung 183 zu berechnen, verwenden wir (6.1). Zuerst muss jenes Intervall bestimmt werden, in dem die Ausprägung liegt. Dann lassen sich Untergrenze, Obergrenze, relative Häufigkeit dieses Intervalls und der Wert der Verteilungsfunktion an der Untergrenze des Intervalls aus Tabelle 6.8 (Seite 86) ablesen:

$$e_{i-1} = 180 \qquad e_i = 185 \qquad p_i = 0,108 \qquad p(x \leq 180) = 0,831$$

$$F(183) = p(x \leq 183) = 0,831 + 0,108 \cdot \frac{183 - 180}{185 - 180} = 0,896$$

89,6 Prozent der Personen sind höchstens 183 cm groß. Dieses Ergebnis kann auch näherungsweise aus dem Abbild der Verteilungsfunktion abgelesen werden. Dazu geht man in Abbildung 6.21 (Seite 88) von der Ausprägung 183 vertikal nach oben, bis man auf die Funktion trifft, dann geht man horizontal nach links und kann an der y-Achse den Wert der Verteilungsfunktion näherungsweise ablesen.

Rechnen mit der Verteilungsfunktion

- $p(x \leq x_i) = p(x \leq x_{i-1}) + p_i$ rekursive Berechnung
- $p(x > x_i) = 1 - p(x \leq x_i)$ Zuverlässigkeitsfunktion
- $p(a < x \leq b) = p(x \leq b) - p(x \leq a)$
- Für $j \in (e_{i-1}; e_i]$ gilt
  $p(x \leq j) = p(x \leq e_{i-1}) + p_i \cdot \frac{j - e_{i-1}}{e_i - e_{i-1}}$

6.4.3 Die empirische Verteilungsfunktion in EXCEL

In EXCEL lassen sich die Werte der Verteilungsfunktion am besten mit der rekursiven Formel berechnen. Das Abbild der Verteilungsfunktion wird mittels Diagrammtyp *Punkt (XY)*, Diagrammuntertyp *Punkte mit Linien* oder *Punkte mit Linien ohne Datenpunkte* erstellt. Für das stetige Merkmal genügen die Datenpunkte als Eingabe, für das diskrete Merkmal müssen alle Eckpunkte der Treppenfunktion eingegeben werden (vgl. dazu Kapitel 6.3.7).

6.4.4 Die empirische Verteilungsfunktion in SPSS

In SPSS werden die kumulierten Häufigkeiten standardmäßig bei der Erstellung einer Häufigkeitstabelle mitgeliefert (vgl. Kapitel 6.1.2). Für die Erstellung der Grafik wird EXCEL empfohlen.

Übungsaufgaben

6.1. Kariöse Zähne

Bei $N = 140$ Kindern einer Jahrgangsstufe werden die Zähne auf Karies untersucht. Bei jedem Kind wird das Merkmal Anzahl der kariösen Zähne erhoben, wobei sich folgende Ergebnisse einstellen:

4	0	0	3	1	5	1	2	2	0	5	0	5	2	1	0	1	0	0	4
0	1	1	3	0	1	1	1	3	1	0	1	4	2	0	3	1	1	7	2
0	2	1	3	0	0	0	0	6	1	1	2	1	0	1	0	3	0	1	3
0	5	2	3	0	2	4	0	1	1	3	0	6	2	1	5	1	1	2	2
0	3	0	1	0	1	0	0	0	5	0	4	1	2	2	7	1	3	1	5
1	0	1	0	0	4	0	3	1	1	7	2	1	0	3	0	1	3	2	2
2	7	1	3	1	5	1	0	0	0	2	1	0	3	1	4	0	2	1	1

Welcher Anteil an Kindern hat

a) mehr als 2 aber höchstens 4 kariöse Zähne?
b) mindestens 2 aber höchstens 4 kariöse Zähne?
c) mindestens 2 aber weniger als 4 kariöse Zähne?
d) mehr als 2 aber weniger als 4 kariöse Zähne?

6.2. Körpergröße von Studierenden

Bei $N = 65$ Studierenden einer Jahrgangsstufe wurde die Körpergröße in cm erhoben. Diese Erhebung führte zu folgendem Ergebnis:

178	182	166	162	181	168	170	164	171	170	165	164	176
169	165	165	175	180	164	188	170	194	171	185	168	164
180	174	183	193	172	178	165	162	174	174	162	163	170
172	176	168	160	170	170	171	166	165	160	175	183	182
189	162	168	160	178	175	168	158	172	163	172	183	164

a) Welcher Anteil an Studierenden ist höchstens 190 cm groß?
b) Welcher Anteil an Studierenden ist größer als 158 cm?
c) Welcher Anteil an Studierenden ist größer als 173 cm, aber höchstens 186 cm groß?

6.3. Altersverteilung

In der Bevölkerung wurde nachstehende Häufigkeitsverteilung des Merkmals Alter erhoben:

Alter	Personen
$0 < x \leq 15$	1.333.505
$15 < x \leq 30$	1.495.740
$30 < x \leq 45$	2.002.259
$45 < x \leq 60$	1.526.110
$60 < x \leq 75$	1.151.122
$75 < x \leq 100$	609.018

a) Berechnen Sie die relativen Häufigkeiten der einzelnen Altersintervalle.
b) Stellen Sie die Häufigkeitsverteilung in einem Histogramm dar.
c) Berechnen Sie die Werte der Verteilungsfunktion und der Zuverlässigkeitsfunktion für die Ausprägungen 12, 35 und 60 und interpretieren Sie Ihre Ergebnisse.

6.4. BundespräsidentInnenwahl
Die österreichische BundespräsidentInnenwahl vom 25.4.2004 brachte folgendes Ergebnis: Dr. Benita Ferrero-Waldner erhielt 1.969.326 Stimmen und Dr. Heinz Fischer erhielt 2.166.690 Stimmen. Berechnen Sie die relativen Häufigkeiten der beiden KandidatInnen.

6.5. Nationalratswahl
Die erste Nationalratswahl der 2. Republik Österreich am 25.11.1945 ergab nachstehende Verteilung der gültigen Stimmen auf die damals zur Wahl stehenden Parteien.

Partei	Stimmen
SPÖ	1.434.898
ÖVP	1.602.227
KPÖ	174.257
Sonstige	5.972

Berechnen Sie die relativen Häufigkeiten der einzelnen Parteien.

6.6. TV-Geräte
Eine Umfrage in 425 Haushalten ergab folgende Verteilung für das Merkmal Anzahl an TV-Geräten:

Anzahl an TV-Geräten	h_j
0	48
1	156
2	193
3	21
4	7

a) Stellen Sie diese Häufigkeitsverteilung in einem Stab- und einem Kreisdiagramm dar.
b) Berechnen Sie die Werte der Verteilungsfunktion und stellen Sie diese grafisch dar.

7

Maßzahlen für eindimensionale Verteilungen

Häufigkeitsverteilungen geben einen ersten Überblick über Informationen aus der Urliste. Oft ist man aber an Informationen in sehr komprimierter Form interessiert. Dabei soll möglichst viel Information über die Daten in einer einzigen Zahl verarbeitet werden. Dies gelingt bis zu einem gewissen Grad mit Lagemaßzahlen (z.B dem Mittelwert), die das Zentrum einer Verteilung widerspiegeln. Dieses Zentrum kann auf unterschiedliche Arten festgestellt werden, daher gibt es auch unterschiedliche Lagekennzahlen. Welche Lagemaßzahl für den konkreten Fall geeignet ist, entscheidet unter anderem das Skalenniveau des Merkmals, aber auch andere Kriterien sind zu beachten.
Die Aussagekraft einer Lagemaßzahl ist begrenzt, daher wird im Normalfall zur besseren Charakterisierung der Datenverteilung noch zusätzlich eine Streuungsmaßzahl erhoben. Diese gibt an, wie weit die Daten von einander oder von einer Lagemaßzahl abweichen.

7.1 Lagemaße

Die wichtigsten Lagemaßzahlen sind das arithmetische Mittel, der Median und der Modus. Für bestimmte Fragestellungen sind diese Kennzahlen aber nicht ausreichend und man muss auf andere Maßzahlen, wie z.B. das geometrische Mittel zurückgreifen. Insbesondere in Hinblick auf die schließende Statistik sind auch sogenannte Quantile wichtige Lagemaßzahlen.

7.1.1 Arithmetisches Mittel

Das arithmetische Mittel ist aus dem Alltagsleben als Mittelwert bekannt. Die meisten Durchschnittswerte, wie beispielsweise das durchschnittliche Einkom-

men, die durchschnittliche Pensionshöhe oder die durchschnittliche Studiendauer sind in statistischer Bezeichnung arithmetische Mittel.

Bezeichnungen

N	Untersuchungsumfang
r	Anzahl der verschiedenen Ausprägungen
x_i	Ausprägung des i-ten Objektes bzw. i-te Ausprägung

Arithmetisches Mittel

Liegen N Messdaten zum Merkmal X vor, dann ist

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{7.1}$$

das arithmetische Mittel (oder der Mittelwert) des Merkmals X.

Liegt eine Häufigkeitsverteilung zum Merkmal X mit r verschiedenen Ausprägungen vor, dann ist

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{r} x_i h_i = \sum_{i=1}^{r} x_i p_i \tag{7.2}$$

Alle Formeln liefern das gleiche Ergebnis. Formel (7.1) verwendet man, wenn die Urliste der Daten vorliegt. Liegt hingegen die Häufigkeitsverteilung in Tabellenform vor, so ist es wesentlich effizienter, eine der beiden Formeln (7.2) zu verwenden.

Beispiel 7.1. Durchschnittsgewicht von Gepäckstücken
Von vier Gepäckstücken soll das durchschnittliche Gewicht erhoben werden. Die gemessenen Gewichte betragen 20 kg, 20 kg, 30 kg und 10 kg. Ausgehend von dieser Urliste lässt sich der Mittelwert mit (7.1) berechnen als

$$\bar{x} = \frac{1}{4} \cdot (10 + 20 + 20 + 30) = 20$$

Man könnte aber auch in einem ersten Schritt eine Häufigkeitstabelle anlegen und aus dieser den Mittelwert berechnen:

x_i	h_i	p_i
10	1	0,25
20	2	0,50
30	1	0,25

Daraus ergibt sich mit (7.2) das arithmetische Mittel als

$$\bar{x} = \frac{1}{4} \cdot (10 \cdot 1 + 20 \cdot 2 + 30 \cdot 1) = 20$$

oder als

$$\bar{x} = 10 \cdot 0,25 + 20 \cdot 0,5 + 30 \cdot 0,25 = 20$$

Bei Intervallskalierung eines Merkmals werden für die Berechnung des Mittelwertes die jeweiligen Intervallmitten anstelle der Ausprägungen verwendet.

Beispiel 7.2. Körpergröße von Studierenden
Fortsetzung von Beispiel 6.2, Seite 69

$$\bar{x} = \frac{1}{65} \cdot (157,5 \cdot 4 + 152,2 \cdot 16 + \ldots + 192,5 \cdot 2) \approx 170,7$$

Die durchschnittliche Körpergröße beträgt 170,7 cm.

Das arithmetische Mittel als Durchschnittswert ist ausschließlich für metrische Merkmale zulässig. Ein besonderer Anwendungsfall ergibt sich, wenn man die Berechnungsvorschrift auf dichotome Merkmale anwendet, deren Ausprägungen mit 0 bzw. 1 kodiert sind. In diesem speziellen Anwendungsfall erhält man als arithmetisches Mittel nicht den Durchschnittswert, sondern den Anteil der 1-Kodierungen an der Untersuchungsgesamtheit.

Beispiel 7.3. Geschlecht von LVA-TeilnehmerInnen
(vgl. Beispiel 6.3, Seite 71) An einer Lehrveranstaltung nehmen 25 Personen teil, davon sind 10 Personen weiblich. Es ist der Frauenanteil zu berechnen.

Geschlecht	Kodierung x_i	Anzahl h_i
männlich	0	15
weiblich	1	10
Summe		25

Das arithmetische Mittel beträgt

$$\bar{x} = \frac{1}{25} \cdot (0 \cdot 15 + 1 \cdot 10) = 0,4$$

Damit beträgt der Frauenanteil in dieser Lehrveranstaltung 40%.

Bei dieser Anwendung ist darauf zu achten, dass diejenige Ausprägung, an der man interessiert ist, mit 1 kodiert ist.

Mittelwert, arithmetisches Mittel

- Bezeichnung $\bar{x}$
- Interpretation: Durchschnitt
- Zulässigkeit: Ausschließlich für metrische Merkmale geeignet. Ungeeignet für nominale und ordinale Merkmale.
- Bei intervallskalierten Merkmalen werden als Ausprägungen die Intervallmitten verwendet.
- Die Berechnungsvorschrift des Mittelwertes angewendet auf dichotome Merkmale ergibt den Anteil der 1-Kodierungen.

7.1.2 Median

Der Median wird durch folgende Eigenschaft charakterisiert: Mindestens 50% der Objekte haben eine Ausprägung, die höchstens so groß ist wie der Median und mindestens 50% der Objekte haben eine Ausprägung, die mindestens so groß ist wie der Median. Man kann sich den Median als mittleren Wert einer geordneten Datenreihe vorstellen.

Zur Bestimmung des Medians werden die Daten der Urliste der Größe nach sortiert. Besteht der Datensatz aus einer ungeraden Anzahl von Objekten, so gibt es genau eine Ausprägung, die im geordneten Datensatz in der Mitte steht. Diese Ausprägung ist der Median. Bei einem Datensatz mit einer geraden Anzahl von Objekten gibt es zwei Ausprägungen in der Mitte. Der Median errechnet sich als arithmetisches Mittel dieser beiden Ausprägungen.

Bezeichnungen

N	Untersuchungsumfang
$x_{(i)}$	Ausprägung an der i-ten Stelle der geordneten Datenreihe

Median

Der Wert

$$\widetilde{x}_{0,5} = \begin{cases} x_{\left(\frac{N+1}{2}\right)} & \text{wenn N ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}\right) & \text{wenn N gerade} \end{cases}$$

der geordneten Merkmalsausprägungen heißt Median des Merkmals X.

Median

- Bezeichnung $\widetilde{x}_{0,5}$
- Interpretation: Mindestens 50% der Objekte haben eine Ausprägung, die höchstens so groß ist wie der Median und mindestens 50% der Objekte haben eine Ausprägung, die mindestens so groß ist wie der Median.
- Zulässigkeit: Für ordinale und metrische Merkmale geeignet. Ungeeignet für nominale Merkmale.

Beispiel 7.4. Seile
Vier Seile werden auf ihre Länge untersucht. Die gemessenen Seillängen betrugen 50m, 60m, 30m und 40m. Bevor der Median bestimmt werden kann, muss diese Urliste nach den Ausprägungen geordnet werden, also 30m, 40m, 50m, 60m. Die Anzahl der untersuchten Objekte ist gerade (N = 4), daher müssen wir das arithmetische Mittel jener Ausprägungen berechnen, die an den Stellen $\frac{N}{2} = 2$ und $\frac{N}{2} + 1 = 3$ der geordneten Datenreihe stehen. In unserem Fall ergibt sich $(40+50)/2 = 45$. Der Median liegt demnach bei 45 Meter. Mindestens 50% der Seile sind höchstens 45 Meter lang und mindestens 50% der Seile sind mindestens 45 Meter lang.
In einer zweiten Untersuchung werden fünf Seile untersucht mit den Ergebnissen 40m, 30m,50m, 30m und 60m. Auch hier muss die Urliste in einem ersten Schritt in eine geordnete Datenreihe umgewandelt werden: 30m, 30m, 40m, 50m, 60m. Die Anzahl der untersuchten Objekte ist ungerade (N = 5), daher steht an der Stelle $\frac{N+1}{2} = 3$ der geordneten Datenreihe der Median. In unserem Beispiel liegt der Median demnach bei 40 Meter. Mindestens 50% der untersuchten Seile sind höchstens 40 Meter lang und mindestens 50% der untersuchten Seile sind mindestens 40 Meter lang.

Ein anderer Weg zur Bestimmung des Medians ist das **Ablesen aus der Verteilungsfunktion**. Für **diskrete Merkmale** können zwei Fälle auftreten:

1. Es gibt eine Ausprägung, für die der Wert der Verteilungsfunktion exakt 0,5 ist. Dieser Fall kann nur dann eintreten, wenn eine gerade Zahl N von Objekten vorliegt und die Ausprägungen an den Stellen $(\frac{N}{2})$ und $(\frac{N}{2}+1)$ der geordneten Datenreihe unterschiedlich sind. In diesem Fall errechnet man den Median als arithmetisches Mittel von jener Ausprägung, für die der Wert der Verteilungsfunktion exakt 0,5 ist, und der nächst größeren Ausprägung (vgl. Beispiel 7.4, erste Untersuchung).

2. Es gibt keine Ausprägung, für die der Wert der Verteilungsfunktion exakt 0,5 ist. Dann ist der Median jene Ausprägung, für die der Wert der Verteilungsfunktion das erste Mal den Wert 0,5 überschreitet (vgl. Beispiel 7.4, zweite Untersuchung).

Beispiel 7.5. Kariöse Zähne von Kindern
Fortsetzung von Beispiel 6.1, Seite 61
Es wurden $N = 140$ Kinder untersucht, der Median berechnet sich als arithmetisches Mittel der Ausprägungen, die an den Stellen 70 bzw. 71 der geordneten Datenreihe stehen. Ordnet man die Ausprägungen der Größe nach (zuerst alle Kinder mit 0 kariösen Zähnen, dann alle mit einem kariösen Zahn usw.) dann findet man als Median $\frac{1+1}{2} = 1$. Mindestens 50% der Kinder haben mindestens einen kariösen Zahn und mindestens 50% der Kinder haben höchstens einen kariösen Zahn.

Intervallskalierte Merkmale

Für intervallskalierte Merkmale wird der Median als jene Ausprägung bestimmt, für die der Wert der empirischen Verteilungsfunktion 0,5 ist. Diese Ausprägung kann aus der grafischen Darstellung der empirischen Verteilungsfunktion abgelesen werden. Man kann den Median auch mit Formel (6.1) berechnen, die für diesen Zweck etwas umgeformt werden muss (vgl. Seite 90). Formel (6.1) berechnet für eine vorgegebene Ausprägung den Wert der Verteilungsfunktion. Nun benötigen wir die Umkehrung, weil zum Wert der Verteilungsfunktion 0,5 die zugehörige Ausprägung gesucht werden muss.

Das Intervall $(e_{i-1}; e_i]$ sei jenes Intervall, bei dem der Wert der Verteilungsfunktion an der Obergrenze des Intervalls das erste Mal den Wert 0,5 erreicht oder überschreitet. Dann ist der Median jene Ausprägung j, für die gilt:

$$0,5 = p(x \leq e_{i-1}) + p_i \cdot \frac{j - e_{i-1}}{e_i - e_{i-1}}$$

Auflösung dieser Gleichung nach j liefert

$$j = (0,5 - p(x \leq e_{i-1})) \cdot \frac{e_i - e_{i-1}}{p_i} + e_{i-1}$$

Berechnung des Medians bei intervallskalierten Merkmalen

Das Intervall $(e_{i-1}; e_i]$ sei jenes Intervall, bei dem der Wert der Verteilungsfunktion an der Obergrenze des Intervalls das erste Mal den Wert 0,5 erreicht oder überschreitet. Dann ist der Median gegeben durch

$$\widetilde{x}_{0,5} = (0,5 - p(x \leq e_{i-1})) \cdot \frac{e_i - e_{i-1}}{p_i} + e_{i-1}$$

Beispiel 7.6. Körpergröße von Studierenden
Fortsetzung von Beispiel 6.2, Seite 69
Die Verteilungsfunktion übersteigt den Wert 0,5 zum ersten Mal an der Obergrenze des Intervalls (165; 170] (vgl. Tabelle 6.8, Seite 86). Dieses Intervall bildet unseren Ausgangspunkt. Daher sind folgende Werte zu verwenden:

$$p(x \leq e_{i-1}) = 0,308 \qquad e_{i-1} = 165 \qquad e_i = 170 \qquad p_i = 0,215$$

Damit ergibt sich:

$$\widetilde{x}_{0,5} = (0,5 - 0,308) \cdot \frac{170 - 165}{0,215} + 165 \approx 169,5$$

Mindestens 50% der Personen sind höchstens 169,5 cm groß und mindestens 50% der Personen sind mindestens 169,5 cm groß.

7.1.3 Modus

Diejenige Ausprägung, welche die größte Häufigkeit aufweist, bezeichnet man als Modus oder als Modalwert. Tritt die größte vorkommende Häufigkeit bei nur einer Ausprägung auf, bezeichnet man die Verteilung als **unimodal**. Bimodale Verteilungen besitzen zwei Modi, die größte Häufigkeit tritt demnach bei zwei Ausprägungen auf. Bei intervallskalierten Merkmalen bezeichnet man das Intervall, welches die größte Häufigkeit aufweist, als **modale Klasse**. Der Modus ist für alle Merkmalstypen geeignet, spielt aber in der Praxis eine untergeordnete Rolle.

Modus, Modalwert

- Bezeichnung x_{mod}
- Interpretation: Ausprägung mit der größten Häufigkeit
- Zulässigkeit: für alle Merkmalstypen zulässig
- Bei intervallskalierten Merkmalen Bezeichnung als modale Klasse
- Unimodale Verteilungen besitzen einen Modus.
- Bimodale Verteilungen besitzen zwei Modi.

Beispiel 7.7. Kariöse Zähne von Kindern
Fortsetzung von Beispiel 6.1, Seite 61
Wie aus Tabelle 6.1 (Seite 62) ersichtlich, ist die größte vorkommende Häufigkeit 42. Demnach ist der Modus dieser Verteilung ein kariöser Zahn, es liegt eine unimodale Verteilung vor. Die meisten Kinder haben genau einen kariösen Zahn.

Beispiel 7.8. Körpergröße von Studierenden
Fortsetzung von Beispiel 6.2, Seite 69
Aus Tabelle 6.2 (Seite 69) kann die größte Häufigkeit in diesem Beispiel abgelesen werden. Die meisten Personen sind zwar größer als 160 cm, aber höchstens 165 cm groß.

7.1.4 Geometrisches Mittel

Das geometrische Mittel verwendet man zur Durchschnittsberechnung von Wachstumsfaktoren. Ein Wachstumsfaktor gibt an, um welchen Faktor sich eine interessierende Größe, beispielsweise der Wert einer Aktie, geändert hat.

Bezeichnungen

I_t	Wert der betrachteten Größe zur Zeit t
I_{t-1}	Wert zur Zeit t-1, also zur Vorperiode
g_t	Wachstumsfaktor zur Zeit t
p_t	Wachstumsrate zur Zeit t

Ein **Wachstumsfaktor** zur Zeit t errechnet sich als Quotient aus den Werten der jeweiligen Periode und der Vorperiode:

$$g_t = \frac{I_t}{I_{t-1}}$$

Die **Wachstumsrate** ist das prozentuelle Wachstum:

$$p_t = (g_t - 1) \cdot 100\%$$

Geometrisches Mittel

Der Durchschnitt von k aufeinander folgenden Wachstumsfaktoren $g_1, g_2, \ldots g_k$ ist

$$g = \sqrt[k]{g_1 \cdot g_2 \cdot \ldots \cdot g_k} = \sqrt[k]{I_k / I_0}$$

und heißt das **geometrische Mittel** dieser Wachstumsfaktoren.

Der Wert

$$p = (g - 1) \cdot 100\%$$

ist die **durchschnittliche Wachstumsrate** in Prozent.

Beispiel 7.9. Aktienrenditen
Eine Aktie mit dem Wert 100 steigt im ersten Jahr um 50%, im zweiten Jahr fällt sie um 50%. Wie hoch ist die durchschnittliche Wachstumsrate pro Jahr?

Ausgehend vom Wert 100 hat die Aktie nach dem ersten Jahr den Wert 150 und nach dem zweiten Jahr den Wert 75. Damit gilt:

$$g_1 = \frac{150}{100} = 1,5 \qquad g_2 = \frac{75}{150} = 0,5$$

und damit

$$g = \sqrt{1,5 \cdot 0,5} = \sqrt{0,75} \approx 0,866$$

Die durchschnittliche Wachstumsrate in Prozent $p = (0,866 - 1) \cdot 100\% = -13,4\%$ besagt, dass der durchschnittliche Wertverlust der Aktie 13,4% pro Jahr beträgt.

Die Berechnung mittels arithmetischem Mittel würde in Beispiel 7.9 zu einem völlig falschen Ergebnis führen. Der Grund dafür liegt in der Tatsache, dass bei der Verzinsung im zweiten Jahr nicht nur das Kapital verzinst wird, sondern auch die bereits angefallenen Zinsen. Diese Zinseszinsen sind in der Durchschnittsberechnung zu berücksichtigen, was das arithmetische Mittel im Gegensatz zum geometrischen Mittel nicht leisten kann. Aus diesem Grund sind Durchschnittswerte, die einen Zeitverlauf und damit das Problem der Zinseszinsen beinhalten, generell mit dem geometrischen Mittel zu berechnen. Dieses Problem kann auch bei Fragestellungen auftreten, die keine monetären Größen betrachten, wie das folgende Beispiel zeigt:

Beispiel 7.10. Bevölkerungswachstum
Man möchte berechnen, wie hoch die durchschnittliche Wachstumsrate in der österreichischen Bevölkerung ist. Dazu wurden die Bevölkerungsdaten der Jahre 1998 bis 2003 erhoben (Quelle: Statistik Austria, Statistisches Jahrbuch 2005, Seite 40) und die zugehörigen Wachstumsfaktoren berechnet.

Tabelle 7.1. Bevölkerung in Österreich

Jahr	Bevölkerung	Wachstumsfaktor
1998	7.976.789	
1999	7.992.323	1,0019
2000	8.011.566	1,0024
2001	8.043.046	1,0039
2002	8.083.797	1,0051
2003	8.117.754	1,0042

Als geometrisches Mittel der Wachstumsfaktoren erhält man

$$g = \sqrt[5]{1,0019 \cdot 1,0024 \cdot \ldots \cdot 1,0042} = \sqrt[5]{1,0177} = 1,0035$$

Die durchschnittliche Wachstumsrate errechnet man mit

$$p = (1,0035 - 1) \cdot 100\% = 0,35\%$$

Die Zahl der österreichische Bevölkerung ist in den Jahren 1998 bis 2003 durchschnittlich um 3,5 Promille pro Jahr gestiegen.

7.1.5 Quantile

Quantile sind Ausprägungen von quantitativen Variablen, die geordnete Datenreihen in Gruppen unterteilen, so dass ein bestimmter Prozentsatz über und ein bestimmter Prozentsatz unter dem Quantil liegt. Quantile werden auch als Perzentile oder Fraktile bezeichnet.

Das α-Quantil ist jener Wert $\widetilde{x}_\alpha$, für den mindestens der Anteil α der Daten kleiner oder gleich x_α und mindestens der Anteil $1 - \alpha$ der Daten größer oder gleich $\widetilde{x}_\alpha$ ist. Zusammen mit der Definition der Verteilungsfunktion ergibt sich daher das α-Quantil als jene Ausprägung, dessen Verteilungsfunktion gerade α ist.

Quartile
Quartile sind besondere Quantilswerte, weil sie die Daten in vier gleichgroße Gruppen teilen. Das untere Quartil ist somit nichts anderes als ein 0,25-Quantil, das mittlere Quartil entspricht dem Median bzw. dem 0,5-Quantil und das obere Quartil ist das 0,75-Quantil.

In Analogie zum Median lauten die Berechnungsvorschriften für Quantile:

α-Quantil

Der Wert

$$\widetilde{x}_\alpha = \begin{cases} x_{(k)} & \text{wenn } N \cdot \alpha \text{ keine ganze Zahl ist} \\ & \text{k ist dann die auf } N \cdot \alpha \text{ folgende ganze Zahl} \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right) & \text{wenn } N \cdot \alpha \text{ eine ganze Zahl ist} \\ & \text{dann ist } k = N \cdot \alpha \end{cases}$$

heißt α-Quantil des Merkmals X, wobei $x_{(k)}$ die Ausprägung an der k-ten Stelle der geordneten Datenreihe bezeichnet.

Berechnung des α-Quantils bei intervallskalierten Merkmalen

Das Intervall $(e_{i-1}; e_i]$ sei jenes Intervall, bei dem der Wert der Verteilungsfunktion an der Obergrenze des Intervalls das erste Mal den Wert α erreicht oder überschreitet. Dann ist das α-Quantil gegeben durch

$$\widetilde{x}_\alpha = (\alpha - p(x \leq e_{i-1})) \cdot \frac{e_i - e_{i-1}}{p_i} + e_{i-1}$$

Beispiel 7.11. Kariöse Zähne von Kindern
Fortsetzung von Beispiel 6.1, Seite 61
Wieviele kariöse Zähnen haben 25% der Kinder höchstens? Zur Beantwortung dieser Frage ist das 0,25-Quantil bei N = 140 zu berechnen.

$$\widetilde{x}_{0,25} = \frac{1}{2}(x_{(35)} + x_{(36)}) = \frac{1}{2}(0 + 0) = 0$$

Demnach haben 25% der Kinder (höchstens) keinen kariösen Zahn.

Beispiel 7.12. Körpergröße von Studierenden
Fortsetzung von Beispiel 6.2, Seite 69
Welche Körpergröße wird von 25% der Personen überschritten? Hier muss das 0,75-Quantil berechnet werden, weil diese Frage äquivalent ist zu der Frage, welche Körpergröße von 75% der Personen höchstens erreicht wird. Die Verteilungsfunktion übersteigt den Wert 0,75 zum ersten Mal an der Obergrenze des Intervalls $(175; 180]$. Dieses Intervall bildet unseren Ausgangspunkt. Daher sind folgende Werte zu verwenden:

$$p(x \leq e_{i-1}) = 0,723 \qquad e_{i-1} = 175 \qquad e_i = 180 \qquad p_i = 0,108$$

Damit ergibt sich:

$$\widetilde{x}_{0,75} = (0,75 - 0,723) \cdot \frac{180 - 175}{0,108} + 175 \approx 176,3$$

Damit sind 75% der untersuchten Personen höchstens 176,3 cm groß und 25% der Personen sind größer.

7.1.6 Lagekennzahlen in EXCEL

EXCEL stellt für die Berechnung von Lagekennzahlen eigene Funktionen zur Verfügung. Diese Funktionen benötigen nur den Bezug zum Datensatz, lediglich beim Quantil muss zusätzlich noch der Wert für α eingegeben werden. Berechnet man den Modus einer Verteilung, die mehrere Modi aufweist, so

wird als Ergebnis derjenige Modus ausgegeben, der zuerst im Datensatz gefunden wird. Die Berechnung des Quantils erfolgt in EXCEL mit Interpolation, allerdings wird nicht dokumentiert, wie interpoliert wird. Die Funktion zur Berechnung des Quantils sollte nur bei einem ausreichend großen Datensatz verwendet werden, weil in diesem Fall der Interpolationsfehler relativ gering ist. Bei Quantilen bzw. beim Modus können daher in EXCEL Ergebnisse auftreten, die von anders ermittelten Ergebnissen (händisch oder SPSS) abweichen. Die richtigen Modi lassen sich am besten aus der Häufigkeitstabelle ablesen, und die Quantile bei kleinen Datensätzen (bis n = 50) sollte man händisch ermitteln.

Sind die Daten nicht als Urliste, sondern als Häufigkeitsverteilung vorhanden, so müssen die Kennzahlen in Anlehnung an die zugrunde liegenden Formeln berechnet bzw. erhoben werden. Die EXCEL-Funktionen liefern nur bei Vorliegen der Urliste richtige Ergebnisse.

Beispiel 7.13. Kariöse Zähne von Kindern
Fortsetzung von Beispiel 6.1, Seite 61
Die Berechnung von Mittelwert, Median, Modus und 0,25-Quantil erfolgt über die Anweisungen

=MITTELWERT('Datensatz EXCEL'!$A$1:$T$7)

=MEDIAN('Datensatz EXCEL'!$A$1:$T$7)

=MODALWERT('Datensatz EXCEL'!$A$1:$T$7)

=QUANTIL('Datensatz EXCEL'!$A$1:$T$7;0,25)

Die Ergebnisse sind $\bar{x} = 1,7$; $\widetilde{x}_{0,5} = 1$; $x_{\text{mod}} = 1$ und $\widetilde{x}_{0,25} = 0$.

Beispiel 7.14. Körpergröße von Studierenden
Fortsetzung von 6.2, Seite 69
Die Berechnungen von Mittelwert, Median, Modus und 0,25-Quantil über die Anweisungen mit Bezug auf den Datensatz liefern die Ergebnisse $\bar{x} = 171,6\,cm$, $\widetilde{x}_{0,5} = 170\,cm$, $x_{\text{mod}} = 170\,cm$, $\widetilde{x}_{0,75} = 176\,cm$. Diese Ergebnisse weichen von unseren händisch ermittelten Ergebnissen deswegen ab, weil wir für die händische Berechnung immer auf die intervallskalierten Daten zurückgegriffen haben, für die EXCEL-Anweisungen aber den Originaldatensatz verwenden.

Beispiel 7.15. Bevölkerungswachstum
Fortsetzung von 7.10, Seite 103
Die Berechnung des geometrischen Mittels erfolgt über die Anweisung

=GEOMITTEL(1,0019;1,0024;1,0039;1,0051;1,0042)

und liefert das Ergebnis $g = 1,0035$. Die Umrechnung des durchschnittlichen Wachstumsfaktors auf das durchschnittliche prozentuelle Wachstum muss zusätzlich durchgeführt werden.

Lagemaße mit EXCEL

MITTELWERT(Datenbezug)

MEDIAN(Datenbezug)

MODALWERT(Datenbezug)

QUANTIL(Datenbezug; α) mit $0 \leq \alpha \leq 1$

GEOMITTEL(Datenbezug)

7.1.7 Lagekennzahlen in SPSS

Unter *Analysieren* → *Deskriptive Statistiken* → *Häufigkeiten* findet man in der Option *Statistik* die verschiedenen Kennzahlen. Bei den Lagemaßen im rechten Fensterbereich können *Mittelwert*, *Median* und *Modalwert* ausgewählt werden. Das Quantil kann im linken Fensterbereich über *Perzentile* angefordert werden, wobei die Eingabe für α in Prozent erfolgt, also eine Zahl zwischen 0 und 100 ist. Berechnet man den Modus einer Verteilung mit mehreren Modi, so wird als Ergebnis der kleinste Modus ausgegeben. Im Bedarfsfall wird auf diese Vorgehensweise hingewiesen. Bei kleinen Datensätzen kann die Berechnung des Quantils in SPSS zu Ergebnissen führen, die von den händisch ermittelten Ergebnissen abweichen. Dies liegt daran, dass SPSS für die Berechnung des Quantils interpoliert, allerdings wird diese Interpolation anders durchgeführt als in EXCEL. Daher ist für kleine Datensätze (bis n = 50) die händische Ermittlung des Quantils zu empfehlen.

Das geometrische Mittel findet man unter dem Menüpunkt *Analysieren* → *Berichte* → *Fälle zusammenfassen* ebenfalls in der Option *Statistik*, dabei ist zu beachten dass die Daten Wachstumsfaktoren beinhalten müssen. Das Ergebnis zeigt den durchschnittlichen Wachstumsfaktor, den man in das prozentuelle Wachstum umrechnen muss.

Lagemaße mit SPSS

- *Analysieren* → *Deskriptive Statistiken* → *Häufigkeiten*, Option *Statistik* ermöglicht die Berechnung von Mittelwert, Median, Modus und Quantilen
- α-Quantil = 100α-Perzentil
- *Analysieren* → *Berichte* → *Fälle zusammenfassen*, Option *Statistik* ermöglicht die Berechnung des geometrischen Mittels

7.2 Streuungsmaße

Mit Hilfe der Lagemaßzahlen haben wir versucht, den gesamten Datensatz durch eine einzige Kennzahl zu beschreiben. Dadurch bleiben oft sehr wichtige Informationen unberücksichtigt. Angenommen wir hätten einerseits einen Datensatz bestehend aus den Ausprägungen 1 und 9. Die Berechnung des Mittelwertes ergibt als durchschnittliche Ausprägung 5. Ein anderer Datensatz mit den Ausprägungen 5 und 5 würde zum gleichen Mittelwert führen, obwohl der Datensatz ganz anders ist als der erste.

Streuungsmaße sollen ausdrücken, wie weit die einzelnen Datenpunkte voneinander bzw. von einem festen Bezugspunkt entfernt sind.

Spannweite

Liegen N Messdaten $x_1, x_2, \ldots, x_N$ zum Merkmal X vor, dann ist der Abstand zwischen der größten und der kleinsten vorkommenden Ausprägung

$$R = x_{max} - x_{min}$$

die Spannweite von x.

Eine Spannweite ist für metrische Merkmale geeignet, nicht aber für nominale oder ordinale Merkmale.

Die Spannweite ist einfach zu berechnen, wird aber in der Praxis sehr selten verwendet. Im Normalfall werden eher Minimum und Maximum der Ausprägungen getrennt angegeben.

Die wichtigste Streuungskennzahl ist die Varianz, die das arithmetische Mittel der quadrierten Abstände der Datenpunkte zum Mittelwert ist. Ausgehend von der Varianz werden weitere Streuungsmaße wie die Standardabweichung oder der Variationskoeffizient berechnet. Varianz, Standardabweichung und Variationskoeffizient sind ausschließlich für metrische Merkmale geeignet. Dies ist unmittelbar einsichtig, weil diese Kennzahlen auf das arithmetische Mittel der Daten zurückgreifen.

Liegt ein Merkmal in Intervallskalierung vor, so werden wie beim Berechnen des arithmetischen Mittels die Intervallmitten als Ausprägungen x_i verwendet.

Varianz

Liegen N Messdaten zum Merkmal X vor, dann ist

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{7.3}$$

die Varianz des Merkmals X.

Liegt eine Häufigkeitsverteilung zum Merkmal X mit r verschiedenen Ausprägungen vor, dann ist

$$s^2 = \frac{1}{N}\sum_{i=1}^{r}(x_i - \bar{x})^2 h_i = \sum_{i=1}^{r}(x_i - \bar{x})^2 p_i \tag{7.4}$$

Auch an diesen Formeln kann man die inhaltliche Bedeutung der Varianz als mittlere quadratische Abweichung der Daten vom Mittelwert erkennen, wenn man die Formeln zur Varianzberechnung denen für die Mittelwertsberechnung gegenüberstellt.

Standardabweichung, Variationskoeffizient

$s = +\sqrt{s^2}$ ist die Standardabweichung von X.

$V = \frac{s}{\bar{x}}$ ist der Variationskoeffizient von X.

Liegt ein Merkmal mit physikalischen Einheiten vor (z.B. Körpergröße in cm) so ist bei der Berechnung der Streuungsparameter folgendes zu beachten: Die Maßeinheit der Varianz ist quadratisch (cm^2), die Standardabweichung und die Spannweite werden in der gleichen Maßeinheit wie die Messwerte angegeben, der Variationskoeffizient besitzt keine Maßeinheit, ist also dimensionslos. Der Variationskoeffizient setzt die Streuung in Relation zum Mittelwert und wird daher auch als relatives Streuungsmaß bezeichnet. Der Variationskoeffizient ist insbesondere zum Vergleich zweier Verteilungen geeignet, die in unterschiedlichen Dimensionen gemessen wurden (z.B. Preise in unterschiedlichen Währungen).

Beispiel 7.16. Kariöse Zähne von Kindern
Fortsetzung von Beispiel 6.1, Seite 61 und 106
Der Mittelwert betrug 1,7, für die Berechnung der Varianz verwenden wir Formel (7.4).

Tabelle 7.2. Varianzberechnung zu Beispiel 6.1

x_i	p_i	$p_i(x_i - \bar{x})^2$
0	0,293	0,846
1	0,300	0,147
2	0,143	0,013
3	0,114	0,193
4	0,050	0,265
5	0,057	0,622
6	0,014	0,264
7	0,029	0,803
Summe	1,000	3,153

Die Varianz beträgt demnach 3,153, daraus ergibt sich die Standardabweichung 1,776 und der Variationskoeffizient von 1,044. Daraus kann lediglich abgelesen werden, dass die Daten vom Mittelwert abweichen, eine detailliertere Interpretation ist vorerst nicht möglich.

Beispiel 7.17. Körpergröße von Studierenden
Fortsetzung von 6.2, Seite 69 und 97
Als Mittelwert wurde 170,7 cm (händisch) berechnet.

Tabelle 7.3. Varianzberechnung zu Beispiel 6.2

Intervallmitte x_i	p_i	$p_i(x_i - \bar{x})^2$
157,7	0,062	10,773
162,5	0,246	16,676
167,5	0,215	2,248
172,5	0,200	0,626
177,5	0,108	4,935
182,5	0,108	14,917
187,5	0,031	8,653
192,5	0,031	14,582
Summe	≈ 1	73,408

Die Varianz beträgt demnach 73,4 cm^2, daraus ergibt sich die Standardabweichung 8,6 cm und der Variationskoeffizient von 0,05.

Für die praktische Berechnung der Varianz wird meist auf die Formel

$$s^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2$$

zurückgegriffen, die sich durch einfache Umformung aus (7.3) ergibt:

$$\begin{aligned} s^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \\ &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - 2\frac{1}{N}\sum_{i=1}^{N} x_i\bar{x} + \frac{1}{N}\sum_{i=1}^{N}\bar{x}^2 = \\ &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - 2\bar{x}\frac{1}{N}\sum_{i=1}^{N} x_i + \bar{x}^2 = \\ &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2 \end{aligned}$$

7.3 Eigenschaften von Lage- und Streuungsmaßen

Die Auswahl geeigneter Lage- und Streuungsmaßzahlen hängt in erster Linie vom Skalenniveau der betrachteten Merkmale ab. In manchen Situationen sind aber nicht alle zulässigen Kennzahlen auch wirklich zweckmäßig. Die Kenntnis über wichtige Eigenschaften der Maßzahlen erleichtert die Entscheidung, welche Kennzahlen für ein vorliegendes Problem adäquat sind.

7.3.1 Maßeinheiten

Lagekennzahlen werden in der gleichen Maßeinheit angegeben wie die Ausprägungen. Auch die Standardabweichung und die Spannweite sind in der gleichen Einheit, die Varianz weist immer quadratische Einheit auf und der Variationskoeffizient ist dimensionslos.
In manchen Fällen soll die Maßeinheit der Merkmalsausprägungen geändert werden, z.B. um die Temperatur in Grad Fahrenheit statt in Grad Celsius anzugeben. Alle Ausprägungen sind für diese konkrete Änderung der Maßeinheit gemäß

$$^\circ F = 1,8 \cdot {}^\circ C + 32$$

umzurechnen.

Alle Lagemaße ändern sich in gleicher Weise wie die einzelnen Ausprägungen, d.h. auch die ursprünglichen Lagemaße können einfach transformiert werden. Die Auswirkung auf die Streuungsmaßzahlen ist unterschiedlich. Für Varianz und Standardabweichung bleibt eine Addition eines Summanden unberücksichtigt, ein Faktor bleibt für die Standardabweichung als Faktor vorhanden, für die Varianz wird der Faktor quadriert.

Transformationsregeln

X und Y seien zwei Merkmale, und sei $Y = a \cdot X + b$ eine lineare Transformation von X dann gilt:

$$\bar{y} = a \cdot \bar{x} + b$$
$$\widetilde{y}_{0,5} = a \cdot \widetilde{x}_{0,5} + b$$
$$y_{\text{mod}} = a \cdot x_{\text{mod}} + b$$
$$s_y^2 = a^2 \cdot s_x^2$$
$$s_y = |a| \cdot s_x$$

Für $b = 0$ gilt weiters

$$|V_y| = |V_x|$$

Betrachtet man die Umkehrung dieser Transformation, so lässt sich auch das Merkmal X als lineare Transformation von Y darstellen:

$$X = \frac{Y - b}{a}$$

Wählt man nun für a die Standardabweichung s_Y und für b den Mittelwert $\bar{y}$, so weist das transformierte Merkmal X den Mittelwert 0 und die Standardabweichung 1 auf. Man nennt diese spezielle lineare Transformation die **Standardisierung** des Merkmals Y.

7.3.2 Minimaleigenschaften

Der Mittelwert ist jener Wert a, für den die Summe der quadratischen Abweichungen der Datenpunkte zu diesem Wert $s^2(a) = \sum_{i=1}^{N}(x_i - a)^2$ am kleinsten ist.

Der Median ist jener Wert a, für den die Summe der Abweichungsbeträge $D(a) = \sum_{i=1}^{N} |x_i - a|$ minimal wird.

7.3.3 Robustheit

Der Mittelwert reagiert sehr empfindlich auf so genannte Ausreißer. Ein Ausreißer ist eine Ausprägung, die auffällig weit von allen anderen Ausprägungen abweicht. Manchmal sind solche Ausreißer auf einen Fehler in der Datenerhebung zurückzuführen. Es kann aber genauso gut sein, dass dieser Ausreißer zu Recht in den Daten vorhanden ist.
Wesentlich unempfindlicher gegenüber solchen Ausreißern sind der Modus und der Median, man nennt diese Lagekennzahlen daher robust.
Die Standardabweichung, die Varianz und die Spannweite reagieren wegen ihrer Abhängigkeit vom Mittelwert ebenfalls sehr empfindlich auf Ausreißer.

7.4 Auswahl geeigneter Lagemaßzahlen

Zur Auswahl geeigneter Lagemaßzahlen bestimmt man zunächst die zulässigen Kennzahlen, die sich aus dem Skalenniveau des Merkmals ergeben. Damit können für nominale Merkmale der Modus, für ordinale Merkmale der Modus, der Median und andere Quantile und für metrische Merkmale alle Lagemaßzahlen berechnet werden. Aus der Fragestellung geht hervor, ob ein arithmetisches oder ein geometrisches Mittel (Berücksichtigung von Zinseszinsen, vgl. Kapitel 7) zu berechnen ist.

Nicht immer sind alle zulässigen Kennzahlen auch tatsächlich geeignet, um Informationen über den Datensatz komprimiert darzustellen. Dies kann am Datensatz liegen, wenn beispielsweise einzelne Ausreißer das arithmetische Mittel unbrauchbar machen, kann aber auch andere Ursachen haben. Der Modus aus einem Datensatz von Körpergrößen ist sicherlich eine zulässige aber ziemlich unnütze Kennzahl, sofern nicht die modale Klasse des intervallskalierten Merkmals bestimmt wird.

Es ist unmöglich, eine genaue Anweisung zu geben, in welchen Fällen welche Kennzahlen sinnvoll sind. Damit bleibt es den LeserInnen überlassen, an geeigneter Stelle über die Aussagekraft und Zweckmäßigkeit verschiedener Kennzahlen nachzudenken.

7.5 Maßzahlen der Schiefe und Wölbung

Die Gestalt einer Verteilung kann zusätzlich über die Schiefe oder die Wölbung einer Verteilung beschrieben werden. Zur Berechnung geeigneter Kennzahlen benötigt man die Momente dritter und vierter Ordnung.

Momente der Ordnung k in Bezug auf den Punkt a

Liegen N Messdaten $x_1, x_2, \ldots, x_N$ zum Merkmal X vor, dann nennt man die Größen

$$m_k(a) = \frac{1}{N}\sum_{i=1}^{N}(x_i - a)^k \qquad \text{für} \qquad k = 1, 2, 3, \ldots$$

die Momente der Ordnung k in Bezug auf den Punkt a.

Ist $a = 0$, so nennt man diese Größen **gewöhnliche Momente**, ist $a = \bar{x}$, so nennt an sie **zentrale Momente**.

