Lehr- und Handbücher der Wirtschafts- und Sozialwissenschaften

Lieferbare Titel:

Bamberg, Baur, Krapp: Statistik
Brümmerhoff: Finanzwissenschaft
Bühner: Betriebswirtschaftliche Organisationslehre
Domschke: Logistik: Rundreisen und Touren
Domschke, Drexl: Logistik: Standorte
Domschke: Logistik: Transport
May: Ökonomie für Pädagogen
Oechsler: Personal und Arbeit – Einführung in die Personalwirtschaft
Peters: Betriebswirtschaftslehre – Einführung

Statistik

von
Prof. em. Dr. Dr. h.c. Günter Bamberg
PD Dr. Franz Baur
Prof. Dr. Michael Krapp

17., überarbeitete Auflage

Oldenbourg Verlag München

Bibliografische Information der Deutschen Nationalbibliothek

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

© 2012 Oldenbourg Wissenschaftsverlag GmbH
Rosenheimer Straße 145, D-81671 München
Telefon: (089) 45051-0
www.oldenbourg-verlag.de

Das Werk einschließlich aller Abbildungen ist urheberrechtlich geschützt. Jede Verwertung außerhalb der Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Bearbeitung in elektronischen Systemen.

Lektorat: Thomas Ammon
Herstellung: Constanze Müller
Einbandgestaltung: hauser lacour
Gesamtherstellung: Beltz Bad Langensalza GmbH, Bad Langensalza

Dieses Papier ist alterungsbeständig nach DIN/ISO 9706.

ISBN 978-3-486-71651-1
eISBN 978-3-486-71747-1

Inhaltsverzeichnis

Vorwort . . . IX
Liste wichtiger Symbole . . . XI

1. Einführung . . . 1
   1.1 Zweierlei Bedeutung des Begriffs Statistik . . . 1
   1.2 Auswahl des Stoffes . . . 1

Teil I: Deskriptive Statistik **3**

2. Grundbegriffe der Datenerhebung . . . 5
   2.1 Merkmal, Merkmalsausprägung, Merkmalsträger, statistische Masse . . . 5
   2.2 Verschiedene Typen statistischer Massen . . . 5
   2.3 Verschiedene Typen von Merkmalen, Skalierung, Klassierung . . . 6
   2.4 Verschiedene Typen statistischer Erhebungen . . . 8
   2.5 Kritische Zusammenfassung, Literaturhinweise . . . 10

3. Auswertungsmethoden für eindimensionales Datenmaterial . . . 11
   3.1 Häufigkeitsverteilungen . . . 11
      3.1.1 Absolute und relative Häufigkeitsverteilung . . . 11
      3.1.2 Histogramm, Hinweise zur Klassenbildung . . . 13
      3.1.3 Kumulierte Häufigkeiten, empirische Verteilungsfunktion . . . 14
   3.2 Lageparameter . . . 15
      3.2.1 Modalwert, Median, arithmetisches und geometrisches Mittel . . . 16
      3.2.2 Eigenschaften der Lageparameter und Vergleich . . . 17
   3.3 Streuungsparameter . . . 20
      3.3.1 Spannweite, durchschnittliche und mittlere quadratische Abweichung, Standardabweichung, Variationskoeffizient . . . 20
      3.3.2 Eigenschaften der mittleren quadratischen Abweichung und der Standardabweichung . . . 21
   3.4 Konzentrationsmaße . . . 22
      3.4.1 Lorenzkurve . . . 23
      3.4.2 Gini-Koeffizient . . . 25
      3.4.3 Weitere Konzentrationsmaße . . . 27
   3.5 Kritische Zusammenfassung, Literaturhinweise . . . 28

4. Auswertungsmethoden für mehrdimensionales Datenmaterial . . . 29
   4.1 Kontingenztabelle, Streuungsdiagramm . . . 29
   4.2 Korrelationsrechnung . . . 33
      4.2.1 Bravais-Pearson-Korrelationskoeffizient . . . 33
      4.2.2 Rangkorrelationskoeffizient von Spearman . . . 35
      4.2.3 Kontingenzkoeffizient . . . 36
   4.3 Regressionsrechnung . . . 39
      4.3.1 Lineare Regression . . . 39
      4.3.2 Nichtlineare Regression . . . 43
   4.4 Berücksichtigung von mehr als zwei Merkmalen . . . 45
   4.5 Kritische Zusammenfassung, Literaturhinweise . . . 46

5. Verhältniszahlen und Indexzahlen . . . 49
5.1 Klassifikation der Verhältniszahlen . . . 49
5.2 Allgemeine Bemerkungen über Preisindizes . . . 50
5.3 Spezielle Preisindizes . . . 51
5.3.1 Die Preisindizes von Laspeyres und Paasche . . . 51
5.3.2 Weitere Preisindizes . . . 54
5.4 Mengenindizes . . . 54
5.5 Umbasierung, Verkettung und Verknüpfung von Indexwerten . . . 55
5.5.1 Umbasierung . . . 56
5.5.2 Verkettung . . . 56
5.5.3 Verknüpfung . . . 56
5.6 Kritische Zusammenfassung, Literaturhinweise . . . 57

6. Zeitreihenzerlegung und Saisonbereinigung . . . 59
6.1 Das additive Zeitreihenmodell . . . 59
6.2 Zur Ermittlung der Zeitreihenkomponenten . . . 60
6.3 Gleitende Durchschnitte . . . 61
6.4 Saisonbereinigung bei konstanter Saisonfigur . . . 63
6.5 Saisonbereinigung bei variabler Saisonfigur . . . 66
6.6 Kritische Zusammenfassung, Literaturhinweise . . . 67

Teil II: Wahrscheinlichkeitsrechnung 69

7. Zufallsvorgänge, Ereignisse und Wahrscheinlichkeiten . . . 71
7.1 Zufallsvorgänge . . . 71
7.2 Ereignisse und ihre Darstellung . . . 72
7.3 Wahrscheinlichkeit von Ereignissen . . . 74
7.3.1 Die Axiome der Wahrscheinlichkeitsrechnung . . . 74
7.3.2 Der klassische Wahrscheinlichkeitsbegriff . . . 75
7.3.3 Häufigkeitsinterpretation des Wahrscheinlichkeitsbegriffs . . . 77
7.3.4 Regeln für Wahrscheinlichkeiten . . . 78
7.3.5 Bedingte Wahrscheinlichkeiten . . . 79
7.3.6 Unabhängigkeit von Ereignissen . . . 82
7.4 Kritische Zusammenfassung, Literaturhinweise . . . 83

8. Zufallsvariablen und Verteilungen . . . 87
8.1 Verschiedene Typen von Zufallsvariablen . . . 87
8.1.1 Eindimensionale Zufallsvariablen . . . 87
8.1.2 Mehrdimensionale Zufallsvariablen . . . 88
8.1.3 Unabhängigkeit von Zufallsvariablen . . . 89
8.2 Die Verteilungsfunktion einer eindimensionalen Zufallsvariablen . . . 89
8.3 Eindimensionale diskrete Zufallsvariablen . . . 91
8.4 Wichtige diskrete Verteilungen . . . 92
8.4.1 Binomialverteilung . . . 92
8.4.2 Hypergeometrische Verteilung . . . 94
8.4.3 Poisson-Verteilung . . . 96
8.5 Eindimensionale stetige Zufallsvariablen . . . 97

8.6 Wichtige stetige Verteilungen . . . 100
8.6.1 Gleichverteilung . . . 100
8.6.2 Exponentialverteilung . . . 100
8.6.3 Normalverteilung . . . 102
8.7 Verteilung mehrdimensionaler Zufallsvariablen . . . 105
8.7.1 Die gemeinsame Verteilungsfunktion . . . 105
8.7.2 Mehrdimensionale diskrete bzw. stetige Zufallsvariablen . . . 106
8.7.3 Randverteilung und bedingte Verteilung . . . 108
8.7.4 Unabhängigkeit von Zufallsvariablen . . . 110

9. Verteilungsparameter . . . 111
9.1 Lageparameter: Modus, Median, Erwartungswert . . . 111
9.2 Streuungsparameter: Varianz und Standardabweichung . . . 114
9.3 Erwartungswerte und Varianzen wichtiger Verteilungen . . . 115
9.4 Weitere Aussagen über Erwartungswert und Varianz . . . 116
9.5 Kovarianz und Korrelation zweier Zufallsvariablen . . . 117
9.6 Kritische Zusammenfassung, Literaturhinweise . . . 119

10. Gesetz der großen Zahlen und zentraler Grenzwertsatz . . . 121
10.1 Gesetz der großen Zahlen . . . 121
10.2 Zentraler Grenzwertsatz . . . 122

Teil III: Induktive Statistik 125

11. Grundlagen der induktiven Statistik . . . 127
11.1 Grundgesamtheit und uneingeschränkte Zufallsauswahl, Verteilung der Grundgesamtheit, Stichprobenvariable und einfache Stichprobe . . . 127
11.2 Stichprobenraum, Stichprobenfunktion, Testverteilungen . . . 129
11.2.1 Bezeichnungen . . . 129
11.2.2 Wichtige Stichprobenfunktionen . . . 130
11.2.3 Testverteilungen . . . 132
11.2.4 Verteilungen von Stichprobenfunktionen . . . 135

12. Punkt-Schätzung . . . 137
12.1 Erwartungstreue und wirksamste Schätzfunktionen . . . 137
12.2 Konsistente Schätzfunktionen . . . 140
12.3 Das Prinzip der kleinsten Quadrate . . . 141
12.4 Das Maximum-Likelihood-Prinzip . . . 143
12.5 Bayes-Schätzfunktionen . . . 145
12.6 Kritische Zusammenfassung, Literaturhinweise . . . 147

13. Intervall-Schätzung . . . 149
13.1 Symmetrische Konfidenzintervalle für den Erwartungswert μ . . . 150
13.1.1 Normalverteilte Grundgesamtheit mit bekannter Varianz . . . 150
13.1.2 Normalverteilte Grundgesamtheit mit unbekannter Varianz . . . 152
13.1.3 Beliebig verteilte, insbesondere dichotome Grundgesamtheit . . . 153
13.2 Symmetrische Konfidenzintervalle für die Varianz σ^2 bei normalverteilter Grundgesamtheit . . . 155
13.3 Kritische Zusammenfassung, Literaturhinweise . . . 157

14. Signifikanztests . . . 159
14.1 Einführungsbeispiel: Einstichproben-Gaußtest . . . 159
14.2 Aufbau und Interpretation von Signifikanztests . . . 164
14.3 Binomialtest . . . 168
14.4 Klassifikation der Signifikanztests . . . 170
14.5 Einstichproben-t-Test, approximativer Gaußtest, Differenzentests . . . 174
14.6 Chi-Quadrat-Test für die Varianz . . . 177
14.7 Zweistichproben-Tests . . . 178
14.7.1 Vergleich zweier Erwartungswerte . . . 178
14.7.2 Vergleich zweier Varianzen . . . 181
14.8 Einfache Varianzanalyse . . . 182
14.9 Chi-Quadrat-Anpassungstest . . . 184
14.10 Kontingenztest . . . 188
14.11 Korrelationstest . . . 190
14.12 Gütefunktion . . . 192
14.13 Kritische Zusammenfassung, Literaturhinweise . . . 195

Teil IV: Überblick über weitere wichtige Teilgebiete der Statistik 199

15. Zeitreihenanalyse und Prognoserechnung . . . 201
15.1 Exponentielles Glätten . . . 201
15.2 Parametrische Zeitreihenmodelle, Box-Jenkins-Modelle . . . 203
15.3 Idee der Spektralanalyse . . . 206

16. Ökonometrie und multiple Regressionsrechnung . . . 209
16.1 Ökonometrische Eingleichungsmodelle . . . 209
16.2 Ökonometrische Mehrgleichungsmodelle . . . 212

17. Multivariate Verfahren . . . 213
17.1 Einteilung der multivariaten Verfahren . . . 213
17.2 Standardisierte Datenmatrix und Korrelationsmatrix . . . 214
17.3 Faktorenanalytisches Modell und Faktorextraktion . . . 215

18. Stichprobenplanung . . . 221
18.1 Arten von Stichprobenplänen . . . 221
18.2 Geschichtete Stichproben . . . 224

19. Statistische Entscheidungstheorie . . . 229
19.1 Grundlegende Daten . . . 229
19.2 Bayes-Verfahren . . . 232

20. Statistik-Software . . . 235
20.1 Qualitätskriterien für Statistik-Pakete . . . 236
20.2 Angebotsüberblick und weiterführende Informationsquellen . . . 241

Lösungen der Aufgaben . . . 245
Tabellenanhang . . . 265
Abkürzungsverzeichnis . . . 291
Literaturverzeichnis . . . 293
Personenverzeichnis . . . 307
Sachverzeichnis . . . 313

Vorwort

Vorwort zur 17. Auflage

Zahlreiche positive Rückmeldungen zur 16. Auflage haben uns darin bestärkt, den eingeschlagenen Weg einer Weiterentwicklung ohne Strukturbruch fortzusetzen. Dementsprechend wurde die Grundkonzeption des Buches – trotz etlicher Veränderungen gegenüber der letzten Auflage – unverändert gelassen. Lesern, welche mit einer der älteren Auflagen vertraut sind, wird vermutlich primär das durch den Einsatz einer Schmuckfarbe moderner wirkende Erscheinungsbild auffallen. Daneben ließen sich noch die obligatorische Aktualisierung des Literaturverzeichnisses, viele Detailanpassungen sowie einige wenige Fehlerkorrekturen aufzählen; darauf wollen wir an dieser Stelle aber verzichten.