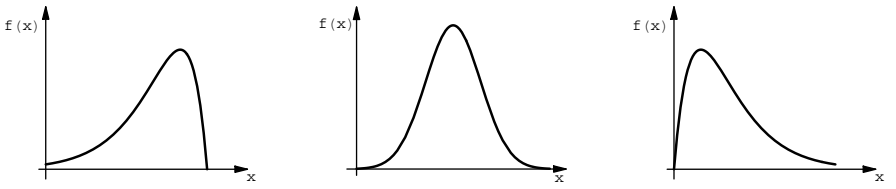
Der Mittelwert entspricht nach dieser Definition dem gewöhnlichen Moment erster Ordnung, die Varianz dem zentralen Moment zweiter Ordnung:

$$\bar{x} = m_1(0) \qquad s^2 = m_2(\bar{x})$$

Die Schiefe einer Verteilung lässt sich über den Momentenkoeffizienten bestimmen:

$$\alpha = \frac{m_3(\bar{x})}{s^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{s^3}$$

Ist $\alpha = 0$, so spricht man von einer um $\bar{x}$ symmetrischen Verteilung. Ist α positiv, liegt eine rechtsschiefe Verteilung vor, bei negativem α haben wir eine linksschiefe Verteilung.



linksschief $\alpha < 0$ | symmetrisch $\alpha = 0$ | rechtsschief $\alpha > 0$

Abb. 7.1. Schiefe einer Verteilung

Das Schiefemaß ist wie der Variationskoeffizient dimensionslos.

Die Schiefe lässt sich auch an der Anordnung der Lagekennzahlen Mittelwert, Median und Modus ablesen: Für symmetrische Verteilungen fallen alle drei Kennzahlen zusammen, für linksschiefe Verteilungen gilt $x_{\text{mod}} > \widetilde{x}_{0.5} > \bar{x}$ und für rechtsschiefe Verteilungen entsprechend $x_{\text{mod}} < \widetilde{x}_{0.5} < \bar{x}$. Von diesen Ungleichungsketten muss zumindest eine Relation als echte Ungleichung erfüllt sein, für die zweite wäre auch eine Gleichung zulässig, also z.B. $x_{\text{mod}} = \widetilde{x}_{0.5} > \bar{x}$ für linksschiefe Verteilungen.

Schiefe

Liegen N Messdaten $x_1, x_2, \ldots, x_N$ zum Merkmal X vor, dann ist

$$\alpha = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{s^3}$$

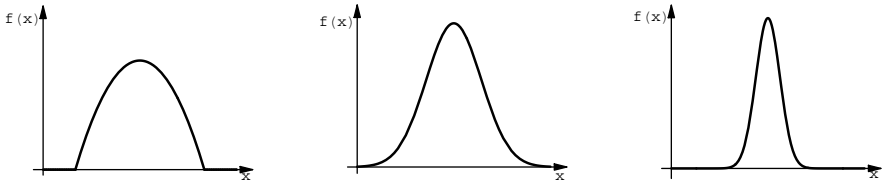
der Momentenkoeffizient der Schiefe dieser Verteilung.

$\alpha < 0 \ldots$ linksschiefe Verteilung
$\alpha = 0 \ldots$ symmetrische Verteilung
$\alpha > 0 \ldots$ rechtsschiefe Verteilung

Das letzte Gestaltmerkmal, dem wir uns zuwenden, ist die Wölbung. Auch der Wölbungskoeffizient wird mit Hilfe der Momente berechnet:

$$\gamma = \frac{m_4(\bar{x})}{s^4} - 3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{s^4} - 3$$

Dieser Wölbungskoeffizient trifft eine Aussage über die Wölbung der Verteilung im Vergleich zur Wölbung einer Normalverteilung (vgl. Kapitel 12.5).



breit gewölbt
$\gamma < 0$

normal gewölbt
$\gamma = 0$

schmal gewölbt
$\gamma > 0$

Abb. 7.2. Wölbung einer Verteilung

Demnach ist für $\gamma = 0$ die Verteilung gewölbt wie eine Normalverteilung, eine Verteilung mit positivem γ ist stärker gewölbt (also spitzer und schmäler) als die Normalverteilung und eine Verteilung mit negativem γ ist weniger gewölbt (also flacher und breiter) als die Normalverteilung.

Wölbung, Kurtosis

Liegen N Messdaten $x_1, x_2, \ldots, x_N$ zum Merkmal X vor, dann ist

$$\gamma = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{s^4} - 3$$

der Wölbungskoeffizient dieser Verteilung.

$\gamma < 0 \ldots$ breit gewölbt
$\gamma = 0 \ldots$ normal gewölbt
$\gamma > 0 \ldots$ spitz gewölbt

Beispiel 7.18. Kariöse Zähne von Kindern
Fortsetzung von Beispiel 6.1, Seite 61, 106 und 109
Die Schiefe und die Wölbung dieser Verteilung ist zu berechnen und zu interpretieren. Bisherige Berechnungen ergaben $\widetilde{x}_{0.5} = 1$, $\bar{x} = 1,7$, $s^2 = 3,153$ und $s = 1,776$.

Tabelle 7.4. Momente zu Beispiel 6.1

x_i	p_i	$p_i(x_i - \bar{x})^3$	$p_i(x_i - \bar{x})^4$
0	0,293	-1,439	2,446
1	0,300	-0,103	0,072
2	0,143	0,004	0,001
3	0,114	0,251	0,326
4	0,050	0,608	1,399
5	0,057	2,054	6,777
6	0,014	1,136	4,884
7	0,029	4,254	22,544
Summe	1,000	6,765	38,450

Als Momentenkoeffizient der Schiefe ergibt sich $\alpha = 1,2$. Demnach liegt eine rechtsschiefe Verteilung vor. Die Rechtsschiefe hätte man auch an der Tatsache ablesen können, dass der Median kleiner als der Mittelwert ist. Der Wölbungskoeffizient beträgt $\gamma = 0,868$. Somit liegt eine Verteilung vor, die stärker gewölbt (also spitzer) ist als die Normalverteilung. Aus dem Stabdia-

gramm in Abbildung 6.4 (Seite 72) ist die Rechtsschiefe deutlich erkennbar, die Wölbung ist aus der Grafik leider nicht so leicht zu entnehmen.

7.6 Streuung, Schiefe und Wölbung in EXCEL

EXCEL stellt für die Berechnung dieser Kennzahlen eigene Funktionen zur Verfügung. Diese Funktionen benötigen nur den Bezug zum Datensatz. Sind die Daten nicht als Urliste vorhanden, sondern als Häufigkeitsverteilung, so müssen die Kennzahlen in Anlehnung an die zugrunde liegenden Formeln berechnet bzw. erhoben werden.

Beispiel 7.19. Kariöse Zähne von Kindern
Fortsetzung von Beispiel 6.1, Seite 61
Die Berechnung von Varianz, Standardabweichung, Schiefe und Wölbung erfolgt über die Anweisungen

=VARIANZEN('Datensatz EXCEL'!$A$1:$T$7)

=STABWN('Datensatz EXCEL'!$A$1:$T$7)

=SCHIEFE('Datensatz EXCEL'!$A$1:$T$7)

=KURT('Datensatz EXCEL'!$A$1:$T$7)

und man erhält die Ergebnisse $s^2 = 3,153$, $s = 1,776$, $\alpha = 1,221$ und $\gamma = 0,944$.

Für die Varianzberechnung stellt EXCEL die Funktionen *Varianz(Datensatz)*, *Varianza(Datensatz)*, *Varianzen(Datensatz)* und *Varianzena(Datensatz)* bereit. Die beiden Funktionen *Varianza(Datensatz)* und *Varianzena(Datensatz)* berücksichtigen in der Berechnung auch Wahrheitswerte, wobei *Wahr* als 1 und *Falsch* als 0 in die Berechnung eingeht. Auch als erfahrene Statistikerin ist mir die Sinnhaftigkeit einer solchen Berechnung bislang verborgen geblieben. Die Funktion *Varianzen(Datensatz)* berechnet die Varianz wie in Abschnitt 7.2 beschrieben, die Funktion *Varianz(Datensatz)* berechnet die sogenannte **korrigierte Varianz** gemäß

$$s^2_{\text{korr}} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{N}{N-1}s^2$$

Auf die Bedeutung der korrigierten Varianz werden wir in der schließenden Statistik zurückkommen. Auch die Funktionen *Schiefe* und *Kurt* liefern korrigierte und damit zahlenmäßig leicht abweichende Koeffizienten. Die exakten

Formeln für die korrigierten Kennzahlen können aus der EXCEL-Hilfe entnommen werden. Nachdem wir aber bei der Varianz nur Größenordnungen und bei Schiefe und Wölbung nur das Vorzeichen verwenden, sind die Abweichungen in den Ergebnissen unbedeutend.

Streuung, Schiefe und Wölbung in EXCEL

VARIANZEN(Datenbezug)	
VARIANZ(Datenbezug)	korrigierte Varianz
STABWN(Datenbezug)	
STABW(Datenbezug)	korrigierte Standardabweichung
SCHIEFE(Datenbezug)	korrigierte Schiefe
KURT(Datenbezug)	korrigierte Wölbung

7.7 Streuung, Schiefe und Wölbung in SPSS

Unter *Analysieren* → *Deskriptive Statistiken* → *Häufigkeiten* findet man in der Option *Statistik* die gewünschten Kennzahlen, wobei die Kurtosis nichts anderes als die Wölbung ist. Auch in SPSS werden die jeweils korrigierten Maßzahlen berechnet.

Streuung, Schiefe und Wölbung in SPSS

- *Analysieren* → *Deskriptive Statistiken* → *Häufigkeiten*, Option *Statistik* ermöglicht die Berechnung von Spannweite, Standardabweichung, Varianz, Schiefe und Wölbung (Kurtosis)
- Es werden die jeweils korrigierten Maßzahlen berechnet

Übungsaufgaben

7.1. Kariöse Zähne
Bei $N = 60$ Kindern einer Jahrgangsstufe werden die Zähne auf Karies untersucht, wobei sich folgende Ergebnisse einstellen:

0	1	1	3	0	1	1	1	3	1	0	1	4	2	0	3	1	1	7	2
0	5	2	3	0	2	4	0	1	1	3	0	6	2	1	5	1	1	2	2
1	0	1	0	0	4	0	3	1	1	7	2	1	0	3	0	1	3	2	2

a) Wie viele kariöse Zähne haben 50% der Kinder höchstens?
b) Wie viele kariöse Zähne haben 50% der Kinder mindestens?
c) Wie viele kariöse Zähne haben 10% der Kinder mindestens?
d) Berechnen Sie Mittelwert und Median.
e) Berechnen Sie Schiefe und Wölbung der Verteilung.

7.2. Kinderzahl
In einer Stadt wurde von der Abteilung für Soziales die Zahl der Kinder pro Haushalt erhoben:

Kinderzahl	Anzahl der Haushalte
0	110
1	140
2	90
3	45
4	10
5	5

a) Berechnen Sie Mittelwert, Modus und Median.
b) Berechnen Sie Varianz, Standardabweichung und Variationskoeffizient.
c) Berechnen Sie Schiefe und Wölbung.

7.3. Körpergröße von Studierenden
Bei $N = 60$ Studierenden einer Jahrgangsstufe wurde die Körpergröße in cm erhoben. Diese Erhebung führte zu folgendem Ergebnis:

178	182	166	162	181	168	170	164	171	170	165	164
169	165	165	175	180	164	188	170	194	171	185	168
180	174	183	193	172	178	165	162	174	174	162	163
172	176	168	160	170	170	171	166	165	160	175	183
189	162	168	160	178	175	168	158	172	163	172	183

a) Welche Größe haben 40% der Studierenden höchstens?
b) Welche Größe haben 70% der Studierenden mindestens?
c) Berechnen Sie Lage- und Streuungsmaßzahlen.

7.4. Altersverteilung
In der Bevölkerung wurde folgende Verteilung des Merkmals Alter erhoben:

Alter	Personen
$0 < x \leq 15$	1.333.505
$15 < x \leq 30$	1.495.740
$30 < x \leq 45$	2.002.259
$45 < x \leq 60$	1.526.110
$60 < x \leq 75$	1.151.122
$75 < x \leq 100$	609.018

a) Bestimmen Sie Mittelwert, Median und modale Klasse.
b) Berechnen Sie Varianz, Schiefe und Wölbung.
c) Berechnen Sie das 0,25- und das 0,75-Quantil der Altersverteilung.
d) Interpretieren Sie Ihre Ergebnisse.

7.5. TV-Geräte
Eine Untersuchung in 100 Haushalten ergab folgende Urliste über die Anzahl an TV-Geräten:

1	0	1	0	0	2	0	3	1	1	3	2	1	0	3	0	1	3	2	2
0	2	1	3	0	0	0	0	6	1	1	2	1	0	1	0	3	0	1	3
1	0	1	0	0	3	0	3	1	1	1	2	1	0	3	0	1	3	2	2
0	0	0	0	3	0	1	3	2	1	3	2	1	0	3	0	1	3	2	2
2	7	1	3	1	5	1	0	0	0	2	1	0	3	1	4	0	2	1	1

a) Berechnen Sie Mittelwert, Median und Modus.
b) Berechnen Sie Varianz, Schiefe und Wölbung.
c) Interpretieren Sie Ihre Ergebnisse.

7.6. Klausurergebnisse
Bei einer Statistikklausur wurden folgende Ergebnisse erreicht

Note	Studierende
1	15
2	25
3	25
4	35
5	40

Berechnen Sie alle zulässigen Lage- und Streuungsmaßzahlen.

8

Multivariate deskriptive Statistik

In den meisten Anwendungsfällen ist man nicht nur an einem Merkmal interessiert, sondern an mehreren Merkmalen und insbesondere an den Zusammenhängen zwischen diesen Merkmalen. Beispielsweise soll die Frage beantwortet werden, ob bei Kindern die sportliche Aktivität die Schlafdauer beeinflusst oder Ähnliches. Zu diesem Zweck müssen von den Erhebungseinheiten mehrere Merkmale gleichzeitig erhoben und einer gemeinsamen Analyse unterzogen werden. Im einfachsten Fall sollen zwei Merkmale gemeinsam analysiert werden.

8.1 Zweidimensionale Häufigkeitsverteilungen

Am besten lassen sich zweidimensionale Häufigkeitsverteilungen mittels Kontingenztabellen darstellen. Dazu ist es notwendig, dass die Merkmale nur wenige Ausprägungen besitzen. Dies kann durch Zusammenfassen von Ausprägungen immer erreicht werden.

Beispiel 8.1. Einfluss von Strategietraining
In einer Studie über N = 235 zufällig ausgewählte Führungskräfte wird der Einfluss von Strategietraining auf den Unternehmenserfolg untersucht. Das Ergebnis der Untersuchung kann aus folgender Tabelle entnommen werden:

	kein Erfolg	Erfolg	Summe
kein Training	40	75	115
mit Training	30	90	120
Summe	70	165	235

Folgende Fragen sollen beantwortet werden: Wie hoch war die Teilnahmequote am Training, wie hoch war die Erfolgsquote? Konnte durch die Teilnahme am Training die Erfolgsquote erhöht werden?

Bei einer zweidimensionalen Häufigkeitsverteilung mit den Merkmalen X und Y verwendet man folgende Bezeichnungen:

Bezeichnungen

h_{ij}	absolute Häufigkeit der Kombination X = i und Y = j
$p_{ij} = h_{ij}/N$	relative Häufigkeit der Kombination X = i und Y = j
$P_{ij} = p_{ij} \cdot 100$	relative Häufigkeit der Kombination X = i und Y = j in Prozent
$h_{i+}(p_{i+})$	Zeilensummen, Randhäufigkeiten des Merkmals X
$h_{+j}(p_{+j})$	Spaltensummen, Randhäufigkeiten des Merkmals Y

Damit weist die Kontingenztabelle zu Beispiel 8.1 folgende allgemeine Form auf:

Tabelle 8.1. Kontingenztabelle

	$Y = 1$	$Y = 2$	Summe
X = 1	h_{11}	h_{12}	h_{1+}
X = 2	h_{21}	h_{22}	h_{2+}
Summe	h_{+1}	h_{+2}	N

Beispiel 8.2. Einfluss von Strategietraining
(Fortsetzung von Beispiel 8.1) Die entsprechende Tabelle mit den relativen Häufigkeiten hilft bei der Beantwortung der Fragen.

	kein Erfolg	Erfolg	Summe
kein Training	0,170	0,319	0,489
mit Training	0,128	0,383	0,511
Summe	0,298	0,702	1,000

Insgesamt wurden 235 Personen in die Untersuchung einbezogen, davon haben 120 Personen ein Training absolviert. Die Teilnahmequote beträgt daher $120/235 \approx 0,511$, also etwa 51%. Insgesamt hatten 165 Personen Erfolg, damit beträgt die Erfolgsquote $165/235 \approx 0,702$, also etwa 70%. Vorerst offen bleibt die Frage, ob das Training die Erfolgschancen verbessert hat.
Aus der Tabelle können folgende weitere Informationen abgelesen werden: 17% der Führungskräfte haben nicht am Training teilgenommen und hatten keinen Erfolg. 31,9% der Personen haben zwar nicht am Training teilgenommen, hatten aber trotzdem Erfolg. 12,8% hatten trotz Trainingsteilnahme keinen Erfolg und 38,3% der Personen haben am Training teilgenommen und hatten Unternehmenserfolg.

8.2 Randverteilungen

Eine Randverteilung gibt Auskunft über die Verteilung eines Merkmals, ohne das andere Merkmal zu berücksichtigen. Liegt eine zweidimensionale Verteilung in Form einer Kontingenztabelle vor, können die Randverteilungen an den Zeilen- bzw. Spaltensummen abgelesen werden.

Beispiel 8.3. Einfluss von Strategietraining
(Fortsetzung von Beispiel 8.1) Die Randverteilung des Merkmals Training kann an den Zeilensummen abgelesen werden (in absoluter oder relativer Form), die Randverteilung des Merkmals Erfolg kann an den Spaltensummen abgelesen werden.

	kein Erfolg	Erfolg	Summe
kein Training	0,170	0,319	0,489
mit Training	0,128	0,383	0,511
Summe	0,298	0,702	1,000

Insgesamt haben 48,9% der Führungskräfte kein Training absolviert und 51,1% haben am Training teilgenommen. 70,2% der befragten Personen hatten Erfolg, der Rest hatte keinen Erfolg.

8.3 Bedingte Verteilung

Mit der zweidimensionalen Verteilung und den beiden Randverteilungen kann noch keine Aussage über den Zusammenhang getroffen werden, aber meist ist dieser Zusammenhang von großem Interesse. Bezogen auf Beispiel 8.1 ist die Kernfrage, ob die Trainingsteilnahme die Erfolgsquote erhöht hat. Man möchte wissen, ob die Erfolgsquoten der Trainingsteilnehmer höher ist als die Erfolgsquote der Trainingsverweigerer.

In statistischer Ausdrucksweise interessiert uns im Beispiel 8.1 die bedingte Verteilung des Merkmals Erfolg, gegeben das Merkmal Training. Wir berechnen die bedingte Verteilung des Merkmal Erfolges bei den Trainingsteilnehmern und bei den Trainingsverweigerern.

Randverteilung - bedingte Verteilung

Für Randverteilungen wird das andere Merkmal vollkommen ausgeblendet, für bedingte Verteilungen unterteilt man die Untersuchungsgesamtheit anhand der Ausprägungen des bedingenden Merkmals in mehrere Teilgesamtheiten und betrachtet dann diese Teilgesamtheiten getrennt.

Bezeichnung	
$h_{ij}/h_{i+} = p_{ij}/p_{i+}$	bedingte relative Häufigkeit der Ausprägung j des Merkmals Y bei gegebener Ausprägung i des Merkmals X

Beispiel 8.4. Einfluss von Strategietraining
(Fortsetzung von Beispiel 8.1) Die bedingten Verteilungen des Merkmals Erfolg bei den Teilnehmern und bei den Verweigerern lassen sich aus folgender Tabelle ablesen:

	kein Erfolg	Erfolg	Summe
kein Training	0,348	0,652	1,000
mit Training	0,250	0,750	1,000

Die Erfolgsquote in der Teilgesamtheit der Trainingsteilnehmer liegt wegen $90/120 = 0,75$ bei 75%, die Erfolgsquote der Trainingsverweigerer liegt hingegen bei ca. 65% ($75/115 = 0,652$). Daraus kann abgelesen werden, dass das Training die Erfolgsquote erhöht hat, dass es also einen Zusammenhang zwischen Training und Erfolg gibt.

In der Praxis liegt die Schwierigkeit meist in der Festlegung, welches der beiden Merkmale das Untersuchungsmerkmal und welches das bedingende Merkmal ist. Manchmal lässt sich diese Frage lösen, indem man sich überlegt welcher Art der Zusammenhang zwischen den Merkmalen ist. Lässt sich eindeutig feststellen, welches die unabhängige Variable (Ursache) und welches die abhängige Variable (Wirkung) ist, so ist leicht nachvollziehbar, dass die Ursache als bedingende Variable verwendet werden sollte und die Wirkung zu untersuchen ist. In unserem Beispiel ist die Ursache das Training und die Wirkung der Erfolg, daher sollte man den Erfolg unter dem Aspekt Training untersuchen.

Sind die bedingten Verteilungen erstellt, so kann man wie für alle anderen Verteilungen Lage- oder Streuungsmaßzahlen berechnen. Diese bedingten Maßzahlen treffen dann ebenfalls nur Aussagen über Teilgesamtheiten.

Mit bedingten Verteilungen lassen sich auch genderspezifische Fragestellungen untersuchen, z.B. um das mittlere Einkommen von Frauen und Männern zu vergleichen. Dazu wird die Verteilung des Merkmals Einkommen bei gegebenen Geschlecht erhoben, also die Einkommensverteilung der Frauen und jene der Männer, und für beide Verteilungen das arithmetische Mittel berechnet.

8.4 Maße für den Zusammenhang zweier Merkmale

Man kann, wie im vorhergehenden Abschnitt beschrieben, über die bedingten Verteilungen Erkenntnisse über den Zusammenhang von Merkmalen gewinnen. Wünschenswert wäre aber eine eigene Kennzahl, die eine Aussage über den Zusammenhang ermöglicht. Es gibt diese Kennzahlen, allerdings ist bei der Anwendung darauf zu achten, dass diese an das Skalenniveau der Variablen gebunden sind. Im Folgenden werden die Kennzahlen und ihre Aussagemöglichkeiten beschrieben.

8.4.1 Zusammenhang zweier nominaler Merkmale

Zur Messung des Zusammenhangs zwischen zwei nominalen Merkmalen kann das **Assoziationsmaß Chi-Quadrat χ^2** oder das **Cramersche Assoziationsmaß V** („Cramers V“) verwendet werden. Ausgangspunkt für beide Maßzahlen ist der Vergleich zwischen tatsächlich beobachteten Häufigkeiten und jenen Häufigkeiten, die man bei Unabhängigkeit der beiden Merkmale erwarten würde.

Bezeichnungen

h_{ij}^o ... beobachtete (= **o**bserved) absolute Häufigkeit der Kombination X = i und Y = j

h_{ij}^e ... bei Unabhängigkeit von X und Y erwartete (= **e**xpected) absolute Häufigkeit dieser Kombination

Dabei gilt $$h_{ij}^e = \frac{h_{i+} \cdot h_{+j}}{N}$$

Hinweis: Bei der Berechnung der erwarteten Häufigkeiten werden im Normalfall Dezimalzahlen entstehen. Zur weiteren Berechnung sollte man diese Dezimalzahlen verwenden und nicht die gerundeten Häufigkeiten, weil das Ergebnis andernfalls durch die entstehenden Rundungsfehler stark verzerrt wird.

Das **Assoziationsmaß Chi-Quadrat χ^2** mit

$$\chi^2 = \sum_i \sum_j \frac{(h_{ij}^o - h_{ij}^e)^2}{h_{ij}^e}$$

misst den Zusammenhang zwischen zwei nominalen Merkmalen.

Eigenschaften des Assoziationsmaßes Chi-Quadrat χ^2

Es gilt $\chi^2 \geq 0$

Interpretation:
$\chi^2 = 0$ kein Zusammenhang
$\chi^2 > 0$ Zusammenhang

Wie aus der Formel leicht nachvollziehbar gilt immer $\chi^2 \geq 0$. Der Fall $\chi^2 = 0$ kann nur dann auftreten, wenn die beobachteten Häufigkeiten den bei Unabhängigkeit erwarteten Häufigkeiten entsprechen. Dies ist gleichbedeutend damit, dass die Merkmale unabhängig sind, also keinen Zusammenhang aufweisen.

Das Assoziationsmaß kann effizienter mit der Formel

$$\chi^2 = N \cdot \left(\sum_i \sum_j \frac{{h_{ij}^o}^2}{h_{i+} \cdot h_{+j}} - 1 \right)$$

berechnet werden.

Der Betrag des Assoziationsmaßes χ^2 ist abhängig vom Untersuchungsumfang und der Anzahl der Ausprägungen. Man kann also nicht direkt von der Größe χ^2 auf die Stärke des Zusammenhanges schließen. Daher verwendet man zur Messung des Zusammenhanges das Cramersche Assoziationsmaß, welches auf χ^2 aufbaut.

Das **Cramersche Assoziationsmaß V**

$$V = \sqrt{\frac{\chi^2}{N \cdot (\min(r, s) - 1)}}$$

misst den Zusammenhang zwischen zwei nominalen Merkmalen.

r ... Anzahl der Merkmalsausprägungen von X.
s ... Anzahl der Merkmalsausprägungen von Y.

Es gilt $0 \leq V \leq 1$

Interpretationshilfe für das Cramersche Assoziationsmaß V

$V = 0$	kein Zusammenhang
$0 < V \leq 0,3$	schwacher Zusammenhang
$0,3 < V \leq 0,7$	mittlerer Zusammenhang
$0,7 < V < 1$	starker Zusammenhang
$V = 1$	vollständiger Zusammenhang

Die angegebenen Zahlenwerte für die Interpretation sollen als Hilfe dienen, sind aber lediglich als Anhaltspunkt, nicht als fixe Grenzen zu verstehen. Generell gilt: Je näher das Cramersche Assoziationsmaß bei 0 liegt, desto schwächer ist der Zusammenhang und je näher es bei 1 liegt, desto stärker ist der Zusammenhang.

Beispiel 8.5. Einfluss von Strategietraining
(Fortsetzung von Beispiel 8.1) Die bei Unabhängigkeit erwarteten Häufigkeiten sind:

	kein Erfolg	Erfolg	Summe
kein Training	34,3	80,7	115
mit Training	35,7	84,3	120
	70,0	165,0	235

Daraus ergibt sich:

$$\chi^2 = N \cdot \left(\sum\sum \frac{{h_{ij}^o}^2}{h_{i+} \cdot h_{+j}} - 1 \right)$$
$$= 235 \cdot \left(\frac{34,3^2}{115 \cdot 70} + \frac{80,7^2}{115 \cdot 165} + \frac{35,7^2}{120 \cdot 70} + \frac{84,3^2}{120 \cdot 165} - 1 \right) = 2,69$$

$$V = \sqrt{\frac{\chi^2}{N \cdot (\min(r,s) - 1)}} = \sqrt{\frac{2,69}{235 \cdot (\min(2,2) - 1)}} = \sqrt{\frac{2,69}{235}} = 0,107$$

Demnach besteht ein schwacher Zusammenhang zwischen Training und Erfolg.

Aus den Assoziationsmaßen ist nicht erkennbar, was Ursache und was Wirkung ist und auch eine nähere Beschreibung des Zusammenhanges (Teilnahme am Training bedeutet höhere Erfolgchancen) ist damit nicht möglich. Diese differenziertere Aussage kann lediglich anhand der bedingten Verteilungen getroffen werden.

8.4.2 Zusammenhang zweier ordinaler Merkmale

Zur Messung des Zusammenhanges zwischen zwei ordinalen Merkmalen werden den Ausprägungen aus der Urliste zuerst Rangzahlen zugeordnet, d.h. die kleinste Ausprägung erhält den Rang 1, die größte Ausprägung den Rang N. Vereinfachend gehen wir vorerst davon aus, dass keine Ausprägung zweimal vorkommt, dass also die Zuordnung von Rängen in eindeutiger Weise möglich ist (***„ohne Bindungen"***). Jede Erhebungseinheit weist somit zwei Ränge r_i und s_i hinsichtlich der beiden zu untersuchenden Merkmale auf. Als Kennzahl zur Berechnung des Zusammenhanges dient der Spearmansche Rangkorrelationskoeffizient.

Spearmansche Rangkorrelationskoeffizient ohne Bindungen

Der Spearmansche Rangkorrelationskoeffizient ρ_s wird berechnet mittels

$$\rho_s = 1 - \frac{6 \cdot \sum d_i^2}{N \cdot (N^2 - 1)}$$

r_i, s_i ... Ränge
d_i ... Rangzahlendifferenz $r_i - s_i$ der i-ten Erhebungseinheit

Für die Interpretation ist einerseits das Vorzeichen wichtig, andererseits der Betrag. Aus dem Vorzeichen ist die Richtung des Zusammenhanges ablesbar. Ein gleichsinniger Zusammenhang (eine niedrige Rangziffer hinsichtlich des einen Merkmals geht einher mit einer niedrigen Rangziffer des anderen Merkmals) führt auf einen positiven Rangkorrelationskoeffizienten, ein gegensinniger Zusammenhang (eine niedrige Rangziffer hinsichtlich des einen Merkmals geht einher mit einer hohen Rangziffer des anderen Merkmals) ergibt einen negativen Rangkorrelationskoeffizienten. Sind die Merkmale unabhängig, so erhält man einen Korrelationskoeffizienten von 0. Aus dem Betrag ist die Stärke des Zusammenhanges ablesbar, denn umso stärker der Zusammenhang, desto näher liegt der Betrag bei 1.

Spearmansche Rangkorrelationskoeffizient

Es gilt $-1 \leq \rho_s \leq 1$

Interpretation:

$\rho_s < 0$	gegensinniger Zusammenhang
$\rho_s = 0$	kein Zusammenhang
$\rho_s > 0$	gleichsinniger Zusammenhang

Je stärker der Zusammenhang, desto näher liegt $|\rho_s|$ bei 1.

Beispiel 8.6. Weinverkostung
Sechs Weine wurden von zwei Expertinnen nach ihrer Qualität geordnet.

Wein	A	B	C	D	E	F
Expertin 1	1	2	4	5	6	3
Expertin 2	1	3	4	6	5	2

Stimmen die Expertinnen in der Beurteilung weitgehend überein? Zur Beantwortung dieser Frage berechnen wir den Spearmanschen Rangkorrelationskoeffizienten.

Wein		A	B	C	D	E	F	Summe
Expertin 1	r_i	1	2	4	5	6	3	
Expertin 2	s_i	1	3	4	6	5	2	
	d_i	0	-1	0	-1	1	1	
	d_i^2	0	1	0	1	1	1	4

$$\rho_s = 1 - \frac{6 \cdot \sum d_i^2}{N \cdot (N^2 - 1)} = \rho_s = 1 - \frac{6 \cdot 4}{6 \cdot 35} = 0,886$$

Zwischen den beiden Reihungen besteht ein starker gleichsinniger Zusammenhang. Von einer Expertin als qualitativ hoch eingeschätzte Weine werden auch von der anderen Expertin als qualitativ hochwertig eingestuft, beide Expertinnen haben eine ähnliche Beurteilung der Weinqualität.

Liegen Bindungen vor, ist also eine Zuordnung von Rängen nicht in eindeutiger Weise möglich, so muss zur Berechnung des Spearmanschen Rangkorrelationskoeffizienten eine etwas aufwändigere Formel herangezogen werden.

Spearmansche Rangkorrelationskoeffizient mit Bindungen

Der Spearmansche Rangkorrelationskoeffizient ρ_s berechnet sich bei N Rangpaaren nach

$$\rho_s = \frac{\sum\limits_i (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum\limits_i (r_i - \bar{r})^2 \sum\limits_i (s_i - \bar{s})^2}}$$

r_i, s_i ... (Durchschnitts-)Ränge, $i = 1, \ldots, N$

$\bar{r} = \bar{s} = \frac{1}{N} \sum\limits_{i=1}^{N} r_i = \frac{1}{N} \sum\limits_{i=1}^{N} i = \frac{N+1}{2}$ ... mittlere Ränge

Weisen mehrere Erhebungseinheiten die gleiche Ausprägung auf, so werden Durchschnittsränge vergeben, die als arithmetisches Mittel der in Frage kommenden Ränge berechnet werden. Alle Erhebungseinheiten mit derselben Ausprägung erhalten somit denselben Rang, die Rangsumme über alle Erhebungseinheiten bleibt gleich.

Die Interpretation ist völlig analog zu dem Fall ohne Bindungen.

Beispiel 8.7. Weinverkostung
Sechs Weine wurden von zwei Expertinnen nach ihrer Qualität geordnet. Expertin 1 hat die Weine D und E gleich gut bewertet, aber beide Weine schlechter als alle anderen. Diese Weine wären demnach auf den Rängen 5 und 6, also erhalten beide Weine den Durchschnittsrang 5,5.

Wein	A	B	C	D	E	F
Expertin 1	1	2	4	5.5	5.5	3
Expertin 2	1	3	4	6	5	2

Stimmen die Expertinnen in der Beurteilung weitgehend überein? Zur Beantwortung dieser Frage berechnen wir den Spearmanschen Rangkorrelationskoeffizienten (für Merkmale mit Bindungen).
Mit $\bar{r} = \bar{s} = 3,5$ erhält man

$$\rho_s = \frac{16}{\sqrt{17 \cdot 17,5}} = 0,928$$

Zwischen den beiden Reihungen besteht ein starker gleichsinniger Zusammenhang. Von einer Expertin als qualitativ hoch eingeschätzte Weine werden auch von der zweiten Expertin tendenziell als qualitativ hochwertig eingestuft. Beide Expertinnen haben eine ähnliche Beurteilung der Weinqualität.

8.4.3 Zusammenhang zweier metrischer Merkmale

Zur Messung des Zusammenhanges zwischen zwei metrischen Merkmalen ist der Korrelationskoeffizient von Bravais-Pearson geeignet. Dieser wird kurz als Korrelationskoeffizient bezeichnet, falls aus dem Zusammenhang keine Verwechslung mit dem Rangkorrelationskoeffizienten möglich ist.

Ausgangspunkt zur Berechnung bildet die **Kovarianz**, die - wie der Name bereits andeutet - ähnlich wie die Varianz aufgebaut ist. Der Unterschied liegt darin, dass zur Berechnung der Varianz nur ein Merkmal herangezogen wird, zur Berechnung der **Ko**-varianz aber zwei. Man kann sich die Kovarianz quasi als zweidimensionales Streuungsmaß vorstellen.

Die geometrische Bedeutung der Kovarianz ist aus Abbildung 8.1 ersichtlich. Zu den zweidimensionalen Daten wird der Datenschwerpunkt berechnet, dessen Koordinaten die Mittelwerte der beiden Merkmale sind $(\bar{x}, \bar{y})$. Nun kann

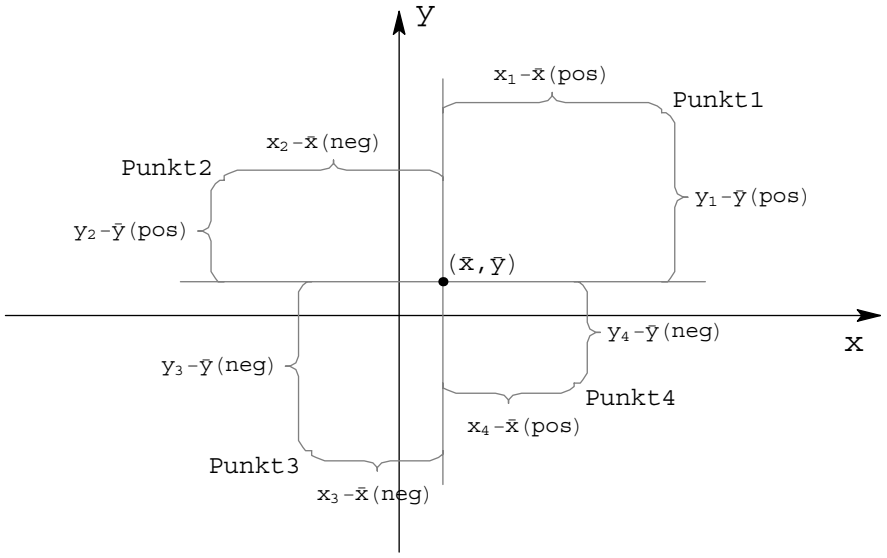


Abb. 8.1. Geometrische Darstellung der Kovarianz

zwischen jedem einzelnen Datenpunkt und dem Schwerpunkt ein Rechteck konstruiert werden. Die Kovarianz ist dann nichts anderes als das arithmetische Mittel der Rechtecksflächen, wobei je nach Vorzeichen der Abweichungen diese Flächen auch mit negativem Vorzeichen in die Mittelwertsberechnung eingehen können. Die Flächen der Punkte 1 und 3 würden in die Berechnung der Kovarianz mit positivem Vorzeichen einfließen, die der Punkte 2 und 4 mit negativem Vorzeichen.

Kovarianz

Liegen zu den Merkmalen X und Y zweidimensionale, metrische Daten $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ vor, dann ist

$$s_{XY} = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$
$$= \frac{1}{N} \cdot \sum_{i=1}^{N} x_i \cdot y_i - \bar{x} \cdot \bar{y}$$

die Kovarianz zwischen den Merkmalen X und Y.

Es gilt $-\infty \leq s_{XY} \leq +\infty$

Aus der Kovarianz können folgende Informationen abgelesen werden:

- Sind die Merkmale X und Y unabhängig, so ist die Kovarianz gleich Null.
- Ein gegensinniger Zusammenhang zwischen den Merkmalen X und Y führt zu einem negativen Vorzeichen, ein gleichsinniger Zusammenhang führt zu einem positiven Vorzeichen.

Die Stärke des Zusammenhanges kann aus der Kovarianz nicht abgelesen werden. Diese lässt sich durch die Berechnung des Korrelationskoeffizienten ermitteln.

Bravais-Pearson-Korrelationskoeffizient

Der Korrelationskoeffizient zur Messung des **linearen** Zusammenhanges zwischen X und Y ist

$$\rho = \frac{s_{XY}}{s_Y \cdot s_Y}$$

mit

s_X ... Standardabweichung des Merkmals X
s_Y ... Standardabweichung des Merkmals Y
s_{XY} ... Kovarianz der Merkmale X und Y

Es gilt $-1 \leq \rho \leq +1$

Interpretation:
$\rho < 0$ gegensinniger linearer Zusammenhang
$\rho = 0$ kein linearer Zusammenhang
$\rho > 0$ gleichsinniger linearer Zusammenhang

Je stärker der lineare Zusammenhang, desto näher liegt $|\rho|$ bei 1.

Besonders wichtig ist der Hinweis darauf, dass der Korrelationskoeffizient lediglich den *linearen* Zusammenhang misst. Würden alle Datenpunkte exakt auf einer Geraden liegen, so wäre $|\rho| = 1$. Je näher die Daten an einer Geraden liegen, desto näher liegt der Betrag von ρ bei eins. Ein positives Vorzeichen deutet auf eine steigende Gerade, ein negatives Zeichen auf eine fallende Gerade (vgl. grafische Darstellungen in Kapitel 8.5).

Beispiel 8.8. Schlafverhalten
Ein Kinderpsychologe will überprüfen, ob sich sportliche Aktivität positiv auf die Schlafdauer von Kindern auswirkt. Es werden neun zufällig ausgewähl-

te Kinder gleichen Alters ausgewählt und ihre Schlafphasen (in h) gemessen. Außerdem wird beobachtet, wie viel Sport das Kind betrieben hat (ebenfalls in h). Es ergeben sich folgende Daten:

Kind	1	2	3	4	5	6	7	8	9
Sport	1,1	0,8	1,3	0,3	1,0	0,9	0,7	1,2	0,2
Schlafdauer	7,9	7,6	8,1	7,6	7,9	7,5	7,5	7,7	7,0

Nach Berechnung der Hilfsgrößen $\bar{x} = 0,83$, $\bar{y} = 7,64$, $s_X^2 = 0,129$ und $s_Y^2 = 0,089$ erhält man

$$s_{XY} = \frac{1}{9}\left(1,1 \cdot 7,9 + \ldots + 0,2 \cdot 7,0\right) - 0,83 \cdot 7,64 = 0,087$$

$$\rho = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{0,087}{\sqrt{0,129}\sqrt{0,089}} = 0,815$$

Man findet einen starken gleichsinnigen linearen Zusammenhang zwischen Sportdauer und Schlafdauer. Das bedeutet je mehr Sport das Kind betreibt, desto höher ist die Schlafdauer.

Das folgende Beispiel soll illustrieren, dass der Korrelationskoeffizient als Maßzahl ausschließlich für lineare Zusammenhänge geeignet ist.

Beispiel 8.9.
Für die Merkmale X und Y wurden folgende Messwerte erhoben:

Messung	1	2	3	4	5	6	7	8	9
Merkmal X	-4	-3	-2	-1	0	1	2	3	4
Merkmal Y	16	9	4	1	0	1	4	9	16

Aus der Datentabelle ist ersichtlich, dass die Merkmale X und Y einen funktionalen Zusammenhang besitzen, denn es gilt $Y = X^2$.
Die Berechnung des Korrelationskoeffizienten erfolgt über $\bar{x} = 0$, $\bar{y} = 6,67$, $s_X^2 = 6,667$ und $s_Y^2 = 34,222$ und man erhält

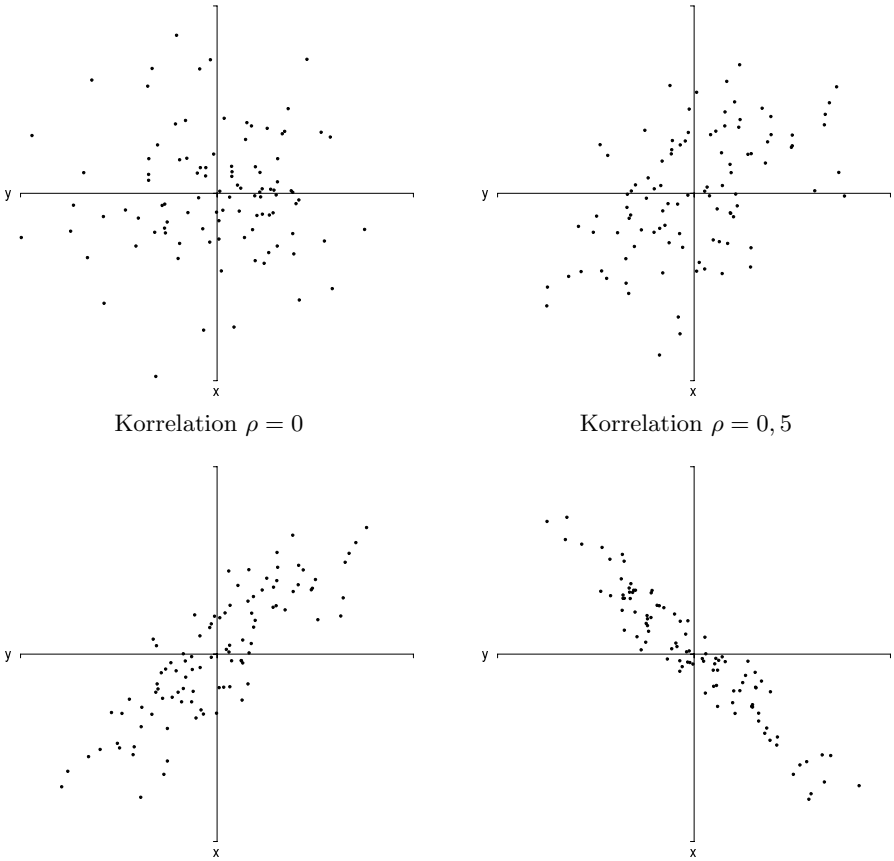
$$s_{XY} = \frac{1}{9}\left(-4 \cdot 16 + \ldots + 4 \cdot 16\right) - 0,00 \cdot 6,67 = 0$$

$$\rho = \frac{s_{XY}}{s_Y \cdot s_Y} = \frac{0}{\sqrt{6,667}\sqrt{34,222}} = 0$$

Obwohl also ein exakter quadratischer Zusammenhang zwischen den Merkmalen besteht, kann der Korrelationskoeffizient diesen nicht entdecken, weil er eben nur lineare Zusammenhänge messen kann. Zwischen den Merkmalen X und Y gibt es keinen linearen Zusammenhang.

8.5 Grafische Darstellung zweidimensionaler metrischer Merkmale

Zweidimensionale metrische Merkmale lassen sich sehr gut in Streudiagrammen darstellen, dazu wird jedem Datenpunkt ein Punkt in einem Koordinatensystem zugeordnet. Oft ist schon an den Streudiagrammen erkennbar, ob die Daten einen linearen Zusammenhang aufweisen.



Korrelation $\rho = 0$ | Korrelation $\rho = 0,5$

Korrelation $\rho = 0,85$ | Korrelation $\rho = -0,95$

Abb. 8.2. Streudiagramme verschiedener Korrelationen

Unkorrelierte Daten ($\rho = 0$) verursachen Streudiagramme, in denen die Datenpunkte relativ unsystematisch angeordnet sind. Je näher der Betrag von ρ bei 1 liegt, desto besser ist der lineare Zusammenhang zwischen den Merkmalen ausgeprägt und die Punktewolke weist ein ellipsenförmiges Bild auf. Diese

Ellipse wird mit steigendem Betrag von ρ immer schmäler, bis die Punkte für $|\rho| = 1$ exakt auf einer Geraden liegen.

Streudiagramm

Ein Streudiagramm ist eine grafische Darstellung eines zweidimensionalen metrischen Merkmals. Dabei wird jeder Erhebungseinheit der zugehörige Datenpunkt in einem Koordinatensystem zugeordnet. Streudiagramme erleichtern das Auffinden von Zusammenhängen.

Daneben lässt sich aus einem Streudiagramm auch die Richtung des Zusammenhanges ablesen. Bei einem gleichsinnigen Zusammenhang muss die Punktewolke bzw. die Gerade ansteigend sein, bei einem gegensinnigen Zusammenhang ist die Punktewolke bzw. die Gerade fallend.

8.6 Korrelation und Kausalität

Bei den einzelnen Maßzahlen zur Berechnung des Zusammenhanges wurde bereits auf den Umstand hingewiesen, dass aus der Kennzahl selbst nicht abgelesen werden kann, was Ursache und was Wirkung ist. Es ist nicht einmal sicher, ob es überhaupt eine Ursache-Wirkungsbeziehung zwischen den beiden Merkmalen gibt.

In der Statistik unterscheidet man zwischen einer statistischen Korrelation und einem kausalen Zusammenhang. Kennzahlen können nur messen, ob die Daten eine statistische Korrelation aufweisen, aber niemals, ob es auch tatsächlich einen kausalen Zusammenhang gibt. Kausale Zusammenhänge sind generell nicht durch eine Berechnung zu finden, hier hilft nur Sachkompetenz und Hausverstand.

Weisen Daten eine statistische Korrelation auf, für die es keine inhaltliche Rechtfertigung gibt, dann spricht man von einer Scheinkorrelation. Als klassisches Beispiel wird meist die starke positive Korrelation zwischen der Anzahl an Störchen und der Geburtenzahl angeführt. Das folgende Beispiel zeigt einen ähnlichen Fall:

Beispiel 8.10. Scheinkorrelation
In fünf aufeinanderfolgenden Jahren entwickelten sich die Anzahl der gemeldeten Aidsfälle und die Anzahl der Mobiltelefon-BenutzerInnen (in Tausend) in der Schweiz gemäß nachstehender Tabelle: (Quellen: www.bakom.ch und www.bag.admin.ch)

Jahr	1995	1996	1997	1998	1999
Aidsfälle	736	542	565	422	262
Mobiltelefon-BenutzerInnen (Tsd.)	447	663	1.044	1.699	3.058

Die Berechnung des Korrelationskoeffizienten führt auf $\rho = -0,94$, und verweist damit auf eine starke gegensinnige Korrelation zwischen Aidsfällen und Anzahl der HandynutzerInnen. Mit dem kausalen Zusammenhang ist es etwas schwieriger, denn Handys dürften wohl kaum als neues Mittel gegen Aids verwendbar sein. Die Variable Zeit spielt uns hier einen bösen Streich, denn diese hat sowohl die Zahl der Aidsfälle beeinflusst, als auch die Zahl der Mobiltelefon-BenutzerInnen.

Scheinkorrelationen werden meist durch eine zusätzliche Einflussgröße verursacht, die in der Berechnung der Korrelation nicht berücksichtigt wurde. Im Beispiel 8.10 wurde beispielsweise die Einflussgröße Zeit nicht beachtet, ein Fehler, der übrigens sehr oft vorkommt.

Bleibt ein entscheidendes Merkmal unberücksichtigt, kann auch der umgekehrte Effekt auftreten, dass statistisch keine Korrelation feststellbar ist, obwohl ein Zusammenhang existiert, wenn ein weiteres Merkmal berücksichtigt wird. In diesem Fall spricht man in der Statistik von verdeckten Korrelationen.

Korrelation und Kausalität

- Scheinkorrelation: statistische Korrelation bei fehlendem direkten Zusammenhang
- Verdeckte Korrelationen: Zusammenhang bei fehlender statistischer Korrelation

Die Ursache liegt bei nicht berücksichtigten weiteren Merkmalen.

8.7 Zweidimensionale Merkmale in EXCEL

Randverteilungen und bedingte Verteilungen können in EXCEL nur über Eingabe der zugehörigen Formeln berechnet werden, ebenso die Assoziationsmaße χ^2 und Cramers V.

Für den Bravais-Pearson-Korrelationskoeffizienten stellt EXCEL die Funktion *Korrel(Datenbezug X;Datenbezug Y)* zur Verfügung, für die Kovarianz die Funktion *Kovar(Datenbezug X;Datenbezug Y)*.

Der Rangkorrelationskoeffizient steht nicht als Funktion zur Verfügung. Liegen die ordinalen Merkmale in geeigneter Form vor (d.h. Rangzahlen zwischen 1 und N, bei gleicher Ausprägung wird der Durchschnittsrang vergeben, die

Rangsumme über alle Erhebungseinheiten bleibt dabei gleich) dann kann die Funktion *Korrel(Datenbezug1;Datenbezug2)* auch zur Berechnung des Rangkorrelationskoeffizienten verwendet werden.

Das Streudiagramm wird über den Diagrammassistenten (vgl. Kapitel 6.3.6) mit dem Diagrammtyp *Punkt (XY)* erstellt.

8.8 Zweidimensionale Merkmale in SPSS

In SPSS sind alle beschriebenen Kennzahlen zur Messung von Zusammenhängen unter dem Menüpunkt *Analysieren* → *Deskriptive Statistiken* → *Kreuztabellen* zu finden.

Der Menüpunkt öffnet zuerst das Variablenfenster (vgl. Abbildung 8.3), in dem man die gewünschten Variablen auswählt, wobei ein Merkmal als *Spalte* und das andere als *Zeile* festgelegt wird. Im Bereich *Schicht 1 von 1* könnte man bei Bedarf Gruppierungen vornehmen, wenn man beispielsweise den Datensatz getrennt nach Geschlecht betrachten möchte.

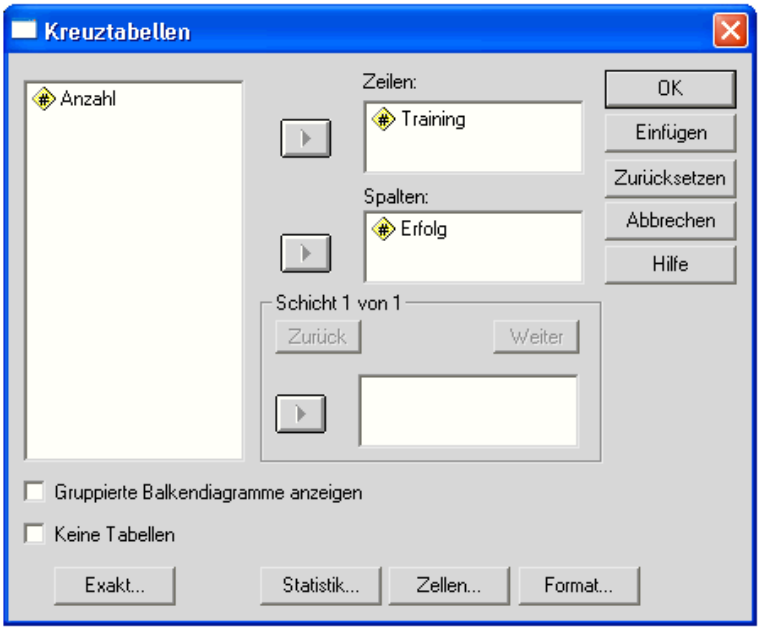


Abb. 8.3. SPSS: Kreuztabellen - Eingabe der Variablen

In der Option *Zellen* (vgl. Abbildung 8.4) ist standardmäßig im Bereich *Häufigkeiten* die Option *Beobachtet* ausgewählt.

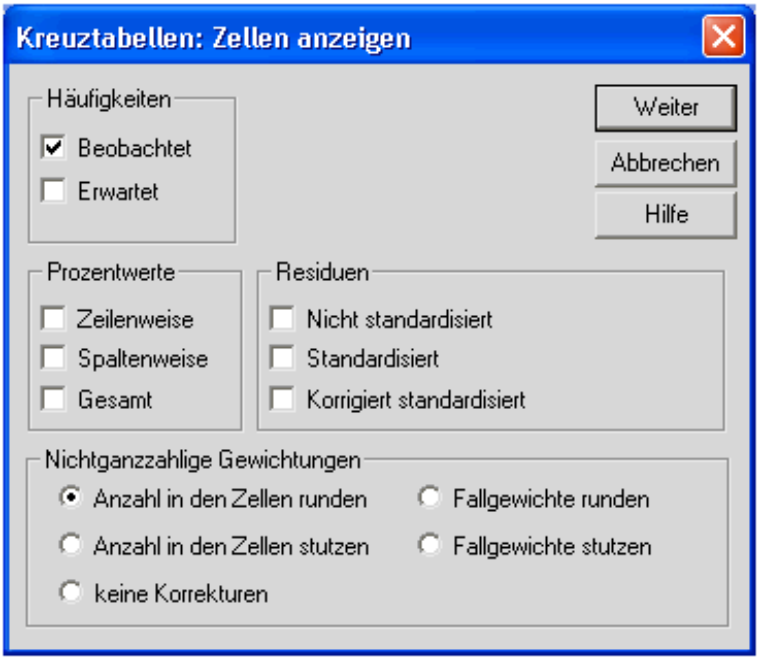


Abb. 8.4. SPSS: Kreuztabellen - Option Zellen

Dort kann man auch die (bei Unabhängigkeit der Merkmale) *erwarteten Häufigkeiten* auswählen, die wir bei den Assoziationsmaßen (vgl. Kapitel 8.4.1) kennen gelernt haben.

Im Abschnitt *Prozentwerte* liefert die Option *Gesamt* die zweidimensionale Verteilung der beiden Merkmale (vgl. Tabelle aus Beispiel 8.2). Die Optionen *Zeilenweise* und *Spaltenweise* ermöglichen die Berechnung der bedingten Verteilungen. Bei *Zeilenweise* wird die Zeilenvariable als bedingendes Merkmal ausgewählt, d.h. die Ausprägungen dieser Variable zerlegen die Untersuchungsgesamtheit in Teilgesamtheiten, die dann hinsichtlich der Spaltenvariable untersucht werden. Nachdem wir als Zeilenvariable *Training* ausgewählt haben, würde uns die Option *Zeilenweise* die Verteilung des Merkmals Erfolg bei den Trainingsteilnehmern und bei den Trainingsverweigerern liefern. Die anderen Auswahlmöglichkeiten in der Option *Zellen* sind für uns vorerst nicht wichtig.

Die Maßzahlen selbst verbergen sich hinter der Option *Statistik* (vgl. Abbildung 8.5). Hier findet man die Auswahlmöglichkeit *Chi-Quadrat* für das

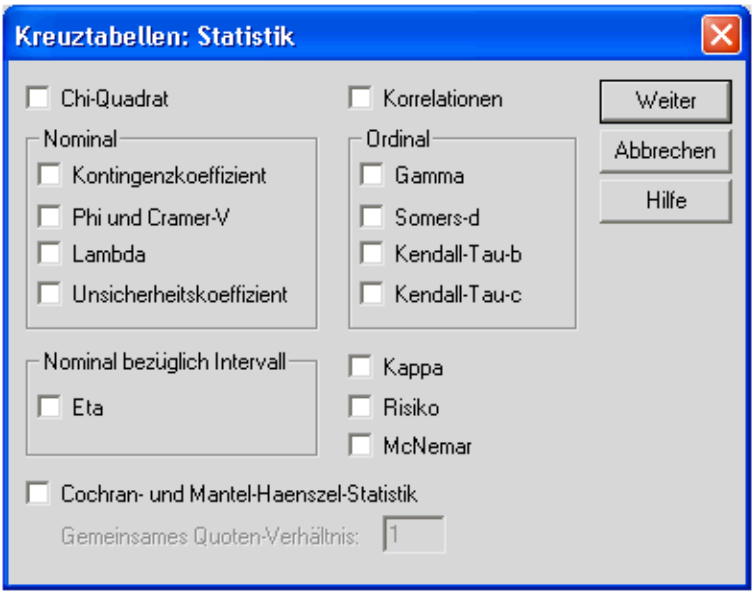


Abb. 8.5. SPSS: Kreuztabellen - Option Statistik

Assoziationsmaß χ^2, im Bereich *Nominal* die Option *Phi und Cramer-V* für das Cramersche Assoziationsmaß V und die Option *Korrelationen* für den Rangkorrelationskoeffizienten nach Spearman und den Bravais-Pearson-Korrelationskoeffizienten. Die Verwendung der Daten aus Beispiel 8.1 (Training und Erfolg) und der Optionen *Chi-Quadrat* und *Phi und Cramer-V* liefert die in Abbildung 8.6 ersichtlichen Ergebnisse.

In der Tabelle *Chi-Quadrat-Tests* findet man in der Zeile *Chi-Quadrat nach Pearson* in der Spalte *Wert* das Assoziationsmaß $\chi^2 = 2,687$. Das zugehörige Ergebnis für das Cramersche Assoziationsmaß $V = 0,107$ ist in der Tabelle *Symmetrische Maße* in Zeile *Cramer-V* und Spalte *Wert* zu finden.

Die Option *Korrelationen* (vgl. Abbildung 8.5) liefert standardmäßig sowohl den Rangkorrelationskoeffizienten nach Spearman, als auch den Korrelationskoeffizienten nach Bravais-Pearson. Wird der Rangkorrelationskoeffizient benötigt, weil ordinale Merkmale vorliegen, so findet man das Ergebnis nach der Berechnung in der Tabelle *Symmetrische Maße* in der Zeile *Korrelation nach Spearman* in der Spalte *Wert*. Der Korrelationskoeffizient für metrische Merkmale ist in derselben Spalte eine Zeile höher bei *Pearson-R* abzulesen.

Die Verwendung der Daten aus Beispiel 8.8 (Sport und Schlafdauer) und der Optionen *Korrelationen* liefert die in Abbildung 8.7 ersichtlichen Ergebnisse.

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	2,687[b]	1	,101		
Kontinuitätskorrektur[a]	2,240	1	,135		
Likelihood-Quotient	2,693	1	,101		
Exakter Test nach Fisher				,117	,067
Zusammenhang linear-mit-linear	2,676	1	,102		
Anzahl der gültigen Fälle	235				

a. Wird nur für eine 2x2-Tabelle berechnet

b. 0 Zellen (,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 34,26.

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	,107	,101
	Cramer-V	,107	,101
Anzahl der gültigen Fälle		235	

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

Abb. 8.6. SPSS: Ergebnis zu Beispiel 8.1

Symmetrische Maße

		Wert	Asymptotischer Standardfehler[a]	Näherungsweises T[b]	Näherungsweise Signifikanz
Intervall- bzgl. Intervallmaß	Pearson-R	,815	,117	3,728	,007[c]
Ordinal- bzgl. Ordinalmaß	Korrelation nach Spearman	,836	,103	4,023	,005[c]
Anzahl der gültigen Fälle		9			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

c. Basierend auf normaler Näherung

Abb. 8.7. SPSS: Ergebnis zu Beispiel 8.8

Daraus lässt sich der Korrelationskoeffizient nach Bravais-Pearson ablesen ($\rho = 0,815$).

Der Menüpunkt *Grafiken* → *Streudiagramm* öffnet eine Dialogbox, welche einfache und überlagerte Streudiagramme, sowie Streudiagramme in Matrixform und in dreidimensionaler Form anbietet. Für EinsteigerInnen ist das einfache Streudiagramm zu empfehlen, in der nun erscheinenden Dialogbox (vgl. Abbildung 8.8) können weitere Spezifikationen vorgenommen werden.

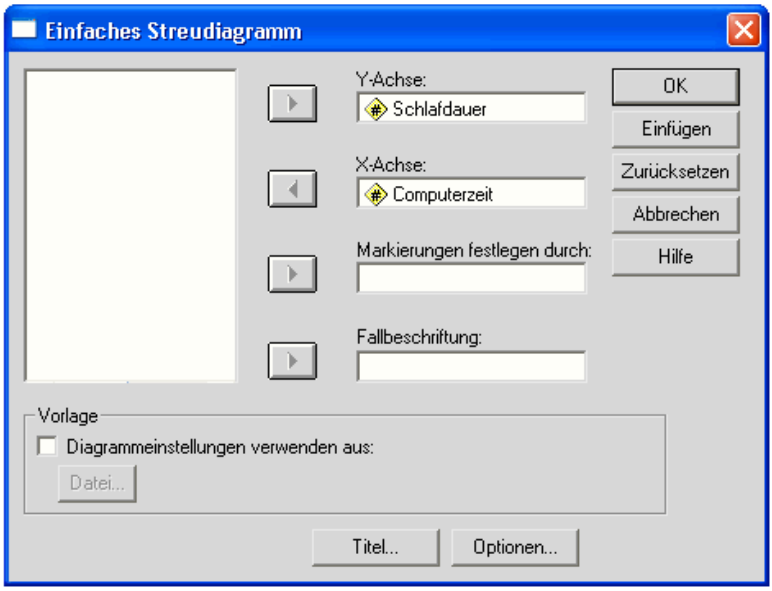


Abb. 8.8. SPSS: Eingabemaske für ein Streudiagramm

Zweidimensionale Merkmale bearbeiten in SPSS

- *Analysieren* → *Deskriptive Statistiken* → *Kreuztabellen*
- Ein Merkmal als Spalte und das andere als Zeile auswählen
- Option *Zellen*
  - beobachtete Häufigkeiten
  - erwartete Häufigkeiten
  - Prozentwerte für zweidimensionale Verteilung (*Gesamt*)
  - Randverteilungen (*Gesamt*)
  - bedingte Verteilungen (*Zeilenweise*, bzw. *Spaltenweise*)
- Option *Statistik* für Maßzahlen des Zusammenhanges
  - *Chi-Quadrat*
  - *Phi und Cramer V*
  - *Korrelationen* (auch für Rangkorrelationskoeffizient)

8.9 Tipps und Tricks

In diesem Kapitel wurden Maßzahlen zur Messung des Zusammenhangs beschrieben, die bei zwei Merkmalen gleichen Skalenniveaus verwendet werden können. In der Praxis kommen oft unterschiedliche Skalenniveaus, z.B. Geschlecht (nominal) und höchste abgeschlossene Schulbildung (ordinal) vor. Es gibt zwar spezielle Maßzahlen für solche Fälle, aber es hilft auch folgende Überlegung: Aufgrund der hierarchischen Anordnung der Skalenniveaus (vgl. Kapitel 2.2.1) sind für ein bestimmtes Niveau auch alle Verfahren zulässig, die im darunter liegenden Niveau zulässig sind. Ein ordinales Merkmal darf also als nominales Merkmal behandelt werden, daher kann man den Zusammenhang zwischen Geschlecht und höchster abgeschlossener Schulbildung mit den Assoziationsmaßen χ^2 und Cramers V messen.

Übungsaufgaben

8.1. Interesse an Sportübertragung
In einer Lehrveranstaltung wurden die dort anwesenden Studierenden gefragt, ob sie sich für Sportübertragungen im TV interessieren. Die 240 befragten Personen verteilten sich folgendermaßen auf dem zweidimensionalen Merkmal Geschlecht und Interesse.

	Interesse	kein Interesse	Summe
männlich	60	30	90
weiblich	70	80	150
Summe	130	110	240

a) Berechnen Sie die zweidimensionalen relativen Häufigkeiten und interpretieren Sie Ihr Ergebnis.
b) Berechnen Sie die bedingten Verteilungen zur Beantwortung folgender Frage: Unterscheiden sich Frauen und Männer bezüglich dem Interesse an Sportübertragungen.
c) Gibt es einen Zusammenhang zwischen Geschlecht und Interesse an Sportübertragungen? Berechnen Sie geeignete Kennzahlen.

8.2. Inflationsrate und Staatsschulden/BIP
Bei 4 Staaten liegen folgende Ausprägungen der Merkmale Inflationsrate und Staatsschulden/BIP vor:

Staat	Inflationsrate (in %)	Staatsschulden (in %)
A	1,9	53,1
B	3,2	55,7
C	2,8	74,4
D	1,4	76,6

Berechnen Sie eine geeignete Maßzahl zur Messung der Abhängigkeit zwischen den Ausprägungen der beiden Merkmale.

8.3. Körpergröße und Gewicht

Bei einer Stichprobe von 10 Personen wurden Körpergröße K und Gewicht G gemessen:

Person	1	2	3	4	5	6	7	8	9	10
K	175	175	184	180	173	173	184	179	168	183
G	75	73	74	82	77	70	88	68	60	82

Zeichnen Sie ein Streudiagramm und berechnen Sie eine geeignete Maßzahl für den Zusammenhang zwischen den Ausprägungen dieser beiden Merkmale.

8.4. Leistung und Drehzahl

Bei einem Gleichstrommotor seien folgende Daten über die Leistung (in PS) und die Drehzahl (in Umdrehungen pro Sekunde) bekannt.

PS	20	30	40	50	60
U/sec	35,2	43,8	51,3	59,2	67,9

Berechnen Sie eine geeignete Maßzahl für die Messung des Zusammenhanges dieser beiden Merkmale.

8.5. Abfahrtslauf

An einem Abfahrtslauf nahmen 8 Personen (A-H) teil. In der nachfolgenden Tabelle sind die Ergebnisse dargestellt.

Bestimmen Sie die Stärke des Zusammenhanges zwischen den Ausprägungen der Merkmale Startnummer und Platzierung und interpretieren Sie das Ergebnis.

Name	Startnummer	Zeit (in min.sec.)
A	5	1.58.90
B	8	2.01.34
C	7	2.00.30
D	1	1.59.60
E	6	2.00.14
F	2	2.00.41
G	3	1.59.62
H	4	1.57.48

8.6. Lehrveranstaltung
Eine Lehrveranstaltungsleiterin hat beim Betrachten der Ergebnisse ihrer Übung festgestellt, dass die beste Klausur von der Studentin mit dem besten hinterlassenen Eindruck in der Übung und die schlechteste Klausur von jener mit dem schlechtesten Eindruck geschrieben wurde. Sie vermutet deshalb einen Zusammenhang zwischen den Rangfolgen bei der Klausur und ihren persönlichen Eindrücken:

Studierende	A	B	C	D	E	F	G
Rang Klausur	1	6	7	5	2	4	3
Rang Eindruck	1	2	7	3	4	5	6

Berechnen Sie zur Messung des Zusammenhanges zwischen den Ausprägungen dieser beiden Merkmale den Spearmanschen Korrelationskoeffizienten.

8.7. Freude an der Schule
Bei einer Befragung von insgesamt 3220 Kindern ergab eine Auswertung nach dem zweidimensionalen Merkmal Geschlecht und Freude an der Schule folgende Verteilung.

	große Freude	geringe Freude	Summe
männlich	1224	226	1450
weiblich	1674	96	1770
Summe	2898	322	3220

a) Berechnen Sie die zweidimensionalen relativen Häufigkeiten und interpretieren Sie Ihr Ergebnis.
b) Berechnen Sie die bedingten Verteilungen zur Beantwortung folgender Frage: Unterscheiden sich Mädchen und Buben bezüglich der Freude an der Schule?
c) Gibt es einen Zusammenhang zwischen Geschlecht und Freude an der Schule? Berechnen Sie geeignete Kennzahlen.

9

Die Regressionsanalyse

Mit dem Begriff Regressionsanalyse werden Verfahren bezeichnet, die den Einfluss von einer oder mehreren Variablen auf eine Zielgröße untersuchen. Der Ursache-Wirkungs-Zusammenhang soll mit Hilfe einer mathematischen Funktion modelliert werden. Es gibt eine Vielfalt an Regressionsverfahren, die sich hinsichtlich der Anzahl der erklärenden Variablen, der mathematischen Funktionen und der Skalenniveaus der Merkmale unterscheiden. Im einfachsten Fall gibt es nur eine erklärende Variable, es wird eine lineare Funktion verwendet und die Merkmale sind metrisch. Dieser Fall der linearen Einfachregression wird hier vorgestellt.

9.1 Die lineare Einfachregression

Ausgangspunkt für die lineare Einfachregression ist das Vorliegen von N zweidimensionalen, metrischen Messdaten $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. Aus der Berechnung des Korrelationskoeffizienten ρ lässt sich ablesen, ob ein linearer Zusammenhang zwischen den Merkmalen X und Y besteht. Existiert ein linearer Zusammenhang, so sollte dieser mit Hilfe einer linearen Gleichung der Form

$$y = a + b \cdot x$$

modelliert werden. Mit diesem Modell wird die abhängige Variable Y durch die unabhängige Variable X erklärt. In der Praxis wird dieses Modell dann für Prognosen verwendet.

Lineare Einfachregression

Idee: Bildung eines Modells für den linearen Zusammenhang in Form einer Geradengleichung
Ziel: Verwendung des Modells zu Prognosezwecken

Die Regressionsgerade

$$f(x) = y = a + b \cdot x$$

soll alle Datenpunkte *möglichst gut* wiedergeben, dazu sind die Parameter a und b der Geradengleichung geeignet zu wählen. Was *möglichst gut* bedeutet, muss noch näher spezifiziert werden.

Setzt man von einem bekannten Datenpunkt (x_i, y_i) den Wert x_i in die gefundene Gleichung ein, so erhält man eine Schätzung für den Wert y_i, der auch als prognostizierter y_i-Wert bezeichnet wird. Diesen Prognosewert wollen wir mit $\widehat{y}_i$ bezeichnen. Die Abweichung $\widehat{e}_i = y_i - \widehat{y}_i$ zwischen dem tatsächlichen Wert und dem prognostizierten Wert sollte möglichst gering sein. Diese Forderung stellen wir an alle Datenpunkte, daher werden wir den Mittelwert der Abweichungen betrachten. Damit sich positive und negative Abweichungen nicht aufheben können, ist es ratsam, den Mittelwert der quadrierten Abweichungen zu betrachten. Damit erhalten wir als Ansatz eine Funktion Q, die von den Parametern a und b abhängig ist und folgendermaßen angeschrieben werden kann:

$$Q(a,b) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2 = \frac{1}{N}\sum_{i=1}^{N}[y_i - (a + b \cdot x_i)]^2$$

Diese Funktion soll ein Minimum annehmen. Anders ausgedrückt sind jene Parameter a und b zu bestimmen, für die diese Funktion minimal wird. Dieses Verfahren zur Bestimmung der Regressionsgerade wird als Methode der kleinsten Quadrate bezeichnet und die beiden gefundenen Schätzwerte $\widehat{a}$ und $\widehat{b}$ für die Parameter nennt man Kleinste-Quadrate-Schätzer.

Zur Bestimmung der Schätzer ist die Funktion Q(a,b) einmal nach a und einmal nach b partiell zu differenzieren.

$$\frac{\partial Q(a,b)}{\partial a} = -\frac{2}{N}\sum_{i=1}^{N}[y_i - (a + b \cdot x_i)]$$

$$\frac{\partial Q(a,b)}{\partial b} = -\frac{2}{N}\sum_{i=1}^{N}[y_i - (a + b \cdot x_i)]x_i$$

Beide Ableitungen sind null zu setzen. Dadurch erhält man zwei Gleichungen mit zwei Unbekannten (nämlich $\widehat{a}$ und $\widehat{b}$), die sich mit elementaren Methoden lösen lassen:

$$\frac{1}{N}\sum_{i=1}^{N} y_i - \widehat{a} - \widehat{b}\frac{1}{N}\sum_{i=1}^{N} x_i = 0$$

und

$$\frac{1}{N}\sum_{i=1}^{N} y_i x_i - \widehat{a}\frac{1}{N}\sum_{i=1}^{N} x_i - \widehat{b}\frac{1}{N}\sum_{i=1}^{N} x_i^2 = 0$$

Aus der oberen Gleichung erhält man

$$\widehat{a} = \bar{y} - \widehat{b} \cdot \bar{x}$$

Dies eingesetzt in die untere Gleichung ergibt

$$\frac{1}{N}\sum_{i=1}^{N} y_i x_i - \bar{y}\frac{1}{N}\sum_{i=1}^{N} x_i + \widehat{b}\bar{x}\frac{1}{N}\sum_{i=1}^{N} x_i - \widehat{b}\frac{1}{N}\sum_{i=1}^{N} x_i^2 = 0$$

Dies lässt sich umformen zu

$$\frac{1}{N}\sum_{i=1}^{N} y_i x_i - \bar{y}\bar{x} = \widehat{b}\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2\right)$$

Unter Verwendung von Varianz s_X^2 und Kovarianz s_{XY} lässt sich dieser Zusammenhang elegant anschreiben als

$$s_{XY} = \widehat{b} \cdot s_X^2$$

Damit erhalten wir für die Schätzer $\widehat{a}$ und $\widehat{b}$ zwei einfache Formeln.

Lineare Einfachregression

Liegen N zweidimensionale metrische Messdaten $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ vor, dann nennt man das Modell

$$y = a + b \cdot x$$

zur Beschreibung des linearen Zusammenhanges die lineare Einfachregression, wobei a den Achsenabschnitt und b die Steigung der Regressionsgeraden bezeichnet.

Die **Kleinste-Quadrate-Schätzer** für a und b sind gegeben durch

$$\widehat{b} = \frac{s_{XY}}{s_X^2}$$

und

$$\widehat{a} = \bar{y} - \widehat{b} \cdot \bar{x}$$

Residuen
Die Abweichungen zwischen den tatsächlichen y-Werten und den prognostizierten $\widehat{y}$-Werten werden als Residuen $\widehat{e}$ bezeichnet

$$\widehat{e}_i = y_i - \widehat{y}_i \qquad \text{mit} \qquad \widehat{y}_i = \widehat{a} + \widehat{b}x_i \qquad \text{für} \qquad i = 1, \ldots, N$$

Beispiel 9.1. Schlafverhalten
(Fortsetzung von Beispiel 8.8, Seite 132) Der Korrelationskoeffizient in diesem Beispiel betrug $\rho = 0,815$ und verwies damit auf einen starken gleichsinnigen linearen Zusammenhang zwischen Sportdauer und Schlafdauer. Nun wollen wir die Schlafdauer der Kinder aufgrund der sportlichen Aktivität prognostizieren. Wir kennen bereits die Hilfsgrößen $\bar{x} = 0,83$, $\bar{y} = 7,64$, $s_X^2 = 0,129$ und $s_{XY} = 0,087$.

Damit erhalten wir für die Kleinste-Quadrate-Schätzer

$$\widehat{b} = \frac{s_{xy}}{s_x^2} = \frac{0,087}{0,129} = 0,678$$

und

$$\widehat{a} = \bar{y} - \widehat{b} \cdot \bar{x} = 7,64 - 0,678 \cdot 0,83 = 7,079$$

Die Regressionsgerade lautet daher

$$y = 7,079 + 0,678 \cdot x$$

Die Residuen für die einzelnen Datenpunkte erhält man durch Einsetzen der x_i-Werte in die Geradengleichung.

Kind		1	2	3	4	5	6	7	8	9
Sportdauer	x_i	1,1	0,8	1,3	0,3	1,0	0,9	0,7	1,2	0,2
Schlafdauer	y_i	7,90	7,60	8,10	7,60	7,90	7,50	7,50	7,70	7,00
prognostizierte Schlafdauer	$\hat{y}_i$	7,83	7,62	7,96	7,28	7,76	7,69	7,55	7,89	7,21
Residuen	$\hat{e}_i$	0,07	-0,02	0,14	0,32	0,14	-0,19	-0,05	-0,19	-0,21

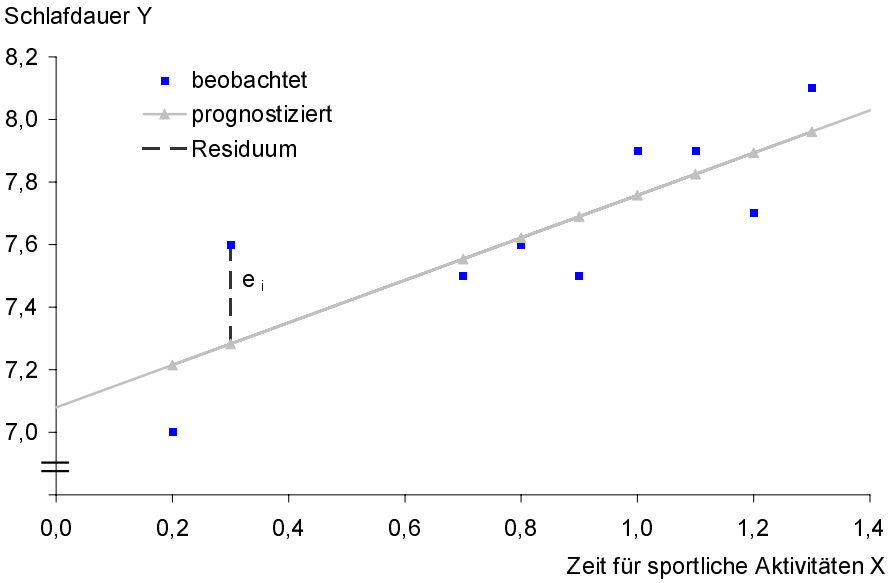


Abb. 9.1. EXCEL: Streudiagramm mit Regressionsgerade

Ergänzt man das Streudiagramm der Datenpunkte (x_i, y_i) mit der neu berechneten Regressionsgerade, so kann man jeweils senkrecht über bzw. unter den Datenpunkten auf der Regressionsgerade die zugehörigen Prognosepunkte entnehmen (vgl. Abbildung 9.1). Der vertikale Abstand zwischen dem Datenpunkt und dem Prognosepunkt entspricht dem Residuum. Diese Residuen sollten möglichst klein sein, und waren deswegen auch Ausgangspunkt unserer Überlegungen. Die Funktion Q(a,b) ist nichts anderes als das arithmetische Mittel der quadrierten Residuen, die es zu minimieren gilt.

Will man die Regressionsgerade zu Prognosen heranziehen, so stellt sich die Frage nach der Qualität dieser Prognose.

Güte der lineare Einfachregression

Die Güte des Modells wird geschätzt durch das **Bestimmtheitsmaß**

$$B = \rho^2$$

$$0 \leq B \leq 1$$

Je größer der Wert des Bestimmtheitsmaßes, desto besser bildet das Modell den Datensatz ab und desto besser ist die gefundene Geradengleichung zu Prognosezwecken geeignet. Dies ist auch unmittelbar einleuchtend, wenn man sich die Berechnung des Bestimmtheitsmaßes näher ansieht.
Ein Korrelationskoeffizient nahe 1 bzw. nahe -1 deutet auf einen starken linearen Zusammenhang hin. Je stärker der lineare Zusammenhang, desto besser kann man diesen auch modellieren und desto aussagekräftiger ist die Regressionsfunktion.

Aus dem Bestimmtheitsmaß lassen sich noch genauere Erkenntnisse gewinnen. Das Merkmal Y weist eine bestimmte Varianz auf, einen Teil dieser Varianz weisen auch die prognostizierten Werte $\widehat{y}$ auf. Dieser Teil der Varianz des Merkmals Y kann also durch das Regressionsmodell erklärt werden, der Rest bleibt unerklärt. Das Bestimmtheitsmaß gibt nun an, welcher Anteil der Varianz des Merkmals Y durch das Modell erklärt werden kann.

Im Beispiel 9.1 würde man $B = 0,665$ erhalten. Das lineare Regressionsmodell weist eine mittlere Güte auf. Weiters lässt sich daraus ablesen, dass das Modell der linearen Einfachregression auf die Sportdauer 66,5% der Varianz des Merkmals Schlafdauer erklären kann, den Rest nicht.

Praxistipp

Auch für Merkmale mit geringer Korrelation lässt sich eine Regressionsgerade berechnen. Sinnvoll ist diese Berechnung meistens nicht, weil die Güte dann so niedrig ist, dass man das gefundene Modell nicht für Prognosen verwenden kann.

9.2 Regressionsanalyse in EXCEL

EXCEL stellt folgende Funktionen zur Schätzung der Regressionsparameter bereit:

- *ACHSENABSCHNITT(Datenbezug Y;Datenbezug X)* berechnet $\widehat{a}$
- *STEIGUNG(Datenbezug Y;Datenbezug X)* berechnet $\widehat{b}$

Die Güte des Modells wird mit der Funktion

BESTIMMTHEITSMASS(Datenbezug Y;Datenbezug X)

bestimmt.

Der prognostizierte $\widehat{y}_i$-Wert für einen bestimmten x_i-Wert wird mit der Anweisung

SCHÄTZER(x_i-Wert;Datenbezug Y;Datenbezug X)

errechnet.

Das Streudiagramm wird mit Hilfe des Diagrammassistenten mit der Auswahl *Punkt (XY)* als Diagrammtyp (vgl. Kapitel 6.3.2) erstellt. Die Regressionsgerade lässt sich ebenfalls mit diesem Diagrammtyp erstellen, allerdings wird dafür der Untertyp *Punkte mit Linien* oder *Punkte mit Linien ohne Datenpunkte* verwendet. Es besteht auch die Möglichkeit, sich die Regressionsgerade von EXCEL direkt erstellen zu lassen. Dazu wird im Streudiagramm zuerst die Datenreihe der beobachteten Werte markiert. Die rechte Maustaste bietet dann ein Untermenü an, aus dem der Punkt *Trendlinie hinzufügen* ausgewählt wird. Die neu geöffnete Dialogbox enthält die Auswahl für die lineare Regression als Standardeinstellung, diese Auswahl muss bestätigt werden.

9.3 Regressionsanalyse in SPSS

Unter dem Menüpunkt *Analysieren* → *Regression* → *Linear* öffnet sich eine Dialogbox zur Dateneingabe. Wir wollen Beispiel 8.8 mit SPSS lösen. Wir wählen daher das Merkmal Schlafdauer als *Abhängige Variable* und als *Unabhängige Variable(n)* das Merkmal Sportdauer. Die Voreinstellungen in *Diagramme* und *Optionen* belassen wir.

In der Option *Statistik* wählen wir neben den bereits standardmäßig ausgewählten Möglichkeiten *Schätzer Regressionskoeffizienten* und *Anpassungsgüte des Modells* zusätzlich im Bereich *Residuen* die *Fallweise Diagnose* für *Alle Fälle* (vgl. Abbildung 9.2).

Die Ergebnisse dieser Anweisung entnehmen wir den Abbildungen 9.3 und 9.4.

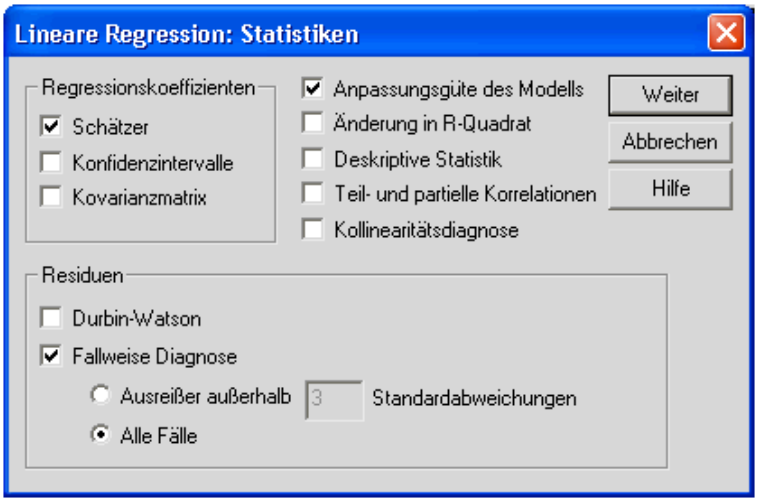


Abb. 9.2. SPSS: Lineare Regression - Option Statistik

In der Tabelle *Modellzusammenfassung* (vgl. Abbildung 9.3) entspricht der Wert in der Spalte *R-Quadrat* dem Bestimmtheitsmaß.

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,815[a]	,665	,617	,19594

a. Einflussvariablen : (Konstante), Sportdauer

b. Abhängige Variable: Schlafdauer

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standard-fehler	Beta	T	Signifikanz
1	(Konstante)	7,079	,165		42,886	,000
	Sportdauer	,678	,182	,815	3,728	,007

a. Abhängige Variable: Schlafdauer

Abb. 9.3. SPSS: Lineare Regression - Bestimmtheitsmaß und Parameter

Die Schätzer für die Parameter der Regressionsgeraden sind der Tabelle *Koeffizienten* zu entnehmen (vgl. Abbildung 9.3). Beide Ergebnisse findet man in der Spalte *Nicht standardisierte Koeffizienten - B*, den Achsenabschnitt $\hat{a}$ in der Zeile *(Konstante)*, die Steigung in der Zeile *Sportdauer*.

Fallweise Diagnose[a]

Fall-nummer	Standardisierte Residuen	Schlafdauer	Nicht standardisierter vorhergesagter Wert	Nicht standardisierte Residuen
1	,381	7,90	7,8253	,07471
2	-,111	7,60	7,6218	-,02184
3	,710	8,10	7,9609	,13908
4	1,619	7,60	7,2828	,31724
5	,727	7,90	7,7575	,14253
6	-,968	7,50	7,6897	-,18966
7	-,276	7,50	7,5540	-,05402
8	-,986	7,70	7,8931	-,19310
9	-1,097	7,00	7,2149	-,21494

a. Abhängige Variable: Schlafdauer

Abb. 9.4. SPSS: Lineare Regression - Residuen

Die Tabelle *Fallweise Diagnose* (vgl. Abbildung 9.4) listet alle Datenpunkte mit den zugehörigen prognostizierten $\widehat{y}_i$-Werten und den Residuen $\widehat{e}_i$ auf.

SPSS bietet die Möglichkeit, die prognostizierten $\widehat{y}_i$-Werte und die Residuen $\widehat{e}_i$ in der Datenmatrix abzuspeichern. Die Auswahlmöglichkeit dazu findet man in der Option *Speichern* der Dialogbox zur Regression (*Analysieren* → *Regression* → *Linear*). Im Bereich *Vorhergesagte Werte* liefert die Auswahl *Nicht standardisiert* die prognostizierten $\widehat{y}_i$-Werte, die Residuen erhält man im Bereich *Residuen* ebenfalls mit der Option *Nicht standardisiert*. Die gewünschten Daten werden als zusätzliche Spalten am Ende der Datenmatrix eingefügt.

Das Streudiagramm wird unter dem Menüpunkt *Grafiken* → *Streudiagramm* erstellt (vgl. Kapitel 8.8). Ein Doppelklick auf das Streudiagramm öffnet dieses im *Diagramm-Editor*, in dem Grafiken layoutiert werden können. Durch Markieren der Datenpunkte wird in der Menüleiste des Diagramm-Editors die Option zum Einfügen einer Anpassungslinie aktiviert. Die Auswahl dieser Option öffnet ein Dialogfenster, in dem eine Anpassungslinie angefordert werden kann, wobei aus verschiedenene Möglichkeiten der Anpassung ausgewählt werden kann (vgl. Abbildung 9.5). Die Anpassungsmethode *Linear* entspricht der Regressionsgeraden, diese Auswahl ist Voreinstellung (vgl. Abbildung 9.6).

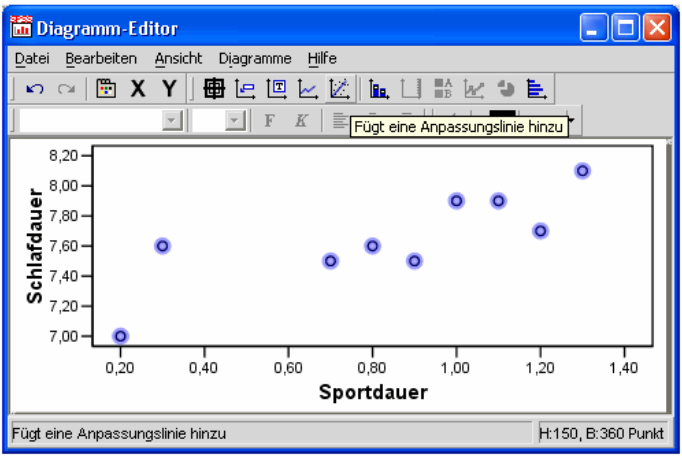


Abb. 9.5. SPSS: Streudiagramm im Diagramm-Editor

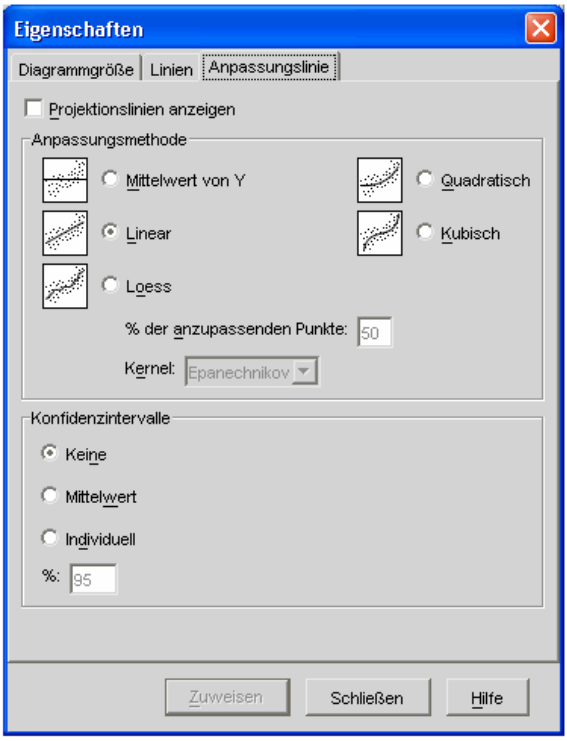


Abb. 9.6. SPSS: Streudiagramm Eigenschaften

In der SPSS-Version 11 erhält man die Regressionsgerade mit einer anderen Vorgehensweise. Auch hier muss das Streudiagramm zuerst mittels Doppelklick im Diagramm-Editor geöffnet werden. Unter dem Menüpunkt *Diagramme* → *Optionen* öffnet sich eine Dialogbox, in der die Anpassungslinie angefordert werden kann.

Regressionsanalyse mit SPSS

- *Analysieren* → *Regression* → *Linear*
- *Abhängige Variable* und *Unabhängige Variable(n)* festlegen
- Option *Statistik*
  - *Schätzer Regressionskoeffizienten* für die Schätzer $\widehat{a}$ und $\widehat{b}$
  - *Anpassungsgüte des Modells* für das Bestimmtheitsmaß B
  - Im Bereich *Residuen* die *Fallweise Diagnose* für *Alle Fälle* für die Residuen $\widehat{e}_i$
- Option *Speichern*
  - Im Bereich *Vorhergesagte Werte* Auswahl *Nicht standardisiert* für die Prognosewerte $\widehat{y}_i$
  - Im Bereich *Residuen* Auswahl *Nicht standardisiert* für die Residuen $\widehat{e}_i$
  - Gespeicherte Werte werden am Ende der Datenmatrix angefügt.
- Streudiagramm unter *Grafiken* → *Streudiagramm*
- Einfügen der Regressionsgerade im Diagramm-Editor möglich

Übungsaufgaben

9.1. Körpergröße und Gewicht

Bei einer Stichprobe von 10 Personen wurden Körpergröße K und Gewicht G gemessen:

Person	1	2	3	4	5	6	7	8	9	10
K	175	175	184	180	173	173	184	179	168	183
G	75	73	74	82	77	70	88	68	60	82

a) Berechnen Sie die Regressionsgerade zur Prognose des Gewichts und tragen Sie diese in das Streudiagramm ein.
b) Berechnen Sie die Residuen der Regression.
c) Berechnen Sie das Bestimmtheitsmaß.
d) Wie würden Sie in diesem Beispiel die Güte einer Vorhersage beurteilen?
e) Prognostizieren Sie das zu erwartende Gewicht bei einer Körpergröße von 180 cm.

9.2. Einkommen und Ausgaben von Haushalten

Von 20 Haushalten wurde das monatliche Einkommen und die monatlichen Ausgaben für Konsum erhoben. Man erhielt folgendes Ergebnis (Angaben in Euro):

Haushalt	1	2	3	4	5	6	7	8	9	10
Einkommen	1100	2400	2100	2250	1700	1200	2000	1500	1200	1800
Ausgaben	960	2070	1780	2120	1250	890	1600	1400	1110	1540

Haushalt	11	12	13	14	15	16	17	18	19	20
Einkommen	1750	1300	2300	2200	2100	2250	1370	2310	1200	1310
Ausgaben	1470	1180	1900	1830	1690	1610	1170	1750	960	1170

a) Berechnen Sie die Regressionsgerade zur Prognose der Ausgaben und tragen Sie diese in das Streudiagramm ein.
b) Berechnen Sie die Residuen der Regression.
c) Berechnen Sie das Bestimmtheitsmaß.
d) Wie würden Sie in diesem Beispiel die Güte einer Vorhersage beurteilen?
e) Prognostizieren Sie die zu erwartenden Haushaltsausgeben bei einem Einkommen von 1200 Euro.

Teil III

Wahrscheinlichkeitsrechnung

10

Wahrscheinlichkeitsrechnung

Bisher haben wir Datensätze *beschrieben*, uns also mit der deskriptiven Statistik beschäftigt. Diese Datensätze können entweder alle interessierenden Objekte umfassen (Vollerhebung einer Grundgesamtheit) oder nur einen Auszug aus dieser Grundgesamtheit (Stichprobe). In Praxis wird der Datensatz meistens eine Stichprobe beinhalten, weil die Vollerhebung zu teuer oder nur schwer möglich ist. Die schließende (induktive) Statistik stellt Methoden bereit, um Rückschlüsse von einer Stichprobe auf die Grundgesamtheit zu ermöglichen. Diese Methoden basieren auf wahrscheinlichkeitstheoretischen Grundlagen. Daher ist es notwendig, sich mit den Grundbegriffen und Denkweisen der Wahrscheinlichkeitsrechnung vertraut zu machen.

10.1 Exkurs: Mengenlehre

Da die Mengenschreibweise für manche Überlegungen in der Wahrscheinlichkeitsrechnung sinnvoll ist, wiederholen wir einige Begriffe der Mengenlehre.