Gemäß einem empirisch gut untermauerten Gesetz enthält jedes Buch Druckfehler. Unter http://www.wiwi.uni-augsburg.de/bwl/krapp, Rubrik „Druckfehler“, werden wir die Liste der nach und nach entdeckten Fehler im Internet veröffentlichen. Wir hoffen auf eine kurze Liste. Dozenten können darüber hinaus einen Foliensatz sowie eine Aufgabensammlung, welche 80 Klausuraufgaben inkl. Musterlösungen umfasst, über die Homepage des Oldenbourg-Verlages (http://www.oldenbourg-wissenschaftsverlag.de) beziehen. Weitere Unterlagen, wie z. B. nicht im Foliensatz enthaltene Abbildungen und Tabellen, stellen wir für den Einsatz in Lehrveranstaltungen sehr gerne zur Verfügung. Bitte kontaktieren Sie uns in diesem Fall per E-Mail (michael.krapp@wiwi.uni-augsburg.de).

Hinweise auf Fehler in vorangegangenen Auflagen sowie wertvolle Anregungen erhielten wir von zahlreichen Studenten und Kollegen, insbesondere von den Herren Prof. Dr. S. Huschens, Prof. Dr. U. Kockelkorn, Prof. Dr. H. Locarek-Junge, Prof. Dr. H.-J. Mittag, Prof. Dr. O. Opitz, Prof. Dr. F. Schmid, Prof. Dr. K. Schürger, Prof. Dr. J.-U. Stiehr und Prof. Dr. D. Stoyan. Bei der Erstellung der vorliegenden Auflage haben uns Frau stud. rer. pol. M. Haberbosch, Herr Dipl.-Kfm. J. Kraus und Herr Dipl.-Kfm. J. Nebel tatkräftig unterstützt. In bewährter Weise professionell und zugleich zuvorkommend wurden wir von unserem Lektor, Herrn Th. Ammon, verlagsseitig betreut. Ihnen allen gebührt unser herzlicher Dank.

Augsburg, im April 2012

G. Bamberg
F. Baur
M. Krapp

Vorwort zur ersten Auflage

Statistische Methoden haben in den letzten Jahrzehnten ständig an Bedeutung gewonnen. Eine Reihe von handelsüblichen elektronischen Taschenrechnern ermöglicht bereits die Anwendung einfacherer statistischer Verfahren per Tastendruck. Der erleichterte Zugang zu größeren EDV-Anlagen hat dazu geführt, dass der mit komplexeren statistischen Verfahren verbundene Rechenaufwand nicht mehr prohibitiv wirkt. Von besonderer Wichtigkeit sind deshalb eine sichere Kenntnis der Prämissen, auf denen die statistischen Verfahren beruhen, sowie die Fähigkeit, sich mit dem Ergebnis und der Interpretation einer statistischen Analyse kritisch auseinanderzusetzen. Im vorliegenden Lehrbuch werden trotz der Betonung dieses Aspekts

wichtige statistische Verfahren in „Rezeptkästchen“ zusammengefasst. Grundlegende Begriffe und Verfahren werden durch Beispiele erläutert und können anhand von Aufgaben, deren Lösungen separat am Ende des Buches zu finden sind, zur Selbstkontrolle „erprobt“ werden.

Die Teile I bis III führen in die deskriptive Statistik, die Wahrscheinlichkeitsrechnung und die induktive Statistik ein. Sie stellen – bis auf gewisse Akzentverlagerungen – das Normalprogramm dar, das sich mittlerweile an den meisten Hochschulen für die wirtschafts- und sozialwissenschaftlichen Studiengänge eingebürgert hat. Zur Lektüre sind die Vorkenntnisse in mathematischer Propädeutik ausreichend, die in allen wirtschafts- und sozialwissenschaftlichen Fakultäten im Grundstudium vermittelt werden. Teil IV enthält einen Ausblick auf einige weitere wichtige Teilgebiete der Statistik.

Für die kritische Durchsicht des Manuskripts und für wertvolle Anregungen möchten wir unseren Freunden und Kollegen Dipl.-Math. Dr. Volker Firchau, Dipl.-oec. Alois Herbein, Dipl.-Ing. Dr. Henning Paul, Prof. Dr. U. K. Schittko, Dipl.-Math. Dr. Jürgen Sommer herzlich danken. Darüber hinaus danken wir Frau B. Emmrich und Frau U. Rook, die mit viel Geduld und Sorgfalt das endgültige Manuskript und die verschiedenen Vorlagen geschrieben haben. Schließlich gilt unser Dank Herrn M. Weigert und dem Oldenburg Verlag für die verständnisvolle Zusammenarbeit und die rasche Drucklegung.

G. Bamberg
F. Baur

Liste wichtiger Symbole

Symbole für Teil I

		def. auf S.
X bzw. Y	Merkmale	11
x_i bzw. y_i	Beobachtungswerte von X bzw. Y	11
$a_1, \ldots, a_k$	realisierte Ausprägungen des Merkmals X	11
$h(a_j)$	absolute Häufigkeit der Ausprägung a_j	11
$f(a_j)$	relative Häufigkeit von a_j	11
$H(x)$	kumulierte absolute Häufigkeitsverteilung	14
$F(x)$	kumulierte relative Häufigkeitsverteilung	15
$\bar{x}$	arithmetisches Mittel	17
s^2	mittlere quadratische Abweichung	20
s	Standardabweichung	20
r	Bravais-Pearson-Korrelationskoeffizient	34

Symbole für Teil II (falls abweichend von Teil I)

X bzw. Y	Zufallsvariablen (ggf. auch zugehörige Merkmale)	87
Ω	Ergebnismenge	71
P	Wahrscheinlichkeitsmaß	74
$P(A\|B)$	bedingte Wahrscheinlichkeit	79
$f(x)$	Wahrscheinlichkeitsfunktion bzw. -dichte	91 bzw. 98
$F(x)$	Verteilungsfunktion	89
$B(n;p)$	Binomialverteilung	93
$N(\mu;\sigma)$	Normalverteilung	102
$N(0;1)$	Standardnormalverteilung	102
$\Phi(x)$	Verteilungsfunktion der Standardnormalverteilung	103
$\mathrm{E}(X)$ bzw. μ	Erwartungswert der Zufallsvariablen X	111
$\mathrm{Var}(X)$ bzw. σ^2	Varianz der Zufallsvariablen X	114

Symbole für Teil III

G	Grundgesamtheit	127
X_i	i-te Stichprobenvariable	128
$\bar{X}$	Stichprobenmittel	131
S^2	Stichprobenvarianz	131
S	Stichproben-Standardabweichung	131
s	Realisation der Stichproben-Standardabweichung	152
$(x_1, \ldots, x_n)$	Stichprobenergebnis	129
$f(x_1, \ldots, x_n\|\vartheta)$	Likelihoodfunktion	130
V_u bzw. V_o	untere bzw. obere Grenze eines Konfidenzintervalls	149
V bzw. v	Testfunktion bzw. ihre Realisation	165
H_0 bzw. H_1	Nullhypothese bzw. Alternativhypothese	164
α	Signifikanzniveau, Irrtumswahrscheinlichkeit	165
B	Verwerfungsbereich eines Tests	165
$g(\vartheta)$	Gütefunktion	192

1. Kapitel:

Einführung

1.1 Zweierlei Bedeutung des Begriffs Statistik

Vielleicht denken Sie beim Stichwort „Statistik“ zunächst an umfangreiche Zahlenkolonnen, an das Statistische Jahrbuch für die Bundesrepublik Deutschland oder an Aussprüche wie:

„Nichts lügt so sehr wie die Statistik“,
„Traue keiner Statistik, die du nicht selbst gefälscht hast“.

Der Begriff „Statistik“ ist mehrdeutig. Zum einen beschreibt er die tabellarische oder grafische Darstellung eines konkret vorliegenden Datenmaterials, wie etwa die Statistik der Verkehrsunfälle für einen bestimmten Autobahnabschnitt, die Statistik der Einfuhren eines Landes, aufgegliedert nach bestimmten Warengruppen, Herkunftsländern usw. Zum anderen beschreibt der Begriff „Statistik“ die Gesamtheit der Methoden, die für die Gewinnung und Verarbeitung empirischer Informationen relevant sind.

Die statistische Methodenlehre, im Folgenden meist kurzerhand als Statistik bezeichnet, bildet den Hauptgegenstand des vorliegenden Lehrbuchs. Bloße Feststellungen über in Zahlen messbare Fakten, z. B. „Am 27.05.1970 betrug die Wohnbevölkerung Bayerns 10 479 400 Personen“, sind relativ unproblematisch und natürlich nicht Gegenstand der statistischen Methodenlehre. Obige Aussprüche haben dagegen einen stärkeren Bezug zur statistischen Methodenlehre; denn durch sie wird immerhin angedeutet, wie problematisch es sein kann, aus Daten Erkenntnisse zu gewinnen. Wir kommen auf die „statistische Lüge“ noch mehrfach zurück.[1]

1.2 Auswahl des Stoffes

Statistische Methoden können im Prinzip überall dort zum Einsatz kommen, wo empirische Daten erhoben und ausgewertet werden, gleichgültig um welchen Wissenschafts- oder Anwendungsbereich es sich handelt. Besonders häufig wird die Statistik allerdings in den Gebieten Betriebswirtschaft, Volkswirtschaft, Psychologie, Soziologie, Pädagogik, Medizin, Biologie, Landwirtschaft, Technik und Meteorologie angewandt. Während die Statistik in diesen Gebieten vorwiegend die Rolle einer nützlichen Hilfswissenschaft spielt, kommt ihr in den neu entwickelten „Metrien“, wie z. B. der Biometrie, der Demometrie, der Ökonometrie, der Psychometrie und der Soziometrie, eine tragende Rolle zu. Trotz dieses weit gefächerten Anwendungsspektrums wollen wir uns bei der Auswahl des Stoffs, der Beispiele und der Aufgaben auf den wirtschafts- und sozialwissenschaftlichen Bereich beschränken.

In Teil I wird die **deskriptive Statistik** behandelt. Hier kommen die von der Wirtschaftspraxis bzw. der Wirtschaftspresse her geläufigen Begriffe wie Indexzahl, Saisonbereinigung, Trendermittlung, Korrelation usw. zur Sprache.

In Teil II wird die **Wahrscheinlichkeitsrechnung** insoweit dargestellt, als sie für das Verständnis der nachfolgenden statistischen Verfahren erforderlich ist.

---

[1] Wegen informativer Beispiele sei insbesondere auf Krämer (2011) und die dort zitierte Literatur verwiesen.

Teil III beschäftigt sich mit denjenigen Anwendungsfällen, bei denen eine vollständige Datenerhebung undurchführbar oder unwirtschaftlich ist. So ist es beispielsweise sinnlos, die gesamte Produktion eines Massenartikels auf die Lebensdauer der einzelnen Artikel hin zu untersuchen; es bliebe kein verkäufliches Stück übrig. Bei Marktuntersuchungen können ebenfalls oft nicht alle potenziellen Marktteilnehmer befragt werden, da eine derartige Totalerhebung zu Zeit raubend und teuer wäre; dasselbe gilt natürlich auch für Probleme der Meinungsforschung. An Stelle einer Totalerhebung führt man eine Teilerhebung durch, indem einige Untersuchungsobjekte auf Grund von bestimmten Zufallsgesetzen ausgewählt werden; man zieht also eine Stichprobe. Der Rückschluss von der Stichprobe auf die Grundgesamtheit ist die Aufgabe der **induktiven Statistik**. Die klassischen Teilbereiche der induktiven Statistik, nämlich die Punkt-Schätzung, die Intervall-Schätzung und das Testen von Hypothesen, bilden den Inhalt von Teil III.

Die Einbeziehung von Teil IV, der einen knappen **Überblick über weitere Teilgebiete der Statistik** liefert, trägt der Tatsache Rechnung, dass synchron mit der Expansion des Anwendungsspektrums eine beträchtliche Erweiterung des Arsenals statistischer Methoden stattgefunden hat. Begünstigt durch die wachsende Leistungsfähigkeit der EDV-Anlagen und das steigende Informationsbedürfnis, das mit sparsamstem Aufwand befriedigt werden sollte, dürfte dieser Trend auch für die Zukunft gelten. Teilgebiete wie etwa die Prognoserechnung, die multivariaten Verfahren, die statistische Entscheidungstheorie, die statistische Software, um nur einige zu nennen, lassen sich trotz ihrer Anwendungsträchtigkeit leider nicht gleichzeitig in einem statistischen Grundkurs unterbringen. Die Verlagerung in Hauptstudiumskurse führt erfahrungsgemäß dazu, dass nur Spezialisten unter den Studenten mit derartigen Gebieten in Berührung kommen. Die Tendenz, im späteren Berufsleben nur das anzuwenden, was vom Studium her vertraut oder zumindest dem Namen nach geläufig ist, spricht jedoch dafür, hier einige dieser wichtigen Teilgebiete wenigstens in ihren Grundzügen vorzustellen.