Menge

Eine **Menge** ist eine Zusammenfassung von Objekten zu einem Ganzen, die einzelnen Objekte werden **Elemente** genannt. Mengen werden üblicherweise mit Großbuchstaben bezeichnet und entweder durch Aufzählung der Elemente oder durch Angabe einer charakterisierenden Eigenschaft festgelegt.

Beispiel 10.1. Mengen
Sei A die Menge der natürlichen Zahlen von eins bis sechs, so lässt sich die Menge A anschreiben als
$A = \{1, 2, 3, 4, 5, 6\}$ oder $A = \{x | x \text{ ist eine natürliche Zahl und } 1 \leq x \leq 6\}$
(zu lesen als: A ist die Menge aller x, für die gilt ...).

Einige Notationen:

- $x \in A$ bedeutet x ist ein Element aus A; sonst $x \notin A$.
- $A \subset B$ bedeutet A ist eine Teilmenge von B, d.h. jedes Element aus A ist auch in B enthalten; sonst $A \not\subset B$.
- $A \cap B$ bezeichnet die Schnittmenge (den Durchschnitt) von A und B. Diese Schnittmenge enthält alle Elemente, die sowohl in A als auch in B enthalten sind.
- $A \cup B$ bezeichnet die Vereinigungsmenge (die Vereinigung) von A und B. Diese Vereinigungsmenge enthält alle Elemente, die in A oder in B oder in beiden enthalten sind.
- $A \backslash B$ bezeichnet die Differenzmenge A ohne B. Diese enthält alle Elemente, die in A, nicht aber in B enthalten sind.
- Sei $A \subset \Omega$, dann bezeichnet $\overline{A}$ oder A^C die Komplementärmenge von A. Diese enthält alle Elemente aus Ω, die nicht in A enthalten sind.
- $|A|$ bezeichnet die Kardinalzahl oder Mächtigkeit der Menge A. Diese gibt an, wie viele Elemente in der Menge A enthalten sind.
- Mit $\emptyset$ oder $\{\}$ bezeichnet man die leere Menge, diese enthält kein einziges Element.

10.2 Grundbegriffe der Wahrscheinlichkeitsrechnung

In der Wahrscheinlichkeitsrechnung betrachtet man Experimente mit ungewissem Ausgang und versucht, ihre Gesetzmäßigkeiten zu beschreiben.

Zufallsexperiment
Ein Zufallsexperiment ist ein Vorgang, bei dem ein nicht vollständig vorhersehbarer Ausgang aus einer Menge prinzipiell möglicher Ausgänge realisiert wird. Weiters muss ein Zufallsexperiment unter gleichen Bedingungen wiederholbar sein. Zur mathematischen Beschreibung solcher Zufallsexperimente bedient man sich häufig der Mengenlehre. Beispiele sind das Werfen eines Würfels oder einer Münze, die Lottoziehung und Ähnliches.

Zufallsvariable
Das Merkmal X, das den Ausgang eines Zufallsexperimentes beschreibt, nennt man zufälliges Merkmal oder Zufallsvariable.

Für die folgenden Begriffe betrachten wir das Zufallsexperiment „Werfen eines Würfels" mit der Zufallsvariable „Geworfene Augenzahl".

Wertebereich
Die Gesamtheit der für diese Zufallsvariable X möglichen Ausprägungen ist der Wertebereich Ω_X ($\Omega_X = \{1, 2, 3, 4, 5, 6\}$).

Versuchsausgang
Jede einzelne Ausprägung der Zufallsvariable X ist ein möglicher Versuchsausgang des betrachteten Zufallsexperimentes (z.B. die Augenzahl 1).

Ereignis
Jede Teilmenge E des Wertebereiches Ω_X entspricht einem Ereignis (z.B. die geraden Augenzahlen, also die Menge $E = \{2, 4, 6\}$). Ein Ereignis tritt dann ein, wenn ein möglicher Versuchsausgang realisiert wurde. (Das Ereignis $E = \{2, 4, 6\}$ tritt ein, wenn beispielsweise die Augenzahl 2 geworfen wurde.)

Elementarereignis
Jede einelementige Teilmenge des Wertebereiches Ω_X nennt man ein Elementarereignis (z.B. die Augenzahl 1, also die Menge $E = \{1\}$).

Unmögliches Ereignis
Ein Ereignis, das nicht eintreten kann, wird als unmögliches Ereignis bezeichnet(z.B. die Augenzahl 7). Unmögliche Ereignisse werden als leere Menge angeschrieben $\emptyset$, $\{\ \}$.

Sicheres Ereignis
Ein Ereignis, das auf jeden Fall eintritt, nennt man ein sicheres Ereignis. Der Wertebereich Ω_X ist immer ein sicheres Ereignis.

Disjunkte Ereignisse
Zwei Ereignisse E_1 und E_2 heißen disjunkt oder elementfremd, wenn $E_1 \cap E_2 = \{\}$, wenn also der Durchschnitt der beiden Mengen die leere Menge ist (z.B. sind $E_1 = \{2, 4, 6\}$ und $E_2 = \{1\}$ disjunkt; $E_1 = \{2, 4, 6\}$ und $E_2 = \{2\}$ sind hingegen nicht disjunkt).

Paarweise disjunkte Ereignisse
Mehrere Ereignisse E_i heißen paarweise disjunkt, wenn alle möglichen Paare von Ereignissen disjunkt sind (z.B. sind $E_1 = \{2, 4\}$, $E_2 = \{1, 3\}$ und $E_3 = \{5, 6\}$ paarweise disjunkt; $E_1 = \{2, 4\}$, $E_2 = \{1, 3\}$ und $E_3 = \{4, 6\}$ sind hingegen nicht paarweise disjunkt).

Komplementärereignis
Das Komplementärereignis E^C tritt genau dann ein, wenn das Ereignis E nicht eintritt. Es umfasst also alle Elemente, die zwar in Ω_X, nicht aber in E enthalten sind (Das Komplementärereignis zu $E = \{2, 4, 6\}$ ist $E^C = \{1, 3, 5\}$). Wie man sich leicht überlegen kann, sind Ereignis und Komple-

mentärereignis stets disjunkt. Die Vereinigung aus Ereignis und Komplementärereignis ist der Wertebereich Ω_X.

Zerlegung
Mehrere Ereignisse E_i heißen Zerlegung des Wertebereiches Ω_X, wenn die Ereignisse E_i paarweise disjunkt sind und die Vereinigung aller Ereignisse wieder den Wertebereich ergibt (z.B. bilden $E_1 = \{2, 4\}$, $E_2 = \{1, 3\}$ und $E_3 = \{5, 6\}$ eine Zerlegung von Ω_X).

10.3 Denkmodelle für den Wahrscheinlichkeitsbegriff

Es gibt verschiedene Denkmodelle, die zur Erfassung des Begriffes Wahrscheinlichkeit hilfreich sein können. Wir wollen zwei dieser Denkmodelle kurz betrachten.

10.3.1 Wahrscheinlichkeit als Anteil

Wir betrachten ein Zufallsexperiment, welches folgende Eigenschaften aufweist:

- Das Experiment besitzt nur endlich viele Ausgänge.
- Alle Ausgänge können als gleichwahrscheinlich betrachtet werden, z.B. aufgrund von Symmetrieüberlegungen (Würfel).

Dann ist es plausibel einem Ereignis E den Quotienten

$$Pr(E) = \frac{\text{Anzahl der zu } E \text{ gehörenden Ereignisse}}{\text{Anzahl aller möglichen Ereignisse}} = \frac{|E|}{|\Omega|}$$

als Wahrscheinlichkeit von E zuzuordnen. Man nennt dies **Laplace-Wahrscheinlichkeit**, klassische Wahrscheinlichkeit oder **Abzählregel**. Die Abzählregel - in der leicht merkbaren Kurzfassung „Günstige durch Mögliche" - ist aber nur bei jenen Zufallsexperimenten anwendbar, bei denen alle Elementarereignisse gleichwahrscheinlich sind.

Abzählregel

$$Pr(E) = \frac{|E|}{|\Omega|} = \frac{\text{„Günstige"}}{\text{„Mögliche"}}$$

$Pr(E)$ ... Probability, Wahrscheinlichkeit von E

Beispiel 10.2. Würfel
Das Werfen eines Würfels erfüllt die genannten Voraussetzungen, denn es gibt nur endlich viele Ausgänge $\Omega = \{1, 2, 3, 4, 5, 6\}$ und alle sind (zumindest bei einem idealen Würfel) gleichwahrscheinlich. Daher kann die Wahrscheinlichkeit für das Ereignis $E = \{3\}$ berechnet werden als $Pr(E) = |E|/|\Omega| = 1/6$.

10.3.2 Wahrscheinlichkeit als relative Häufigkeit

Wir betrachten ein Zufallsexperiment, welches beliebig oft wiederholt wird, wobei die Wiederholungen einander nicht beeinflussen. Tritt bei n Versuchen das Ereignis k-mal ein, so kann man mittels $p_E = k/n$ die relative Häufigkeit des Ereignisses E unter diesen n Versuchen berechnen. Könnte man diesen Versuch unendlich oft durchführen, so würde sich die relative Häufigkeit bei einer Art Grenzwert stabilisieren, den man als Wahrscheinlichkeit bezeichnet. Allerdings ist dies kein Grenzwert im strengen Sinne der Mathematik. Dafür würde man verlangen, dass für jede beliebige Genauigkeit eine Anzahl von Versuchen bestimmt werden kann, ab der die relative Häufigkeit nicht weiter vom Grenzwert abweicht als die gewünschte Genauigkeit. Bei unserer Art Grenzwert wächst nur die Wahrscheinlichkeit des Erreichens der Genauigkeit mit wachsendem Stichprobenumfang, die sichere Einhaltung der gewünschten Genauigkeit kann aber nicht garantiert werden.

Dieses Denkmodell hat den Vorteil, dass man sich Wahrscheinlichkeit als etwas Ähnliches wie eine relative Häufigkeit vorstellen kann. Es birgt aber die Gefahr in sich, dass dieser Grenzwertbegriff tatsächlich als echter Grenzwert angesehen wird, was aber falsch ist und zu Fehlinterpretationen führen kann. Eine solche fehlerhafte Überlegung wäre etwa die Folgende: Jetzt muss doch endlich ein 6er kommen, weil alle anderen Zahlen schon so oft gekommen sind und insgesamt alle Zahlen gleich oft kommen müssen.

10.4 Rechnen mit Wahrscheinlichkeiten

Grundlage für das Rechnen mit Wahrscheinlichkeiten sind die Axiome von Kolmogorov. Das Wort Axiom bedeutet Grundwahrheit, in der Mathematik meint man damit Aussagen, die keinen Beweis benötigen. Aus diesen Axiomen lassen sich dann weitere Aussagen ableiten, deren Gültigkeit allerdings zu beweisen ist.

10.4.1 Axiome von Kolmogorov

Die Axiome von Kolmogorov beschreiben in mathematischer Form die Eigenschaften einer Wahrscheinlichkeitsverteilung. Alle Wahrscheinlichkeitsverteilungen erfüllen diese drei Axiome.

Axiome von Kolmogorov

1. $0 \leq Pr(E) \leq 1$ für alle Ereignisse $E \subseteq \Omega$
2. $Pr(\{\}) = 0$ und $Pr(\Omega) = 1$
3. $Pr(E_1 \cup E_2) = Pr(E_1) + Pr(E_2)$ für disjunkte Ereignisse E_1 und $E_2 \subseteq \Omega$

Verbal ausgedrückt bedeuten diese Axiome Folgendes:

1. Für alle Ereignisse liegt die Wahrscheinlichkeit des Eintreffens immer zwischen 0 und 1.
2. Das unmögliche Ereignis tritt mit der Wahrscheinlichkeit Null ein, und das sichere Ereignis tritt mit der Wahrscheinlichkeit 1, also 100%, ein.
3. Sind zwei Ereignisse disjunkt, so kann die Wahrscheinlichkeit dafür, dass das Ereignis 1 oder das Ereignis 2 eintritt, als Summe der beiden Einzelwahrscheinlichkeiten berechnet werden.

Die Sinnhaftigkeit der ersten beiden Axiome ist augenscheinlich, ein Beispiel zeigt sehr schnell die Intention des dritten Axioms.

Beispiel 10.3. Würfel
Wir betrachten das Zufallsexperiment „Werfen eines Würfels" und interessieren uns für die Zufallsvariable „Geworfene Augenzahl". Sei nun $E_1 = \{1\}$ und $E_2 = \{2\}$ mit den Wahrscheinlichkeiten $Pr(E_1) = 1/6$ und $Pr(E_2) = 1/6$. Möchte man nun die Wahrscheinlichkeit für das Ereignis „1 oder 2" berechnen, so ist das die Summe aus den beiden Einzelwahrscheinlichkeiten.

$$Pr(\{1,2\}) = Pr(E_1 \cup E_2) = Pr(E_1) + Pr(E_2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Sind die Ereignisse hingegen nicht elementfremd, so kann die Wahrscheinlichkeit nicht als Summe der einzelnen Wahrscheinlichkeiten berechnet werden. Die Ereignisse $E_1 = \{2,4\}$ und $E_2 = \{1,2,3\}$ haben die Wahrscheinlichkeiten $Pr(E_1) = 2/6$ und $Pr(E_2) = 3/6$. Damit folgt:

$$Pr(E_1 \cup E_2) = Pr(\{1,2,3,4\}) = \frac{4}{6} \quad \neq \quad Pr(E_1) + Pr(E_2) = \frac{5}{6}$$

Aus den Axiomen von Kolmogorov lassen sich weitere Rechenregeln ableiten:

Rechenregeln

1. $Pr(E^C) = 1 - Pr(E)$
2. $Pr(E_1 \cup E_2) = Pr(E_1) + Pr(E_2) - Pr(E_1 \cap E_2)$
3. $Pr(\bigcup_{i=1}^{k} E_i) = \sum_{i=1}^{k} Pr(E_i)$ für paarweise disjunkte Ereignisse E_i.
4. $Pr(E_1 \backslash E_2) = Pr(E_1) - Pr(E_1 \cap E_2)$

Anmerkungen zu diesen Rechenregeln:

1. $Pr(E^C)$ wird als **Gegenwahrscheinlichkeit** des Ereignisses E bezeichnet.
2. Dieser **Additionssatz** ist eine Erweiterung des dritten Axioms auf beliebige (disjunkte und nicht disjunkte) Ereignisse.
3. Dies ist eine Erweiterung des dritten Axioms auf eine beliebige Anzahl von disjunkten Ereignissen.
4. Dies ist eine Erweiterung der Gegenwahrscheinlichkeit, für $E_1 = \Omega$ erhält man die erste Rechenregel.

10.4.2 Bedingte Wahrscheinlichkeiten

Als einführendes Beispiel zu den bedingten Wahrscheinlichkeiten betrachten wir das Zufallsexperiment Werfen eines Würfels. Wir können bereits die Wahrscheinlichkeit für das Ereignis „Gerade Zahl" und für das Ereignis „Augenzahl 2" berechnen. Wie groß ist aber die Wahrscheinlichkeit für die Augenzahl 2, wenn man bereits weiß, dass eine gerade Zahl gekommen ist? Man könnte nun das Zufallsexperiment an die neuen Gegebenheiten anpassen, dann wäre $\widetilde{\Omega} = \{2, 4, 6\}$, $\widetilde{A} = \{2\}$ und damit $Pr(\widetilde{A}) = 1/3$.
Diese Wahrscheinlichkeit lässt sich aber auch ohne Anpassung des Zufallsexperimentes berechnen. Sei $\Omega = \{1, 2, 3, 4, 5, 6\}$, $A = \{2\}$ und $B = \{2, 4, 6\}$. Mit $Pr(A|B)$ bezeichnet man die Wahrscheinlichkeit für das Ereignis A unter der Bedingung, dass B bereits eingetreten ist. Durch die zusätzliche Information (das Ereignis „Gerade Zahl" ist eingetreten) kann sich die Wahrscheinlichkeit für das interessierende Ereignis verändern.

Bedingte Wahrscheinlichkeit

Für Ereignisse $A, B \subseteq \Omega$ mit $Pr(B) > 0$ gilt:

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

In unserem Beispiel ist die Wahrscheinlichkeit für das gemeinsame Eintreten der Ereignisse A und B (Es kommt eine gerade Zahl und es kommt die Augenzahl 2) gleich $Pr(A \cap B) = Pr(\{2\}) = 1/6$ und daher

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{1/6}{3/6} = \frac{1}{3}$$

.

Aus der Definition der bedingten Wahrscheinlichkeit lässt sich durch Umformung die Produktregel ableiten.

Produktregel

Für Ereignisse $A, B \subseteq \Omega$ mit $Pr(B) > 0$ gilt:

$$Pr(A \cap B) = Pr(A|B) \cdot Pr(B)$$

10.4.3 Stochastisch unabhängige Ereignisse

Zwei Ereignisse sind **stochastisch unabhängig**, wenn der Ausgang des einen Ereignisses die Wahrscheinlichkeit für das Eintreten des anderen Ereignisses nicht beeinflusst. Wird beispielsweise ein Würfel zweimal hintereinander geworfen, so tritt unabhängig davon, welche Augenzahl beim ersten Wurf geworfen wurde, beim zweiten Wurf jede Augenzahl mit der Wahrscheinlichkeit 1/6 ein.

Multiplikationsregel

Für stochastisch unabhängige Ereignisse $A, B \subseteq \Omega$ gilt:

$$Pr(A \cap B) = Pr(A) \cdot Pr(B)$$

Von einem unmöglichen Ereignis ist per Definition jedes Ereignis unabhängig. Aus der Multiplikationsregel folgt für stochastisch unabhängige Ereignisse auch $Pr(A|B) = Pr(A)$ und $Pr(B|A) = Pr(B)$.

Beispiel 10.4. Würfel
Ein Würfel wird zweimal hintereinander geworfen. Wie hoch ist die Wahrscheinlichkeit für das Ereignis „Beide Male kommt die Augenzahl 6"?
Man könnte dies als neues Zufallsexperiment mit dem Wertebereich $\widetilde{\Omega} = \{(1,1),(1,2),\ldots,(6,6)\}$ mit $|\widetilde{\Omega}| = 36$ betrachten, in dem man am Ereignis $\widetilde{A} = \{(6,6)\}$ interessiert ist. Die Wahrscheinlichkeit $Pr(\widetilde{A})$ lässt sich mit der Abzählregel berechnen als $Pr(\widetilde{A}) = 1/36$.
Ebenso kann man von zwei stochastisch unabhängigen Ereignissen des Zufallsexperimentes „Werfen eines Würfels" mit $\Omega = \{1,2,3,4,5,6\}$ ausgehen, wobei A das Ereignis „Beim ersten Werfen kommt die Augenzahl 6" und B das Ereignis „Beim zweiten Werfen kommt die Augenzahl 6" bezeichnet. Aus dem Multiplikationssatz ergibt sich damit:

$$Pr(A \cap B) = Pr(A) \cdot Pr(B) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

10.4.4 Das Theorem von Bayes

In manchen Aufgabenstellungen kann es passieren, dass man Informationen über bedingte Ereignisse hat, aber die Wahrscheinlichkeit für das Eintreten des Ereignisses ohne Bedingung vorerst unbekannt ist. Um diese zu berechnen, benötigen wir den bereits eingeführten Begriff der Zerlegung und den Satz von der totalen Wahrscheinlichkeit.

Satz von der totalen Wahrscheinlichkeit

Die Ereignisse $E_1,\ldots,E_r$ seien eine Zerlegung des Wertebereiches Ω. Dann gilt für $A \subseteq \Omega$

$$Pr(A) = \sum_{i=1}^{r} Pr(A|E_i) \cdot Pr(E_i)$$

Ein klassisches Beispiel für diese Aufgabenstellung bietet die Medizin.

Beispiel 10.5. Medizinische Untersuchung
Eine Untersuchungsmethode zur Früherkennung einer bestimmten Krankheit liefert bei einer kranken Person mit einer Wahrscheinlichkeit von 98% einen positiven Befund. Allerdings liefert diese Methode auch bei gesunden Menschen mit einer Wahrscheinlichkeit von 1% einen positiven Befund. Darüber hinaus ist bekannt, dass die Krankheit bei 3% der Bevölkerung auftritt. Man ist interessiert an der Wahrscheinlichkeit, dass eine zufällig ausgewählte Person einen positiven Befund hat.

Mit den Bezeichnungen

$$\begin{aligned} &\text{A} = \{\text{positiver Befund}\} \\ &E_1 = \{\text{Person krank}\} \text{ und} \\ &E_2 = \{\text{Person gesund}\} \end{aligned}$$

haben wir aus der Angabe folgende Informationen (Hinweis: E_1 und E_2 bilden eine Zerlegung):

$$\begin{aligned} &Pr(A|E_1) = 0,98 \\ &Pr(A|E_2) = 0,01 \\ &Pr(E_1) = 0,03 \\ &Pr(E_2) = 1 - Pr(E_1) = 0,97 \end{aligned}$$

Damit gilt nach dem Satz von der totalen Wahrscheinlichkeit:

$$\begin{aligned} Pr(A) &= Pr(A|E_1) \cdot Pr(E_1) + Pr(A|E_2) \cdot Pr(E_2) = \\ &= 0,98 \cdot 0,03 + 0,01 \cdot 0,97 = 0,0391 \end{aligned}$$

Die Wahrscheinlichkeit dafür, dass eine zufällig ausgewählte Person einen positiven Befund erhält, beträgt 3,9%.

Beispiel 10.5 wirft zugleich eine weitere Frage auf. Personen, die einen positiven Befund erhalten, möchten sicherlich wissen, wie hoch die Wahrscheinlichkeit ist, krank zu sein. Unser nächstes Ziel ist es, in der bedingten Wahrscheinlichkeit Bedingung und bedingtes Ereignis quasi zu tauschen. Zur Beantwortung dieser Frage benötigen wir die Definition der bedingten Wahrscheinlichkeit

$$Pr(E_1|A) = \frac{Pr(E_1 \cap A)}{Pr(A)}$$

Den Zähler stellen wir mit dem Produktsatz dar als

$$Pr(E_1 \cap A) = Pr(A|E_1) \cdot Pr(E_1)$$

für den Nenner benutzen wir den Satz der totalen Wahrscheinlichkeit

$$Pr(A) = \sum_{i=1}^{r} Pr(A|E_i) \cdot Pr(E_i)$$

und erhalten insgesamt einen Zusammenhang, der als Satz von Bayes bezeichnet wird:

Satz von Bayes

Die Ereignisse $E_1, \ldots, E_r$ seien eine Zerlegung des Wertebereiches Ω. Für mindestens ein i gilt $P(E_i) > 0$ und $P(A|E_i) > 0$. Dann gilt:

$$Pr(E_i|A) = \frac{Pr(A|E_i) \cdot Pr(E_i)}{Pr(A)} = \frac{Pr(A|E_i) \cdot Pr(E_i)}{\sum_{i=1}^{r} Pr(A|E_i) \cdot Pr(E_i)}$$

$Pr(E_i)$	a-priori Wahrscheinlichkeit
$Pr(E_i\|B)$	a-posteriori Wahrscheinlichkeit

Beispiel 10.6. Medizinische Untersuchung
Fortsetzung von Beispiel 10.5.
Mit dem Satz von Bayes lässt sich jetzt auch die Frage beantworten, mit welcher Wahrscheinlichkeit eine Person mit einem positiven Befund auch tatsächlich krank ist.

Mit den Bezeichnungen und Angaben gilt nach dem Satz von Bayes:

$$Pr(E_1|A) = \frac{Pr(A|E_1) \cdot Pr(E_1)}{Pr(A|E_1) \cdot Pr(E_1) + Pr(A|E_2) \cdot Pr(E_2)} = \\ = \frac{0,98 \cdot 0,03}{0,98 \cdot 0,03 + 0,01 \cdot 0,97} = 0,75$$

Die Wahrscheinlichkeit dafür, dass eine Person mit positivem Befund tatsächlich krank ist, beträgt 75%.

Übungsaufgaben

10.1. Roulette
Ein Zufallsexperiment bestehe aus dem Werfen einer Kugel in einen Trichter. Das Ergebnis des Zufallsexperimentes ist eine ganze Zahl zwischen 0 und 36. Geben Sie an:

a) den Wertebereich
b) zwei mögliche Ereignisse
c) zwei mögliche Elementarereignisse
d) zwei unmögliche Ereignisse
e) zwei disjunkte Ereignisse
f) das sichere Ereignis

10.2. Roulette
Fortsetzung von Aufgabe 10.1: Es sei A das Ereignis, das aus dem Eintreffen der geraden Ausgänge (0 ist keine gerade Zahl) besteht, B bestehe aus der Menge der Zahlen im 1. Drittel (1 bis 12) und C aus der Menge der Zahlen von 10 bis 15. Geben Sie folgende Ereignisse durch Aufzählen ihrer Elementarereignisse an:

a) $A \cap B$ und $A \cup B$
b) A^C
c) $A \cap B \cap C$
d) $A \cup B \cup C$ und $(A \cup B \cup C)^C$
e) Sind die Mengen A, B und C paarweise disjunkt?
f) Sind die Mengen $A \cup B \cup C$ und $(A \cup B \cup C)^C$ disjunkt?

10.3. Roulette
Fortsetzung von Aufgabe 10.2: Berechnen Sie die Wahrscheinlichkeiten folgender Ereignisse:

a) $Pr(A), Pr(B)$ und $Pr(C)$
b) $Pr(A \cap B)$ und $Pr(A \cup B)$
c) $Pr(A^C)$
d) $Pr(A \cap B \cap C)$
e) $Pr(A \cup B \cup C)$ und $Pr((A \cup B \cup C)^C)$

10.4. Roulette
Fortsetzung von Aufgabe 10.2: Berechnen Sie die Wahrscheinlichkeiten folgender Ereignisse:

a) Zweimal hintereinander das Ereignis C
b) C, gegeben beim vorherigen Versuch ist B eingetroffen
c) „2", gegeben eine gerade Zahl ist eingetroffen
d) C, gegeben „0" ist eingetroffen
e) C, gegeben „10" ist eingetroffen

10.5. Rubbellos
Bei einem Rubbellos verteilt sich eine Serie von 10.000.000 Losen bei einem Lospreis von 2 Euro folgendermaßen auf die Auszahlungskategorien:

Anzahl pro Serie	Auszahlung in Euro
10	30.000
50	3.000
150	1.000
10.000	100
200.000	10
850.000	4
1.750.000	2

Berechnen Sie die Wahrscheinlichkeiten dafür, dass das Los einen Auszahlungsbetrag von

a) 30.000 Euro anzeigt
b) mindestens 1.000 Euro anzeigt
c) höchstens 2 Euro anzeigt
d) mindestens 1.000 Euro anzeigt, gegeben es wurde keine Niete gezogen
e) genau 2 Euro anzeigt, gegeben es wurde keine Niete gezogen

10.6. Zwei Würfel
Ein Zufallsexperiment bestehe aus dem Werfen zweier Würfel. Das Ergebnis des Zufallsexperimentes ist die geworfene Augenzahl. Geben Sie an:

a) einige Elementarereignisse
b) den Ereignisraum (Wertebereich) Ω
c) einige Ereignisse (durch Aufzählen ihrer Elemente)
d) einige unmögliche Ereignisse
e) einige disjunkte (elementfremde) Ereignisse
f) das sichere Ereignis

10.7. Zwei Würfel
Fortsetzung von Aufgabe 10.6: Berechnen Sie die Wahrscheinlichkeiten folgender Ereignisse:

a) alle Elementarereignisse
b) „6", gegeben eine gerade Zahl
c) „6", gegeben eine ungerade Zahl
d) eine gerade Zahl, gegeben höchstens „4"
e) eine gerade Zahl, gegeben mindestens „4"
f) Berechnen Sie den Erwartungswert der Summe der Augenzahlen beim Würfeln zweier Würfel.

10.8. Zwei Würfel
Ein Zufallsexperiment bestehe im Werfen zweier Würfel. Das Ergebnis ist die

Summe der beiden Augenzahlen. Es sei A die Menge der geraden Ausgänge, B die Menge der Ausgänge unter 7 und C die Menge der Ausgänge über 7. Geben Sie folgende Mengen durch Aufzählen ihrer Elemente an:

a) $A \cap B$ und $A \cup B$
b) A^C (die zu A komplementäre Menge)
c) $A \cap B \cap C$ und $A \cup B \cup C$
d) $(A \cup B \cup C)^C$
e) Gibt es unter den Mengen A, B und C disjunkte Mengen?
f) Sind $A \cup B \cup C$ und $(A \cup B \cup C)^C$ disjunkte Mengen?

10.9. Zwei Würfel
Fortsetzung von Aufgabe 10.8: Berechnen Sie folgende Wahrscheinlichkeiten:

a) $Pr(A), Pr(B), Pr(C)$
b) $Pr(A \cap B)$, $Pr(A \cup B)$
c) $Pr(A^C)$
d) $Pr(A \cap B \cap C)$, $Pr(A \cup B \cup C)$
e) $Pr(A \cup B \cup C)^C$
f) $Pr(A \cup D)$ mit D = $\{7\}$.

10.10. Zwei Würfel
Ein Zufallsexperiment bestehe im Werfen zweier Würfel. Das Ergebnis sind die beiden Augenzahlen (z.B. (1 und 3)). Wie wahrscheinlich ist es, dass

a) der erste Würfel eine „1" und der zweite eine „3" anzeigt
b) beide Würfel eine „6" anzeigen
c) beide Würfel einen „Pasch", das sind zwei gleiche Augenzahlen, anzeigen
d) der erste Würfel eine gerade, der zweite eine ungerade Zahl anzeigt
e) der erste Würfel eine höhere Zahl als der zweite anzeigt.

10.11. Kinder
Ein Zufallsexperiment bestehe im Feststellen des Geschlechtes von 2 Kindern. Die Ausgänge sind die verschiedenen möglichen Kombinationen. Mit welcher Wahrscheinlichkeit sind beide Kinder einer Familie Mädchen,

a) wenn angenommen wird, dass Mädchen- und Knabengeburten gleich wahrscheinlich sind
b) wenn man weiß, dass das erste ein Mädchen ist
c) wenn man weiß, dass mindestens ein Kind ein Mädchen ist?

11

Diskrete Wahrscheinlichkeitsverteilungen

Bei Zufallsexperimenten mit diskreten Zufallsvariablen besteht der Ereignisraum bei entsprechender Kodierung aus einer Teilmenge der natürlichen Zahlen.

11.1 Dichte und Verteilungsfunktion

Gegeben sei eine diskrete Zufallsvariable X mit dem Wertebereich Ω. Man nennt jene Funktion $f(x)$, die jedem Elementarereignis $i \in \Omega$ seine Wahrscheinlichkeit $Pr(X = i)$ zuordnet, die Dichte einer diskreten Zufallsvariable.

Dichte einer diskreten Zufallsvariable

$$f(x) = \begin{cases} Pr(X = i) & \text{für } x = i \quad (\in \Omega) \\ 0 & \text{sonst} \end{cases}$$

Vorsicht: Bei stetigen Zufallsvariablen ist der Begriff Dichte anders definiert.

Alternativ werden auch die vereinfachten Schreibweisen f(i), Pr(i) verwendet.

Eigenschaften der Dichte

$$f(i) = Pr(X = i) \geq 0 \qquad \text{Nichtnegativität}$$

$$\sum_{i \in \Omega} f(i) = \sum_{i \in \Omega} Pr(X = i) = 1 \qquad \text{Normierung}$$

Jene Funktion $F(i)$, die jedem Elementarereignis i die Wahrscheinlichkeit dafür zuordnet, dass bei einem Versuch ein Ausgang $x \leq i$ beobachtet wird, nennt man die Verteilungsfunktion der Wahrscheinlichkeitsverteilung. Die Verteilungsfunktion ist stets nichtnegativ und monoton steigend.

Verteilungsfunktion einer diskreten Zufallsvariable

$$F(i) = Pr(X \leq i) = \sum_{j=1}^{i} Pr(X = j)$$

Eigenschaften der Verteilungsfunktion

$F(i) = Pr(x \leq i)$	≥ 0	$\forall\, i \in \Omega$	Nichtnegativität
$F(i) \leq F(i+1)$			monoton steigend

Es gilt

$$Pr(X = i) = F(i) - F(i-1)$$

und damit weiters:

$$\begin{aligned}
&Pr(a < x \leq b) = F(b) - F(a)\\
&Pr(a < x < b) = Pr(a < x \leq b-1) = F(b-1) - F(a)\\
&Pr(a \leq x < b) = Pr(a-1 < x \leq b-1) = F(b-1) - F(a-1)\\
&Pr(a \leq x \leq b) = Pr(a-1 < x \leq b) = F(b) - F(a-1)
\end{aligned}$$

Das Abbild der Verteilungsfunktion einer diskreten Zufallsvariable ist eine monoton von 0 nach 1 ansteigende Treppenfunktion.
Hinweis: Bei diskreten Zufallsvariablen ist die Unterscheidung zwischen *mindestens* ($\geq$) und *mehr als* ($>$) beziehungsweise zwischen *höchstens* ($\leq$) und *weniger als* ($<$) sehr wichtig.

Beispiel 11.1. Würfel
Wir werfen einmal einen Würfel und betrachten die geworfene Augenzahl X, demnach ist unser Wertebereich $\Omega_X = \{1, 2, 3, 4, 5, 6\}$. Dichte und Verteilungsfunktion können angeschrieben werden als:

$$f(i) = Pr(x = i) = \frac{1}{6} \qquad F(i) = Pr(x \leq i) = \frac{i}{6} \qquad i = 1, \ldots, 6$$

11.2 Lage- und Streuungsparameter

Für Zufallsvariablen lassen sich Lage- und Streuungsparameter bestimmen, die vergleichbar sind mit den Lage- und Streuungsmaßzahlen empirischer Verteilungen. Die Berechnung ist für diskrete Zufallsvariablen äquivalent zur Berechnung der Maßzahlen empirischer Verteilungen, allerdings ist die Interpretation unterschiedlich. Bei Häufigkeitsverteilungen entspricht das arithmetische Mittel ($\overline{x}$) der durchschnittlichen Ausprägung, die Varianz (s^2) ist die durchschnittliche quadratische Abweichung der Daten vom Mittelwert.

Bei Zufallsvariablen spricht man von einem Erwartungswert (bezeichnet mit $E(X)$ oder μ oder μ_x) statt von einem Mittelwert. Der Erwartungswert zeigt die Lage der theoretischen Verteilung und entsteht nicht aus konkreten Realisationen eines Zufallsexperimentes. Man berechnet also den Erwartungswert ohne das Zufallsexperiment konkret ausgeführt zu haben. Analog dazu gibt die theoretische Varianz (bezeichnet mit *Var(X)* oder σ^2) Auskunft über die erwartete quadratische Abweichung zum Erwartungswert.

Die Formeln zur Berechnung haben bei diskreten Zufallsvariablen die gleiche Bauart wie die Formeln zur Berechnung von Lagemaßzahlen von Merkmalen aus Häufigkeitsverteilungen, wie der folgende Vergleich zeigt:

Mittelwert und Varianz empirischer Häufigkeitsverteilungen für ein Merkmal mit r Ausprägungen werden berechnet mittels

$$\overline{x} = \sum_{i=1}^{r} x_i p_i$$

$$s^2 = \sum_{i=1}^{r} (x_i - \overline{x})^2 p_i$$

Erwartungswert und Varianz diskreter Zufallsvariablen

$$E(X) = \mu = \sum_{i=1}^{r} x_i Pr(x_i) \tag{11.1}$$

$$Var(X) = \sigma^2 = \sum_{i=1}^{r} (x_i - E(X))^2 Pr(x_i) \tag{11.2}$$

Der Vergleich der Formeln mit denen der empirischen Häufigkeitsverteilungen zeigt, dass auch hier das Denkmodell, Wahrscheinlichkeiten als eine Art Grenzwert zu betrachten, wieder wertvolle Dienste leisten kann.

Ein Beispiel zeigt, dass der Begriff Erwartungswert für Mittelwerte von Zufallsvariablen durchaus passend ist.

Beispiel 11.2. Rubbellos
Bei einem Rubbellos verteilt sich eine Serie von 10.000.000 Losen bei einem Lospreis von 2 Euro folgendermaßen auf die Auszahlungskategorien:

Anzahl pro Serie	Auszahlung in Euro
10	30.000
50	3.000
150	1.000
10.000	100
200.000	10
850.000	4
1.750.000	2

Berechnen Sie den durchschnittlich zu erwartenden Gewinn.

Bevor mit der Berechnung des Erwartungswertes begonnen werden kann, sollten zuerst einige grundsätzliche Überlegungen angestellt werden. Die Tabelle gibt den Auszahlungsbetrag an, wir hingegen sind am Gewinn interessiert, also ist der Auszahlungsbetrag um je 2 € zu reduzieren. Außerdem ist die Tabelle nicht vollständig, denn es fehlt die Angabe der Nieten. Die Wahrscheinlichkeit eines bestimmten Gewinnes wird mit Hilfe der Abzählregel berechnet (z.B. 10/10.000.000 bei einem Losgewinn von 29.998 €). Demnach kann der Erwartungswert folgendermaßen berechnet werden:

Gewinn in Euro	Wahrscheinlichkeit	$i \cdot Pr(X = i)$
29.998	0,000001	0,030
2.998	0,000005	0,015
998	0,000015	0,015
98	0,001000	0,098
8	0,020000	0,160
2	0,085000	0,170
0	0,175000	0,000
-2	0,718979	-1,438
	Summe	-0,950

Dieses Ergebnis bedeutet, dass man beim Kauf eines Rubbelloses einen durchschnittlichen Verlust von 0,95 € erwarten muss.

11.3 Spezielle diskrete Verteilungen

Für viele diskrete Verteilungen kann ein sogenanntes Urnenexperiment als Denkmodell für das Zufallsexperiment dienen. In der Praxis auftretende Fragestellungen können oft als Analogie zu einem Urnenexperiment betrachtet werden.

11.3.1 Alternativverteilung

Der Alternativverteilung entspricht folgendes Urnenexperiment: Gegeben sei eine Urne mit N Kugeln, von denen A Kugeln markiert sind. Man zieht eine Kugel aus dieser Urne und interessiert sich für die Wahrscheinlichkeit, eine nicht markierte bzw. eine markierte Kugel zu erhalten.

Alternativverteilung A(p)

N	Anzahl der Objekte in der Urne
A	Anzahl der markierten Objekte in der Urne
i	Anzahl der gezogenen markierten Objekte ($i = 0$ oder $i = 1$)
p	Erfolgswahrscheinlichkeit $p = A/N$

Dichte

$f(0) = Pr(i = 0) = 1 - p$ kein markierte Objekt gezogen
$f(1) = Pr(i = 1) = p$ ein markiertes Objekt gezogen

Verteilungsfunktion

$F(0) = f(0) = 1 - p$
$F(1) = f(0) + f(1) = 1$

Erwartungswert

$E(X) = p$

Varianz

$Var(X) = p(1 - p)$

Eine konkrete Alternativverteilung ist durch Angabe der „Erfolgswahrscheinlichkeit" p vollständig bestimmt, das bedeutet, dass weder A noch N explizit bekannt sein müssen. Diese Erfolgswahrscheinlichkeit p ist der **Parameter** (= Bestimmungsgröße) der Alternativverteilung.

Ein anderes Denkmodell für die Alternativverteilung ist das Folgende: Wir betrachten ein Zufallsexperiment, bei dem es genau zwei Alternativen gibt,

nämlich ein Ereignis E tritt ein oder es tritt nicht ein. Auch dieses Experiment lässt sich mit einer Alternativverteilung modellieren, der Parameter p ergibt sich aus der Wahrscheinlichkeit für das Eintreten des Ereignisses E.

Beispiel 11.3. Münzwurf
Sind wir bei einem einmaligem Münzwurf an der Wahrscheinlichkeit für „Kopf" interessiert, so kann man dies als Alternativverteilung formulieren. In diesem speziellen Fall wäre dann $p = 0,5$.

Beispiel 11.4. Würfel
Wollen wir bei einmaligem Würfeln die Wahrscheinlichkeit für das Ereignis „Augenzahl 6" wissen, so kann diese mit Hilfe der Alternativverteilung berechnet werden. In diesem speziellen Fall wäre dann $p = \frac{1}{6}$. Anders formuliert sind wir interessiert an der Anzahl der geworfenen 6er und deren Wahrscheinlichkeiten.

Möchte man Erwartungswert und Varianz berechnen, so könnte man natürlich auch die Formeln (11.1) und (11.2) verwenden, allerdings ist bei Vorliegen einer speziellen Verteilung die Verwendung der spezifischen Formel für Erwartungswert und Dichte meist effizienter.

11.3.2 Diskrete Gleichverteilung

Das Urnenexperiment der diskreten Gleichverteilung besteht im einmaligen Ziehen aus einer Urne mit N von 1 bis N durchnummerierten Objekten.

Diskrete Gleichverteilung G(N)

N	Anzahl der durchnummerierten Objekte
i	Nummer des gezogenen Objektes

Dichte

$$f(i) = Pr(x = i) = \frac{1}{N} \qquad i = 1, \ldots, N$$

Verteilungsfunktion

$$F(i) = Pr(x \leq i) = \frac{i}{N} \qquad i = 1, \ldots, N$$

Diskrete Gleichverteilung G(N)

Erwartungswert

$$E(X) = \frac{N+1}{2}$$

Varianz

$$Var(X) = \frac{N^2-1}{12}$$

Eine diskrete Gleichverteilung ist durch Angabe des Parameters N vollständig bestimmt. Verlässt man das Denkmodell der Urne, so könnte N auch ganz allgemein für eine Anzahl von möglichen Ausprägungen einer Zufallsvariable stehen, die alle gleichwahrscheinlich sind.

Beispiel 11.5. Würfel
Wir werfen einmal einen Würfel. Die geworfene Augenzahl unterliegt einer Gleichverteilung mit $N = 6$. Wir sind im Gegensatz zur Alternativverteilung nun nicht an der Anzahl der geworfenen Sechser interessiert, sondern generell an der geworfenen Augenzahl. An diesem Beispiel wird deutlich, dass Zufallsexperimente genau beschrieben werden müssen und dass die Statistik nach einer sehr genauen Ausdrucksweise verlangt.

11.3.3 Binomialverteilung

Wir erweitern nun unsere einfachen Experimente und ziehen mehrmals aus einer Urne. Nun ist es entscheidend, ob eine bereits gezogene Kugel nach der Ziehung wieder in die Urne zurückgelegt wird oder nicht. Wir betrachten vorerst den Fall, dass die Kugel nach jedem Zug wieder in die Urne zurückgelegt wird. Dies hat zur Folge, dass die Urne vor dem Entnehmen der nächsten Kugel wieder in den Anfangszustand zurückversetzt wird. Die Wahrscheinlichkeit, auch beim nächsten Ziehen wieder eine markierte Kugel zu ziehen, ist damit unabhängig von den Ergebnissen vorhergehender Ziehungen.

Ausgangspunkt unserer Überlegung ist eine mit Kugeln gefüllte Urne, von denen ein Anteil p markiert ist. Natürlich lässt sich ein solcher Anteil auch jederzeit aus der Anzahl der Kugeln N und der Anzahl der markierten Kugeln A über $p = \frac{A}{N}$ berechnen. Wir sind interessiert an der Wahrscheinlichkeit, nach n-maligem **Ziehen mit Zurücklegen** genau i markierte Kugeln gezogen zu haben.

Diesem Urnenexperiment entspricht die Binomialverteilung.

Binomialverteilung B(n,p)

p Anteil der markierten Objekte in der Urne
n Anzahl der (mit Zurücklegen) gezogenen Objekte
i Anzahl der gezogenen, markierten Objekte, $i = 0, \ldots, n$

Dichte

$$Pr(x = i) = \binom{n}{i} \cdot p^i \cdot (1-p)^{(n-i)}$$

mit Binomialkoeffizient $\binom{n}{i} = \frac{n!}{i!(n-i)!} \quad i = 0, \ldots, n$

Verteilungsfunktion

$$F(i) = Pr(x \leq i)$$

Erwartungswert

$$E(X) = np$$

Varianz

$$Var(X) = np(1-p)$$

Die Binomialverteilung besitzt zwei Parameter, nämlich n und p. Für $n = 1$ kommen wir wieder zur Alternativverteilung zurück, wie man leicht nachrechnen kann.

Beispiel 11.6. Münzwurf
Eine Münze werde dreimal hintereinander geworfen. Wie hoch ist die Wahrscheinlichkeit dafür, dass genau zweimal Kopf kommt? Mit $n = 3$ (entsprechend den 3 Würfen) und $p = 0,5$ (entspricht der Wahrscheinlichkeit für Kopf bei einem Wurf) erhält man für $i = 2$ (Anzahl von Kopf in den Würfen)

$$\begin{aligned} Pr(x = 2) &= \binom{3}{2} 0,5^2 \cdot (1 - 0,5)^{(3-2)} \\ &= \frac{3!}{2!(3-2)!} 0,5^2 \cdot 0,5^1 \\ &= 3 \cdot 0,5^3 = 0,375 \end{aligned}$$

Von EXCEL werden für die Berechnung von Dichten und Verteilungsfunktionen mancher Verteilungen Funktionen bereitgestellt. Die Funktionsanweisung in EXCEL für die Binomialverteilung ist:

= BINOMVERT(AnzahlErfolge; Versuche; ErfolgsWahrscheinlichkeit; Kumuliert)

Kumuliert ist ein Wahrheitswert, der bewirkt dass entweder die Dichte (*Falsch*) oder die Verteilungsfunktion (*Wahr*) berechnet wird. Anstatt im Feld *Kumuliert* die Wörter Falsch und Wahr auszuschreiben, kann auch für Falsch der Wert 0 und für Wahr der Wert 1 verwendet werden. Damit ergibt sich mit der von uns verwendeten Notation:

Binomialverteilung B(n,p) in EXCEL

Dichte $f(i)$ = BINOMVERT(i; n; p; 0)
Verteilungsfunktion $F(i)$ = BINOMVERT(i; n; p; 1)

Beispiel 11.7. Münzwurf, Umsetzung in EXCEL
Fortsetzung von Beispiel 11.6.
Die konkreten Zahlen $i = 2$, $n = 3$ und $p = 0,5$ führen auf die Ergebnisse $f(2) = 0,375$ und $F(2) = 0,875$.
Das bedeutet, dass die Wahrscheinlichkeit, bei dreimaligem Werfen einer Münze genau zweimal Kopf zu erhalten, 37,5 % beträgt. Die Wahrscheinlichkeit, *höchstens* zweimal Kopf zu erhalten, beträgt 87,5 %.

11.3.4 Hypergeometrische Verteilung

Als Anfangszustand haben wir wieder eine Urne mit N Kugeln, von denen A Kugeln markiert sind. Man zieht nun n Kugeln **ohne Zurücklegen** aus dieser Urne und ist interessiert an der Wahrscheinlichkeit, unter diesen n Kugeln insgesamt i markierte Kugeln zu haben.

Der wesentliche Unterschied zum Experiment der Binomialverteilung ist die Tatsache, dass die Kugel nach dem Ziehen *nicht* in die Urne zurückgelegt wird. Damit ändert sich nach jeder Entnahme einer Kugel die Wahrscheinlichkeit, bei der nächsten Ziehung eine markierte Kugel zu ziehen.

Die Hypergeometrische Verteilung besitzt die drei Parameter N, A und n.

Hypergeometrische Verteilung H(N,A,n)

N	Anzahl der Objekte in der Urne
A	Anzahl der markierten Objekte in der Urne
n	Anzahl der gezogenen Objekte (ohne Zurücklegen)
i	Anzahl der gezogenen, markierten Objekte, $i = 0, \ldots, n$

Dichte

$$Pr(x = i) = \frac{\binom{A}{i}\binom{N-A}{n-i}}{\binom{N}{n}} \qquad i = 0, \ldots, n$$

Verteilungsfunktion

$$F(i) = Pr(x \leq i)$$

Erwartungswert

$$E(X) = n \cdot \frac{A}{N}$$

Varianz

$$Var(X) = n \cdot \frac{A}{N} \cdot \frac{N-A}{N} \cdot \frac{N-n}{N-1}$$

Beispiel 11.8. Befragung
Von zwanzig VerkäuferInnen eines Geschäfts sind vier mit längeren Ladenöffnungszeiten einverstanden. Ein Journalist befragt für einen Beitrag drei zufällig ausgewählte Angestellte. Wie groß ist die Wahrscheinlichkeit, dass zwei der Befragten bereit sind, länger zu arbeiten?
Bei einer Befragung soll vermieden werden, dass dieselbe Person zweimal befragt wird. Daher entspricht eine Befragung dem Urnenmodell „Ziehen ohne Zurücklegen" und damit der hypergeometrischen Verteilung.
Mit $N = 20$, $A = 4$, $n = 3$ und $i = 2$ ergibt sich:

$$Pr(x = 2) = \frac{\binom{4}{2}\binom{16}{1}}{\binom{20}{3}} = \frac{6 \cdot 16}{1140} = 0,084$$

Demnach beträgt die Wahrscheinlichkeit dafür, dass zwei der Befragten mit längeren Ladenöffnungszeiten einverstanden sind, 8,4 %.

Für die Hypergeometrische Verteilung stellt EXCEL lediglich eine Funktion zur Berechnung der Dichte bereit. Die Verteilungsfunktion muss über Aufsummieren selbst berechnet werden. Die Syntax in EXCEL lautet:

= HYPGEOMVERT(Erfolge_S; Umfang_S; Erfolge_G; Umfang_G)

Hypergeometrische Verteilung H(N,A,n) in EXCEL

Dichte $f(i)$ = HYPGEOMVERT(i; n; A; N)
Verteilungsfunktion $F(i)$ = Summe der Dichten

Beispiel 11.9. Befragung, Umsetzung in EXCEL
Fortsetzung von Beispiel 11.8.
Für unser Beispiel ergibt sich somit f(2) = HYPGEOMVERT(2; 3; 4; 20) = 0,084. Möchte man die Verteilungsfunktion mit EXCEL berechnen, so muss man für alle benötigten Ausprägungen die Dichte berechnen und diese Teilergebnisse summieren. Die Möglichkeit, wie bei der Binomialverteilung direkt über eine EXCEL-Anweisung die Verteilungsfunktion zu berechnen, gibt es bei der hypergeometrischen Verteilung nicht.

11.3.5 Poissonverteilung

Bei der Binomialverteilung bzw. der Hypergeometrischen Verteilung haben wir Zufallsvariablen betrachtet, die einen Zählvorgang modellieren. Dabei war die Anzahl markierter Objekte nach n-maligen Ziehen aus einer Urne (mit bzw. ohne Zurücklegen) von Interesse. Damit war der Wertebereich der Zufallsvariablen in beiden Fällen nach oben beschränkt. Auch die Poissonverteilung eignet sich zum Modellieren von Zählvorgängen, allerdings ist nun der Wertebereich nach oben unbeschränkt und es wird das Auftreten von Ereignissen in bestimmten Bezugseinheiten, meistens Zeitintervalle, gezählt. Ist dieses Zeitintervall (sehr) klein, so ist auch die Wahrscheinlichkeit für das Eintreten eines Ereignisses in diesem Zeitintervall sehr klein. Deswegen bezeichnet man die Poissonverteilung auch als Verteilung der seltenen Ereignisse. Der Parameter λ der Poissonverteilung wird oft als Intensitätsrate oder als Ankunftsrate bezeichnet.

Poissonverteilung P(λ)

λ	durchschnittliche Anzahl an Ereignissen in einer Bezugseinheit (z. B. durchschnittliche Fehler pro 12 Stunden)
i	Anzahl der Ereignisse in dieser Bezugseinheit $i \geq 0$
e	eulersche Zahl, Exponentialfunktion

Dichte

$$Pr(x = i) = \frac{\lambda^i}{i!} \cdot e^{-\lambda}$$

Verteilungsfunktion

$$F(i) = Pr(x \leq i)$$

Erwartungswert

$$E(X) = \lambda$$

Varianz

$$Var(X) = \lambda$$

Der Parameter λ und die Anzahl i müssen sich auf dieselbe Bezugseinheit beziehen, im Bedarfsfall ist λ an die interessierende Bezugseinheit anzupassen.

Beispiel 11.10. Telefonanrufe
Die Anzahl der Telefonanrufe, die in einer Telefonvermittlung innerhalb einer Minute ankommen, sei poissonverteilt mit dem Parameter $\lambda = 1$. Bestimmen Sie die Wahrscheinlichkeit dafür, dass in einer Minute ein Anruf ankommt. Bestimmen Sie auch die Wahrscheinlichkeit dafür, dass in fünf Minuten sechs Anrufe ankommen.

Für den ersten Teil setzt man $i = 1$ und erhält

$$Pr(X = 1) = \frac{1}{1!} \cdot e^{-1} = \frac{1}{e} = 0,368$$

Für den zweiten Teil muss der Parameter λ an das interessierende Zeitintervall angepasst werden, daher ist nun $\lambda = 5 \cdot 1 = 5$ und mit $i = 6$ erhält man

$$Pr(X = 6) = \frac{5^6}{6!} \cdot e^{-5} = 0,146$$

Beispiel 11.11. Telefonanrufe, Umsetzung in EXCEL
Fortsetzung von Beispiel 11.10.
Für die Poissonverteilung ist eine Funktion implementiert, mit der die Dichte oder die Verteilungsfunktion berechnet werden kann. Die Syntax lautet

= POISSON(x; Mittelwert; Kumuliert)

Damit erhalten wir für unser Beispiel:
Pr(X=1)= POISSON(1; 1 ; 0) = 0,368 und
Pr(X=6)= POISSON(6; 5 ; 0) = 0,146

Poissonverteilung P(λ) in EXCEL

Dichte $f(x)$ = POISSON(i; λ ; 0)
Verteilungsfunktion $F(x)$ = POISSON(i; λ ; 1)

11.4 Rechnen mit diskreten Verteilungen

Unter gewissen Umständen ist es zulässig, zur Berechnung von Wahrscheinlichkeiten einer Zufallsvariable eine Verteilung zu verwenden, die auf den ersten Blick nicht passend erscheint.
Zwischen Binomialverteilung und Hypergeometrischer Verteilung lag der einzige Unterschied im Denkmodell in der Frage des Zurücklegens. Wir stellen uns nun eine Urne mit sehr vielen Kugeln vor (z.B. eine Million Kugeln), aus der man nur wenige Kugeln herausnimmt (z.B. 2 Kugeln). In diesem speziellen Fall wird es für die Berechnung der Wahrscheinlichkeit, eine markierte Kugel zu ziehen, nahezu keinen Unterschied machen, ob mit oder ohne Zurücklegen gezogen wurde. Dem entsprechend könnte man, obwohl ohne Zurücklegen gezogen wurde, die Formel für Ziehen mit Zurücklegen verwenden und der Fehler im Ergebnis wäre verschwindend klein.
Mathematisch ausgedrückt hat man die Hypergeometrische Verteilung durch die Binomialverteilung **approximiert** (= angenähert). Diese Approximation ist allerdings nur dann zulässig, wenn gewisse Voraussetzungen erfüllt sind. In diesem Fall muss gewährleistet sein, dass der Stichprobenumfang n in Relation zur Größe der Grundgesamtheit N klein ist.

Bei genügend großem Stichprobenumfang n kann eine Binomialverteilung mit kleinem Anteil p (seltene Ereignisse) durch eine Poissonverteilung approximiert werden. Daraus abgeleitet kann auch eine Hypergeometrische Verteilung durch eine Poissonverteilung approximiert werden, wenn die Voraussetzungen erfüllt sind.

Approximationsregeln

$\mathbf{H}_{N,A,n} \approx \mathbf{B}_{n,p}$	mit $p = \frac{A}{N}$	wenn $n \leq \frac{N}{10}$
$\mathbf{B}_{n,p} \approx \mathbf{P}_{\lambda}$	mit $\lambda = np$	wenn $n \geq 10$ und $p \leq 0,1$
$\mathbf{H}_{N,A,n} \approx \mathbf{P}_{\lambda}$	mit $\lambda = \frac{nA}{N}$	wenn $10 \leq n \leq \frac{N}{10}$ und $\frac{A}{N} \leq 0,1$

Beispiel 11.12. Uhren
Eine Lieferung von 1000 Armbanduhren enthält 50 Einheiten mit wertmindernden Oberflächenfehlern. Zur Qualitätskontrolle wird eine Stichprobe von 20 Uhren gezogen und geprüft. Wie groß ist die Wahrscheinlichkeit, dass die Stichprobe genau eine beschädigte Einheit enthält?
Dieses Problem lässt sich mit einer Hypergeometrischen Verteilung modellieren mit den Parametern $N = 1000$, $A = 50$ und $n = 20$. Damit ergibt sich mit $i = 1$

f(1)= HYPGEOMVERT(1; 20; 50; 1000)= 0,381

Die Voraussetzungen für eine Approximation durch die Binomialverteilung sind erfüllt und man erhält mit $p = A/N = 0,05$

f(1)= BINOMVERT(1; 20; 0,05; 1)= 0,377

Da weiters auch die Voraussetzung für eine Approximation durch die Poissonverteilung erfüllt ist, erhält man mit $\lambda = n \cdot p = 1$

f(1) = POISSON(1; 1 ; 0) = 0,368

Zuletzt soll die Frage beantwortet werden, wozu diese Approximationen überhaupt notwendig sind, wenn man die vorliegende Verteilung kennt und z.B. über EXCEL die gewünschten Dichten berechnen kann. Kleine Änderungen im Beispiel 11.12 veranschaulichen die Notwendigkeit von Approximationen.

Beispiel 11.13. Uhren
Eine Lieferung von 100.000 Uhren enthält 40.000 Einheiten mit wertmindernden Fehlern. Zum Zweck einer statistischen Qualitätskontrolle wird eine Stichprobe von 1.000 Uhren gezogen und geprüft. Wie groß ist die Wahrscheinlichkeit, dass die Stichprobe genau 400 beschädigte Einheiten enthält?
Eingabe dieser Zahlenwerte in die Funktion für die Hypergeometrische Verteilung in EXCEL liefert lediglich
HYPGEOMVERT(400; 1.000; 40.000; 100.000) = „#ZAHL!"

EXCEL (Version 2002) gibt eine Fehlermeldung aus, weil bei der Berechnung Zahlen benötigt werden, die für eine Verarbeitung zu groß sind. Um trotzdem zu einem richtigen Ergebnis zu kommen, muss man daher auf die Binomialverteilung mit $p = 0{,}4$ ausweichen und erhält als Wahrscheinlichkeit 2,6%. In neueren EXCEL-Versionen kann auch für diese Zahlenkonstellation die Hypergeometrische Verteilung berechnet werden, ein Ausweichen wäre dann nicht notwendig.

Übungsaufgaben

11.1. Würfel
Sie werfen einmal einen Würfel und beobachten die Anzahl der Einser. Mit welcher Wahrscheinlichkeit befindet sich in diesem Versuch

a) kein Einser
b) ein Einser
c) höchstens ein Einser

11.2. Urne mit Zurücklegen
In einer Urne liegen 10 Kugeln, davon sind 4 rot. Man zieht mit Zurücklegen 3 Kugeln heraus. Mit welcher Wahrscheinlichkeit befinden sich unter den gezogenen Kugeln

a) keine einzige rote Kugel
b) eine rote Kugel
c) zwei rote Kugeln
d) alles rote Kugeln
e) Wie viele rote Kugeln befinden sich durchschnittlich in der Stichprobe?

11.3. Urne ohne Zurücklegen
In einer Urne befinden sich 10 Kugeln, von denen 4 rot sind. Man zieht ohne Zurücklegen 3 Kugeln heraus. Mit welcher Wahrscheinlichkeit befindet sich unter den gezogenen Kugeln

a) keine einzige rote Kugel
b) eine rote Kugel
c) zwei rote Kugeln
d) alles rote Kugeln
e) Wie viele rote Kugeln befinden sich durchschnittlich in der Stichprobe?

11.4. Karten
Aus einem Kartenstapel von 13 Karten (von „2" bis „As") wird eine Karte

gezogen. Wie wahrscheinlich ist es,

a) eine bestimmte Karte (z.B. das As) zu ziehen
b) eine Karte mit einem Wert von höchstens „5" zu ziehen
c) beim zweiten Ziehen eine höhere Karte wie beim ersten Mal zu ziehen, wenn beim ersten Mal ein „Bube" gezogen wurde und die erste Karte nicht zurückgelegt wird?

11.5. Joker
Betrachten Sie die österreichische Jokerziehung (dabei werden 6 Zahlen jeweils aus dem Zahlenbereich der ganzen Zahlen zwischen 0 und 9 gezogen; z.B.: 562876).

a) Welchem Urnenmodell entspricht die Jokerziehung?
b) Wie groß ist Ihre Gewinnchance für den Joker, d.h. dafür, dass alle 6 gezogenen Zahlen mit Ihrer sechsstelligen Lottoscheinnummer übereinstimmen?
c) Wie viele Zahlen einer Lottoscheinnummer stimmen durchschnittlich mit den Jokerzahlen überein?
d) Wie wahrscheinlich ist es, dass Sie bei der Jokerziehung einen „Dreier"-Gewinn „einstreifen" (d.h.: die letzten drei Ziffern der Ziehung stimmen mit jenen auf ihrem Lottoschein überein, nicht aber die Viertletzte! Z.B.: Ihre Nummer ist 562876 und gezogen wurde 233876)?

11.6. Münze
Eine Münze wird sechsmal geworfen. Mit welcher Wahrscheinlichkeit kann jemand, der rät,

a) alle sechs Ausgänge richtig vorhersagen?
b) mindestens einen Ausgang richtig vorhersagen?
c) Wenn nun 512 Personen versuchen, das Ergebnis der sechs Münzwürfe zu erraten, wie wahrscheinlich ist es, dass niemand alle sechs Versuche richtig vorhergesagt hat? Wie viele Personen sagen im Durchschnitt alle Münzwürfe richtig voraus?

11.7. Glühbirne
Die Wahrscheinlichkeit dafür, dass eine Glühbirne länger als 100 Stunden brennt, beträgt 0,2. Wie wahrscheinlich ist es, dass von 10 Glühbirnen

a) genau eine länger als 100 Stunden
b) mindestens eine länger als 100 Stunden brennt?

11.8. Lotto
Die Wahrscheinlichkeit für die Anzahl an „Sechsern" in einer normalen Lot-

torunde (keine Jackpotrunde), ist (annähernd) poissonverteilt mit einem Mittelwert von $\lambda = 1{,}2$. Wie wahrscheinlich ist es, dass es in einer Runde

a) einen Jackpot gibt (also kein „Sechser" erreicht wurde)
b) mindestens zwei „Sechser" gibt?

11.9. Autolackiererei
In einer Autolackiererei werden Kotflügel zunächst mit einer Grundierung und danach mit einem Deckglanzlack lackiert. Im Durchschnitt werden bei der Grundierung vier Staubpartikel je Kotflügel eingeschlossen.

a) Wie groß ist die Wahrscheinlichkeit, nach der Grundierung auf einem Kotflügel zwei Staubpartikel zu finden?
b) Wie groß ist die Wahrscheinlichkeit, insgesamt (auf vier Kotflügel) acht Staubpartikel zu finden?

11.10. Urne
Aus einer sehr großen Urne mit einem Anteil von 20% an schwarzen Kugeln werden ohne Zurücklegen $n = 30$ Kugeln gezogen.

a) Wie wahrscheinlich ist es, dass darunter höchstens 2 schwarze Kugeln sind?
b) Wie viele schwarze Kugeln befinden sich durchschnittlich unter 30 zufällig gezogenen Kugeln?
c) Überprüfen Sie, ob auch die Approximation durch eine Poissonverteilung zulässig ist.

11.11. Qualitätskontrolle
Eine Lieferung von $N = 5000$ Mikroprozessoren enthält 1 % Ausschuss. Im Rahmen einer statistischen Qualitätskontrolle werden $n = 200$ Einheiten entnommen. Wie groß ist die Wahrscheinlichkeit dafür, dass

a) mehr als 2 %
b) weniger als 0,5 %

fehlerhafte Einheiten in der Stichprobe gefunden werden?

12

Stetige Wahrscheinlichkeitsverteilungen

Ausgangspunkt ist ein Zufallsexperiment mit stetigem Ausgang, dem entsprechend umfasst der Wertebereich ein Intervall $[a, b] \subset \mathbb{R}$. Als Realisierung einer in $[a, b]$ stetigen Zufallsvariablen kann jeder beliebige Wert aus diesem Intervall auftreten. Es gibt somit überabzählbar unendlich viele mögliche Realisationen.

12.1 Dichte und Verteilungsfunktion

Als einführendes Beispiel betrachten wir ein Glücksrad, welches in vier gleich große Sektoren unterteilt ist (vgl. Abbildung 12.1). Das Glücksrad wird gedreht und nach zufälliger Zeit angehalten, ein Pfeil markiert bei Stillstand einen der Sektoren und den damit verbundenen Gewinn.

Statistisch gesprochen genügt die Zufallsvariable „Markierter Sektor" einer diskreten Gleichverteilung mit vier Ausprägungen (entsprechend den vier Sektoren). Demnach beträgt die Wahrscheinlichkeit für die Auswahl eines Sektors jeweils 25%. Betrachtet man nun nicht die Sektoren, sondern die zugehörigen Zentriwinkel, so würde der Sektor 1 einem Zentriwinkel zwischen 0° und 90° entsprechen. Bezogen auf den Zentriwinkel erhalten wir eine stetige Zufallsvariable, die in intervallskalierter Form vorliegt. In Kapitel 6.3.2 haben wir als Möglichkeit der grafischen Darstellung für intervallskalierte Merkmale das Histogramm kennen gelernt. Wesentlich bei einem Histogramm war die Darstellung der relativen Häufigkeiten als Flächen, was durch die Berechnung der Dichte erreicht wurde. Jetzt verwenden wir die gleiche Vorgehensweise für eine Zufallsvariable und erstellen ein Wahrscheinlichkeitshistogramm (vgl. Tabelle 12.1).

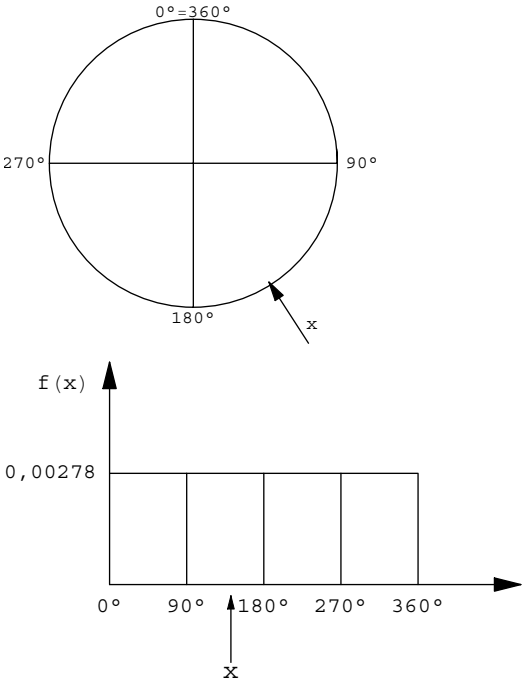


Abb. 12.1. Glücksrad mit 4 Sektoren, Wahrscheinlichkeitshistogramm

Tabelle 12.1. Glücksrad mit 4 Sektoren, Dichte

Winkel	Sektor j	$Pr(x = j)$	Intervallbreite d_j	Dichte $f_j = Pr(x = j)/d_j$
0° - 90°	1	0,25	90	0,00278
90° - 180°	2	0,25	90	0,00278
180° - 270°	3	0,25	90	0,00278
270° - 360°	4	0,25	90	0,00278

Im nächsten Schritt erhöhen wir die Sektorenzahl und erhalten damit eine feinere Intervallskalierung.

Abbildung 12.2 zeigt ein Glücksrad mit acht Sektoren, die nun jeweils mit einer Wahrscheinlichkeit von $12,5\%$ ausgewählt werden. Im zugehörigen Wahrscheinlichkeitshistogramm wird sichtbar, dass sich zwar die Anzahl der Flächen und die zugehörigen Wahrscheinlichkeiten geändert haben, nicht jedoch die Dichte.

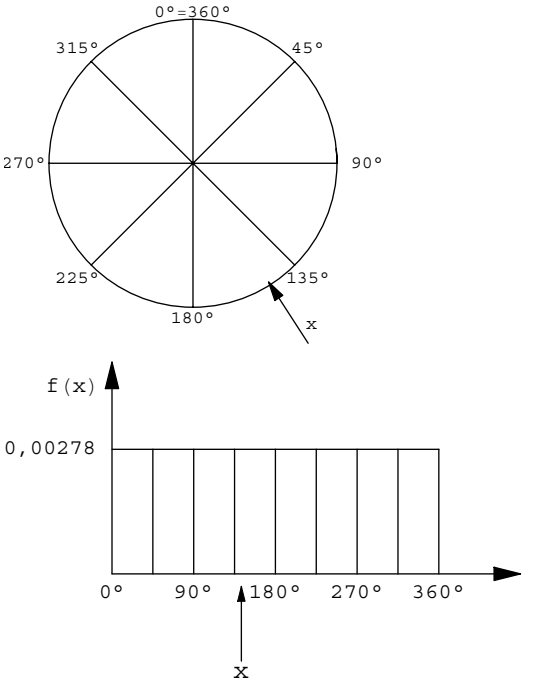


Abb. 12.2. Glücksrad mit 8 Sektoren, Wahrscheinlichkeitshistogramm

Tabelle 12.2. Glücksrad mit 8 Sektoren, Dichte

Winkel	Sektor j	$Pr(x = j)$	Intervallbreite d_j	Dichte $f_j = Pr(x = j)/d_j$
0° - 45°	1	0,125	45	0,00278
45° - 90°	2	0,125	45	0,00278
90° - 135°	3	0,125	45	0,00278
135° - 180°	4	0,125	45	0,00278
180° - 225°	5	0,125	45	0,00278
225° - 270°	6	0,125	45	0,00278
270° - 315°	7	0,125	45	0,00278
315° - 360°	8	0,125	45	0,00278

Man könnte nun die Anzahl der Sektoren immer weiter erhöhen und damit die Intervallskalierung immer feiner wählen. Dies würde aber weder die Dichte verändern, noch die Tatsache, dass die Wahrscheinlichkeit als Fläche unter der Dichtefunktion zu finden ist.

Ist die Dichte nicht (wie in diesem Beispiel) eine konstante Funktion, sondern eine beliebige stetige Funktion, so muss zur Berechnung der Fläche unter der

Dichtefunktion die Integralrechnung herangezogen werden. Damit erhalten wir die Definition einer Dichte für stetige Zufallsvariablen.

Dichte einer stetigen Zufallsvariable

Eine Zufallsvariable X heißt stetig, wenn es eine Funktion $f(x) \geq 0$ gibt, sodass für jedes Intervall [a,b]

$$Pr(a \leq x \leq b) = \int_a^b f(x)dx$$

gilt. Die Funktion $f(x)$ wird als Dichte bezeichnet.

Bei stetigen Zufallsvariablen entspricht die Dichte an der Stelle x nicht der Wahrscheinlichkeit des Ereignisses x, wie es bei diskreten Zufallsvariablen der Fall ist. Die Wahrscheinlichkeit von Ereignissen kann nur über das Integral der Dichte berechnet werden.

Für die Berechnung von Wahrscheinlichkeiten stetiger Zufallsvariablen ist es unerheblich, ob das interessierende Intervall an den Intervallgrenzen offen oder geschlossen ist. Die Wahrscheinlichkeit für einen einzelnen Punkt a ist gleich Null, da ein Punkt als Intervall [a,a] betrachtet werden kann und das Integral demnach gleich Null wäre. Ein einzelner Versuchsausgang besitzt somit zwar eine Dichte, aber keine von Null verschiedene Wahrscheinlichkeit.

Für stetige Zufallsvariablen gilt:

- $Pr(a \leq x \leq b) = Pr(a \leq x < b) = Pr(a < x \leq b) = Pr(a < x < b)$
- $Pr(X = x) = 0$ für alle $x \in \mathbb{R}$

Eigenschaften der Dichte

- Positivität: $f(x) \geq 0$ für alle $x \in \mathbb{R}$
- Normierung: $\int_{-\infty}^{+\infty} f(x)dx = 1$

Verteilungsfunktion einer stetigen Zufallsvariable

Die Funktion $F(a) = Pr(x \leq a)$ nennt man die Verteilungsfunktion der Wahrscheinlichkeitsverteilung von X

$$F(a) = Pr(x \leq a) = \int_{-\infty}^{a} f(x)dx$$

$F(a)$ gibt die Wahrscheinlichkeit an, eine Ausprägung kleiner oder gleich a zu beobachten.

Eigenschaften der Verteilungsfunktion:

- $F(a)$ ist stetig und monoton wachsend mit Werten im Intervall $[0, 1]$
- $F(-\infty) = 0$ und $F(\infty) = 1$
- $Pr(a \leq x \leq b) = F(b) - F(a)$ und $Pr(x \geq a) = 1 - F(a)$
- Für alle Werte x, für die $f(x)$ stetig ist, ist die Dichte die Ableitung der Verteilungsfunktion $F'(x) = f(x)$

12.2 Unabhängigkeit zweier stetiger Zufallsvariablen

Nachdem die Wahrscheinlichkeit einzelner Elementarereignisse bei stetigen Zufallsvariablen gleich Null ist, muss die Definition von unabhängigen Zufallsvariablen (vgl. Kapitel 10.4.3) modifiziert werden. In der Definition der Unabhängigkeit für stetige Zufallsvariablen wird die Wahrscheinlichkeit einzelner Ereignisse x durch die Wahrscheinlichkeit spezieller Ereignisse der Form $x \leq a$ ersetzt. Damit erhalten wir folgende Definition:

Unabhängigkeit stetiger Zufallsvariablen

Seien $F_X(a)$ und $F_Y(b)$ die Verteilungsfunktionen der beiden stetigen Zufallsvariablen X und Y. Dann heißen X und Y unabhängig, wenn für alle $a \in \mathbb{R}$ und $b \in \mathbb{R}$ gilt:

$$Pr(x \leq a, y \leq b) = Pr(x \leq a) \cdot Pr(y \leq b)$$

12.3 Lage- und Streuungsparameter

Wir kehren noch einmal zum Glücksrad zurück (Kapitel 12.1). Im Wahrscheinlichkeitshistogramm ergibt sich die Dichte als Quotient aus Wahrscheinlichkeit und Intervallbreite $f_j = Pr(x = j)/d_j$. Umgekehrt ist daher die Wahrscheinlichkeit das Produkt aus Dichte und Intervallbreite.

In Analogie zu diskreten Zufallsvariablen wäre der Erwartungswert

$$E(X) = \sum x_j Pr(x = j) = \sum x_j f_j d_j$$

Verkleinern der Intervalle führt zum Grenzübergang

$$\sum x_j f_j d_j \quad \stackrel{d_j \to 0}{\longrightarrow} \quad \int_{-\infty}^{\infty} x f(x) dx$$

Damit können nun Erwartungswert und Varianz definiert werden:

Erwartungswert und Varianz stetiger Zufallsvariablen

Der Erwartungswert $E(X)$ einer stetigen Zufallsvariablen X mit Dichte $f(x)$ ist

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Die Varianz $Var(X)$ einer stetigen Zufallsvariablen X mit Dichte $f(x)$ ist

$$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Wie für empirische Verteilungen können auch andere Maßzahlen für Wahrscheinlichkeitsverteilungen berechnet werden, wie z.B. Modus, Median oder Quantile. Mit Hilfe dieser Maßzahlen können dann Gestaltmerkmale, wie z.B. die Schiefe, analysiert werden.

Der Modus einer stetigen Zufallsvariable ist jener Wert, für den die Dichte ein (lokales) Maximum annimmt. Analog zu empirischen Verteilungen kann es auch hier Verteilungen mit mehreren Modi geben.

Quantile unterteilen die Daten in Gruppen, so dass ein bestimmter Prozentsatz über und ein bestimmter Prozentsatz unter dem Quantil liegt.

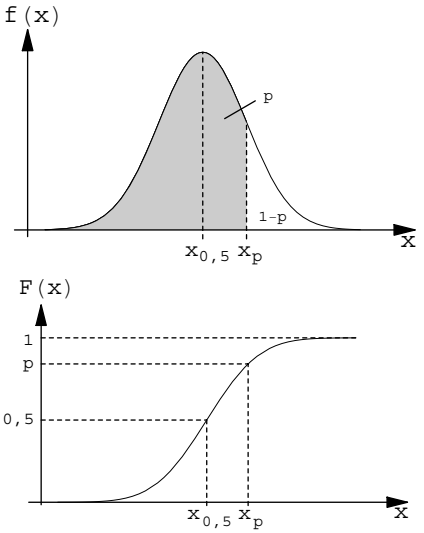


Abb. 12.3. Dichte und Verteilungsfunktion, jeweils mit Median und p-Quantil

Das p-Quantil ist somit jeder Wert x_p, für den mindestens der Anteil p der Daten kleiner oder gleich x_p und mindestens der Anteil $1-p$ der Daten größer oder gleich x_p ist. Das 0,5-Quantil wird als Median bezeichnet.

Zusammen mit der Definition der Verteilungsfunktion ergibt sich daher das p-Quantil als jener Wert, dessen Verteilungsfunktion gerade p ist:

Für das p-Quantil x_p einer quantitativen Variablen gilt

$$F(x_p) = Pr(x \leq x_p) = p$$

Aus dem Vergleich von Modus, Median und Erwartungswert lassen sich Aussagen über die Schiefe einer Verteilung treffen.

Schiefe einer Verteilung

$x_{mod} < x_{0,5} < E(X)$	rechtsschiefe Verteilung
$x_{mod} = x_{0,5} = E(X)$	symmetrische Verteilung
$x_{mod} > x_{0,5} > E(X)$	linksschiefe Verteilung

Abbildung 12.3 zeigt die Dichte und die Verteilungsfunktion einer stetigen Zufallsvariablen. Aus der Dichte ist erkennbar, dass eine unimodale Verteilung

vorliegt. Auch die Symmetrie der Verteilungsfunktion ist ersichtlich, weil der Modus gleichzeitig auch der Median ist. Die Abbildung zeigt weiters, dass ein p-Quantil die Fläche unter der Dichtefunktion in zwei Bereiche trennt, wobei der Bereich links vom Quantil den Anteil p der Gesamtfläche umfasst und rechts vom Quantil der Flächenanteil 1-p beträgt.

12.4 Die stetige Gleichverteilung

Die stetige Zufallsvariable X „Zentriwinkel der Glücksradstelle, auf die der Pfeil zeigt" entspricht einer auf dem Intervall [0, 360] gleichverteilten Zufallsvariablen, weil alle Intervalle gleicher Länge auch die gleiche Wahrscheinlichkeit aufweisen. Die Dichte ist lediglich vom Gesamtintervall abhängig, daher bilden die beiden Intervallgrenzen a und b die Parameter der stetigen Gleichverteilung $G(a, b)$.

Stetige Gleichverteilung G(a,b)

Dichte

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{für } a \le x \le b \\ 0 & \text{sonst} \end{cases}$$

Verteilungsfunktion

$$F(x) = \begin{cases} 0 & \text{für } x < a \\ \dfrac{x-a}{b-a} & \text{für } a \le x \le b \\ 1 & \text{für } x > b \end{cases}$$

Erwartungswert

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \frac{a+b}{2}$$

Varianz

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \frac{(b-a)^2}{12}$$

Aus der grafischen Darstellung der Dichte (Abbildung 12.4) ist ersichtlich, dass die stetige Gleichverteilung symmetrisch ist mit $x_{0,5} = E(X) = (a+b)/2$. Die stetige Gleichverteilung besitzt keinen Modus.

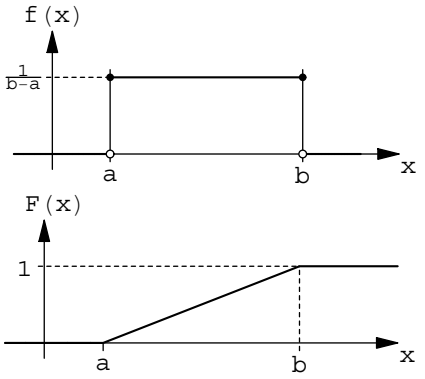


Abb. 12.4. Dichte und Verteilungsfunktion der stetigen Gleichverteilung

12.5 Die Normalverteilung

Die wichtigste und auch bekannteste stetige Verteilung ist die Normalverteilung. Viele Verteilungen lassen sich unter gewissen Voraussetzungen durch eine Normalverteilung approximieren.

Normalverteilung NV(μ, σ^2)

Eine stetige Zufallsvariable X heißt normalverteilt, wenn sie eine Dichte der Form

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad \text{mit } \mu, \sigma \in \mathbb{R} \text{ und } \sigma \geq 0$$

besitzt.

Die Parameter dieser Verteilung sind der Erwartungswert μ und die Varianz σ^2.

In Abbildung 12.5 sind links Dichten von Normalverteilungen dargestellt, die sich hinsichtlich des Erwartungswertes unterscheiden, rechts unterscheiden sich die Verteilungen hinsichtlich der Varianz. Außerdem ist in dieser Abbildung die Symmetrie der Normalverteilung erkennbar, denn für jede Normalverteilung gilt $x_{mod} = x_{0,5} = E(X)$.

Eine besondere Stellung nimmt die Normalverteilung mit den Parametern $\mu = 0$ und $\sigma^2 = \sigma = 1$ ein. Diese spezielle Verteilung wird als **Standardnormalverteilung** bezeichnet.

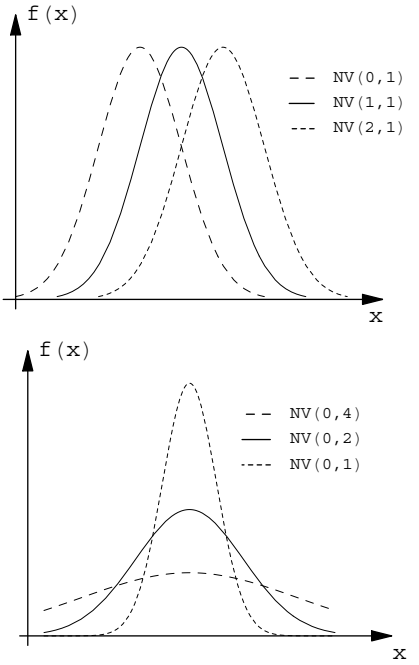


Abb. 12.5. Dichten von Normalverteilungen

Eine normalverteilte Zufallsvariable X wird durch Standardisierung in eine standardnormalverteilte Zufallsvariable U transformiert.

$$u = \frac{x - \mu}{\sigma} \qquad \textbf{Standardisierung}$$

Diese Standardisierung ist insbesondere für die händische Berechnung von Wahrscheinlichkeiten normalverteilter Zufallsvariablen von Bedeutung, denn es gilt:

$$\begin{aligned} Pr(x \leq a) &= Pr\left(\frac{x-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) = Pr\left(u \leq \frac{a-\mu}{\sigma}\right) = \\ &= \Phi\left(\frac{a-\mu}{\sigma}\right) = \Phi(u) \end{aligned} \tag{12.1}$$

Dabei bezeichnet $\Phi(u)$ die Verteilungsfunktion der Standardnormalverteilung an der Stelle u.

Verbal ausgedrückt bedeutet Formel (12.1) Folgendes: Der Wert der Verteilungsfunktion für eine Ausprägung a einer normalverteilten Zufallsvariable

mit den Parametern μ und σ^2 ist gleich dem Wert der Verteilungsfunktion für die standardisierte Ausprägung $(a - \mu)/\sigma$ einer standardnormalverteilten Zufallsvariable.

Die Verteilungsfunktion ist in allgemeiner Form durch

$$F(a) = Pr(x \leq a) = \int_{-\infty}^{a} f(x)dx$$

gegeben. Dieses Integral ist im Falle der Normalverteilung nicht in geschlossener Form lösbar, dies gilt ebenso für die Standardnormalverteilung. Für die Standardnormalverteilung werden diese Integrale durch numerische Verfahren gelöst und die Ergebnisse tabelliert. Es ist völlig ausreichend, die Ergebnisse der Standardnormalverteilung zu tabellieren, weil man mittels Standardisierung jedes Problem einer beliebigen Normalverteilung in ein Problem der Standardnormalverteilung überführen kann.

Beispiel 12.1. Schrauben
Die Länge von Schrauben einer Produktion sei normalverteilt mit Mittelwert $4\,cm$ und Varianz $0,04\,cm^2$. Wie wahrscheinlich ist es, dass eine Schraube aus dieser Produktion höchstens 4,15 cm lang ist?
In statistische Ausdrucksweise übersetzt ist die Zufallsvariable X (Länge der Schrauben) normalverteilt mit $\mu = 4\,cm$ und $\sigma^2 = 0,04\,cm^2$.
Gesucht ist $Pr(x \leq 4,15) = F(4,15)$.
Unter Verwendung der Standardisierung ergibt sich:

$$Pr(x \leq 4,15) = Pr\left(u \leq \frac{4,15 - 4}{0,2}\right) = Pr(u \leq 0,75) = \Phi(0,75)$$

Der Wert $\Phi(0,75)$ ist aus Tabelle 1 im Anhang abzulesen. An den Zeilen- und Spaltenbezeichnungen der Tabelle findet man die u-Werte, im Inneren der Tabelle die zugehörigen Werte der Verteilungsfunktion der Standardnormalverteilung $\Phi(u)$. Für unser Beispiel benötigen wir in der mit „0,7" bezeichneten Zeile den Wert in der Spalte „0,05" und finden somit $\Phi(0,75) = 0,7734$. Damit ist die Wahrscheinlichkeit dafür, dass eine Schraube aus dieser Produktion höchstens $4,15\,cm$ lang ist 77,34%.

Die Wahrscheinlichkeiten normalverteilter Zufallsvariablen lassen sich in EXCEL mit der Anweisung

=NORMVERT(x; Mittelwert; Standabwn; Kumuliert)

berechnen. Kumuliert ist ein Wahrheitswert, der bewirkt, dass entweder die Dichte (Kumuliert = Falsch oder 0) oder die Verteilungsfunktion (Kumuliert = Wahr oder 1) berechnet wird.

Normalverteilung $\mathbf{N}(\mu, \sigma^2)$ in EXCEL

Dichte $f(x)$ = NORMVERT(x; μ; σ; 0)
Verteilungsfunktion $F(x)$ = NORMVERT(x; μ; σ; 1)

Beispiel 12.2. Schrauben, Umsetzung in EXCEL
(Fortsetzung von Beispiel 12.1) Wir benötigen die Verteilungsfunktion $F(4,15)$, daher lautet die EXCEL-Anweisung NORMVERT(4,15; 4; 0,2; wahr). Diese Anweisung liefert ebenfalls das Ergebnis 0,7734.

Folgende Zusammenhänge sind für die Berechnung von Wahrscheinlichkeiten hilfreich:

- $Pr(x \leq a) = Pr(x < a)$ für stetige Verteilungen
- $S(a) = Pr(x > a) = 1 - Pr(x \leq a)$ Zuverlässigkeitsfunktion, Überlebensfunktion
- $\Phi(-u) = 1 - \Phi(u)$
- $u_p = -u_{1-p}$

Im nächsten Beispiel soll die Vorgehensweise zur Berechnung eines Quantils gezeigt werden.

Beispiel 12.3. Schrauben
Die Länge von Schrauben einer Produktion sei normalverteilt mit Mittelwert $4\,cm$ und Varianz $0,04\,cm^2$. Berechnen Sie die Länge jener Schrauben, die mit einer Wahrscheinlichkeit von 95% unterschritten wird.
In statistische Ausdrucksweise übersetzt ist die Zufallsvariable X (Länge der Schrauben) normalverteilt mit $\mu = 4\,cm$ und $\sigma^2 = 0,04\,cm^2$. Gesucht ist das 95%-Quantil $x_{0,95}$. Auch hier ist die Standardisierung wieder hilfreich:

$$Pr(x \leq x_p) = Pr\left(\frac{x-\mu}{\sigma} \leq \frac{x_p-\mu}{\sigma}\right) = Pr\left(u \leq \frac{x_p-\mu}{\sigma}\right) =$$
$$= \Phi\left(\frac{x_p-\mu}{\sigma}\right) = \Phi(u_p) = p$$

Diese Gleichung wird nun von hinten nach vorne bearbeitet. Wir beginnen mit $p = 0,95 = \Phi(u_{0,95})$ und ermitteln das Quantil $u_{0,95}$ aus Tabelle 1. Dazu suchen wir den Wert 0,95 im Inneren der Tabelle, und lesen an den Zeilen-

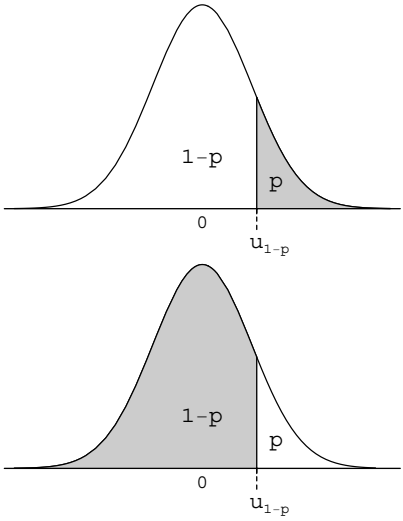


Abb. 12.6. $Pr(x > u_{1-p}) = 1 - Pr(x \leq u_{1-p})$

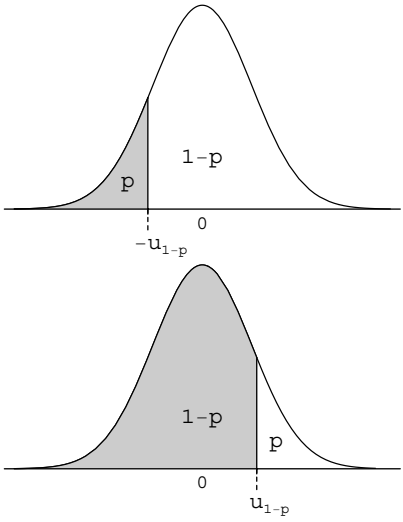


Abb. 12.7. $\Phi(-u_{1-p}) = 1 - \Phi(u_{1-p})$

bzw. Spaltenbezeichnungen das Quantil $u_{0,95}$ ab. Lässt sich der Wert 0,95 nicht exakt finden, so begnügen wir uns mit einem möglichst naheliegenden Wert, in unserem Fall mit 1,64 bzw. 1,65. Tabelle 1a listet besonders wichtige Quantile zusätzlich auf. Aus dieser Tabelle können wir $u_{0,95} \approx 1,645$ ablesen. Dieser Wert ist aber lediglich das Quantil der Standardnormalverteilung. Wir

müssen nun die Standardisierung quasi rückgängig machen, was mittels

$$x = \mu + u \cdot \sigma$$

erreicht wird. Daher ergibt sich als 0,95-Quantil $x_{0,95}$ in unserem Fall

$$x_{0,95} = 4 + 1,645 \cdot 0,2 \approx 4,33.$$

Mit einer Wahrscheinlichkeit von 95% wird eine Länge von 4,33 cm unterschritten.

Für die Berechnung von Quantilen stellt EXCEL die Funktion

NORMINV(Wahrsch;Mittelwert;Standabwn)

bereit.

Normalverteilung N(μ, σ^2) in EXCEL

p-Quantil = NORMINV(p; μ; σ)

Beispiel 12.4. Schrauben, Umsetzung in EXCEL
Fortsetzung von Beispiel 12.3
Für unser Beispiel liefert NORMINV(0,95; 4; 0,2) das Ergebnis 4,33.

Als nächster Schritt sei ein symmetrisches Intervall um den Mittelwert gesucht, welches eine Wahrscheinlichkeit von 95% aufweisen soll. Aus der Symmetrie der Normalverteilung folgt, dass die Länge der Schrauben mit je 2,5%iger Wahrscheinlichkeit unterhalb der unteren bzw. oberhalb der oberen Intervallgrenze liegt. Die obere Intervallgrenze ist demnach nichts anderes als das 0,975-Quantil, die untere Intervallgrenze dem entsprechend das 0,025-Quantil (vgl. Abbildung 12.8).

Beispiel 12.5. Schrauben, symmetrische Intervalle
Die Länge von Schrauben einer Produktion sei normalverteilt mit Mittelwert $\mu = 4\,cm$ und Varianz $\sigma^2 = 0,04\,cm^2$. Berechnen Sie ein um den Mittelwert symmetrisches Intervall, in welchem die Länge der Schrauben mit einer Wahrscheinlichkeit von 95% liegen.

Für die händische Berechnung kann das 0,975-Quantil der Standardnormalverteilung direkt aus der Tabelle 1 oder der Tabelle 1a entnommen werden $u_{0,975} \approx 1,96$. Für die Berechnung des 0,025-Quantils muss man noch zusätzlich den Zusammenhang $\Phi(-u) = 1 - \Phi(u)$ nutzen. Dieser Zusammenhang lässt sich auch anschreiben als $u_p = -u_{1-p}$, das bedeutet, man sucht hier das

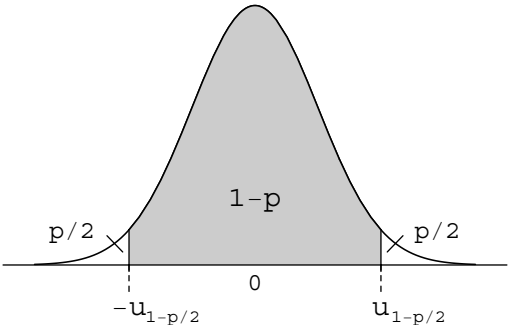


Abb. 12.8. Symmetrische Intervalle

1-0,025=0,975-Quantil und verwendet ein negatives Vorzeichen. Dem entsprechend gilt $u_{0,025} = -u_{0,975} \approx -1,96$. Diese Quantile der Standardnormalverteilung müssen nun in Quantile der vorliegenden Normalverteilung transformiert werden, in dem die Standardisierung rückgängig gemacht wird:

$$x_{0,025} = \mu + u_{0.025} \cdot \sigma = \mu - u_{0,975} \cdot \sigma = 4 - 1,96 \cdot 0,2 = 3,61$$

$$x_{0,975} = \mu + u_{0.975} \cdot \sigma = 4 + 1,96 \cdot 0,2 = 4,39$$

Es hätte natürlich auch genügt, die obere Intervallgrenze zu berechnen und dann aufgrund der Symmetrieüberlegungen die untere Intervalluntergrenze mittels

$$x_{0,025} = \mu - (x_{0,975} - \mu) = 4 - (4,39 - 4) = 3,61$$

zu berechnen. (Die obere Intervallgrenze muss vom Mittelwert gleich weit entfernt sein, wie die untere Intervallgrenze vom Mittelwert entfernt ist.)
Damit erhalten wir folgendes Ergebnis: Mit einer Wahrscheinlichkeit von 95% wird die Länge der Schrauben zwischen 3,61 cm und 4,39 cm betragen.