Teil I:
Deskriptive Statistik

Bei statistischen Erhebungen fallen im Regelfall Tausende von Einzeldaten an. Die Methoden der deskriptiven Statistik (Synonyma sind: beschreibende Statistik, explorative Statistik) zielen darauf ab, diese unüberschaubare Datenmenge durch möglichst wenige – jedoch noch aussagekräftige – Zahlen zu charakterisieren. Im Extremfall wird lediglich eine einzige Zahl zur Charakterisierung der gesamten Datenmenge benutzt. So soll ein Lageparameter die Größenordnung aller Einzeldaten charakterisieren, ein Preisindex die Veränderung aller Preise eines umfangreichen Warenkorbs, ein Korrelationskoeffizient die Abhängigkeit zwischen den Ausprägungen zweier verschiedener Merkmale usw.

Selbstverständlich kann eine derart weit gehende Aufbereitung bzw. Verdichtung nicht nur auf eine einzige Art und Weise erfolgen. Aus diesem Grunde ist eine fundierte Kenntnis wesentlicher Voraussetzungen und Eigenschaften verschiedener Auswertungsmethoden unerlässlich für die begründete Auswahl einer adäquaten Methode und die sachgerechte Interpretation des nummerischen Ergebnisses.

Lernziele:

- Kennenlernen zentraler Grundbegriffe der Statistik.
- Beherrschung wichtiger Lage- und Streuungskenngrößen sowie deren Interpretation.
- Fähigkeit zur Analyse mehrdimensionalen Datenmaterials mithilfe von Zusammenhangsmaßen und deskriptiver Regressionsanalyse.
- Erlangung elementaren Wissens hinsichtlich Bestimmung, Interpretation sowie Transformation von Preisindizes und Verhältniszahlen.
- Erwerb von Kenntnissen über den Umgang mit additiven Zeitreihenmodellen, insbesondere im Hinblick auf Saisonbereinigung.

2. Kapitel:

Grundbegriffe der Datenerhebung

2.1 Merkmal, Merkmalsausprägung, Merkmalsträger, statistische Masse

Sicherlich haben Sie bereits einmal einen Fragebogen ausfüllen und dabei etwa Angaben über Alter, Geschlecht, Familienstand, Wohnort, Stellung im Beruf, Einkommenshöhe, Art des Krankenversicherungsschutzes, Zeitaufwand für den Weg zur Arbeitsstätte usw. machen müssen. Ihren begreiflichen Unmut darüber wollen wir hier nicht weiter erörtern, sondern eine Reihe von Fachausdrücken erläutern. Im statistischen Jargon fungierten Sie als **Merkmalsträger** oder **statistische Einheit**, an der die **Merkmale** Alter, Geschlecht usw. erhoben wurden. Als **Merkmalsausprägungen** bezeichnet man sowohl die verschiedenen Zahlen, die ein **quantitatives** Merkmal (wie etwa das Alter, die Einkommenshöhe) annehmen kann, als auch die verschiedenen Kategorien, die bei einem **qualitativen** Merkmal (wie etwa dem Geschlecht, der Religionszugehörigkeit) registriert werden können. Die für eine statistische Untersuchung relevanten Merkmalsträger fasst man zur **Grundgesamtheit** zusammen. Wird bei einer statistischen Erhebung die Grundgesamtheit vollständig erfasst, so spricht man von einer **Vollerhebung** oder **Totalerhebung**, anderenfalls von einer **Teilerhebung** oder **Stichprobenerhebung**.

Beispiel 2.1:

- Während es sich bei der Volkszählung von 1987 (Stichtag 25.05.1987) um eine Totalerhebung der Wohnbevölkerung der damaligen Bundesrepublik Deutschland handelte, wurde im Rahmen des so genannten „Zensus 2011“ (Stichtag 09.05.2011) – neben einer Vollerhebung für den Bereich Gebäude und Wohnungen – eine Stichprobe gezogen, die ca. 9,6 % der Bevölkerung umfasste.
- Wurde Ihnen der eingangs erwähnte Fragebogen im Rahmen der als Mikrozensus bezeichneten „Repräsentativstatistik der Bevölkerung und des Erwerbslebens“ von einem Interviewer Ihres Statistischen Landesamtes präsentiert, so gehörten Sie zu einem der rund 230 000 Haushalte, die nach einem bestimmten Stichprobenverfahren ausgewählt wurden.
- Bei der Handwerkszählung 1968 wurde eine Stichprobe gezogen, die rund 620 000 der in die Handwerksrolle eingetragenen Unternehmungen umfasste. Als Merkmale wurden u. a. erhoben: Rechtsform, Vorhandensein von Zweigniederlassungen, Anzahl Beschäftigte, Löhne, Gehälter, Sozialaufwendungen, Materialverbrauch, Exporterlöse. ▪

Für die deskriptive Statistik ist es – im Gegensatz zur induktiven Statistik – unerheblich, ob das zu verarbeitende Datenmaterial von einer Vollerhebung oder einer Teilerhebung herrührt. Wenn im Folgenden von einer **statistischen Masse** die Rede ist, so ist darunter entweder die Grundgesamtheit oder eine ihrer Teilgesamtheiten zu verstehen. Falls nur eine Teilgesamtheit betrachtet wird, sind im Rahmen der deskriptiven Statistik nur Aussagen über diese Teilgesamtheit möglich.

2.2 Verschiedene Typen statistischer Massen

Es gibt statistische Massen, die man sinnvollerweise auf einen bestimmten **Zeitpunkt** bezieht. Solche statistischen Massen werden als **Bestandsmassen** bezeichnet; Beispiele sind der Bestand eines Lagers oder die Bevölkerung eines Landes. Andere statistische Massen bezieht

man sinnvollerweise auf eine **Zeitperiode**. Diese statistischen Massen werden als **Bewegungsmassen** oder auch als **Ereignismassen** bezeichnet; Beispiele sind die Verkehrsunfälle eines Jahres oder die Lagerabgänge einer Woche. Bei einer Bewegungsmasse haben die Merkmalsträger eine vernachlässigbar kleine „Lebensdauer", bei einer Bestandsmasse dagegen nicht.