Bezeichnet man die Wahrscheinlichkeit des Intervalls allgemein mit $1 - \alpha$ so lässt sich das symmetrische Intervall um den Mittelwert in folgender Weise darstellen:

Symmetrische Intervalle

Für die normalverteilte Zufallsvariable X (kurz: $X \sim NV(\mu, \sigma^2)$) gilt:

$$Pr(\mu - u_{1-\frac{\alpha}{2}} \cdot \sigma \leq X \leq \mu + u_{1-\frac{\alpha}{2}} \cdot \sigma) = 1 - \alpha$$

12.6 Approximationen durch die Normalverteilung

Viele Verteilungen lassen sich durch eine Normalverteilung approximieren, was vor allem in der schließenden Statistik eine wichtige Rolle spielt. Wir wollen uns in einem ersten Teil die theoretischen Hintergründe dafür ansehen und in einem zweiten Teil verschiedene Approximationsmöglichkeiten näher beleuchten. Die theoretische Basis liefert das Gesetz der großen Zahlen und verschiedene Grenzwertsätze, die wir im Folgenden betrachten wollen.

12.6.1 Gesetz der großen Zahlen und Grenzwertsätze

Bezeichne X eine diskrete oder stetige Zufallsvariable, die eine beliebige Verteilung aufweist und den Erwartungswert μ und die Varianz σ^2 besitzt. Nun wiederholt man das zu X gehörende Zufallsexperiment n-mal hintereinander. Diese Wiederholungen seien darüber hinaus voneinander unabhängig. Jede einzelne Durchführung des Zufallsexperimentes erhält eine neue Bezeichnung $X_1, \ldots, X_n$. Damit sind die Zufallsvariablen $X_1, \ldots, X_n$ voneinander unabhängig und sie besitzen alle dieselbe Verteilung, nämlich genau diejenige, die auch X aufweist. Kürzer formuliert sagt man $X_1, \ldots, X_n$ sind **unabhängig und identisch verteilt** (**i**ndependent and **i**dentically **d**istributed - daher wird diese Bedingung in der Statistik auch kurz iid-Bedingung genannt).

Das arithmetische Mittel dieser Zufallsvariablen

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \left(X_1 + \cdots + X_n \right)$$

ist selbst wieder eine Zufallsvariable mit Erwartungswert μ und Varianz σ^2/n.

Damit ist der Erwartungswert des arithmetischen Mittels gleich dem Erwartungswert der einzelnen Verteilung von X, die Varianz strebt mit wachsendem n gegen Null. Daraus resultiert auch das Gesetz der großen Zahlen:

Gesetz der großen Zahlen

Für beliebig kleine $\varepsilon > 0$ gilt:

$$Pr(|\overline{X}_n - \mu| \leq \varepsilon) \longrightarrow 1 \qquad \text{für} \qquad n \longrightarrow \infty$$

Man sagt dazu $\overline{X}_n$ konvergiert stochastisch gegen μ.

Das Gesetz der großen Zahlen sagt aus, dass die Wahrscheinlichkeit dafür, dass das arithmetische Mittel in einem beliebig kleinen Intervall $[\mu - \varepsilon, \mu + \varepsilon]$ um den Erwartungswert liegt, gegen Eins konvergiert, wenn n unendlich groß wird. Für große n ist damit auch die Wahrscheinlichkeit für $Pr(\mu - \varepsilon \leq \overline{X}_n \leq \mu + \varepsilon)$ nahe Eins.

Wendet man dieses Gesetz der großen Zahlen auf Zufallsvariablen an, die $A(p) = B(1, p)$-verteilt sind, dann entspricht das arithmetische Mittel gerade der relativen Häufigkeit p_A des Eintretens von A. Dieser Spezialfall kommt im Theorem von Bernoulli zum Ausdruck:

Theorem von Bernoulli

Die relative Häufigkeit, mit der ein Ereignis A bei n voneinander unabhängigen Versuchen eintritt, konvergiert stochastisch gegen $Pr(A)$.

Nicht nur über den Erwartungswert und die Varianz von Summen von Zufallsvariablen kann eine Aussage getroffen werden, auch die Verteilung der Summe kann zumindest approximativ bestimmt werden.

Zentraler Grenzwertsatz

Seien $X_1, \ldots, X_n$ unabhängige, identisch verteilte Zufallsvariablen mit $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$. Dann ist die Zufallsvariable $Y = X_1 + \cdots + X_n$ approximativ normalverteilt mit Erwartungswert $n \cdot \mu$ und Varianz $n \cdot \sigma^2$.

$$Y \sim NV(n\mu, n\sigma^2)$$

Wendet man diesen zentralen Grenzwertsatz auf Zufallsvariablen an, die $A(p) = B(1, p)$-verteilt sind, erhält man als Spezialfall den Grenzwertsatz von de Moivre.

Grenzwertsatz von de Moivre

Für große n lässt sich die Binomialverteilung B(n,p) durch eine Normalverteilung mit dem Erwartungswert np und der Varianz $np(1 - p)$ approximieren.

Bei dieser Approximation muss eine sogenannte Stetigkeitskorrektur durchgeführt werden. Diese bewirkt, dass jedem diskreten Ereignis i das stetige Intervall $[i - 0,5; i + 0,5]$ zugeordnet wird. Eine Stetigkeitskorrektur ist immer dann zu berücksichtigen, wenn eine diskrete Verteilung durch eine stetige Verteilung approximiert wird.

12.6.2 Approximationen durch die Normalverteilung

Die im vorhergehenden Abschnitt beschriebenen Grenzwertsätze sind die theoretische Grundlage für Approximationen verschiedener Verteilungen durch die Normalverteilung. Für die Binomialverteilung und die Poissonverteilung werden diese Approximationen hier explizit angegeben:

Approximation der Binomialverteilung

$$B(n,p) \approx NV(\mu = np, \sigma^2 = np(1-p))$$

Voraussetzung: $\min\{np, n(1-p) > 10\}$

Stetigkeitskorrektur

Damit gelten folgende Approximationen:

- $Pr(x \leq i \mid B(n,p)) \approx Pr(x \leq i + 0,5 \mid NV(np, np(1-p)))$
- $Pr(x < i \mid B(n,p)) \approx Pr(x \leq i - 0,5 \mid NV(np, np(1-p)))$
- $Pr(x > i \mid B(n,p)) \approx 1 - Pr(x \leq i + 0,5 \mid NV(np, np(1-p)))$
- $Pr(x \geq i \mid B(n,p)) \approx 1 - Pr(x \leq i - 0,5 \mid NV(np, np(1-p)))$

Approximation der Poissonverteilung

$$P(\lambda) \approx NV(\mu = \lambda, \sigma^2 = \lambda)$$

Voraussetzung: $\lambda > 10$

Stetigkeitskorrektur

Damit gelten folgende Approximationen:

- $Pr(x \leq i \mid P(\lambda)) \approx Pr(x \leq i + 0,5 \mid NV(\lambda, \lambda))$
- $Pr(x < i \mid P(\lambda)) \approx Pr(x \leq i - 0,5 \mid NV(\lambda, \lambda))$
- $Pr(x > i \mid P(\lambda)) \approx 1 - Pr(x \leq i + 0,5 \mid NV(\lambda, \lambda))$
- $Pr(x \geq i \mid P(\lambda)) \approx 1 - Pr(x \leq i - 0,5 \mid NV(\lambda, \lambda))$

Ein Beispiel soll diese Anhäufung von Formeln etwas leichter verdaulich machen.

Beispiel 12.6. Uhren
(vgl. Beispiel 11.12, Seite 186) Eine Lieferung von 100.000 Uhren enthält 40.000 Einheiten mit wertmindernden Fehlern. Zum Zweck einer statistischen Qualitätskontrolle wird eine Stichprobe von 1.000 Uhren gezogen und geprüft. Wie groß ist die Wahrscheinlichkeit, dass die Stichprobe höchstens 400 beschädigte Einheiten enthält? Wie groß ist die Wahrscheinlichkeit, dass die Stichprobe genau 400 beschädigte Einheiten enthält? Dieses Problem lässt sich mit einer Hypergeometrischen Verteilung modellieren. Die Voraussetzungen für eine Approximation durch die Binomialverteilung sind erfüllt mit $p = A/N = 0,4$. Da weiters auch die Voraussetzung für eine Approximation durch die Normalverteilung erfüllt ist, erhält man mit $\mu = np = 400$ und $\sigma^2 = np(1-p) = 240$

$$\begin{aligned} Pr(x \leq 400) &= Pr(x \leq 400 \mid B(1000; 0,4)) = \\ &= Pr(x \leq 400,5 \mid NV(400, 240)) \\ &= \Phi\left(\frac{400,5 - 400}{\sqrt{240}}\right) = \Phi(0,03) = 0,512 \end{aligned}$$

Die Wahrscheinlichkeit dafür, dass die Stichprobe höchstens 400 beschädigte Einheiten enthält, beträgt 51,2%.

$$\begin{aligned} Pr(x = 400) &= Pr(x = 400 \mid B(1000; 0,4)) = \\ &= Pr(x \leq 400,5 \mid NV(400, 240)) - Pr(x \leq 399,5 \mid NV(400, 240)) = \\ &= \Phi\left(\frac{400,5 - 400}{\sqrt{240}}\right) - \Phi\left(\frac{399,5 - 400}{\sqrt{240}}\right) = \\ &= \Phi(0,03) - \Phi(-0,03) = 0,512 - (1 - 0,512) = 0,024 \end{aligned}$$

Die Wahrscheinlichkeit dafür, dass die Stichprobe genau 400 beschädigte Einheiten enthält, beträgt 2,4%.

Übungsaufgaben

12.1. Schrauben
Die Länge von Schrauben einer Produktion ist normalverteilt mit Mittelwert 4 cm und Varianz $0,04\,cm^2$. Berechnen Sie die Wahrscheinlichkeiten dafür, dass eine Schraube

a) höchstens 4,28 cm lang ist.
b) höchstens 3,8 cm lang ist.
c) mindestens 4,21 cm lang ist.
d) mindestens 3,91 cm lang ist.
e) zwischen 3,85 cm und 4,1 cm lang ist.

12.2. Schrauben
Fortsetzung von 12.1.

a) Berechnen Sie jene Länge der Schrauben, die mit einer Wahrscheinlichkeit von 0,90 unterschritten wird.
b) Geben Sie das um den Mittelwert symmetrische Intervall an, in welchem die Länge der Schrauben mit einer Wahrscheinlichkeit von 0,90 liegt.
c) Berechnen Sie jene Länge der Schrauben, die mit einer Wahrscheinlichkeit von 0,90 überschritten wird.

12.3. Produktion
Eine Maschine produziert fehlerfreie Stücke mit einer gleichbleibenden Wahrscheinlichkeit von 0,98. Mit welcher Wahrscheinlichkeit befinden sich in einer Serie vom Umfang 600

a) höchstens sechs defekte Stücke?
b) genau sechs defekte Stücke?

12.4. Lotto
Die Wahrscheinlichkeit für die Anzahl an „Sechsern" in einer Lottorunde ist poissonverteilt mit einem Mittelwert von 1,2. Berechnen Sie die Wahrscheinlichkeit dafür, dass in 100 normalen Lottorunden

a) höchstens 125 Sechser gezogen werden.
b) mindestens 100 Sechser gezogen werden.

12.5. Niederschlagsmenge
Langjährige Aufzeichnungen ergeben für die Niederschlagsmenge im April in Linz eine Normalverteilung mit $\mu = 75\ mm$ und $\sigma = 7\ mm$. Berechnen Sie

aus diesen Daten

a) das 0,05-Quantil dieser Verteilung.
b) das 0,95-Quantil dieser Verteilung.

12.6. Waschpulver
Eine Maschine paketiert Waschpulver mit einem Normgewicht von 1000 g. Das Merkmal „Gewicht pro Paket" ist in guter Näherung normalverteilt. In 90% aller Fälle haben die Pakete ein Gewicht zwischen 990 g und 1010 g. Welchen Wert muss das Normgewicht annehmen, wenn bei gleichbleibender Varianz nur mehr 1 % der Pakete einen Inhalt von weniger als 990 g haben dürfen?

12.7. Radioaktivität
Einer der bekanntesten Poissonprozesse in der Natur ist die Absonderung von Teilchen bei radioaktivem Material. Rutherford und Geiger untersuchten die Radioaktivität eines Poloniumpräparats, das durchschnittlich 150 Teilchen pro Sekunde emittierte. Wie groß ist die Wahrscheinlichkeit dafür, innerhalb einer Sekunde weniger als 125 abgestrahlte Teilchen zu registrieren?

12.8. Brücke
Das Gewicht von Bundesheersoldaten ist normalverteilt mit $\mu = 85\,\text{kg}$ und einer Standardabweichung von 8 kg. Ein Bataillon von 1170 Mann überquert eine Brücke mit einer Tragkraft von 100 Tonnen.

a) Mit welcher Wahrscheinlichkeit wird die Tragkraft der Brücke überschritten?
b) Welche Tragkraft müsste die Brücke aufweisen, damit diese nur mehr mit einer Wahrscheinlichkeit von 0,1 % überschritten wird?

12.9.
Eine diskrete Zufallsvariable X besitzt folgende Verteilung:

$X:$	-2	-1	0	1	2
$Pr(x):$	0,05	0,35	0,25	0,20	0,15

Bestimmen Sie die Wahrscheinlichkeit dafür, dass die Summe $Y = \sum_{i=1}^{40} x_i$

a) größer als null ist.
b) genau null ist.

Teil IV

Schließende Statistik

13

Die Gedankenwelt der schließenden Statistik

In der Praxis ist es oft nicht möglich, alle interessierenden Objekte der Grundgesamtheit zu untersuchen. Man muss sich dann mit einer Auswahl aus dieser Grundgesamtheit, der Stichprobe, begnügen. Ziel der schließenden Statistik ist es nun, Rückschlüsse von der Stichprobe auf die Grundgesamtheit zu ermöglichen.

Die schließende Statistik umfasst die beiden Teilbereiche Schätzen und Testen. Grundlage der Analyse ist in beiden Fällen eine Zufallsstichprobe aus der Grundgesamtheit (vgl. Kapitel 1.3). Alle hier vorgestellten Formeln und Verfahren beruhen auf dem Vorliegen einer einfachen Zufallsauswahl und dürfen daher nicht angewendet werden, wenn diese Voraussetzung verletzt ist.

Schließende Statistik

Wesentliche Voraussetzung für die Verfahren der schließenden Statistik ist das Vorliegen einer Zufallsstichprobe. Die schließende Statistik stellt Methoden bereit, die einen Rückschluss von einer Stichprobe auf die Grundgesamtheit zulassen.

13.1 Stichprobenverteilung

Wir betrachten ein Beispiel bestehend aus einer Grundgesamtheit von 6 Personen. Davon stimmen 4 Personen einer bestimmten Behauptung zu (z.B. die Personen 1, 3, 4 und 5). Dem entsprechend ist der Anteil der Zustimmung in der Grundgesamtheit $p = 0,667$. Wir ziehen nun alle möglichen Stichproben (ohne Zurücklegen) vom Umfang $n = 3$ aus dieser Grundgesamtheit und eruieren in jeder Stichprobe den Stichprobenanteil $\hat{p}$.

Tabelle 13.1. Stichproben vom Umfang n = 3

Nr	Ausgewählte Person						Anteil
	1	2	3	4	5	6	$\hat{p}$
1	ja	nein	ja				0,667
2	ja	nein		ja			0,667
3	ja	nein			ja		0,667
4	ja	nein				nein	0,333
5	ja		ja	ja			1,000
6	ja		ja		ja		1,000
7	ja		ja			nein	0,667
8	ja			ja	ja		1,000
9	ja			ja		nein	0,667
10	ja				ja	nein	0,667
11		nein	ja	ja			0,667
12		nein	ja		ja		0,667
13		nein	ja			nein	0,333
14		nein		ja	ja		0,667
15		nein		ja		nein	0,333
16		nein			ja	nein	0,333
17			ja	ja	ja		1,000
18			ja	ja		nein	0,667
19			ja		ja	nein	0,667
20				ja	ja	nein	0,667

Insgesamt gibt es 20 verschiedene Möglichkeiten, 3 Personen aus 6 Personen auszuwählen. Man sieht nun, dass einige Stichproben denselben Anteil an Zustimmung aufweisen wie die Grundgesamtheit, andere Stichproben jedoch zum Teil erheblich vom tatsächlichen Anteil abweichen. Die Variable „Stichprobenanteil" ist somit selbst eine Zufallsvariable, deren Verteilung in Tabellenform darstellbar ist:

Tabelle 13.2. Stichprobenverteilung des Anteils, n = 3

Ausprägung $\hat{p}$	*Anzahl*	$Pr(\hat{p})$
0,333	4	0,2
0,667	12	0,6
1,000	4	0,2
Summe	20	1,0

Berechnet man den Erwartungswert dieser Verteilung, so stellt man fest, dass der Erwartungswert der Stichprobenanteile gerade dem Anteil in der Grundgesamtheit entspricht.

$$E(\hat{p}) = \sum \hat{p} Pr(\hat{p}) = p$$

Die Berechnung der Varianz ergibt in diesem Fall $Var(\hat{p}) \approx 0,044$.
Nun betrachten wir den Fall $n = 5$.

Tabelle 13.3. Stichprobenverteilung des Anteils, n = 5

Nr	Ausgewählte Person						Anteil
	1	2	3	4	5	6	$\hat{p}$
1	ja	nein	ja	ja	ja		0,800
2	ja	nein	ja	ja		nein	0,600
3	ja	nein	ja		ja	nein	0,600
4	ja	nein		ja	ja	nein	0,600
5	ja		ja	ja	ja	nein	0,800
6		nein	ja	ja	ja	nein	0,600

Ausprägung $\hat{p}$	*Anzahl*	$Pr(\hat{p})$
0,600	4	0,667
0,800	2	0,333
Summe	6	1,000

Die Berechnung von Erwartungswert und Varianz zeigt, dass sich der Erwartungswert nicht verändert hat, die Varianz aber kleiner geworden ist ($Var(\hat{p}) \approx 0,009$).

Erhöht man den Stichprobenumfang auf $n = 6$, so liegt eine Vollerhebung vor. Daher muss der Stichprobenanteil exakt dem Anteil der Grundgesamtheit entsprechen und die Varianz muss Null sein.

Allgemein kann gezeigt werden, dass der Erwartungswert des Stichprobenanteils immer dem Anteil der Grundgesamtheit entspricht und die Varianz mit zunehmendem Stichprobenumfang kleiner wird.

13.2 Parameterschätzung

Ein wesentlicher Teilbereich der schließenden Statistik ist das Schätzen von Parametern. Wir haben in Kapitel 11 und 12 verschiedene Verteilungen betrachtet und festgestellt, dass alle Wahrscheinlichkeitsverteilungen einen oder mehrere Parameter als Bestimmungsgrößen haben. In den Beispielen waren bislang diese Parameter bekannt und wir haben Wahrscheinlichkeiten dafür berechnet, dass in der Stichprobe gewisse interessierende Ereignisse auftreten. Nun haben wir eine Stichprobe vorliegen, und wollen aus dieser Kenntnisse über den unbekannten Parameter erlangen, wir schätzen also den unbekannten Parameter. Diese geschätzten Parameter sollen zusätzlich gewisse Gütekriterien erfüllen.

Die beiden wesentlichsten Gütekriterien haben wir schon in Abschnitt 13.1 kennen gelernt. Wir haben festgestellt, dass der Erwartungswert des Stichprobenanteils dem Anteil der Grundgesamtheit entspricht. Schätzer, die diese Eigenschaft aufweisen, nennt man erwartungstreu. Weiters haben wir festgestellt, dass mit zunehmendem Stichprobenumfang die Varianz des Schätzers immer kleiner wird. Diese Eigenschaft wird als Konsistenz bezeichnet.

Gütekriterien für Schätzer

- **Erwartungstreue**
  Der Erwartungswert des Schätzers entspricht dem gesuchten Parameter.
  $$E(\hat{\theta}) = \sum \hat{\theta} Pr(\hat{\theta}) = \theta$$
- **Konsistenz**
  Mit zunehmendem Stichprobenumfang wird die Varianz des Schätzers kleiner.
  $$\lim_{n \to \infty} Var(\hat{\theta}_n) = 0$$

In Kapitel 13.1 haben wir den Stichprobenanteil als Schätzer für den Anteil der Grundgesamtheit herangezogen. Wir haben weiters festgestellt, dass der Stichprobenanteil ein erwartungstreuer und konsistenter Schätzer des Anteils der Grundgesamtheit ist. Dieser Schätzer nimmt für eine bestimmte Stichprobe einen einzigen Wert an und wird daher auch als Punktschätzer bezeichnet. Ziehen wir eine zweite Zufallsstichprobe und errechnen erneut den Punktschätzer, so kann dieser natürlich vom ersten Schätzer abweichen. Der Punktschätzer kann auch vom wahren Anteil in der Grundgesamtheit abweichen. Der Nachteil von Punktschätzern liegt darin, dass man wenig Information über die Qualität der Schätzung hat.

Wiederholt man diese Stichprobenziehung sehr oft und betrachtet die verschiedenen Ergebnisse, so bildet der Punktschätzer selbst wieder eine Zufallsvariable, die einer gewissen Verteilung, der so genannten Stichprobenverteilung, gehorcht. Manche Stichprobenverteilungen können bei genügend großem Stichprobenumfang gut durch eine Normalverteilung approximiert werden. Die Stichprobenverteilung des Anteils kann beispielsweise durch eine Normalverteilung mit Erwartungswert p und Varianz $p(1-p)/n$ approximiert werden.

Damit kann man Intervalle bestimmen, die mit vorgegebener Wahrscheinlichkeit den Stichprobenanteil enthalten, sofern der Anteil der Grundgesamtheit bekannt ist.

$$Pr\left(p - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} \leq \widehat{p} \leq p + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

Dieses Wissen ist nur bedingt brauchbar, weil ja dieser Anteil der Grundgesamtheit in Wahrheit nicht bekannt ist. Allerdings kann man zeigen, dass diese Behauptung auch in der Form

$$Pr\left(\widehat{p} - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq p \leq \widehat{p} + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) = 1 - \alpha$$

gilt. Somit erhalten wir ein Intervall, welches den gesuchten Parameter mit der Wahrscheinlichkeit $1 - \alpha$ überdeckt. Wir haben auf diese Weise einen Bereichschätzer für den Anteil konstruiert. Diese Bereichschätzer werden in der Statistik als Konfidenzintervalle bezeichnet. Übliche α-Werte für die Konstruktion von Konfidenzintervallen sind etwa 0,05 oder 0,01.

Parameterschätzung

- **Punktschätzer**
  Der Schätzer ist ein einzelner Wert, der aufgrund des Stichprobenergebnisses errechnet wird.
- **Bereichschätzer - Konfidenzintervalle**
  Mit Hilfe der Stichprobenverteilung werden Intervalle konstruiert, die den gesuchten Parameter mit einer Sicherheit von $1 - \alpha$ überdecken. Übliche α-Werte sind $\alpha = 0,05$ oder $\alpha = 0,01$.

Man kann sich diese Sicherheit etwa folgendermaßen vorstellen: Zieht man sehr viele verschiedene Stichproben und bestimmt zu jeder dieser Stichproben das Konfidenzintervall beispielsweise zur Sicherheit $1-\alpha = 0,95$, dann werden 95% der Konfidenzintervalle den Parameter der Grundgesamtheit überdecken. Daneben gibt es aber leider auch 5% der Konfidenzintervalle, die den Parameter nicht überdecken. Das heißt, ein kleines Restrisiko des Irrtums (nämlich 5%) bleibt. Diese Irrtumswahrscheinlichkeit kann nicht völlig ausgelöscht werden, aber über die Wahl von α reduziert werden.
Reduziert man diese Irrtumswahrscheinlichkeit bei gleichbleibendem Stichprobenumfang, so wird das zugehörige Konfidenzintervall etwas breiter werden, also etwas ungenauer sein.

13.3 Schätzen von Anteilen

Das Schätzen von Anteilen haben wir bereits als einführendes Beispiel betrachtet, daher begnügen wir uns mit der Zusammenfassung der Erkenntnisse.

Bezeichnungen

p	**wahrer Anteil** einer Eigenschaft in der Grundgesamtheit
$\widehat{p}$	**Schätzwert** aus der Stichprobe für den unbekannten Anteil p erwartungstreuer, konsistenter Punktschätzer
$u_{1-\frac{\alpha}{2}}$	$(1-\alpha/2)$ - Quantil der Standardnormalverteilung

Konfidenzintervall für Anteile

Ein Intervall der Form $[\underline{p}, \overline{p}]$ mit

$$\overline{p} = \widehat{p} + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

$$\underline{p} = \widehat{p} - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

wird als Konfidenzintervall zur Sicherheit $1-\alpha$ bezeichnet.

Ein Konfidenzintervall zur Sicherheit $1-\alpha$ überdeckt den wahren Parameter p mit einer Wahrscheinlichkeit von $100(1-\alpha)\%$.

Beispiel 13.1. XPÖ-WählerInnen
Das Umfrageergebnis einer Zufallsstichprobe von $n = 2000$ Personen aus der Grundgesamtheit aller Wahlberechtigten liegt vor. Es gaben 910 Befragte an, die XPÖ wählen zu wollen. Berechnen Sie

- das 99%-Konfidenzintervall
- das 95%-Konfidenzintervall
- das 90%-Konfidenzintervall

für den Anteil der XPÖ-WählerInnen in der Gesamtbevölkerung.

Für das 99%-Konfidenzintervall gilt:

$$[\underline{p}, \overline{p}] = \widehat{p} \pm u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 0,455 \pm 2,576 \cdot \sqrt{\frac{0,455(1-0,455)}{2000}} =$$
$$= [0,426; 0,484]$$

Dies bedeutet, dass der Anteil der XPÖ-WählerInnen in der Grundgesamtheit mit einer Sicherheit von 99% von dem Intervall $[0,426; 0,484]$ überdeckt wird.

Begnügt man sich mit weniger Sicherheit, so entstehen folgende Intervalle:

$$[\underline{p}, \overline{p}] = 0,455 \pm 1,960 \cdot \sqrt{\frac{0,455(1-0,455)}{2000}} = [0,433; 0,477]$$

Der Anteil der XPÖ-WählerInnen in der Grundgesamtheit wird mit einer Sicherheit von 95% von dem Intervall $[0,433; 0,477]$ überdeckt.

$$[\underline{p}, \overline{p}] = 0,455 \pm 1,645 \cdot \sqrt{\frac{0,455(1-0,455)}{2000}} = [0,437; 0,473]$$

Mit einer Sicherheit von 90% wird der Anteil der XPÖ-WählerInnen in der Grundgesamtheit von dem Intervall $[0,437; 0,473]$ überdeckt.

Aus den Ergebnissen wird deutlich sichtbar, dass die Intervalle bei geringerer Sicherheit kleiner werden. Kleinere Intervalle bei gleichbleibender Sicherheit erhält man, indem man den Stichprobenumfang erhöht.

13.4 Schätzen von Mittelwerten

Ausgangspunkt unserer Überlegungen ist ein normalverteiltes Merkmal, dessen Varianz bekannt ist und dessen Mittelwert geschätzt werden soll.

Es bietet sich an, den theoretischen Mittelwert (= Erwartungswert) durch den empirischen Mittelwert (= arithmetisches Mittel aus den Stichprobendaten) zu schätzen. Auch dieser Schätzer ist erwartungstreu und konsistent.

Wie beim Punktschätzer von Anteilswerten ist auch hier ein Bereichschätzer dem Punktschätzer auf jeden Fall vorzuziehen, weil dieser mehr Information bietet.

Bezeichnungen

μ	**wahrer Mittelwert** eines normalverteilten Merkmals in der Grundgesamtheit
σ^2	bekannte Varianz des normalverteilten Merkmals in der Grundgesamtheit
$\bar{x}$	arithmetisches Mittel der Stichprobenwerte erwartungstreuer und konsistenter **Punktschätzer** für den unbekannten Mittelwert μ
$u_{1-\frac{\alpha}{2}}$	$(1-\alpha/2)$ - Quantil der Standardnormalverteilung

Konfidenzintervall für Mittelwerte bei bekannter Varianz σ^2

Als Unter- bzw. Obergrenze eines Konfidenzintervalls zur Sicherheit $1-\alpha$ für den Parameter μ bestimmt man

$$\overline{\mu} = \bar{x} + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma^2}{n}}$$

$$\underline{\mu} = \bar{x} - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma^2}{n}}$$

In Praxis mag es noch realistisch sein anzunehmen, dass man über ein interessierendes Merkmal weiß, dass es normalverteilt ist, aber die Kenntnis der Varianz ist eher unwahrscheinlich. Weil wir an dieser Stelle ein Eingehen auf die tatsächlichen mathematischen Hintergründe vermeiden wollen, betrachten wir das Problem intuitiv. Wenn man die (wahre) Varianz aus der Grundgesamtheit nicht kennt, dann wird diese geschätzt. Dies erfolgt mit der sogenannten korrigierten Varianz

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

Vergleicht man diese Formel mit der Varianz, die wir in der deskriptiven Statistik (vgl. Kapitel 7.2) kennen gelernt haben, so fällt der veränderte Nenner auf. Dieser wird verwendet, um auch beim Schätzer für die Varianz die Erwartungstreue zu gewährleisten.

Durch die Tatsache, dass die Varianz nicht bekannt ist, sondern geschätzt werden muss, erhöht sich die Unsicherheit der Aussage über den Parameter μ. Um diese Unsicherheit zu kompensieren, muss das Konfidenzintervall etwas breiter angelegt werden. Dies wird erreicht, indem man nicht die Quantile der Standardnormalverteilung $u_{1-\alpha/2}$ verwendet, sondern die Quantile $t_{n-1;1-\alpha/2}$ der so genannten **Student-Verteilung** (auch t-Verteilung). Diese Quantile hängen nicht nur vom Niveau α ab, sondern auch vom Stichprobenumfang n.

Konfidenzintervall für Mittelwerte bei unbekannter Varianz σ^2

Als Unter- bzw. Obergrenze eines Konfidenzintervalls zur Sicherheit $1-\alpha$ für den Parameter μ bestimmt man

$$[\underline{\mu}, \overline{\mu}] = \bar{x} \pm t_{n-1;1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{s}^2}{n}}$$

$t_{n-1;1-\alpha/2}$ $(1-\alpha/2)-$Quantil der Student-Verteilung mit $n-1$ Freiheitsgraden

Auch die Quantile der Student-Verteilung sind im Anhang tabelliert. Ein Vergleich der Quantile der Student-Verteilung mit jenen der Standardnormalverteilung zeigt:

- Bei gleicher Wahrscheinlichkeit sind die Quantile der Student-Verteilung stets größer als die der Standardnormalverteilung.
- Mit wachsendem Stichprobenumfang nähern sich die Quantile der Student-Verteilung denen der Standardnormalverteilung an.

Nachdem sich die Student-Verteilung der Standardnormalverteilung annähert, kann man ab n = 30 wieder die Standardnormalverteilung zur Berechnung des Konfidenzintervalls verwenden, auch wenn die Varianz aus der Grundgesamtheit nicht bekannt ist und geschätzt werden muss.

Beispiel 13.2. Konfidenzintervalle für Mittelwerte
Eine Zufallsstichprobe vom Umfang n = 50 aus einer Grundgesamtheit ergab hinsichtlich eines normalverteilten Merkmals X nachstehendes Ergebnis:

948,0	924,6	958,4	961,4	934,8	978,8	978,8	955,3	962,5	959,4
953,3	970,6	945,1	998,2	953,3	958,4	976,8	957,3	954,3	940,0
962,5	930,8	969,6	987,0	943,0	972,7	979,8	925,6	928,7	967,6
949,2	969,6	971,7	969,6	980,9	969,6	978,8	933,8	956,3	953,3
925,6	961,4	935,9	982,9	941,0	966,5	960,4	938,9	912,7	955,3

Bestimmen Sie das 95%-Konfidenzintervall für den Mittelwert μ, wenn

a) die Standardabweichung in der Grundgesamtheit $\sigma = 18,7$ ist
b) die Standardabweichung in der Grundgesamtheit unbekannt ist und durch die Standardabweichung der Stichprobe geschätzt werden muss.

Als Punktschätzer aus der Stichprobe erhält man $\bar{x} = 957,0$.
Fall a)

$$\underline{\overline{\mu}} = \bar{x} \pm u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma^2}{n}} = \bar{x} \pm u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 957,0 \pm 1,960 \cdot \frac{18,7}{\sqrt{50}} = [951,8; 962,2]$$

Mit 95%-iger Sicherheit überdeckt das Intervall $[951,8; 962,2]$ den Mittelwert der Grundgesamtheit.

Fall b)

$$\underline{\overline{\mu}} = \bar{x} \pm t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{\hat{s}}{\sqrt{n}} \qquad \text{mit} \qquad \hat{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Verwendet man die Tabelle der Student-Verteilung, so stellt man fest, dass 49 Freiheitsgrade nicht tabelliert sind. Man verwendet daher einen möglichst nahe liegenden Wert, in diesem Fall das Quantil für 50 Freiheitsgrade. Demnach ergibt sich:

$$\underline{\mu} = 957,0 - 2,009 \cdot \frac{18,7}{\sqrt{50}} = 951,7 \qquad \overline{\mu} = 957,0 + 2,009 \cdot \frac{18,7}{\sqrt{50}} = 962,3$$

Mit 95%-iger Sicherheit überdeckt das Intervall $[951,7; 962,3]$ den Mittelwert der Grundgesamtheit.

Stammen die Daten nicht aus einer normalverteilten Grundgesamtheit, so entspricht trotzdem die Stichprobenverteilung des Mittelwertes bei ausreichend großem Stichprobenumfang n zumindest approximativ einer Normalverteilung. Diese Aussage basiert auf dem zentralen Grenzwertsatz. Daher kann bei genügend großem Stichprobenumfang (ab n = 30) auch in diesem Fall wieder mit den genannten Konfidenzintervallen gearbeitet werden.

13.5 Konfidenzintervalle in EXCEL

Konfidenzintervalle werden am besten mit Hilfe der angeführten Formeln berechnet. Die Quantile der Standardnormalverteilung werden über die Funktion NORMINV(Wahrsch;Mittelwert;Standabwn) angefordert, jene der Student-Verteilung über TINV(Wahrsch;Freiheitsgrade). Bei der Verwendung der Funktion TINV ist allerdings Vorsicht geboten, da die Verwendung der Parameter ungewöhnlich ist. Um das Quantil $t_{n-1;1-\alpha/2}$ zu erhalten, muss als Wahrscheinlichkeit der Wert α eingegeben werden.

Quantile in EXCEL

$t_{n;1-\alpha/2}$	=	TINV(α; n)
$u_{1-\alpha/2}$	=	NORMINV($1-\alpha/2$; 0; 1)

13.6 Konfidenzintervalle in SPSS

Unter *Analysieren → Mittelwerte vergleichen → T-Test bei einer Stichprobe* können Konfidenzintervalle erstellt werden. Dazu wird die gewünschte Variable als *Testvariable* und als *Testwert* die Standardeinstellung 0 ausgewählt. In

Optionen kann die gewünschte Sicherheit $1-\alpha$ in Prozent angegeben werden. Die Ergebnisse findet man unter der Spaltenbezeichnung *Konfidenzintervall der Differenz*. Genau genommen wird ein Konfidenzintervall für die Differenz der Testvariablen zum Testwert berechnet, daher haben wir als Testwert 0 verwendet.

Eine zweite Möglichkeit findet man unter dem Menüpunkt *Analysieren* → *Deskriptive Statistiken* → *Explorative Datenanalyse*. Die Variable wird als *Abhängige Variable* ausgewählt, in der Option *Statistik* wird das Konfidenzintervall zur gewünschten Sicherheit ausgewählt. Die Ergebnisse findet man in der Spalte *Statistik*.

SPSS berechnet Konfidenzintervalle immer unter Verwendung der Student-Verteilung. Für die Berechnung von Konfidenzintervallen für Anteile wird der Anteil als Mittelwert interpretiert, wie in Beispiel 7.3 (Seite 97) beschrieben. Bei geeigneter Kodierung kann damit das Konfidenzintervall für Mittelwerte auch für Anteile herangezogen werden. Durch die Verwendung der Student-Verteilung kann es im Vergleich zu den händisch oder mit EXCEL ermittelten Ergebnissen zu kleinen Abweichungen im Ergebnis kommen. Da die von SPSS berechneten Intervalle immer größer sind (Verwendung der Student-Verteilung führt zu breiteren Intervallen), haben diese Abweichungen aber keinerlei negative Konsequenzen für die Überdeckungswahrscheinlichkeit des Intervalls.

Konfidenzintervalle in SPSS

- *Analysieren* → *Mittelwerte vergleichen* → *T-Test bei einer Stichprobe*
- *Analysieren* → *Deskriptive Statistiken* → *Explorative Datenanalyse*
- SPSS verwendet Student-Verteilung zur Berechnung
- Konfidenzintervalle für Anteile: Daten werden geeignet kodiert, der Anteil wird dann als Mittelwert behandelt

Übungsaufgaben

13.1. AkademikerInnenquote
Nach den Ergebnissen des Mikrozensus (n = 35.200) aus dem Jahr 2003 haben 5,4% der österreichischen Bevölkerung einen Universitätsabschluss. Berechnen Sie das 95%-Konfidenzintervall für den tatsächlichen AkademikerInnenanteil in der Gesamtbevölkerung.

13.2. Nationalratswahl
Vor der Nationalratswahl 2002 veröffentlichte ein Meinungsforschungsinstitut das Umfrageergebnis einer Zufallsstichprobe von n = 500 Personen aus der Grundgesamtheit aller Wahlberechtigten. Es gaben 36% der Befragten an, die

SPÖ wählen zu wollen. Berechnen Sie und interpretieren Sie

a) das 99%-Konfidenzintervall
b) das 95%-Konfidenzintervall

für den Anteil der SPÖ-WählerInnen zu diesem Zeitpunkt in der Gesamtbevölkerung.

13.3. Mineralwasser
Bei der Abfüllung eines Mineralwassers in Literflaschen ist der Magnesiumgehalt je Liter normalverteilt. 16 Kontrollmessungen ergaben folgende Werte für den Gehalt an Magnesium (in mg/l):

22,6	24,1	22,0	25,4	24,1	26,8	27,0	25,2
24,2	26,9	23,5	27,4	26,0	24,9	28,6	24,6

Berechnen Sie ein 95%-Konfidenzintervall für den mittleren Magnesiumgehalt in der Gesamtproduktion.

13.4. Bier
Beim Ausschank von Bier ist die Füllmenge je Glas normalverteilt. 33 Kontrollmessungen ergaben folgende Werte für die Füllmenge (in ml):

500	503	498	497	502	505	495	499	505	495	499
510	509	499	506	495	490	505	507	506	495	490
495	499	505	507	495	499	505	495	499	503	496

Berechnen Sie ein 95%-Konfidenzintervall für die mittlere Füllmenge pro Glas.

14

Statistisches Testen

Ein statistischer Test ist eine Regel zur Entscheidung bei Unsicherheit. Diese Unsicherheit liegt vor, weil man keine Kenntnisse über die Grundgesamtheit hat, sondern nur über eine Stichprobe. Die Entscheidung ist zwischen zwei Behauptungen zu treffen, die als Hypothesen bezeichnet werden.

Wir wollen uns zuerst mit einigen Begriffen der Testtheorie auseinandersetzen und uns dann einige spezielle Tests genauer ansehen.

14.1 Grundbegriffe der Testtheorie

Beim statistischen Testen bezeichnet man mit H_0 die **Nullhypothese** und mit H_1 die **Alternativhypothese**. Beide Hypothesen beinhalten eine Behauptung über die Grundgesamtheit, wobei die beiden Hypothesen einander ausschließen und ergänzen. Diese Hypothesen können sich beispielsweise auf den Parameter θ einer Verteilung eines Merkmales aus der Grundgesamtheit beziehen.

Wir wollen das Prinzip des statistischen Testens anhand eines Beispiels untersuchen, das eigentlich nichts mit Statistik zu tun hat, nämlich an einer Gerichtsverhandlung.

Statistisches Testen

Statistischer Test	Entscheidungsregel zwischen zwei Hypothesen
Hypothesen	Behauptungen über die Grundgesamtheit H_0 Nullhypothese, H_1 Alternativhypothese schließen einander aus und ergänzen sich

Beispiel 14.1. Gerichtsverhandlung
In Österreich, wie in vielen anderen Rechtsstaaten auch, gilt die sogenannte Unschuldsvermutung. Ein Angeklagter ist somit so lange als unschuldig anzusehen, bis seine Schuld nachgewiesen werden kann. Durch diese Unschuldsvermutung soll auf jeden Fall ausgeschlossen werden, dass ein Unschuldiger verurteilt wird (Justizirrtum). Der Richter muss sich somit zwischen zwei Behauptungen entscheiden, die einander ausschließen und ergänzen. Eine der beiden Hypothesen, sagen wir die Nullhypothese, lautet „Der Angeklagte ist unschuldig", die zugehörige Alternativhypothese lautet demnach „Der Angeklagte ist schuldig". Die Zuordnung der Unschuldsvermutung zur Nullhypothese ist nicht zufällig, sondern gezielt gewählt. Die Begründung für diese Zuordnung erfolgt an späterer Stelle.

Die Entscheidung für eine der beiden Hypothesen ist aufgrund eines Stichprobenergebnisses zu treffen. Damit wird die Entscheidung unter Unsicherheit getroffen und kann daher richtig oder falsch sein. Als Ergebnis eines statistischen Tests formuliert man daher „Entscheidung für die Nullhypothese" oder „Entscheidung für die Alternativhypothese". Unpräzise und daher auf jeden Fall abzulehnen ist eine Formulierung der Art „Die Nullhypothese trifft zu", weil man diese Aussage nur dann treffen könnte, wenn die Grundgesamtheit vollständig bekannt wäre.

Fällt die Entscheidung zugunsten der Alternativhypothese H_1, obwohl in der Grundgesamtheit H_0 richtig ist, dann begeht man einen Fehler 1. Art oder α-Fehler. Ein Fehler 2. Art oder β-Fehler ensteht bei der Entscheidung für H_0, obwohl in der Grundgesamtheit H_1 richtig ist.

Tabelle 14.1. Fehler beim statistischen Testen

		Entscheidung auf	
		H_0	H_1
wahr ist	H_0	kein Fehler	α-Fehler
	H_1	β-Fehler	kein Fehler

Natürlich sollten diese Fehler so gering wie möglich sein. Allerdings sind die Fehler nicht unabhängig voneinander, ein kleinerer α-Fehler führt zu einem größeren β-Fehler und umgekehrt. Der β-Fehler ist aber **nicht** als Gegenwahrscheinlichkeit zum α-Fehler anzusetzen, es gilt also **nicht** $1 = \alpha + \beta$.

Das Ausmaß des α-Fehlers nennt man das **Signifikanzniveau** des Tests (üblich sind $\alpha = 0,05$ oder $\alpha = 0,01$). Dieses Signifikanzniveau wird vor Durchführung des Tests festgelegt. Signifikanztests sind so konstruiert, dass

der Fehler 1. Art maximal $100\alpha\%$ beträgt. Damit hat man den α-Fehler unter Kontrolle, den β-Fehler üblicherweise aber nicht. Hauptaugenmerk liegt demnach auf dem α-Fehler, daher werden die Hypothesen so formuliert, dass der schwerwiegendere Fehler als α-Fehler konzipiert ist.

Fehler beim statistischen Testen

α-Fehler	Verwerfen von H_0, obwohl H_0 richtig ist Signifikanzniveau des Tests üblich sind $\alpha = 0,05$ oder $\alpha = 0,01$
β-Fehler	Beibehalten von H_0, obwohl H_1 richtig ist

Beispiel 14.2. Gerichtsverhandlung
(Fortsetzung von Beispiel 14.1) Durch die Unschuldsvermutung soll auf jeden Fall ein Justizirrtum vermieden werden. Würde man die Gerichtsverhandlung als statistisches Testproblem formulieren, so wäre der Justizirrtum als α-Fehler anzusetzen. Ein α-Fehler entsteht beim irrtümlichen Ablehnen der Nullhypothese, daher lautet in diesem Fall die Nullhypothese „Der Angeklagte ist unschuldig" und die zugehörige Alternativhypothese „Der Angeklagte ist schuldig".

Nun sind die Hypothesen formuliert und wir sind informiert über mögliche Fehlentscheidungen. Der nächste Schritt ist die Entscheidung selbst. Ausgangspunkt ist eine möglichst unvoreingenommene Haltung in Form der Nullhypothese, alle anderen Möglichkeiten sind als Alternativhypothese formuliert. In der Folge wird versucht, in der Stichprobe Indizien dafür zu finden, dass dieser Ausgangspunkt falsch ist und daher verworfen werden muss. Findet man in der Stichprobe genug Indizien, um die Nullhypothese zu verwerfen, dann entscheidet man sich für die Alternativhypothese, ansonsten muss die Nullhypothese beibehalten werden.

Arbeitsweise eines statistischen Tests

Ausgangspunkt ist immer die Nullhypothese. In der Stichprobe wird nach ausreichenden Indizien gesucht, die eine Ablehnung der Nullhypothese ermöglichen.

- Gelingt dies, so kann die Nullhypothese mit Sicherheit $1-\alpha$ verworfen werden. Man erhält ein signifikantes Ergebnis zum Niveau $1 - \alpha$.
- Gelingt dies nicht, so muss (aus Mangel an Beweisen) die Nullhypothese beibehalten werden. Wir erhalten kein signifikantes Ergebnis.

Beispiel 14.3. Gerichtsverhandlung
(Fortsetzung von Beispiel 14.1) Ausgangspunkt ist die Unschuldsvermutung, in der Gerichtsverhandlung werden Hinweise zur Schuld des Angeklagten vorgelegt. Sind die Indizien ausreichend, wird der Angeklagte verurteilt (Entscheidung für Alternativhypothese), ansonsten erfolgt ein Freispruch (Entscheidung für Nullhypothese). Die Beweislast liegt beim Kläger, daher wird der Angeklagte auch im Fall mangelnder Beweise freigesprochen.

Auch beim statistischen Testen entscheidet man sich im Zweifel immer für die Nullhypothese. Die beiden Hypothesen sind daher in ihrer Konsequenz nicht gleichwertig. Lassen sich in der Stichprobe genug Indizien zur Verwerfung der Nullhypothese finden, dann konnte die Alternativhypothese mit Sicherheit $1-\alpha$ nachgewiesen werden. Entscheidungen für die Alternativhypothese werden als signifikante Ergebnisse bezeichnet. Sind nicht genug Indizien in der Stichprobe zu finden, müssen wir uns für die Beibehaltung der Nullhypothese entscheiden. Wir haben diese aber nicht nachgewiesen, sondern wir behalten diese nur wegen mangelnder Beweise bei. Daraus ergeben sich folgende Überlegungen zur **Formulierung von Hypothesen**:

- Als Nullhypothese wird eine möglichst unvoreingenommene Behauptung verwendet.
- Die Hypothesen sind so zu formulieren, dass der α-Fehler den schwerwiegenderen Fehler beinhaltet.
- Die nachzuweisende Behauptung wird als Alternativhypothese formuliert.

Damit lässt sich der allgemeine Ablauf eines statistischen Tests darstellen:

Ablauf eines statistischen Tests

1. Hypothesen formulieren.
2. Signifikanzniveau festlegen ($\alpha = 0{,}01$ oder $0{,}05$).
3. Nach den vorliegenden Regeln aufgrund eines Stichprobenergebnisses eine Entscheidung für eine der beiden Hypothesen treffen.
4. Entscheidung interpretieren.