Umfangreiche Bestandsmassen werden aus Kostengründen häufiger durch Fortschreibung als durch eine neue Erhebung aktualisiert. Bei der **Fortschreibung** wird die statistische Masse des früheren Stichtages um die **korrespondierende Bewegungsmasse** der zwischen den Stichtagen erfolgten Zugänge vermehrt und um die entsprechende korrespondierende Bewegungsmasse der Abgänge vermindert.

Aufgabe 2.2: Handelt es sich um Bestands- oder Bewegungsmassen?

a) Gärbottiche einer Brauerei
b) Arbeitsunfälle in Bayern mit tödlichem Ausgang
c) Neueinstellungen eines Betriebs
d) Mitarbeiter eines Betriebs
e) Filialen einer Bank ▪

2.3 Verschiedene Typen von Merkmalen, Skalierung, Klassierung

Bereits in Abschnitt 2.1 wurde erwähnt, dass diejenigen Merkmale als **quantitativ** bezeichnet werden, deren Merkmalsausprägungen Zahlen sind; alle anderen Merkmale werden als **qualitativ** bezeichnet. Bei qualitativen Merkmalen sind die Merkmalsausprägungen primär durch einen verbalen Ausdruck (wie etwa gut, mittel, schlecht; ledig, geschieden usw.) charakterisiert. Ersetzt man diese verbalen Ausdrücke durch Zahlen, d. h. führt man eine **Quantifizierung** durch, so wird ein qualitatives Merkmal formal zu einem quantitativen.

Beispiel 2.3: Pkw-Farbtöne werden durch die Produktionsnummer quantifiziert, etwa grün = 117, blau = 410 usw. Bei einer statistischen Erhebung wird die Konfession z. B. in der Form: keine = 0, röm.-kath. = 1, evangelisch = 2, sonstige = 3 verschlüsselt. Die Rechtsform einer Unternehmung kann z. B. durch Aktiengesellschaft = 1, GmbH = 2, OHG = 3, KG = 4, Einzelunternehmung = 5, sonstige = 6 quantifiziert werden. ▪

Die Quantifizierung verwischt den Unterschied zwischen einem (echt) quantitativen und einem (qualitativen aber) quantifizierten Merkmal jedoch nur äußerlich. Bei der statistischen Verarbeitung des Datenmaterials ist darauf zu achten, ob eine Quantifizierung vorgenommen wurde oder nicht.

Beispiel 2.4: Wenn zwei Personen die gemäß Beispiel 2.3 quantifizierte Konfession 1 (= röm.-kath.) bzw. 3 (= sonstige) besitzen, so ist es sinnlos, von der „mittleren Konfession" 2 (= evangelisch) zu sprechen. Erhalten die beiden Personen dagegen 1 000 € bzw. 3 000 € Weihnachtsgratifikation, so kann man sehr wohl von der „mittleren Weihnachtsgratifikation" in Höhe von 2 000 € sprechen. ▪

Die Sachlogik, die sich hinter den nummerischen Merkmalsausprägungen verbirgt, erfasst man in der Statistik durch den Begriff der **Skalierung**. Im Folgenden werden die drei wichtigsten Skalierungen kurz erläutert.

Eine **Nominalskala** liegt vor, wenn die Ausprägungen des untersuchten Merkmals durch die zugeordneten Zahlen lediglich unterschieden werden sollen. Allen in Beispiel 2.3 aufgeführten Merkmalen liegt eine Nominalskala zu Grunde. Da die zugeordneten Zahlen nur eine reine Bezeichnungsfunktion haben, können sie durch eine injektive Abbildung willkürlich in andere Zahlen transformiert werden.

Eine **Ordinalskala** oder **Rangskala** liegt vor, wenn die Ausprägungen des untersuchten Merkmals nicht nur unterschieden, sondern auch in eine Rangordnung gebracht werden können und diese Rangordnung durch die zugeordneten Zahlen widergespiegelt wird. Ordinale Skalen liegen der Messung vieler sozialwissenschaftlicher Merkmale wie etwa Aggressivität, Autoritätsgläubigkeit, Intelligenz, sozialer Status usw. zu Grunde. Da an die zugeordneten Zahlen lediglich die Anforderung gestellt wird, dass sie die richtige Rangordnung der Merkmalsausprägungen reproduzieren, können sie durch eine ordnungserhaltende Abbildung willkürlich in andere Zahlen transformiert werden.

Eine **Kardinalskala** oder metrische Skala liegt vor, wenn die Ausprägungen des untersuchten Merkmals nicht nur in eine Rangordnung gebracht werden können, sondern zusätzlich noch bestimmt werden kann, in welchem Ausmaß sich je zwei verschiedene Merkmalsausprägungen unterscheiden.[1] Typische Beispiele für derartige Merkmale sind alle monetären Größen sowie die meisten Variablen der Physik und der Chemie. Viele statistische Auswertungsmethoden sind nur bei kardinalskalierten Merkmalen legitim; wir werden an verschiedenen Stellen darauf hinweisen.

Neben der Untergliederung bzgl. der Skalierung wäre noch die Einteilung in **diskrete** und **stetige** Merkmale zu erwähnen. Ein Merkmal heißt **diskret**, wenn seine möglichen Ausprägungen eine diskrete Zahlenmenge bilden. Typische Beispiele sind die nominalskalierten Merkmale und generell alle Merkmale, denen ein Zählvorgang zu Grunde liegt, wie etwa die Anzahl der Arbeitnehmer einer Unternehmung, die Anzahl der Räume einer Wohnung usw. Bei einem **stetigen Merkmal** können dagegen für je zwei Merkmalsausprägungen auch alle Zwischenwerte realisiert werden. Typische Beispiele sind diejenigen Merkmale, denen ein beliebig genauer Messvorgang zu Grunde liegt, wie es etwa bei der Lebensdauer eines Ventils, dem Durchmesser einer Röhre usw. der Fall sein kann.

Aus Zweckmäßigkeitsgründen behandelt man ein diskretes Merkmal, das sehr viele Ausprägungen annehmen kann, häufig als stetiges Merkmal; man spricht dann der Deutlichkeit halber von einem **quasistetigen** Merkmal. Auch der umgekehrte Vorgang, die Behandlung eines stetigen Merkmals als diskretes Merkmal, kann aus Zweckmäßigkeitsgründen geboten sein. Hierfür unterteilt man – wie in Beispiel 2.5 – die Merkmalsachse in Klassen (= Intervalle) und registriert an Stelle der exakten Merkmalsausprägung jeweils nur noch die Klasse, in die sie fällt; man spricht in diesem Fall von **Klassierung**[2] bzw. von **klassierten Daten**, gelegentlich auch von **gruppierten Daten**.