In der Statistik werden die Testverfahren nach verschiedenen Kriterien in Bereiche zusammengefasst. Eines dieser Kriterien unterscheidet parametrische und nichtparametrische Tests. **Parametrische Tests** benötigen als Voraussetzung Annahmen über den Verteilungstyp in der Grundgesamtheit, **nichtparametrische Tests** hingegen kommen ohne Verteilungsannahmen aus. Ein Beispiel für einen parametrischen Test ist der Test, ob ein Mittelwert eines normalverteilten Merkmals von einem vorgegebenen Sollwert abweicht (T-Test, vgl. Abschnitt 14.3.1). Der Chi-Quadrat-Test auf Unabhängigkeit über-

prüft, ob es zwischen zwei nominalen Merkmalen einen Zusammenhang gibt, und zählt zu den nichtparametrischen Tests (vgl. Kapitel 14.5), weil keine bestimmte Verteilung vorliegen muss.

Eine weitere wichtige Möglichkeit zur Unterscheidung ist aus der konkreten Formulierung der Hypothesen zu entnehmen:

Einseitige und zweiseitige Tests

Die Hypothesenformulierung

$$H_0\text{: „} = \text{”} \qquad H_1\text{: „} \neq \text{”}$$

wird als **zweiseitiges Testproblem** bezeichnet.

Falls die Hypothesen

$$H_0\text{: „} \leq \text{”} \qquad H_1\text{: „} > \text{”}$$

oder

$$H_0\text{: „} \geq \text{”} \qquad H_1\text{: „} < \text{”}$$

lauten, so bezeichnet man dies als **einseitiges Testproblem**.

14.2 Testen von Hypothesen über Anteile

Wir gehen davon aus, dass unsere Stichprobe groß genug ist ($n \geq 30$), um den zentralen Grenzwertsatz anwenden zu dürfen. Ist die vorliegende Stichprobe nicht groß genug, so muss der Hypothesentest über Anteile anders als hier durchgeführt werden.

14.2.1 Testen von zweiseitigen Hypothesen

Ausgangspunkt ist die Behauptung, dass ein Anteil einen ganz bestimmten Wert p_0 annimmt. Als Alternative wird formuliert, dass der Anteil den Wert p_0 nicht annimmt, dabei ist es völlig unerheblich, ob dieser Wert unter- oder überschritten wird.

Hypothesenformulierung

$$H_0\text{: } p = p_0 \qquad H_1\text{: } p \neq p_0$$

Beispiel 14.4. XPÖ-WählerInnen
Man ist interessiert an dem Anteil der XPÖ-WählerInnen in der österreichischen Bevölkerung. Dazu hat man die Nullhypothese, dass sich der Anteil seit der letzten Wahl nicht verändert hat, damals lag der Anteil bei 45%. Die Alternativhypothese lautet dem entsprechend, dass sich der Anteil geändert hat und somit von 45% verschieden ist. Demnach lauten die Hypothesen:

$$H_0\colon p = 0,45 \qquad H_1\colon p \neq 0,45$$

Signifikanzniveau festlegen
Wir wollen einen α-Fehler von 0,05 zugestehen. Fällt die Entscheidung später auf die Alternativhypothese, so könnte dies mit 5%-iger Wahrscheinlichkeit eine falsche Entscheidung sein. Über den β-Fehler treffen wir - wie üblich - keine Aussage.

Entscheidung
Die Entscheidung für dieses Testproblem kann folgendermaßen getroffen werden: Man bestimmt ein Konfidenzintervall zur Sicherheit $1 - \alpha$ (vgl. Kapitel 13.3).

$$[\underline{p}, \overline{p}] = \hat{p} \pm u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Von diesem Konfidenzintervall wissen wir bereits, dass es den wahren Anteil in der Grundgesamtheit mit der Wahrscheinlichkeit $1 - \alpha$ überdeckt. Wäre nun p_0 der tatsächliche Anteil in der Grundgesamtheit, so müsste dieses Intervall auch den Hypothesenwert p_0 überdecken.

Daraus lässt sich folgende **Entscheidungsregel** ableiten:
Gilt $p_0 \in [\underline{p}, \overline{p}]$, dann entscheidet man sich für die Nullhypothese, sonst für die Alternativhypothese.

Beispiel 14.5. XPÖ-WählerInnen
(Fortsetzung von Beispiel 14.4). Von n = 1000 zufällig ausgewählten Personen gaben 430 an, XPÖ-WählerInnen zu sein. Damit ergibt sich als Stichprobenanteil $\hat{p} = 430/1000 = 0,43$ und als Konfidenzintervall mit $\alpha = 0,05$

$$[\underline{p}, \overline{p}] = 0,43 \pm 1,960 \cdot \sqrt{\frac{0,43(1-0,43)}{1000}} = [0,399; 0,461]$$

Der Hypothesenwert $p_0 = 0,45$ wird vom Konfidenzintervall $[0,399; 0,461]$ überdeckt, daher entscheidet man sich für die Nullhypothese.

Interpretation
Man hat sich für die Nullhypothese entschieden, es konnten nicht genug Indizien für die Alternativhypothese gefunden werden. Demnach würde man die Entscheidung folgendermaßen formulieren: Die Nullhypothese muss beibehalten werden, es konnte nicht nachgewiesen werden, dass sich der Anteil der XPÖ-WählerInnen verändert hat.

Beispiel 14.6. XPÖ-WählerInnen
(Fortsetzung von Beispiel 14.4). Wie würde die Entscheidung lauten, wenn der Stichprobenanteil $\hat{p} = 0,4$ betragen hätte? Das zugehörige Konfidenzintervall mit $\alpha = 0,05$ lautet

$$[\underline{p}, \overline{p}] = 0,4 \pm 1,960 \cdot \sqrt{\frac{0,4(1-0,4)}{1000}} = [0,370; 0,430]$$

Der Hypothesenwert $p_0 = 0,45$ wird von dem Konfidenzintervall $[0,370; 0,430]$ nicht überdeckt, daher entscheidet man sich für die Alternativhypothese.

Interpretation
Man hat sich für die Alternativhypothese entschieden, es war also möglich, genug Indizien für die Alternativhypothese zu finden. Die Entscheidung kann folgendermaßen interpretiert werden: Die Alternativhypothese wird akzeptiert, es konnte mit einer Sicherheit von 95% nachgewiesen werden, dass sich der Anteil der XPÖ-WählerInnen in der Grundgesamtheit verändert hat.

Anmerkung
Es braucht etwas Zeit und auch viele Überlegungen, um die ganze Problematik hinter dem Testproblem zu erfassen. Ein wesentlicher Aspekt ist die Tatsache, dass eine Nullhypothese nicht bewiesen werden kann. Die Nullhypothese ist der Ausgangspunkt, und man versucht Indizien dafür zu finden, dass dieser Ausgangspunkt falsch ist. Findet man diese Indizien, so kann mit Sicherheit $1-\alpha$ behauptet werden, dass der Ausgangspunkt falsch war. Findet man diese nicht, so beweist dies jedoch nicht, dass die Nullhypothese richtig ist.

Zum Verständnis kann auch folgende Überlegung hilfreich sein: Die Interpretationen der Konfidenzintervalle lassen sich mit dem Testproblem verknüpfen. Als Entscheidungsgrundlage haben wir ein zweiseitiges Konfidenzintervall für den Parameter p der Grundgesamtheit bestimmt. Damit können wir behaupten, dass das Intervall $[\underline{p}, \overline{p}]$ den Parameter p der Grundgesamtheit mit einer Wahrscheinlichkeit von $1 - \alpha$ überdeckt. Liegt der Hypothesenwert p_0 außerhalb dieses Intervalls, so können wir mit Sicherheit $1-\alpha$ behaupten, dass dieser Wert nicht der Parameter der Grundgesamtheit sein kann. Liegt der Hypothesenwert im Konfidenzintervall, so kann es sein, dass der Wert p_0 tatsächlich der Parameter der Grundgesamtheit ist. Aber das Konfidenzintervall besagt nur, dass der Paramter überdeckt wird, aber nicht, wo genau der Parameter

liegt. Daher könnte genauso jeder andere Wert aus dem Konfidenzintervall der Parameter der Grundgesamtheit sein. Also können wir nicht nachweisen, dass die Nullhypothese richtig ist, auch wenn wir uns für die Nullhypothese entscheiden.

Entscheidungen für die Alternativhypothese werden als signifikant bezeichnet („Der WählerInnenanteil ist von 45% signifikant verschieden."). Als hochsignifikantes Ergebnis wird eine Entscheidung zugunsten der Alternativhypothese auf einem Signifikanzniveau von $\alpha = 0,01$ bezeichnet. Um Unklarheiten auszuschließen, ist es besser, von einer signifikanten Entscheidung auf einem Sicherheitsniveau von 99% zu sprechen.

Testen von zweiseitigen Hypothesen über Anteile

Hypothesen

$$H_0\colon p = p_0 \qquad H_1\colon p \neq p_0$$

Entscheidungsregel

$$[\underline{p}, \overline{p}] = \hat{p} \pm u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Gilt $p_0 \in [\underline{p}, \overline{p}]$, dann wird die Nullhypothese beibehalten, andernfalls verworfen.

14.2.2 Testen von einseitigen Hypothesen

Einseitige Hypothesen behandeln die Fragstellung, ob sich nachweisen lässt, dass ein Parameter einen bestimmten Sollwert unter- oder überschreitet. Wir betrachten zuerst die Frage, ob ein Parameter einen bestimmten Sollwert überschreitet. Gerade in dem Beispiel der XPÖ-WählerInnen wird es für die Partei von Interesse sein, ob sie ihren WählerInnenanteil vergrößern konnte. Die Faustregeln zur Hypothesenbildung beachtend muss die Alternativhypothese lauten: „Der Anteil der XPÖ-WählerInnen ist größer als bei der letzten Wahl (45%)."

Hypothesen formulieren

$$H_0\colon p \leq p_0 \qquad H_1\colon p > p_0$$

Signifikanzniveau festlegen
Auch hier wählen wir wieder $\alpha = 5\%$.

Entscheidungsregel
Beim Testen einseitiger Hypothesen ist eine einseitige Vertrauensschranke für p zur Sicherheit $1-\alpha$ zu bestimmen, sozusagen ein einseitiges Konfidenzintervall. In dem vorliegenden Fall wird die untere Vertrauensschranke zur Sicherheit $1-\alpha$ bestimmt:

$$\underline{p} = \hat{p} - u_{1-\alpha} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Gilt $p_0 < \underline{p}$, dann entscheidet man sich für die Alternativhypothese, sonst behält man die Nullhypothese bei.
Äquivalent dazu ist folgende Überlegung: Die untere Vertrauensschranke wurde berechnet, die Obergrenze des einseitigen Vertrauensintervalls kann durch Überlegen (ein Anteil kann niemals größer als 1 sein) ergänzt werden. Das Konfidenzintervall ist daher von der Form $[\underline{p}; 1]$.

Überdeckt dieses Intervall den Hypothesenwert, so entscheidet man sich für die Nullhypothese, ansonsten für die Alternativhypothese.

Beispiel 14.7. XPÖ-WählerInnen, einseitiger Test
Eine politische Partei will feststellen, ob ihr Stimmanteil über 45% liegt. In einer Umfrage unter $n = 1000$ Wahlberechtigten kam sie auf einen Anteil von 49%.

$$H_0\colon p \leq 0,45 \qquad H_1\colon p > 0,45$$

$$\underline{p} = 0,49 - 1,645 \cdot \sqrt{\frac{0,49(1-0,49)}{1000}} = 0,464$$

Es gilt $0,45 < 0,464$, daher Entscheidung für die Alternativhypothese, die Überlegung $0,45 \notin [0,464; 1]$ führt zu derselben Entscheidung. Mit einer Sicherheit von 95% konnte nachgewiesen werden, dass der WählerInnenanteil über 45% liegt. Der WählerInnenanteil ist signifikant größer als 45%.

Testen von einseitigen Hypothesen über Anteile
Nachweis einer Überschreitung

Hypothesen

$$H_0\colon p \leq p_0 \qquad H_1\colon p > p_0$$

Entscheidungsregel

$$[\underline{p}, \overline{p}] = \left[\hat{p} - u_{1-\alpha} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; 1\right]$$

Gilt $p_0 \in [\underline{p}, \overline{p}]$, dann entscheidet man sich für die Nullhypothese, sonst für die Alternativhypothese.

Will man nachweisen, dass der Anteil einen bestimmten Sollwert p_0 unterschreitet, sind die Hypothesen folgendermaßen zu formulieren:

$$H_0\colon p \geq p_0 \qquad H_1\colon p < p_0$$

Als Entscheidungsgrundlage bestimmen wir eine obere Vertrauensschranke zur Sicherheit $1 - \alpha$:

$$\overline{p} = \hat{p} + u_{1-\alpha} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Gilt $\overline{p} < p_0$, dann wird zugunsten der Alternativhypothese entschieden, sonst behält man die Nullhypothese bei.
Äquivalent dazu ist folgende Überlegung: Die obere Vertrauensschranke wurde berechnet, die Untergrenze des einseitigen Vertrauensintervalls kann ergänzt werden (ein Anteil kann niemals kleiner als 0 sein). Das Vertrauensintervall kann daher angeschrieben werden als $[0; \overline{p}]$.
Überdeckt dieses Intervall den Hypothesenwert, so entscheidet man sich für die Nullhypothese, ansonsten für die Alternativhypothese.

Testen von einseitigen Hypothesen über Anteile
Nachweis einer Unterschreitung

Hypothesen

$$H_0\colon p \geq p_0 \qquad H_1\colon p < p_0$$

Entscheidungsregel

$$[\underline{p}, \overline{p}] = \left[0; \hat{p} + u_{1-\alpha} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Gilt $p_0 \in [\underline{p}, \overline{p}]$, dann entscheidet man sich für die Nullhypothese, sonst für die Alternativhypothese.

14.3 Testen von Hypothesen über einen Mittelwert

Ausgangspunkt für diesen Test, der auch als t-Test bezeichnet wird, ist ein Merkmal, das in der Grundgesamtheit normalverteilt ist. Getestet werden Aussagen über den Mittelwert dieses Merkmals. Die Vorgehensweise ist analog zu jener, die beim Testen von Hypothesen über Anteile verwendet wurde. Lediglich die Konfidenzintervalle, die als Entscheidungsgrundlage dienen, werden anders berechnet. In Übereinstimmung zu den Überlegungen, die wir bei

den Konfidenzintervallen (vgl. Kapitel 13.4) für Mittelwerte angestellt haben, müssen auch beim Testen die beiden Fälle bekannte bzw. unbekannte Varianz der Grundgesamtheit unterschieden werden.

14.3.1 Testen von zweiseitigen Hypothesen

Analog zum zweiseitigen Testproblem der Anteile lauten die Hypothesen:

$$H_0\colon \mu = \mu_0 \qquad H_1\colon \mu \neq \mu_0$$

Als Entscheidungsgrundlage wird ein Konfidenzintervall zur Sicherheit $1-\alpha$ bestimmt. Ist die Varianz der Grundgesamtheit bekannt, werden zur Berechnung die Quantile der Standardnormalverteilung herangezogen, bei geschätzter Varianz werden die Quantile der Student-Verteilung verwendet.

Testen von zweiseitigen Hypothesen über Mittelwerte

Hypothesen

$$H_0\colon \mu = \mu_0 \qquad H_1\colon \mu \neq \mu_0$$

Entscheidungsregel

Fall a) Varianz der Grundgesamtheit bekannt

$$[\underline{\mu}; \overline{\mu}] = \bar{x} \pm u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma^2}{n}}$$

Fall b) Varianz der Grundgesamtheit unbekannt, Schätzer $\hat{s}^2$

$$[\underline{\mu}; \overline{\mu}] = \bar{x} \pm t_{n-1;1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{s}^2}{n}} \qquad \text{mit} \qquad \hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Gilt $\mu_0 \in [\underline{\mu}, \overline{\mu}]$, dann entscheidet man sich für die Nullhypothese, sonst für die Alternativhypothese.

Beispiel 14.8. Zweiseitiger T-Test
(vgl. Beispiel 13.2, Seite 223) Eine Zufallsstichprobe vom Umfang n = 50 aus einer Grundgesamtheit ergab hinsichtlich eines normalverteilten Merkmals X nachstehendes Ergebnis.

Mittels statistischem Test ist die Frage zu beantworten, ob der Mittelwert der Grundgesamtheit von einem Sollwert, der mit 950 angegeben ist, abweicht.

948,0	924,6	958,4	961,4	934,8	978,8	978,8	955,3	962,5	959,4
953,3	970,6	945,1	998,2	953,3	958,4	976,8	957,3	954,3	940,0
962,5	930,8	969,6	987,0	943,0	972,7	979,8	925,6	928,7	967,6
949,2	969,6	971,7	969,6	980,9	969,6	978,8	933,8	956,3	953,3
925,6	961,4	935,9	982,9	941,0	966,5	960,4	938,9	912,7	955,3

Hypothesenformulierung:

$$H_0\colon \mu = 950 \qquad H_1\colon \mu \neq 950$$

Das Sicherheitsniveau legen wir mit $\alpha = 0,05$ fest. Da keine Angabe über die Varianz des Merkmals gemacht wurde, ist die unbekannte Varianz aus der Stichprobe zu schätzen und zur Berechnung des Konfidenzintervalls sind die Quantile der Student-Verteilung heranzuziehen. Bei Verwendung der Tabelle für die Student-Verteilung stellt man fest, dass 49 Freiheitsgrade nicht tabelliert sind. Daher wird ein möglichst nahe liegender Wert verwendet, in diesem Fall das Quantil für 50 Freiheitsgrade. Es ergibt sich mit $\bar{x} = 957,0$ und $\hat{s} = 18,7$

$$[\underline{\mu}; \overline{\mu}] = \bar{x} \pm t_{n-1;1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{s}^2}{n}} = 957,0 \pm 2,009 \cdot \frac{18,7}{\sqrt{50}} = [951,7; 962,3]$$

Das Intervall überdeckt den Hypothesenwert nicht, daher ist die Nullhypothese zu verwerfen. Mit 95%-iger Sicherheit konnte nachgewiesen werden, dass der Mittelwert in der Grundgesamtheit von 950 abweicht. Für die Konstruktion des Konfidenzintervalls hätte man in diesem Fall wegen $n \geq 30$ auch die Quantile der Standardnormalverteilung verwenden dürfen (vgl. Kapitel 13.4).

14.3.2 Testen von einseitigen Hypothesen

Wir versuchen nun zu beweisen, dass der Mittelwert der Grundgesamtheit einen bestimmten Vergleichswert überschreitet. Demnach sind die Hypothesen folgendermaßen zu formulieren:

$$H_0\colon \mu \leq \mu_0 \qquad H_1\colon \mu > \mu_0$$

Als Entscheidungsgrundlage dient die untere Vertrauensschranke zur Sicherheit $1 - \alpha$, die bei bekannter Varianz σ^2 mit

$$\underline{\mu} = \bar{x} - u_{1-\alpha} \cdot \sqrt{\frac{\sigma^2}{n}}$$

und bei unbekannter Varianz σ^2 und Schätzer $\hat{s}^2$ mit

$$\underline{\mu} = \bar{x} - t_{n-1;1-\alpha} \cdot \sqrt{\frac{s^2}{n}}$$

berechnet wird. Gilt $\mu_0 < \underline{\mu}$, dann wird die Nullhypothese verworfen, ansonsten beibehalten.
Äquivalent dazu ist folgende Überlegung: Die untere Vertrauensschranke wurde berechnet, die Obergrenze des einseitigen Vertrauensintervalls wird ergänzt.Damit weist das Vertrauensintervall die Form $[\underline{\mu}; \infty]$ auf.
Überdeckt dieses Intervall den Hypothesenwert, so entscheidet man sich für die Nullhypothese, ansonsten für die Alternativhypothese.

Testen von einseitigen Hypothesen über Mittelwerte
Nachweis einer Überschreitung

Hypothesen

$$H_0\colon \mu \leq \mu_0 \qquad H_1\colon \mu > \mu_0$$

Entscheidungsregel
Fall a) Varianz der Grundgesamtheit bekannt

$$[\underline{\mu}; \overline{\mu}] = \left[\bar{x} - u_{1-\alpha} \cdot \sqrt{\frac{\sigma^2}{n}}; \infty\right]$$

Fall b) Varianz der Grundgesamtheit unbekannt, Schätzer $\hat{s}^2$

$$[\underline{\mu}; \overline{\mu}] = \left[\bar{x} - t_{n-1;1-\alpha} \cdot \sqrt{\frac{\hat{s}^2}{n}}; \infty\right] \qquad \text{mit} \qquad \hat{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Gilt $\mu_0 \in [\underline{\mu}, \overline{\mu}]$, dann wird die Nullhypothese beibehalten, ansonsten wird sie verworfen.

Beispiel 14.9. Einseitiger T-Test
(vgl. Beispiel 14.8) Es ist auf einem Sicherheitsniveau von 95% zu überprüfen, ob der Mittelwert den Sollwert 950 überschreitet.

$$H_0\colon \mu \leq 950 \qquad H_1\colon \mu > 950$$

$$\underline{\mu} = \bar{x} - t_{n-1;1-\alpha} \cdot \sqrt{\frac{\hat{s}^2}{n}} \quad = 957,0 - 1,676 \cdot \frac{18,7}{\sqrt{50}} = 952,6$$

Das Intervall $[952,6; \infty]$ überdeckt den Hypothesenwert nicht, daher ist die

Nullhypothese zu verwerfen. Mit 95%-iger Sicherheit konnte nachgewiesen werden, dass der Mittelwert in der Grundgesamtheit den Sollwert 950 überschreitet.

Im zweiten Fall einer einseitigen Testformulierung soll überprüft werden, ob ein vorgegebener Sollwert unterschritten wird, demnach lauten die Hypothesen

$$H_0\colon \mu \geq \mu_0 \qquad H_1\colon \mu < \mu_0$$

Als Entscheidungsgrundlage wird eine obere Vertrauensschranke zur Sicherheit $1-\alpha$ bestimmt. Bei bekannter Varianz werden zur Berechnung der Vertrauensschranke die Quantile der Standardnormalverteilung verwendet, bei unbekannter Varianz die Quantile der Student-Verteilung. Gilt $\mu_0 > \overline{\mu}$, dann entscheidet man zugunsten der Alternativhypothese, sonst behält man die Nullhypothese bei.
Äquivalent dazu ist folgende Überlegung: Die obere Vertrauensschranke wurde berechnet, die Untergrenze des einseitigen Vertrauensintervalls kann ergänzt werden. Das Vertrauensintervall hat daher die Form $[-\infty; \overline{\mu}]$.
Überdeckt dieses Intervall den Hypothesenwert, so entscheidet man sich für die Nullhypothese, ansonsten für die Alternativhypothese.

Testen von einseitigen Hypothesen über Mittelwerte
Nachweis einer Unterschreitung

Hypothesen

$$H_0\colon \mu \geq \mu_0 \qquad H_1\colon \mu < \mu_0$$

Entscheidungsregel
Fall a) Varianz der Grundgesamtheit bekannt

$$[\underline{\mu}; \overline{\mu}] = \left[-\infty; \bar{x} + u_{1-\alpha} \cdot \sqrt{\frac{\sigma^2}{n}}\right]$$

Fall b) Varianz der Grundgesamtheit unbekannt, Schätzer $\hat{s}^2$

$$[\underline{\mu}; \overline{\mu}] = \left[-\infty; \bar{x} + t_{n-1;1-\alpha} \cdot \sqrt{\frac{\hat{s}^2}{n}}\right] \qquad \text{mit} \qquad \hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Gilt $\mu_0 \in [\underline{\mu}, \overline{\mu}]$, dann wird die Nullhypothese beibehalten, ansonsten wird sie verworfen.

14.4 Testen von Hypothesen in EXCEL und SPSS

In diesem Abschnitt wird die Umsetzung der Tests über Mittelwerte und Anteile in EXCEL bzw. SPSS gezeigt. In EXCEL werden die notwendigen Quantile über Funktionen angefordert und die einseitigen oder zweiseitigen Intervalle mit Hilfe der beschriebenen Formeln berechnet. Die Entscheidung wird dann nach den jeweils gültigen Regeln getroffen.

Quantile in EXCEL

$$
\begin{aligned}
t_{n;1-\alpha/2} &= \text{TINV}(\alpha;\, n) \\
t_{n;1-\alpha} &= \text{TINV}(2\alpha;\, n) \\
u_{1-\alpha/2} &= \text{NORMINV}(1-\alpha/2;\, 0;\, 1) \\
u_{1-\alpha} &= \text{NORMINV}(1-\alpha;\, 0;\, 1)
\end{aligned}
$$

In SPSS können ebenfalls die benötigten Konfidenzintervalle berechnet werden (vgl. Kapitel 13.6). SPSS berechnet dabei standardmäßig immer zweiseitige Intervalle. Falls ein einseitiges Intervall benötigt wird, kann dies über die Anpassung der Sicherheit erreicht werden. Um ein einseitiges $1-\alpha = 95\%$-Intervall zu berechnen, wird ein zweiseitiges Intervall zum Niveau $1-2\alpha = 90\%$ angefordert und je nach Bedarf nur die Obergrenze oder nur die Untergrenze des Intervalls verwendet.

In SPSS werden Entscheidungen beim statistischen Testen im Normalfall mit einer anderen Überlegung getroffen. Unter *Analysieren → Mittelwerte vergleichen → T-Test bei einer Stichprobe* kann der Test für Mittelwerte angefordert werden. Dazu wird die gewünschte Variable als *Testvariable* und als *Testwert* der Hypothesenwert ausgewählt. In *Optionen* kann die gewünschte Sicherheit $1-\alpha$ in Prozent angegeben werden. Als Ergebnis wird ein sogenannter p–Wert ausgegeben, der in diesem Fall in der Spalte *Sig. (2-seitig)* zu finden ist. Dieser p–Wert ist quasi der berechnete α-Fehler der speziellen Stichprobe. Ist dieser kleiner oder gleich dem (von uns vorgegebenen) zulässigen α-Fehler, so wird die Nullhypothese verworfen.

Beispiel 14.10. t-Test in SPSS
(vgl. Beispiel 14.8) Mit der Eingabe von 950 als Testwert liefert SPSS als Ergebnis einen p–Wert von 0,011. Wegen $0,011 \leq 0,05 (= \alpha)$ ist die Nullhypothese zu verwerfen.
Dieser p–Wert ist für eine zweiseitige Fragestellung berechnet. Möchte man einseitig Testen so wird der halbierte p–Wert mit dem vorgegebenen α verglichen.

p–Wert

Ein p–Wert entspricht der Wahrscheinlichkeit, unter der Nullhypothese das beobachtete (oder ein noch extremeres) Ergebnis zu erhalten.
Damit ist der p–Wert quasi der errechnete α-Fehler für die Stichprobe.

t-Test in SPSS

- *Analysieren → Mittelwerte vergleichen → T-Test bei einer Stichprobe*
- Hypothesenwert als *Testwert* verwenden
- Entscheidung treffen
  - zweiseitiger Test:
    Gilt p–Wert $\leq \alpha$, so wird die Nullhypothese verworfen.
  - einseitiger Test:
    Gilt p–Wert$/2 \leq \alpha$, so wird die Nullhypothese verworfen.

14.5 Der Chi-Quadrat-Test auf Unabhängigkeit

Gegeben sind zwei nominale Merkmale mit s bzw. r Ausprägungen. Getestet wird die Abhängigkeit der Merkmalsausprägungen. In Kapitel 8.4.1 haben wir zur Messung des Zusammenhanges zwischen zwei nominalen Merkmalen die Assoziationsmaße χ^2 und Cramers V kennen gelernt. Es wurde festgestellt, dass beide Maßzahlen den Wert 0 annehmen, falls die Merkmale unabhängig voneinander sind. Dem entsprechend lassen sich die Hypothesen für unser Testproblem folgendermaßen ansetzen:

H_0 : Ausprägungen der Merkmale unabhängig, $\chi^2 = 0$

H_1 : Ausprägungen der Merkmale abhängig, $\chi^2 > 0$

Teststrategie:
Man berechnet zuerst den χ^2-Wert mit Hilfe folgender Formel aus den Daten

$$\chi^2_{err} = \sum_{j=1}^{s} \sum_{i=1}^{r} \frac{(h^o_{ij} - h^e_{ij})^2}{h^e_{ij}}$$

mit den Bezeichnungen

h_{ij}^o ... beobachtete absolute Häufigkeit der Kombination X = i und Y = j

h_{ij}^e ... bei Unabhängigkeit von X und Y erwartete absolute Häufigkeit dieser Kombination

Dabei gilt $$h_{ij}^e = \frac{h_{i+} \cdot h_{+j}}{N}$$

Statt eines Konfidenzintervalls wird das $(1-\alpha)$-Quantil der χ^2-Verteilung als Vergleichswert verwendet. Ähnlich wie bei der Student-Verteilung benötigen wir auch hier wieder Freiheitsgrade zur näheren Spezifikation des Quantils. Diese werden allerdings in diesem Fall über $m = (r-1) \cdot (s-1)$ berechnet.

H_0 wird mit Irrtumswahrscheinlichkeit α verworfen, wenn

$$\chi^2_{(r-1)(s-1);1-\alpha} < \chi^2_{err}$$

Damit die Anwendung dieses Tests zulässig ist, muss die erwartete Häufigkeit in jeder Kategorie mindestens 1 betragen und bei höchstens 20% der Kategorien darf die erwartete Häufigkeit unter 5 liegen. Sind diese Voraussetzungen nicht erfüllt, so kann man sich damit behelfen, dass man Ausprägungen zusammenfasst. Dies führt zu einer entsprechenden Reduktion von r bzw. s.

χ^2-Test auf Unabhängigkeit

Ausgangspunkt sind zwei nominale Merkmale mit r bzw. s Ausprägungen.

Hypothesen

H_0: $\chi^2 = 0$ Unabhängigkeit $\qquad$ H_1: $\chi^2 > 0$ Abhängigkeit

Entscheidungsregel

Gilt

$$\chi^2_{(r-1)(s-1);1-\alpha} \quad < \quad \chi^2_{err} = \sum_{j=1}^{s} \sum_{i=1}^{r} \frac{(h_{ij}^o - h_{ij}^e)^2}{h_{ij}^e}$$

dann wird die Nullhypothese verworfen.

Voraussetzungen

- Die erwartete Häufigkeit in jeder Kategorie muss mindestens 1 betragen.
- Bei höchstens 20% der Kategorien darf die erwartete Häufigkeit unter 5 liegen.

Beispiel 14.11. Einfluss von Strategietraining
(vgl. Beispiel 8.1, Seite 121) In einer Studie wird bei N = 235 zufällig ausgewählten Führungskräften der Einfluss von Strategietraining auf den Unternehmenserfolg untersucht. Das Ergebnis der Untersuchung kann aus nachstehender Tabelle entnommen werden.

Kann in der Grundgesamtheit ein Zusammenhang zwischen Trainingsteilnahme und Erfolg nachgewiesen werden?

	kein Erfolg	Erfolg	Summe
kein Training	40	75	115
mit Training	30	90	120
Summe	70	165	235

Die Formulierung der Hypothesen ist vorgegeben, wir wählen als Signifikanzniveau $\alpha = 0,05$. Den χ^2-Wert aus den Daten haben wir bereits berechnet (Seite 127) $\chi^2_{err} = 2,69$. Nachdem beide Merkmale je zwei Ausprägungen aufweisen, haben wir einen Freiheitsgrad und damit als Quantil der χ^2-Verteilung $\chi^2_{(r-1)(s-1);1-\alpha} = 3,84$ (Tabelle 3, Seite 251 und 252).
Da der errechnete Wert das Quantil nicht überschreitet, muss die Nullhypothese beibehalten werden. Es konnte kein Zusammenhang zwischen den Merkmalen Training und Erfolg nachgewiesen werden.

Umsetzung in EXCEL
Die Umsetzung in EXCEL ist eher aufwändig, weil der χ^2-Wert der Stichprobe mit Hilfe der Formel berechnet werden muss. Das Quantil der χ^2-Verteilung wird über die Funktion *CHIINV* ermittelt. Analog dazu kann auch mit der Funktion *CHIVERT* der p–Wert der Stichprobe errechnet werden, der mit dem α-Fehler verglichen wird.

χ^2-Test auf Unabhängigkeit in EXCEL

- $\chi^2_{(r-1)(s-1);1-\alpha}$ = CHIINV(α;$(r-1)(s-1)$)
  Bei $\chi^2_{(r-1)(s-1);1-\alpha} < \chi^2_{err}$ wird die Nullhypothese verworfen.

- p–Wert = CHIVERT(χ^2_{err};$(r-1)(s-1)$)
  Bei p–Wert $\leq \alpha$ wird die Nullhypothese verworfen.

Umsetzung in SPSS
Unter *Analysieren* → *Deskriptive Statistiken* → *Kreuztabellen* findet man den χ^2-Test. In der Option *Statistiken* wird *Chi-Quadrat* ausgewählt. In der Option *Zellen* können die erwarteten Häufigkeiten angefordert werden. Nun wird

ein Merkmal als Spalte und das andere als Zeile ausgewählt und nach Bestätigung erhält man den gewünschten Output.
In der Tabelle *Chi-Quadrat-Test* ist der p–Wert in der ersten Zeile der Spalte *Asymptotische Signifikanz* zu finden. Ist diese kleiner oder gleich dem vorher festgelegten α-Fehler, so muss die Nullhypothese verworfen werden.

χ^2-Test auf Unabhängigkeit in SPSS

- *Analysieren* → *Deskriptive Statistiken* → *Kreuztabellen*
- Option *Statistiken*: *Chi-Quadrat* auswählen
- p–Wert = *Asymptotische Signifikanz*
- Bei p–Wert $\leq \alpha$ wird die Nullhypothese verworfen.

Anmerkung
Der Test überprüft, ob ein Zusammenhang für zwei Merkmale in der Grundgesamtheit nachweisbar ist. Das Cramersche V gibt Auskunft über die Stärke des Zusammenhanges. Diese beiden Erkenntnisse sagen aber nicht das gleiche aus. Es kann durchaus vorkommen, dass ein Zusammenhang zwar in der Grundgesamtheit nachweisbar (also signifikant) ist, aber nur sehr schwach ist z.B. mit $V = 0,01$. Damit ist der Zusammenhang zwar nachweisbar, aber nicht besonders stark. Umgekehrt kann es ebenso vorkommen, dass ein starker Zusammenhang (z.B. $V = 0,8$) nicht signifikant ist.

Übungsaufgaben

14.1. Nationalratswahl

Vor der Nationalratswahl 2002 veröffentlichte ein Meinungsforschungsinstitut das Umfrageergebnis einer Zufallsstichprobe von n = 500 Personen aus der Grundgesamtheit aller Wahlberechtigten. Es gaben 36% der Befragten an, die SPÖ wählen zu wollen. Bei der vorangehenden Wahl 1999 konnte die SPÖ 33% der Stimmen erlangen. Übersteigt der Anteil der SPÖ-WählerInnen den Vergleichswert der letzten Wahl?

14.2. Mineralwasser

Bei der Abfüllung eines Mineralwassers in Literflaschen ist der Magnesiumgehalt je Liter normalverteilt. 16 Kontrollmessungen ergaben folgende Werte für den Gehalt an Magnesium (in mg/l):

22,6	24,1	22,0	25,4	24,1	26,8	27,0	25,2
24,2	26,9	23,5	27,4	26,0	24,9	28,6	24,6

Der Sollwert für den Magnesiumgehalt liegt bei 25mg/l. Mit Hilfe der Messungen soll überprüft werden, ob der mittlere Magnesiumgehalt vom vorgegebenen Sollwert abweicht.

14.3. Bier
Beim Ausschank von Bier ist die Füllmenge je Glas normalverteilt. 33 Kontrollmessungen ergaben folgende Werte für die Füllmenge (in ml):

500	503	498	497	502	505	495	499	505	495	499
510	509	499	506	495	490	505	507	506	495	490
495	499	505	507	495	499	505	495	499	503	496

Der Wirt, bei dem diese Kontrollmessungen stattgefunden haben, wird verdächtigt, zuwenig Bier in die Halbe-Gläser zu füllen. Überprüfen Sie diese Anschuldigung.

14.4. Interesse an Sportübertragungen
In einer Lehrveranstaltung wurden die dort anwesenden Studierenden gefragt, ob sie sich für Sportübertragungen im TV interessieren. Die 240 befragten Personen verteilten sich folgendermaßen auf dem zweidimensionalen Merkmal Geschlecht und Interesse.

	Interesse	kein Interesse	Summe
männlich	60	30	90
weiblich	70	80	150
Summe	130	110	240

Kann ein Zusammenhang zwischen den Merkmalen Geschlecht und Sportübertragung in der Grundgesamtheit nachgewiesen werden?

14.5. Freude an der Schule
Bei einer Befragung von insgesamt 3220 Kindern ergab eine Auswertung nach dem zweidimensionalen Merkmal Geschlecht und Freude an der Schule folgende Verteilung.

	große Freude	geringe Freude	Summe
männlich	1224	226	1450
weiblich	1674	96	1770
Summe	2898	322	3220

Kann ein Zusammenhang zwischen den Merkmalen Geschlecht und Freude an der Schule in der Grundgesamtheit nachgewiesen werden?

Tabellen

Tabelle 1: Verteilungsfunktion der Standardnormalverteilung

Tabelle 2: Quantile der Student-Verteilung

Tabelle 3: Quantile der Chi-Quadrat-Verteilung

Tabelle 1: Verteilungsfunktion der Standardnormalverteilung

$$\Phi(-u) = 1 - \Phi(u)$$

Ablesebeispiel: $\Phi(-1,91) = 1 - \Phi(1,91) = 1 - 0,9719 = 0,0281$

u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	u
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359	0,0
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753	0,1
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141	0,2
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517	0,3
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879	0,4
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224	0,5
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549	0,6
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852	0,7
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8079	0,8106	0,8133	0,8
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389	0,9
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621	1,0
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830	1,1
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015	1,2
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177	1,3
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319	1,4
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441	1,5
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545	1,6
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633	1,7
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706	1,8
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767	1,9
2,0	0,9773	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817	2,0
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857	2,1
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890	2,2
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916	2,3
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936	2,4
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952	2,5
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964	2,6
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974	2,7
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981	2,8
2,9	0,9981	0,9982	0,9983	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986	2,9

Tabelle 1a: Ausgewählte Quantile der Standardnormalverteilung

p	0,8	0,9	0,95	0,975	0,98	0,99	0,995
u_p	0,84162	1,28155	1,6449	1,9600	2,0538	2,3264	2,5758

Tabelle 2: Quantile der Student-Verteilung $t_{n;p}$

	p								
n	80%	90%	95%	97,5%	99%	99,5%	99,9%	99,95%	n
1	1,3764	3,0777	6,3138	12,706	31,821	63,657	318,31	636,62	1
2	1,0607	1,8856	2,9200	4,3027	6,9646	9,9248	22,327	31,599	2
3	0,9785	1,6377	2,3534	3,1825	4,5407	5,8409	10,215	12,924	3
4	0,9410	1,5332	2,1319	2,7765	3,7470	4,6041	7,1732	8,6103	4
5	0,9195	1,4759	2,0151	2,5706	3,3649	4,0321	5,8934	6,8688	5
6	0,9057	1,4398	1,9432	2,4469	3,1427	3,7074	5,2076	5,9588	6
7	0,8960	1,4149	1,8946	2,3646	2,9980	3,4995	4,7853	5,4079	7
8	0,8889	1,3968	1,8596	2,3060	2,8965	3,3554	4,5008	5,0413	8
9	0,8830	1,3830	1,8331	2,2622	2,8214	3,2498	4,2968	4,7809	9
10	0,8791	1,3722	1,8125	2,2281	2,7638	3,1693	4,1437	4,5869	10
11	0,8755	1,3634	1,7959	2,2010	2,7181	3,1058	4,0247	4,4370	11
12	0,8726	1,3562	1,7823	2,1788	2,6810	3,0545	3,9296	4,3178	12
13	0,8702	1,3502	1,7709	2,1604	2,6503	3,0123	3,8520	4,2208	13
14	0,8681	1,3450	1,7613	2,1448	2,6245	2,9768	3,7874	4,1405	14
15	0,8662	1,3406	1,7531	2,1315	2,6025	2,9467	3,7328	4,0728	15
16	0,8647	1,3368	1,7459	2,1199	2,5835	2,9208	3,6862	4,0150	16
17	0,8633	1,3334	1,7396	2,1098	2,5669	2,8982	3,6458	3,9651	17
18	0,8621	1,3304	1,7341	2,1009	2,5524	2,8784	3,6105	3,9217	18
19	0,8610	1,3277	1,7291	2,0930	2,5395	2,8609	3,5794	3,8834	19
20	0,8600	1,3253	1,7247	2,0860	2,5280	2,8453	3,5518	3,8495	20
21	0,8591	1,3232	1,7207	2,0796	2,5177	2,8314	3,5272	3,8193	21
22	0,8583	1,3212	1,7171	2,0739	2,5083	2,8188	3,5050	3,7921	22
23	0,8575	1,3195	1,7139	2,0687	2,4999	2,8073	3,4850	3,7676	23
24	0,8569	1,3178	1,7109	2,0639	2,4922	2,7969	3,4668	3,7454	24
25	0,8562	1,3164	1,7081	2,0595	2,4851	2,7874	3,4502	3,7251	25
26	0,8557	1,3150	1,7056	2,0555	2,4786	2,7787	3,4350	3,7066	26
27	0,8551	1,3137	1,7033	2,0518	2,4727	2,7707	3,4210	3,6896	27
28	0,8547	1,3125	1,7011	2,0484	2,4671	2,7633	3,4082	3,6739	28
29	0,8542	1,3114	1,6991	2,0452	2,4620	2,7564	3,3962	3,6594	29
30	0,8538	1,3104	1,6973	2,0423	2,4573	2,7500	3,3852	3,6460	30
50	0,8489	1,2987	1,6759	2,0086	2,4033	2,6778	3,2614	3,4960	50
100	0,8452	1,2901	1,6602	1,9840	2,3642	2,6259	3,1737	3,3905	100
150	0,8440	1,2872	1,6551	1,9759	2,3515	2,6090	3,1455	3,3566	150
200	0,8434	1,2858	1,6525	1,9719	2,3451	2,6006	3,1315	3,3398	200
500	0,8423	1,2833	1,6479	1,9647	2,3338	2,5857	3,1066	3,3101	500
∞	0,8416	1,2816	1,6449	1,9600	2,3264	2,5758	3,0902	3,2905	∞

Tabelle 3: Quantile der Chi-Quadrat-Verteilung $\chi^2_{n;p}$

	p						
n	0,5%	1%	2,5%	5%	10%	50%	n
1	0,0000	0,0002	0,0010	0,0039	0,0158	0,4549	1
2	0,0100	0,0201	0,0506	0,1026	0,2107	1,3863	2
3	0,0717	0,1148	0,2158	0,3518	0,5844	2,3660	3
4	0,2070	0,2971	0,4844	0,7107	1,0636	3,3567	4
5	0,4117	0,5543	0,8312	1,1455	1,6103	4,3515	5
6	0,6757	0,8721	1,2373	1,6354	2,2041	5,3481	6
7	0,9893	1,2390	1,6899	2,1674	2,8331	6,3458	7
8	1,3444	1,2390	2,1797	2,7326	3,4895	7,3441	8
9	1,7349	1,2390	2,7004	3,3251	4,1682	8,3428	9
10	2,1559	2,5582	3,2470	3,3251	4,8652	9,3418	10
11	2,6032	3,0535	3,8158	4,5748	5,5778	10,3410	11
12	3,0738	3,5706	4,4038	5,2260	6,3038	11,3403	12
13	3,5650	4,1069	5,0088	5,8919	7,0415	12,3398	13
14	4,0747	4,6604	5,6287	6,5706	7,7895	13,3393	14
15	4,6009	5,2294	6,2621	7,2609	8,5468	14,3389	15
16	5,1422	5,8122	6,9077	7,9617	9,3122	15,3385	16
17	5,6972	6,4078	7,5642	8,6718	10,0852	16,3382	17
18	5,6972	7,0149	8,2308	9,3905	10,8649	17,3379	18
19	6,8440	7,6327	8,9065	10,1170	11,6509	18,3377	19
20	7,4338	8,2604	9,5908	10,8508	12,4426	19,3374	20
21	8,0337	8,8972	10,2829	11,5913	13,2396	20,3372	21
22	8,6427	9,5425	10,9823	12,3380	14,0415	21,3370	22
23	9,2604	10,1957	11,6886	13,0905	14,8480	22,3369	23
24	9,8862	10,8564	12,4012	13,8484	15,6587	23,3367	24
25	10,5197	11,5240	13,1197	14,6114	16,4734	24,3366	25
26	11,1602	12,1982	13,8439	15,3792	17,2919	25,3365	26
27	11,8076	12,8785	14,5734	16,1514	18,1139	26,3363	27
28	12,4613	13,5647	15,3079	16,9279	18,9392	27,3362	28
29	13,1212	14,2565	16,0471	17,7084	19,7677	28,3361	29
30	13,7867	14,9535	16,7908	18,4927	20,5992	29,3360	30
40	20,7065	22,1643	24,4330	26,5093	29,0505	39,3353	40
50	27,9908	29,7067	32,3574	34,7643	37,6887	49,3349	50
60	35,5345	37,4849	40,4818	43,1880	46,4589	59,3347	60
70	43,2752	45,4417	48,7576	51,7393	55,3289	69,3345	70
80	51,1719	53,5401	57,1532	60,3915	64,2778	79,3343	80
90	59,1963	61,7541	65,6466	69,1260	73,2911	89,3342	90
100	67,3276	70,0649	74,2219	77,9295	82,3581	99,3341	100

Tabelle 3: Quantile der Chi-Quadrat-Verteilung $\chi^2_{n;p}$ (Fortsetzung)

	p						
n	50%	90%	95%	97,5%	99%	99,5%	n
1	0,4549	2,7055	3,8415	5,0239	6,6349	7,8794	1
2	1,3863	4,6052	5,9915	7,3778	9,2103	10,5966	2
3	2,3660	6,2514	7,8147	9,3484	11,3449	12,8382	3
4	3,3567	7,7794	9,4877	11,1433	13,2767	14,8603	4
5	4,3515	9,2364	11,0705	12,8325	15,0863	16,7496	5
6	5,3481	10,6446	12,5916	14,4494	16,8119	18,5476	6
7	6,3458	12,0170	14,0671	16,0128	18,4753	20,2777	7
8	7,3441	13,3616	15,5073	17,5346	20,0902	21,9550	8
9	8,3428	14,6837	16,9190	19,0228	21,6660	23,5894	9
10	9,3418	15,9872	18,3070	20,4832	23,2093	25,1882	10
11	10,3410	17,2750	19,6751	21,9201	24,7250	26,7569	11
12	11,3403	18,5494	21,0261	23,3367	26,2170	28,2995	12
13	12,3398	19,8119	22,3620	24,7356	27,6883	29,8195	13
14	13,3393	21,0641	23,6848	26,1190	29,1412	31,3194	14
15	14,3389	22,3071	24,9958	27,4884	30,5779	32,8013	15
16	15,3385	23,5418	26,2962	28,8454	31,9999	34,2672	16
17	16,3382	24,7690	27,5871	30,1910	33,4087	35,7185	17
18	17,3379	25,9894	28,8693	31,5264	34,8053	37,1565	18
19	18,3377	27,2036	30,1435	32,8523	36,1909	38,5823	19
20	19,3374	28,4120	31,4104	34,1696	37,5662	39,9969	20
21	20,3372	29,6151	32,6706	35,4789	38,9322	41,4011	21
22	21,3370	30,8133	33,9244	36,7807	40,2894	42,7957	22
23	22,3369	32,0069	35,1725	38,0756	41,6384	44,1813	23
24	23,3367	33,1962	36,4150	39,3641	42,9798	45,5585	24
25	24,3366	34,3816	37,6525	40,6465	44,3141	46,9279	25
26	25,3365	35,5632	38,8851	41,9232	45,6417	48,2899	26
27	26,3363	36,7412	40,1133	43,1945	46,9629	49,6449	27
28	27,3362	37,9159	41,3371	44,4608	48,2782	50,9934	28
29	28,3361	39,0875	42,5570	45,7223	49,5879	52,3356	29
30	29,3360	40,2560	43,7730	46,9792	50,8922	53,6720	30
40	39,3353	51,8051	55,7585	59,3417	63,6907	66,7660	40
50	49,3349	63,1671	67,5048	71,4202	76,1539	79,4900	50
60	59,3347	74,3970	79,0819	83,2977	88,3794	91,9517	60
70	69,3345	85,5270	90,5312	95,0232	100,425	104,215	70
80	79,3343	96,5782	101,879	106,629	112,329	116,321	80
90	89,3342	107,565	113,145	118,136	124,116	128,299	90
100	99,3341	118,498	124,342	129,561	135,807	140,169	100

Lösungen zu den Übungsaufgaben

Lösungen zu Kapitel 1

1.1 Notenverteilung

a) Die Grundgesamtheit umfasst alle Studierenden, die z.B. am 31. Jänner 2005 im Hörsaal 7 an der Johannes Kepler Universität Linz zwischen 15.30 und 17.00 Uhr die Statistik-Klausur geschrieben haben.
b) Die Erhebungseinheit ist eine Studentin aus der Grundgesamtheit.
c) Interessante Merkmale sind unter anderem das Geschlecht, die Studienrichtung und die Abschlussnote in Mathematik, Klausurnote.
d) Geschlecht: männlich, weiblich Note: 1, 2, 3, 4, 5
Studienrichtung: Wirtschaftspädagogik, Statistik, Soziologie, ...