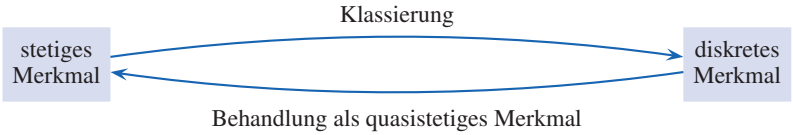

Abb. 2.1: Überführung eines stetigen Merkmals in ein diskretes Merkmal und umgekehrt

[1] Genauer untergliedert man die Kardinalskala in die **Intervallskala** (bei der die Abstände zwischen den Ausprägungen verglichen werden können), die **Verhältnisskala** (bei der zur Intervallskala noch ein natürlicher Nullpunkt hinzukommt) und die **Absolutskala** (bei der zur Verhältnisskala noch eine natürliche Einheit hinzukommt). Beispielsweise ist die Temperatur (wenn sie nicht gerade in Kelvin gemessen wird) intervallskaliert; Preisen, Längen und Gewichten liegt eine Verhältnisskala, Stückzahlen eine Absolutskala zu Grunde.

[2] Diese kann selbstverständlich auch bei quasistetigen oder diskreten Merkmalen vorgenommen werden.

Beispiel 2.5: In den Erhebungsbögen für den Mikrozensus werden (seit 1972) 15 Einkommensgrößenklassen vorgegeben. Dies hat seinen Grund in der Tatsache, dass die Interviewten erstens ihr Einkommen vielfach nicht exakt kennen und zweitens wesentlich leichter dazu zu bewegen sind, sich in vorgegebene Einkommensgrößenklassen einzustufen, als sämtliche Einnahmen centgenau nachzuweisen. Auch im Statistischen Jahrbuch werden Klassenbildungen vorgenommen, wie etwa bei der in Abbildung 2.2 auszugsweise wiedergegebenen Tabelle. ■

Bruttolohn	Lohnsteuerpflichtige mit vermögenswirksamen Leistungen insgesamt		
	Fälle (in 1 000)	Vermögenswirksame Leistungen	Arbeitnehmer-Sparzulage
unter 2 400	146,0	37	11
2 400 bis unter 4 800	378,7	158	47
4 800 bis unter 7 200	424,4	194	58
7 200 bis unter 9 600	366,7	165	50
9 600 bis unter 12 000	468,9	223	67
12 000 bis unter 16 000	1 096,4	555	169
16 000 bis unter 20 000	1 791,7	960	296
20 000 bis unter 25 000	2 774,1	1 518	473
25 000 bis unter 36 000	4 965,7	2 749	821
36 000 bis unter 50 000	3 508,0	1 986	592
50 000 bis unter 75 000	1 090,7	645	171
75 000 bis unter 100 000	102,7	59	14
100 000 und mehr	19,5	11	2
Ingesamt	17 133,5	9 259	2 770

Abb. 2.2: Vermögenswirksame Leistungen und Arbeitnehmer-Sparzulage 1974 (jeweils in Mio. DM) nach dem Dritten Vermögensbildungsgesetz (entnommen dem Statistischen Jahrbuch 1978 für die Bundesrepublik Deutschland, S. 428)

Aufgabe 2.6: Welche der folgenden Merkmale sind stetig bzw. diskret?

a) Beruf
b) Bremsweg eines Pkw
c) Nationalität ■

Aufgabe 2.7: Welche Skalierung liegt den folgenden Merkmalen zu Grunde?

a) Produktionsdauer
b) Schulnote
c) Schwierigkeitsgrad einer Klettertour
d) Kopfumfang einer Person ■

2.4 Verschiedene Typen statistischer Erhebungen

Neben der bereits in Abschnitt 2.1 angesprochenen Einteilung statistischer Erhebungen in Vollerhebungen und Teilerhebungen sind eine Reihe weiterer Unterscheidungsmöglichkeiten zu erwähnen. Nach der Art der Ermittlung des Datenmaterials kann einerseits unterschieden werden zwischen **Befragung** (postalisch/online per Fragebogen oder persönlich durch Interviewer), **Beobachtung** (Verkehrszählung, Messung der Wartezeit von Kunden vor der Ladenkasse

eines Supermarktes) und **Experiment** (Registrierung des Verhaltens von Versuchspersonen in hypothetischen Entscheidungssituationen, Messung der Schlafdauern bei Verabreichung unterschiedlicher Schlafmittel usw.) sowie andererseits zwischen einer

- **Primärerhebung**, bei der das Datenmaterial eigens für die geplante Untersuchung erhoben wird, und einer
- **Sekundärerhebung**, bei der auf vorhandenes (und für andere Zwecke erhobenes) Datenmaterial zurückgegriffen wird, wie es etwa bei der Verwendung von Lohnsteuerkarten für die Untersuchung der Einkommensverteilung der Fall ist.

Für die Durchführung von Teilerhebungen sind neben der willkürlichen Auswahl von Merkmalsträgern, die leider immer noch in der Praxis vorkommt, vor allen Dingen die verschiedenen **Stichprobentechniken**, auf die in Kapitel 18 genauer eingegangen wird, und die Methoden der **geplanten Auswahl** zu nennen. Die beiden bekanntesten Methoden der geplanten Auswahl sind das Abschneideverfahren und die Quotenauswahl.

Beim **Abschneideverfahren**, das die amtliche Statistik häufig benutzt, werden nur die Merkmalsträger mit den größten Beiträgen zu den untersuchten Merkmalen in die Auswahl einbezogen. So verfährt das Statistische Bundesamt beispielsweise bei der kurzfristigen Berichterstattung im produzierenden Gewerbe; nur Betriebe mit einer gewissen Mindestanzahl von Beschäftigten werden in die monatliche Erhebung einbezogen.

Bei der **Quotenauswahl**, die von der amtlichen Statistik selten, von Markt- und Meinungsforschungsinstituten dagegen häufiger angewandt wird, achtet man darauf, dass bestimmte Merkmalsausprägungen in der ausgewählten Teilgesamtheit dieselbe relative Häufigkeit besitzen wie in der Grundgesamtheit. Diese (absoluten oder) relativen Häufigkeiten werden als Quoten bezeichnet und den Interviewern vorgegeben. Im Rahmen der Quoten, die den „repräsentativen Bevölkerungsquerschnitt“ definieren sollen, können sich die Interviewer die zu befragenden Merkmalsträger willkürlich aussuchen.