1.2 Medizinische Studie

a) Die Grundgesamtheit umfasst jene PatientInnen, die mit einem der beiden Medikamente behandelt wurden. Eine räumliche Eingrenzung könnte sich auf ein bestimmtes Land beziehen, daneben ist auch der exakte Zeitraum der Studie festzulegen.
b) Erhebungseinheit ist eine einzelne Patientin bzw. ein Patient.
c) Neben dem Blutdruck sind sicherlich Geschlecht, Alter und die Zugehörigkeit zu Risikogruppen (Übergewicht, RaucherIn) zu erheben.
d) systolischer Blutdruck: 130 mm Hg, 140 mm Hg, ...
diastolischer Blutdruck: 80 mm Hg, 100 mm Hg, ...
Geschlecht: männlich, weiblich
Alter: 36 Jahre, 70 Jahre, ...
RaucherIn: ja, nein
Übergewicht: ja, nein

Lösungen zu Kapitel 2

2.1 Skalenniveaus von Merkmalen

a) nominal, diskret
b) metrisch, stetig
c) metrisch, stetig
d) metrisch, diskret (quasistetig)
e) nominal, diskret
f) metrisch, diskret
g) metrisch, stetig (diskretisiert)
h) ordinal, diskret
i) metrisch, stetig (diskretisiert)
j) ordinal, diskret
k) metrisch, stetig

2.2 Volkszählung

Bundesland: nominal, diskret
Zeitpunkt der Geburt: metrisch, stetig (diskretisiert)
Familienstand: nominal, diskret

Lösungen zu Kapitel 6

6.1 Kariöse Zähne

a) $Pr(2 < x \leq 4) = f(3) + f(4) = F(4) - F(2) = 16,4\%$
b) $Pr(2 \leq x \leq 4) = f(2) + f(3) + f(4) = F(4) - F(1) = 30,7\%$
c) $Pr(2 \leq x < 4) = f(2) + f(3) = F(3) - F(1) = 25,7\%$
d) $Pr(2 < x < 4) = f(3) = F(3) - F(2) = 11,4\%$

6.2 Körpergröße von Studierenden

Bei Intervallskalierung]155,160],]160,165], ...

a) 89,5% der Studierenden sind höchstens 183 cm groß.
b) 96,3% der Studierenden sind größer als 158 cm.
c) 96,9% der Studierenden sind höchstens 190 cm groß.

Ohne Intervallskalierung

a) 92,3% der Studierenden sind höchstens 183 cm groß.
b) 98,5% der Studierenden sind größer als 158 cm.
c) 96,9% der Studierenden sind höchstens 190 cm groß.

6.3 Altersverteilung

a) p_i (0,164; 0,184; 0,247; 0,188; 0,142; 0,075)
b) Dichten f_i (0,011; 0,012; 0,016; 0,013; 0,009; 0,003)
c) $F(12) = 0,131$ $S(12) = 0,869$
$F(35) = 0,431$ $S(35) = 0,569$
$F(60) = 0,783$ $S(60) = 0,217$
13,1% der Bevölkerung sind höchstens 12 Jahre alt, 86,9% sind älter als 12 Jahre.

6.4 BundespräsidentInnenwahl

p_i (0,476; 0,524)

6.5 Nationalratswahl

p_i (0,446; 0,498; 0,054; 0,002)

6.6 TV-Geräte

a) p_i (0,113; 0,367; 0,454; 0,049; 0,016)
b) F_i (0,113; 0,480; 0,934; 0,984; 1); Treppenfunktion

Lösungen zu Kapitel 7

7.1 Kariöse Zähne

a) $\widetilde{x}_{0,5} = 1$
b) $\widetilde{x}_{0,5} = 1$
c) $\widetilde{x}_{0,9} = 4$
d) $\overline{x} = 1,77$
$\widetilde{x}_{0,5} = 1$
e) EXCEL: $\alpha = 1,33$ $\gamma = 1,60$
händisch: $\alpha = 1,29$ $\gamma = 1,37$

7.2 Kinderzahl

a) $\overline{x} = 1,3$ $x_{\text{mod}} = 1$ $\widetilde{x}_{0,5} = 1$
b) $s^2 = 1,29$ $s = 1,13$ $V = 0,87$
c) EXCEL: $\alpha = 0,79$ $\gamma = 0,32$
händisch: $\alpha = 0,78$ $\gamma = 0,30$

7.3 Körpergröße von Studierenden

EXCEL - Verwendung des unveränderten Datensatzes

a) $\widetilde{x}_{0,4} = 168$
b) $\widetilde{x}_{0,3} = 165,7$
c) $\overline{x} = 171,7 \quad x_{\text{mod}} = 165 \quad \widetilde{x}_{0,5} = 170$
$s^2 = 71,3 \quad s = 8,4 \quad V = 0,05$

Händisch nach Klasseneinteilung:

$155 < x \leq 165$
$165 < x \leq 175$
$175 < x \leq 185$
$185 < x \leq 195$

a) $\widetilde{x}_{0,4} = 167,3$
b) $\widetilde{x}_{0,3} = 165,0$
c) $\overline{x} = 170,3 \quad x_{\text{mod}} =]165,175] \quad \widetilde{x}_{0,5} = 169,6$
$s^2 = 76,6 \quad s = 8,7 \quad V = 0,05$

7.4 Altersverteilung

a) $\overline{x} = 40,6 \quad \widetilde{x}_{0,5} = 39,2 \quad x_{\text{mod}} =]30,45]$
b) $s^2 = 536,96 \quad \alpha = 0,30 \quad \gamma = -0,72$
c) $\widetilde{x}_{0,25} = 21,98 \quad \widetilde{x}_{0,75} = 57,35$

7.5 TV-Geräte

a) $\overline{x} = 1,38 \quad \widetilde{x}_{0,5} = 1 \quad x_{\text{mod}} = 0$
b) EXCEL: $s^2 = 1,92 \quad \alpha = 1,29 \quad \gamma = 2,40$
händisch: $s^2 = 1,92 \quad \alpha = 1,27 \quad \gamma = 2,22$

7.6 Klausurergebnisse

$\widetilde{x}_{0,5} = 4 \qquad x_{\text{mod}} = 5$

Lösungen zu Kapitel 8

8.1 Interesse an Sportübertragungen

a) m: j: 0,250 n: 0,125
w: j: 0,292 n: 0,333
25% der Studierenden sind männlich und haben Interesse an Sportübertragungen.

b) m: j: 0,667 n: 0,333
w: j: 0,467 n: 0,533
66,7% der männlichen Studierenden haben Interesse an Sportübertragungen. Bei den Männern ist der Anteil der an Sportübertragungen Interessierten höher als bei den Frauen.

c) $\chi^2 = 9,063$
$V = 0,194$
Es existiert ein eher schwacher Zusammenhang zwischen Geschlecht und Interesse an Sportübertragungen.

8.2 Inflationsrate und Staatsschulden/BIP

$\rho = -0,309$
Inflationsrate und Schuldenanteil weisen einen schwachen bis mittelstarken gegensinnigen linearen Zusammenhang auf. Länder mit einer niedrigeren Verschuldung haben tendenziell höhere Inflationsraten.

8.3 Körpergröße und Gewicht

$\rho = 0,727$
Zwischen Körpergröße und Gewicht herrscht ein mittelstarker bis starker gleichsinniger linearer Zusammenhang. Größere Menschen sind tendenziell auch schwerer.

8.4 Leistung und Drehzahl

$\rho = 0,9997$
Zwischen Leistung und Drehzahl besteht ein fast vollständiger gleichsinniger linearer Zusammenhang. Hohe Leistung bedeutet auch hohe Drehzahl.

8.5 Abfahrtslauf

$\rho_s = 0,405$
Zuerst müssen die Zeiten in Platzierung umkodiert werden, dann kann der Spearmansche Rangkorrelationskoeffizient berechnet werden. Es gibt einen mittelstarken gleichsinnigen Zusammenhang zwischen Startnummer und Platzierung. LäuferInnen mit niedrigeren Startnummern haben bessere Plätze.

8.6 Lehrveranstaltung

$\rho_s = 0,393$
Zwischen Eindruck und Leistung besteht ein mittelstarker gleichsinniger Zusammenhang. Studierende, die einen besseren Eindruck erweckt haben, weisen tendenziell bessere Ergebnisse auf.

8.7 Freude an der Schule

a) m: groß: 0,380 gering: 0,070
w: groß: 0,520 gering: 0,030
38% der Kinder sind männlich und haben große Freude an der Schule.
b) m: groß: 0,844 gering: 0,156
w: groß: 0,946 gering: 0,054
84,4% der Buben und 94,6% der Mädchen haben große Freude an der Schule.
c) $\chi^2 = 91,462$
$V = 0,169$
Es existiert ein schwacher Zusammenhang zwischen Geschlecht und Freude an der Schule.

Lösungen zu Kapitel 9

9.1 Körpergröße und Gewicht

a) Gewicht $= -114,49 + 1,07 \cdot$ Körpergröße
b) (2,66; 0,66; -7,95; 4,32; 6,80; -0,20; 6,05; -8,61; -4,86; 1,12)
c) $R^2 = 0,528$
d) Die Güte der Prognose ist mittelmäßig. 52,8% der Varianz des Merkmals Gewicht kann durch das Modell der linearen Einfachregression erklärt werden.
e) *bei* $K = 180$: $\widehat{G} = 77.68$

9.2 Einkommen und Ausgaben von Haushalten

a) Ausgaben $= 98,72 + 0,78 \cdot$ Einkommen
b) (6,07; 105,36; 48,60; 271,98; -170,41; -141,68; -53,65; 135,08; 78,32; 41,84; 10,72; 70,58; 13,11; 20,86; -41,40; -238,02; 6,15; -144,66; -71,68; 52,80)
c) $R^2 = 0,903$
d) Die Güte der Prognose ist sehr gut. 90,3% der Varianz des Merkmals Haushaltsausgaben kann durch das Modell der linearen Einfachregression erklärt werden.
e) *bei Einkommen* $= 1200$: $\widehat{A} = 1031,68$

Lösungen zu Kapitel 10

10.1 Roulette

a) $\Omega = \{0, 1, 2, ..., 36\}$
b) $E_1 = \{1, 3, 5, ..., 35\}$ „ungerade Zahlen"
$E_2 = \{31, 32, ..., 36\}$ „Zahl größer als 30"
c) $E_1 = \{1\}$, $E_2 = \{13\}$
d) {3,5}, {42}, $\{\pi\}$, $\emptyset$
e) A = {1, 2, ..., 12} „Zahlen im ersten Drittel"
B = {13, 14, ..., 24} „Zahlen im zweiten Drittel"
f) $\Omega = \{0, 1, 2, ..., 36\}$

10.2 Roulette

a) $A \cap B = \{2, 4, 6, 8, 10, 12\}$, $A \cup B = \{1, 2, ..., 12, 14, 16, 18, ..., 36\}$
b) $A^C = \{1, 3, 5, ..., 35\}$
c) $A \cap B \cap C = \{10, 12\}$
d) $A \cup B \cup C = \{1, 2, ..., 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36\}$
$(A \cup B \cup C)^C = \{0, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35\}$
e) nein
f) ja

10.3 Roulette

a) $Pr(A) = 18/37$, $Pr(B) = 12/37$, $Pr(C) = 6/37$
b) $Pr(A \cap B) = 6/37$, $Pr(A \cup B) = 24/37$
c) $Pr(A^C) = 19/37$
d) $Pr(A \cap B \cap C) = 2/37$
e) $Pr(A \cup B \cup C) = 26/37$, $Pr((A \cup B \cup C)^C) = 11/37$

10.4 Roulette

a) $Pr(CC) = (6/37)^2 = 0,026$
b) $Pr(C|B) = 6/37$
c) $Pr(2|geradeZahl) = 1/18$
d) $Pr(C|\{0\}) = 0$
e) $Pr(C|\{10\}) = 1$

10.5 Rubbellos

a) $Pr(30.000\,€) = 10/10.000.000 = 0,000001$
b) $Pr(X \geq 1.000\,€) = (150 + 50 + 10)/10.000.000 = 0,000021$
c) $Pr(X \leq 2\,€) = (1.750.000 + 7.189.790)/10.000.000 = 0,894$
d) $Pr(X \geq 1.000\,€|X > 0) = (10 + 50 + 150)/2.810.210 = 0,00007473$
e) $Pr(X = 2\,€|X > 0) = 1.750.000/2.810.210 = 0,6227$

10.6 Zwei Würfel

a) $E_1 = \{1\}$, $E_2 = \{8\}$
b) $\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$
c) $A = \{2, 4, 6, 8, 10, 12\}$ „gerade Summe"
   $B = \{2, 3, 4, 5, 6\}$ „Summe kleiner Sieben"
d) $\{1\}, \{7, 5\}, \{\pi\}, \{25\}$
e) A = {2, 3, 4, 5, 6} „Summe kleiner Sieben" und
   B = {8, 9, 10, 11, 12} „Summe größer Sieben"
f) $\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

10.7 Zwei Würfel

a) $Pr(X = 2) = 1/36$ $Pr(X = 3) = 2/36$ $Pr(X = 4) = 3/36$,
   $Pr(X = 5) = 4/36$ $Pr(X = 6) = 5/36$ $Pr(X = 7) = 6/36$,
   $Pr(X = 8) = 5/36$ $Pr(X = 9) = 4/36$ $Pr(X = 10) = 3/36$,
   $Pr(X = 11) = 2/36$ $Pr(X = 12) = 1/36$
b) $Pr(6|\text{gerade Zahl}) = 5/18$
c) $Pr(6|\text{ungerade Zahl}) = 0$
d) $Pr(\text{gerade Zahl}|\text{höchstens 4}) = 2/3$
e) $Pr(\text{gerade Zahl}|\text{mindestens 4}) = 17/33$
f) E(X) = 7

10.8 Zwei Würfel

a) $A \cap B = \{2, 4, 6\}$ $A \cup B = \{2, 3, 4, 5, 6, 8, 10, 12\}$
b) $A^C = \{3, 5, 7, 9, 11\}$
c) $A \cap B \cap C = \{\}$ $A \cup B \cup C = \{2, 3, 4, 5, 6, 8, 9, 10, 11, 12\}$
d) $(A \cup B \cup C)^C = \{7\}$
e) B und C sind paarweise disjunkt
f) ja

10.9 Zwei Würfel

a) $Pr(A) = 0,5 \qquad Pr(B) = 5/12 \qquad Pr(C) = 5/12$
b) $Pr(A \cap B) = 0,25 \qquad Pr(A \cup B) = 2/3$
c) $Pr(A^C) = 0,5$
d) $Pr(A \cap B \cap C) = 0 \qquad Pr(A \cup B \cup C) = 5/6$
e) $Pr(A \cup B \cup C)^C = 1/6$
f) $Pr(A \cup D) = 2/3$

10.10 Zwei Würfel

1/36 1/36 1/6 0,25 5/12

10.11 Kinder

0,25 0,5 1/3

Lösungen zu Kapitel 11

11.1 Würfel

$Pr(X = 0) = 5/6 \qquad Pr(X = 1) = 1/6 \qquad Pr(X \leq 1) = 1$

11.2 Urne mit Zurücklegen

$Pr(X = 0) = 0,216 \qquad Pr(X = 1) = 0,432 \qquad Pr(X = 2) = 0,288$
$Pr(X = 3) = 0,064 \qquad E(X) = 1,2$

11.3 Urne ohne Zurücklegen

$Pr(X = 0) = 0,167 \qquad Pr(X = 1) = 0,5 \qquad Pr(X = 2) = 0,3$
$Pr(X = 3) = 0,033 \qquad E(X) = 1,2$

11.4 Karten

$Pr(\text{„}As\text{”}) = 1/13 \quad Pr(X \leq 5) = 4/13 \quad Pr(X > \text{„}Bube\text{”}) = 3/12$

11.5 Joker

Ziehen mit Zurücklegen $\qquad Pr(\text{„}Joker\text{”}) = 0,000001 \qquad E(X) = 0,6$
$Pr(\text{„}3er\text{”}) = \frac{9}{10}\frac{1}{10}\frac{1}{10}\frac{1}{10} = 9/10.000$

11.6 Münze

$Pr(\text{alle richtig}) = 0,015625$ $Pr(\text{mind. einmal richtig}) = 0,984375$
B(512; 0,015625): $Pr(\text{niemand richtig}) = 0,000315$ 8 Personen

11.7 Glühbirne

$Pr(X = 1) = 0,268$ $Pr(X \geq 1) = 0,893$

11.8 Lotto

$Pr(X = 0) = 0,301$ $Pr(X \geq 2) = 0,337$

11.9 Autolackiererei

$Pr(X = 2) = 0,147$ $Pr(X = 8) = 0,012$

11.10 Urne

$Pr(X \leq 2) = 0,044$ (Approximation durch Binomialverteilung)
$E(X) = 6$ Nein, die Voraussetzung ist nicht erfüllt

11.11 Qualitätskontrolle

Approximation durch Binomialverteilung:

$Pr(X > 2\%) = Pr(X > 4) = 0,052$
$Pr(X < 0,5\%) = Pr(X < 1) = 0,134$

Approximation durch Poissonverteilung:

$Pr(X > 2\%) = 0,053$ $Pr(X < 0,5\%) = 0,135$

Lösungen zu Kapitel 12

12.1 Schrauben

$Pr(X \leq 4,28) = 0,919$ $Pr(X \leq 3,8) = 0,159$
$Pr(X \geq 4,21) = 0,147$ $Pr(X \geq 3,91) = 0,674$
$Pr(3,85 \leq X \leq 4,1) = 0,691 - 0,227 = 0,464$

12.2 Schrauben

$Pr(X \leq 4,26) = 0,90$ $\quad Pr(3,67 \leq X \leq 4,33) = 0,90$
$Pr(X \geq 3,74) = 0,90$

12.3 Produktion

$Pr(X \leq 6) = 0,0443$ $\quad Pr(X = 6) = 0,0248$ $\quad$ (Binomialverteilung)
$Pr(X \leq 6) = 0,0544$ $\quad Pr(X = 6) = 0,0254$ $\quad$ (Normalverteilung)

12.4 Lotto

$Pr(X \leq 125) = 0,696$ $\quad Pr(X \geq 100) = 0,972$ $\quad$ (Poissonverteilung)
$Pr(X \leq 125) = 0,692$ $\quad Pr(X \geq 100) = 0,969$ $\quad$ (Normalverteilung)

12.5 Niederschlagsmenge

$x_{0,05} = 63,5$ $\quad x_{0,95} = 86,5$
Mit 5%iger Wahrscheinlichkeit beträgt die Niederschlagsmenge im April in Linz höchstens 63,5mm.

12.6 Waschpulver

$\mu = 1004,14$

12.7 Radioaktivität

$Pr(X < 125) = 0,017$ $\quad$ (Poissonverteilung)
$Pr(X < 125) = 0,019$ $\quad$ (Normalverteilung)

12.8 Brücke

$Pr(Y > 100.000) = 0,022$ $\quad x_{0,999} = 100.296$

12.9

$Pr(Y > 0) = 0,581$ $\quad Pr(Y = 0) = 0,052$

Lösungen zu Kapitel 13

13.1 AkademikerInnenquote

$[\underline{p}, \overline{p}] = [0,052; 0,056]$
Mit 95%iger Wahrscheinlichkeit überdeckt das Intervall $[0,052; 0,056]$ den Anteil der österreichischen Bevölkerung mit einem Universitätsabschluss.

13.2 Nationalratswahl

$[\underline{p}, \overline{p}] = [0,305; 0,415]$
Mit 99%iger Wahrscheinlichkeit überdeckt das Intervall $[0,305; 0,415]$ den Anteil der SPÖ-WählerInnen in der österreichischen Gesamtbevölkerung zu diesem Zeitpunkt.
$[\underline{p}, \overline{p}] = [0,318; 0,402]$

13.3 Mineralwasser

$[\underline{\mu}, \overline{\mu}] = [24,2; 26,2]$
Mit 95%iger Wahrscheinlichkeit überdeckt das Intervall $[24,2; 26,2]$ den mittleren Magnesiumgehalt in der Gesamtproduktion.

13.4 Bier

$[\underline{\mu}, \overline{\mu}] = [498,4; 502,1]$
Mit 95%iger Wahrscheinlichkeit überdeckt das Intervall $[498,4; 502,1]$ die mittlere Füllmenge pro Glas.

Lösungen zu Kapitel 14

14.1 Nationalratswahl

$H_0 : p \leq 0,33 \qquad H_1 : p > 0,33 \qquad \alpha = 0,05$
$[\underline{p}; \overline{p}] = [0,323; 1]$ Nullhypothese beibehalten
Es konnte nicht nachgewiesen werden, dass der derzeitige Stimmanteil der SPÖ-WählerInnen den Anteil der letzten Wahl überschreitet.

14.2 Mineralwasser

$H_0 : \mu = 25 \qquad H_1 : \mu \neq 25 \qquad \alpha = 0,05$

$[\underline{\mu}; \overline{\mu}] = [24,24; 26,17]$ Nullhypothese beibehalten

Es konnte nicht nachgewiesen werden, dass der mittlere Magnesiumgehalt vom Sollwert 25mg/l abweicht.

14.3 Bier

$H_0 : \mu \geq 500 \qquad H_1 : \mu < 500 \qquad \alpha = 0,05$

$[\underline{\mu}; \overline{\mu}] = [-\infty; 501,81]$ Nullhypothese beibehalten

Es konnte nicht nachgewiesen werden, dass der Wirt zuwenig Bier in die Halbe-Gläser füllt.

14.4 Interesse an Sportübertragungen

$H_0 : \chi^2 = 0 \qquad H_1 : \chi^2 > 0 \qquad \alpha = 0,05$

$\chi^2_{(r-1)(s-1);1-\alpha} = 3,8415 \qquad \chi^2_{err} = 9,06$

Nullhypothese verwerfen

Mit 95%iger Sicherheit konnte nachgewiesen werden, dass ein Zusammenhang zwischen dem Geschlecht und dem Interesse an Sportübertragungen besteht.

14.5 Freude an der Schule

$H_0 : \chi^2 = 0 \qquad H_1 : \chi^2 > 0 \qquad \alpha = 0,05$

$\chi^2_{(r-1)(s-1);1-\alpha} = 3,8415 \qquad \chi^2_{err} = 91,46$

Nullhypothese verwerfen

Mit 95%iger Sicherheit konnte nachgewiesen werden, dass ein Zusammenhang zwischen dem Geschlecht und der Freude an der Schule besteht.

Symbolverzeichnis

Symbol	Kurzbeschreibung	Seite
α	α-Fehler, Fehler 1. Art	228
α	Momentenkoeffizient der Schiefe	115
α_i	Zentriwinkel	71
β	β-Fehler, Fehler 2. Art	228
χ^2	Assoziationsmaß Chi-Quadrat	126
χ^2_{err}	Assoziationsmaß Chi-Quadrat der Stichprobe	242
$\chi^2_{n;p}$	p-Quantil der Chi-Quadrat-Verteilung mit n Freiheitsgraden	243
$\emptyset$, $\{\}$	leere Menge	160
γ	Wölbungskoeffizient	116
$\hat{p}$	Stichprobenanteil	215
μ	Erwartungswert	196
$\overline{A}$, A^C	Komplementärmenge von A	160
$\Phi(u)$	Verteilungsfunktion der Standardnormalverteilung	200
ρ	Bravais-Pearson-Korrelationskoeffizient	132
ρ_s	Spearmansche Rangkorrelationskoeffizient	128
σ	Standardabweichung	196
σ^2	Varianz	196
$\underline{\mu}, \overline{\mu}$	Unter- bzw. Obergrenze eines Konfidenzintervalls für einen Mittelwert	222
$\underline{p}, \overline{p}$	Unter- bzw. Obergrenze eines Konfidenzintervalls für einen Anteil	220
$\widehat{a}$	Achsenabschnitt der Regressionsgerade	148
$\widehat{b}$	Steigung der Regressionsgerade	148
$\widehat{e}_i$	Residuum (Regression)	148
$\widehat{y}$	geschätzter Wert (Regression)	148
$\widetilde{x}_{0,5}$	Median	98
$\|A\|$	Kardinalzahl oder Mächtigkeit der Menge A	160

Symbol	Kurzbeschreibung	Seite
$A \backslash B$	Differenzmenge A ohne B	160
$A \cap B$	Schnittmenge (Durchschnitt) von A und B	160
$A \cup B$	Vereinigungsmenge von A und B	160
$A \subset B$	A ist eine Teilmenge von B	160
$A(p)$	Alternativverteilung	177
B	Bestimmtheitsmaß (Regression)	150
$B(n,p)$	Binomialverteilung	180
d_i	Intervallbreite	73
d_i	Rangzahlendifferenz	128
$E(X)$	Erwartungswert	175
e_{i-1}	Untergrenze des i-ten Intervalls, $i = 1, \ldots, r$	68
e_i	Obergrenze des i-ten Intervalls	68
$f(x), f_i$	Dichte	73, 173, 194
$F(x), F(x_i)$	Verteilungsfunktion	85, 174, 195
g	geometrisches Mittel	102
$G(a,b)$	Stetige Gleichverteilung	198
$G(N)$	Diskrete Gleichverteilung	178
g_t	Wachstumsfaktor zur Zeit t	102
$H(N,A,n)$	Hypergeometrische Verteilung	182
$h_{+j}(p_{+j})$	Spaltensummen, Randhäufigkeiten des Merkmals Y	122
$h_{i+}(p_{i+})$	Zeilensummen, Randhäufigkeiten des Merkmals X	122
h_{ij}	zweidimensionale absolute Häufigkeit	122
h_{ij}^e	erwartete zweidimensionale absolute Häufigkeit	125
h_{ij}^o	beobachtete zweidimensionale absolute Häufigkeit	125
H_0, H_1	Nullhypothese, Alternativhypothese	227
h_i	absolute Häufigkeit der Ausprägung x_i	62
$m_k(a)$	Moment der Ordnung k in Bezug auf den Punkt a	114
N	Untersuchungsumfang	62
$NV(\mu, \sigma^2)$	Normalverteilung	199
p	Anteil in der Grundgesamtheit	220
p	durchschnittliche Wachstumsrate	102
$P(\lambda)$	Poissonverteilung	184
$p(x \leq x_i)$	Verteilungsfunktion	85
p_{ij}, P_{ij}	zweidimensionale relative Häufigkeit, $\sim$ in Prozent	122
p_i, P_i	relative Häufigkeit der Ausprägung x_i, $\sim$ in Prozent	62
p_t	Wachstumsrate zur Zeit t	102
$Pr(A\|B)$	bedingte Wahrscheinlichkeit	165
$Pr(E)$	Probability, Wahrscheinlichkeit von E	162
r	Anzahl an verschiedenen Ausprägungen	62
R	Spannweite	108
r_i, s_i	(Durchschnitts-)Ränge	128

Symbol	Kurzbeschreibung	Seite
s	Standardabweichung	109
$S(a)$	Zuverlässigkeitsfunktion, Überlebensfunktion	202
s^2	Varianz	109
$s^2_{\text{korr}}, \hat{s}^2$	korrigierte Varianz	117, 222
s_{XY}	Kovarianz der Merkmale X und Y	131
$t_{n;p}$	p-Quantil der Student-Verteilung mit n Freiheitsgraden	222
u_p	p-Quantil der Standardnormalverteilung	202
V	Assoziationsmaß Cramers V	126
V	Variationskoeffizient	109
$Var(X)$	Varianz	175
$x \in A$	x ist ein Element aus A	160
$\bar{x}$	Mittelwert	96
x_α, x_p	Quantil	104, 197
x_{mod}	Modus	101
$x_{(i)}$	i-te Ausprägung der geordneten Datenreihe	98
x_i	Ausprägung, $i = 1, \ldots, r$ oder $i = 1, \ldots, N$	62

Literaturverzeichnis

[1] Brosius F. (2004) SPSS 12. mitp-Verlag, Bonn.

[2] Bühl A., P. Zöfel (2004) SPSS 12. Einführung in die moderne Datenanalyse unter Windows. Pearson, München.

[3] Fahrmeir L., Tutz G. (2001) Multivariate Statistical Modelling Based on Generalized Linear Models (2. Auflage). Springer-Verlag, New York.

[4] Fahrmeir L., Hammerle A., Tutz G. (1996) Multivariate statistische Verfahren (2. Auflage). de Gruyter, Berlin.

[5] Fahrmeir L., Künstler R., Pigeot I., Tutz G. (2004) Statistik (5. Auflage). Springer-Verlag, Berlin.

[6] Fahrmeir L., Künstler R., Pigeot I., Tutz G., Caputo A., Lang S. (2004) Arbeitsbuch Statistik (4. Auflage). Springer-Verlag, Berlin.

[7] Hafner R. (2000) Statistik für Sozial- und Wirtschaftswissenschaftler, Band 1. Springer-Verlag, Wien.

[8] Hafner R., Waldl H. (2001) Statistik für Sozial- und Wirtschaftswissenschaftler, Band 2. Springer-Verlag, Wien.

[9] Hartung J., B. Elpelt (1999) Multivariate Statistik (6. Auflage). Oldenbourg-Verlag, München.

[10] Hartung J., B. Heine (1996) Statistik-Übungen. Induktive Statistik (3. Auflage). Oldenbourg-Verlag, München.

[11] Hartung J., B. Elpelt, Klösener K.-H. (2002) Statistik (13. Auflage). Oldenbourg-Verlag, München.

[12] Holm K. (Hrsg.) (1991) Die Befragung 1 (4. Auflage). Uni-Taschenbuch 372. Francke Verlag, München.

[13] Jarai H. (2004) Excel 2003/xp. Franzis-Verlag, München.

[14] Kirchhoff S., Kuhnt S., Lipp P., Schlawin S. (2002) Der Fragebogen (3. Auflage). UTB, Stuttgart.

[15] Krämer W. (1991) So lügt man mit Statistik (3. Auflage). Campus, Frankfurt/New York.

[16] Krämer W. (1999) Statistik verstehen. Campus, Frankfurt a.M.

[17] Kreienbrock L. (1993) Einführung in die Stichprobenverfahren (2. Auflage). Oldenbourg-Verlag, München/Wien.

[18] Little J. A., Rubin D. B. (1987) Statistical Analysis with Missing Values. Wiley, New York.

Sachverzeichnis

α-Fehler 228
α-Quantil *siehe* Quantil
β-Fehler 228
χ^2-Test *siehe* Chi-Quadrat-Test

a-posteriori 169
a-priori 169
absolute Häufigkeit 62, 68
absoluter Zellbezug 32
Abzählregel 162
Additionssatz 165
Alternativhypothese 227
Alternativverteilung 177
Ankunftsrate 183
Approximation 185, 186
  Binomialverteilung 186, 208
  Hypergeometrische Verteilung 186
  Normalverteilung 206, 208
  Poissonverteilung 186, 208
arithmetisches Mittel 95, 96, 98
  EXCEL 107
  SPSS 107
Assoziationsmaße 125
  Chi-Quadrat 125, 126
  Cramers V 125–127
  EXCEL 136
  SPSS 141
Ausprägung 8, 62
  Gruppieren 14
Axiome von Kolmogorov 164

Balkendiagramm *siehe* Stabdiagramm
Bearbeitungsleiste 31
bedingte Verteilung 123
  EXCEL 136
  SPSS 141
bedingte Wahrscheinlichkeit 166
Bereichschätzer
  siehe Konfidenzintervall
Bestimmtheitsmaß 150
Bewertungsfrage 16, 17
Bezug *siehe* Zellbezug
bimodal 101
Bindungen 128
Binomialverteilung 180
  Approximation 186, 208
  EXCEL 181
bivariate Statistik 9
Bravais-Pearson-Korrelationskoeffizient
  siehe Korrelation

Checkliste
  Gesamtbericht 22
  Grafik 21
  Tabellen 22
Chi-Quadrat 125, 126
  EXCEL 136
  SPSS 141
Chi-Quadrat-Test 242, 243
  EXCEL 244
  SPSS 245
Cramers V 126, 127
  EXCEL 136
  SPSS 141
Cramersche Assoziationsmaß *siehe*
  Cramers V
Ctrl-Taste 28

Datenansicht 41
Dateneditor 41
Datenerfassung 18
  EXCEL 36
Datenschutz 22
deskriptive Statistik 9
dichotom 14
Dichte 73
  diskrete Zufallsvariable 173
  stetige Zufallsvariable 194
diskret 14
Diskrete Gleichverteilung 178
diskretisiert 14
Drag and Drop 29

Eingabe-Taste 28
einseitiger Test 231
  Anteil 235, 236
  Mittelwert 239, 240
  Unabhängigkeit 243
Elementarereignis 161
Enter-Taste 28
Ereignis
  disjunkte 161
  Elementar- 161
  Komplementär- 161
  paarweise disjunkte 161
  sichere 161
  stochastisch unabhängige 166
  unmögliche 161
Ereignisse 161
Erhebungseinheit 8
Erwartungstreue 218
Erwartungswert
  diskrete Zufallsvariable 175
  stetige Zufallsvariable 196
EXCEL 31
  absoluter Zellbezug 32
  arithmetisches Mittel 107
  Assoziationsmaße 136
  Bearbeitungsleiste 31
  bedingte Verteilung 136
  Binomialverteilung 181
  Chi-Quadrat 136
  Chi-Quadrat-Test 244
  Cramers V 136
  Dateneingabe 36
  Funktionsassistent 37
  geometrisches Mittel 107
  Grafik 81
  Häufigkeitsverteilung 63, 64, 70
  Histogramm 83
  Hypergeometrische Verteilung 183
  Konfidenzintervall 224
  Korrelation 136
  Kovarianz 136
  Lagemaße 105, 107
  Median 107
  Mittelwert 107
  Modus 107
  Normalverteilung 202, 204, 224
  Poissonverteilung 185
  Praxistipp 33, 37
  Quantil 107, 224, 241
  Randverteilung 136
  Rangkorrelationskoeffizient 136
  Regression 151
  relativer Zellbezug 32
  Schiefe 117, 118
  Standardabweichung 118
  Streudiagramm 136
  Symbolleiste 31
  Test
    Unabhängigkeit 244
  Testen
    Anteil 241
    Mittelwert 241
  Varianz 118
  Verteilungsfunktion 91
  Wölbung 117, 118
  Zellen formatieren 33
  Zusammenhang 136
  zweidimensionale Merkmale 136
explorative Datenanalyse 9

Fälle gewichten 54, 56
fehlende Daten 24
fehlende Werte
  SPSS 45, 50
Fehler 1. Art 228
Fehler 2. Art 228
Fragebogen 15

Gegenwahrscheinlichkeit 165
geometrisches Mittel 102
  EXCEL 107
  SPSS 107
Gesamtbericht

Checkliste 22
geschlossene Frage 16, 17
Gesetz der großen Zahlen 206
gewöhnliches Moment 114
Gleichverteilung
Diskrete 178
Stetige 198
Grafik
Anforderungen 80
Auswahl 80
Checkliste 21
EXCEL 81
Häufigkeitsverteilung 71
Histogramm 73, 80
Kreisdiagramm 71, 80
Qualitätskriterien 76
SPSS 84
Stabdiagramm 72, 80
Verteilungsfunktion 86
Grenzwertsatz
von de Moivre 207
Zentraler 207
Grundgesamtheit 8
Gruppieren 14
Güte der Regression 150
Gütekriterien für Schätzer 218

Häufigkeit
absolut 62, 68
eindimensionale Verteilung 61
erwartete 125
kumuliert 85
relativ 62, 68
relativ in Prozent 62, 68
Summe 63
Häufigkeitsverteilung
diskretes Merkmal 61
eindimensional 61
EXCEL 63, 64, 70
Grafik 71
SPSS 66, 67, 70
zweidimensional 121
Histogramm 73, 80
EXCEL 83
Hypergeometrische Verteilung 182
Approximation 186
EXCEL 183
Hypothese
Alternativ- 227
Null- 227

iid-Bedingung 206
induktive Statistik 9
Intensitätsrate 183
intervallskaliert 12
Intervallskalierung 70

kardinalskaliert *siehe* metrisch
Kausalität 136
Kleinste-Quadrate-Schätzer 146
Kodeplan 18
Komplementärereignis 161
Konfidenzintervall 219
EXCEL 224
für den Anteil 220
für den Mittelwert
bekannte Varianz 222
unbekannte Varianz 222
Sicherheit 219
SPSS 225
Konsistenz 218
Kontingenztabelle *siehe* Kreuztabelle
Korrelation
EXCEL 136
Kausalität 136
Scheinkorrelation 136
SPSS 141
Streudiagramm 134, 135
Kovarianz 130, 131
EXCEL 136
SPSS 141
Kreisdiagramm 71, 80
Kreuztabelle 122
Kurtosis 116, *siehe* Wölbung

Lagekennzahl *siehe* Lagemaße
Lagemaße 95
Auswahl geeigneter 113
Eigenschaften 111
EXCEL 105, 107
SPSS 107
Transformationsregeln 112

Median 98, 99
EXCEL 107
intervallskaliertes Merkmal 100
SPSS 107
stetige Zufallsvariable 197

Mehrfachantwort 17
Menge 159
Merkmal 8, 12
  dichotom 14
  diskret 14
    Häufigkeitsverteilung 61
  diskretisiert 14
  Hierarchie 13
  intervallskaliert 12
  metrisch 12
    Intervallskalierung 70
  nominal 13
  ordinal 13
  quasistetig 14
  Skalenniveau 12, 15
  stetig 14
    Häufigkeitsverteilung 68
  verhältnisskaliert 12
Merkmalstyp 12
Messniveau *siehe* Skalenniveau
Methode der kleinsten Quadrate 146
metrisch 12
  Intervallskalierung 70
Mikrozensus 5
Mittel
  arithmetisches 95, 96, 98
  EXCEL 107
  geometrisches 102
Mittelwert 95, 96, 98
modale Klasse *siehe* Modus
Modalwert *siehe* Modus
Modus 101
  EXCEL 107
  SPSS 107
  stetige Zufallsvariable 196
Moment 114
  gewöhnliches 114
  zentrales 114
Multiplikationsregel 166
multivariate Statistik 9

nominal 13
Normalverteilung 199
  Approximation 206, 208
  EXCEL 202, 204, 224
  Rechnen mit der 202
  Standard- 199
  Symmetrische Intervalle 205
Nullhypothese 227

offene Frage 16, 17
ordinal 13

p–Quantil *siehe* Quantil
p–Wert 242
Parameter 177
Poissonverteilung 184
  Approximation 186, 208
  EXCEL 185
Primärstatistik 17
Produktregel 166
Punktschätzer 219

Quantil 104
  EXCEL 107, 224, 241
  intervallskaliertes Merkmal 105
  SPSS 107
  stetige Zufallsvariable 196
quantitativ *siehe* metrisch
Quartil 104
quasistetig 14

Randverteilung 123
  EXCEL 136
  SPSS 141
Rangkorrelationskoeffizient 128, 129
  EXCEL 136
  SPSS 141
Regression 145, 148
  Bestimmtheitsmaß 150
  EXCEL 151
  Güte 150
  Kleinste-Quadrate-Schätzer 146, 148
  linear 147
  Methode der kleinsten Quadrate 146
  Praxistipp 150
  Regressionsgerade 146
  Residuen 148
  SPSS 151, 155
  Ziel 146
Regressionsgerade 146
relative Häufigkeit 62, 68
  in Prozent 62, 68
relativer Zellbezug 32
Residuen 148
Return-Taste 28
Robustheit 113

Satz von Bayes 169
Satz von der totalen Wahrscheinlichkeit 167
Säulendiagramm *siehe* Stabdiagramm
Schätzer
  Bereich- *siehe* Konfidenzintervall
  Erwartungstreue 218
  Gütekriterien 218
  Konsistenz 218
  Punkt- 219
Scheinkorrelation 136
Schiefe 113–115, 197
  EXCEL 117, 118
  SPSS 118
Schließende Statistik 215
Sekundärstatistik 17, 23
Sicherheit 219
Signifikanzniveau 228
Skalenniveau 12, 15
  Hierarchie 13
  SPSS 46
  Zulässige Verfahren 14
Spannweite 108
Spearman *siehe* Rangkorrelationskoeffizient
SPSS 39
  arithmetisches Mittel 107
  Assoziationsmaße 141
  bedingte Verteilung 141
  Chi-Quadrat 141
  Chi-Quadrat-Test 245
  Cramers V 141
  Datenansicht 41
  Dateneditor 41
  Fälle gewichten 54, 56
  fehlende Werte 45, 50
  geometrisches Mittel 107
  Grafik 84
  Häufigkeitsverteilung 66, 67, 70
  Konfidenzintervall 225
  Korrelation 141
  Kovarianz 141
  Lagemaße 107
  Median 107
  Mittelwert 107
  Modus 107
  Öffnen anderer Dateiformate 47
  Quantil 107
  Randverteilung 141
  Rangkorrelationskoeffizient 141
  Regression 151, 155
  Schiefe 118
  Skalenniveau 46
  Standardabweichung 118
  Streudiagramm 141
  Tabelle eingeben *siehe* Fälle gewichten
  Testen
    Anteil 242
    Mittelwert 242
    Unabhängigkeit 242, 245
  Tipps 57
  Variable transformieren 53, 54
  Variable umkodieren 50, 53
  Variablenansicht 41
  Variablenlabel 44
  Variablenname 42
  Variablentyp 42
  Varianz 118
  Verteilungsfunktion 91
  Viewer 49
  Wertelabels 45
  Wölbung 118
  Zusammenhang 141
    Test *siehe* SPSS-Testen-Unabhängigkeit
  zweidimensionale Merkmale 137, 141
Stabdiagramm 72, 80
Standardabweichung 109
  EXCEL 118
  SPSS 118
Standardisierung 112, 200
Standardnormalverteilung 199
Statistischer Test *siehe* Test
stetig 14
Stetige Gleichverteilung 198
Stichprobe 7, 8
  repräsentativ 8
  Zufalls- 8
Stichprobenverteilung 215
Streudiagramm 134, 135
  EXCEL 136
  SPSS 141
Streuungsmaße 108
  Eigenschaften 111
  Transformationsregeln 112
Strg-Taste 28

Student-Verteilung 222
Symbolleiste 31

t-Test *siehe* Test-Mittelwert
t-Verteilung *siehe* Student-Verteilung
Tabelle
  Checkliste 22
Tastenkombination 28
Test 227
  Ablauf 230
  Anteil 234–236
    EXCEL 241
    SPSS 242
  Arbeitsweise 229
  einseitig 231
  Fehler 228
  Hypothesen-Formulierung 230
  Mittelwert 237, 239, 240
    EXCEL 241
    SPSS 242
  nichtparametrisch 230
  parametrisch 230
  Signifikanzniveau 228
  Unabhängigkeit 242
  zweiseitig 231
Theorem von Bernoulli 207
Tortendiagramm *siehe* Kreisdiagramm
Transformationsregeln 112
Transformieren von Variablen 53, 54

Überlebensfunktion 202
Umkodieren von Variablen 50, 53
Umschalt-Taste 28
Unabhängigkeit
  diskrete Zufallsvariable 166
  stetige Zufallsvariable 195
unimodal 101
univariate Statistik 9
Untersuchungsumfang 62
Urliste 62

Variable
  Transformieren in SPSS 53, 54
  Umkodieren in SPSS 50, 53
Variablenansicht 41
Variablenlabel 44
Varianz 109
  diskrete Zufallsvariable 175
  EXCEL 118
  korrigierte 117
  SPSS 118
  stetige Zufallsvariable 196
Variationskoeffizient 109
verhältnisskaliert 12
Versuchsausgang 161
Verteilung
  Alternativ- 177
  bedingte 123
  Binomial- 180
  Diskrete Gleich- 178
  Hypergeometrische 182
  Normal- 199
  Poisson- 184
  Rand- 123
  Standardnormal- 199
  Stetige Gleich- 198
Verteilungsfunktion 85
  diskrete Zufallsvariable 174
  empirische
    Bezeichnungen 85
    Eigenschaften 85
  EXCEL 91
  Grafik 86
  Rechenregeln 91
  rekursive Darstellung 88, 91
  SPSS 91
  stetige Zufallsvariable 195
Viewer 49

Wachstumsfaktor 102
Wachstumsrate 102
Wahrscheinlichkeit
  a-posteriori 169
  a-priori 169
  Abzählregel 162
  Additionssatz 165
  Axiome von Kolmogorov 164
  bedingte 165, 166
  Gegen- 165
  Laplace- 162
  Multiplikationsregel 166
  Produktregel 166
  Rechenregeln 165
  totale 167
  unabhängiger Ereignisse 166
Wahrscheinlichkeitshistogramm 191
Wertebereich 8, 161

Wertelabels 45
Wölbung 113, 115, 116
  EXCEL 117, 118
  SPSS 118

Zellbezug
  absolut 32, 33
  relativ 32, 33
Zentraler Grenzwertsatz 207
zentrales Moment 114
Zentriwinkel 71
Zerlegung 162
Ziehen mit Zurücklegen 179
Ziehen ohne Zurücklegen 181
Zufallsexperiment 160
Zufallsstichprobe 8
Zufallsvariable 160
  diskrete 173
    Dichte 173
    Erwartungswert 175
    Unabhängigkeit 166
    Varianz 175
    Verteilungsfunktion 174
  stetige 191
    Dichte 194
    Erwartungswert 196
    Unabhängigkeit 195
    Varianz 196
    Verteilungsfunktion 195
Zusammenhang
  EXCEL 136
  gegensinnig 128, 132, 135
  gleichsinnig 128, 132, 135
  linear 132
  metrische Merkmale 130
  nominale Merkmale 125
  ordinale Merkmale 128
  SPSS 141
  Streudiagramm 134, 135
  Test *siehe* Test-Unabhängigkeit
  Tipps 142
Zuverlässigkeitsfunktion 88, 91, 202
zweidimensionale Häufigkeitsverteilung 121
zweidimensionale Merkmale 121
  EXCEL 136
  SPSS 137, 141
zweiseitiger Test 231
  Anteil 234
  Mittelwert 237