Beispiel 2.8: Das Institut für Demoskopie Allensbach verwendet folgende Merkmale (in Klammern die prozentualen Quoten der jeweiligen Ausprägungen) als Basis für seine Quotenauswahl aus der Bevölkerung ab 16 Jahre:[3]

- Geschlecht (Männer 49 %; Frauen 51 %)
- Alter (16 bis 29 Jahre 19 %; 30 bis 44 Jahre 24 %; 45 bis 59 Jahre 26 %; 60 Jahre und älter 31 %)
- Berufskreis (Arbeiter 17 %; Angestellte 33 %; Beamte 3 %; Selbständige und freiberuflich Tätige 7 %; Nichterwerbspersonen 40 %)
- Wohnortgröße (unter 5 000 Einwohner 16 %; 5 000 bis unter 20 000 Einwohner 26 %; 20 000 bis unter 100 000 Einwohner 28 %; 100 000 und mehr Einwohner 30 %)
- Region (Schleswig-Holstein, Hamburg, Niedersachsen, Bremen 16 %; Nordrhein-Westfalen 21 %; Hessen, Rheinland-Pfalz, Saarland 13 %; Baden-Württemberg 13 %; Bayern 15 %; Berlin 4 %; Brandenburg, Mecklenburg-Vorpommern, Sachsen-Anhalt 9 %; Sachsen, Thüringen 9 %)

Wenn demnach 1 000 Personen interviewt werden sollen, müssen sich darunter 490 Männer und 510 Frauen, ferner 190 Personen im Alter von 16 bis 29 Jahren, 330 Angestellte usw. befinden. ■

[3] Vgl. Institut für Demoskopie Allensbach (2010, S. 68 f.). Dort werden zur Komplettierung des repräsentativen Bevölkerungsquerschnitts weitere Merkmale wie Familienstand und Haushaltsgröße berücksichtigt. Die Quoten des Instituts für Demoskopie Allensbach werden auf Basis der amtlichen Statistik (in diesem Falle: Mikrozensus 2008) berechnet.

2.5 Kritische Zusammenfassung, Literaturhinweise

In den folgenden Kapiteln wollen wir das vorliegende Datenmaterial nicht weiter problematisieren, sondern lediglich die darauf anzuwendenden Auswertungsmethoden. Insofern sind an dieser Stelle einige kritische Anmerkungen zur Entstehung des Datenmaterials angebracht.

- Die **Festlegung der Untersuchungsmerkmale** stellt bereits eine entscheidende Weichenstellung dar. Ist der Katalog der Untersuchungsmerkmale im Sinne der beabsichtigten Untersuchung unzweckmäßig oder unvollständig, so kann dies selbst durch die ausgeklügeltsten Auswertungsmethoden nicht mehr behoben werden. Will man beispielsweise Informationen darüber gewinnen, welche Faktoren für den Aufstieg eines Angestellten zu einem leitenden Angestellten maßgeblich sind, so liegt es nahe, sich auf Merkmale wie etwa die Ausbildung, den Intelligenzquotienten, das Durchsetzungsvermögen usw. zu konzentrieren. Dennoch besteht keine Gewähr dafür, dass das wichtigste Merkmal in die Erhebung einbezogen wurde; es könnte – das Beispiel ist fiktiv – durchaus auch die unberücksichtigte Augenfarbe sein. Leider gibt es kein Patentrezept für die Aufstellung des adäquaten Merkmalskatalogs. Die nahe liegende Lösung, möglichst viele Merkmale einzubeziehen, ist aus Kostengründen und wegen der Überforderung der Interviewten meist nicht praktikabel.

- Ähnlich verzerrende Einflüsse können aus der bewussten **Antwortverweigerung**, **unklaren Formulierungen im Fragebogen** oder aus dem **Einfluss des Interviewers** resultieren. Die empirische Sozialforschung widmet diesen Problemen zu Recht ihre besondere Aufmerksamkeit. Es sei beispielsweise auf Holm (1975) oder Atteslander (2010) verwiesen.

- Weitere verzerrende Einflüsse können von der **Auswahl „falscher“ Teilgesamtheiten** herrühren. Als drastisches Beispiel sei ein Interviewer erwähnt, der die Meinung der Gesamtbevölkerung zum Thema „Studienbeiträge“ empirisch erkunden soll und zu diesem Zweck ausschließlich Studenten auswählt, weil er diese als besonders auskunftsfreudig kennt und zahlreich in der Mensa antrifft. Weitere Beispiele sind bei Krämer (2011) zu finden.

- Verzerrungen können auch durch **Klassenbildung**, insbesondere durch Bildung unbeschränkter Randklassen, entstehen.

- Ferner ist die **Antwortvariabilität** ein nicht zu unterschätzendes Problem, vgl. dazu Strecker et al. (1997).

Als weitere ergänzende und vertiefende Literatur zu Kapitel 2 seien erwähnt: Ferschl (1991), Krug et al. (2001), Menges/Skala (1973), Pfanzagl (1983), Schwarze (2009a), Wetzel (1971).

3. Kapitel:

Auswertungsmethoden für eindimensionales Datenmaterial

Ein quantitatives oder quantifiziertes Merkmal X werde an den n Merkmalsträgern einer statistischen Masse beobachtet. Die resultierenden Zahlen $x_1, \ldots, x_n$ heißen **Beobachtungswerte** oder **Merkmalswerte**; x_i beschreibt also die beim i-ten Merkmalsträger beobachtete Merkmalsausprägung des Merkmals X. Das n-Tupel $(x_1, \ldots, x_n)$ aller n Merkmalswerte wird als **Urliste** bezeichnet.

3.1 Häufigkeitsverteilungen

Die Urliste umfasst in der Regel so viele Daten, dass die Übersichtlichkeit nicht mehr gewährleistet ist. Selbst für bescheidene Werte von n kann das Bild bereits verwirrend sein.

Beispiel 3.1: In einer Unternehmung wird auf Grund vorhandener Belege festgestellt, dass der Hauptlieferant für die letzten 50 Bestellungen folgende Lieferzeiten (in Tagen) benötigt hatte:

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

Für das Merkmal X (= Lieferzeit in Tagen) wurden also die $n = 50$ Merkmalswerte $x_1 = 4, x_2 = 5, \ldots, x_{50} = 5$ beobachtet. Als Auswertungsmethode drängt sich hier die Bildung einer Häufigkeitsverteilung auf. ■

3.1.1 Absolute und relative Häufigkeitsverteilung

Die in der Urliste vorkommenden Merkmalsausprägungen seien mit $a_1, \ldots, a_k$ bezeichnet. Wir wollen uns diese Ausprägungen bereits der Größe nach geordnet vorstellen:

$$a_1 < a_2 < \cdots < a_k \ .$$

Unter der **absoluten Häufigkeit** von a_j versteht man die Anzahl der Merkmalswerte in der Urliste, die mit der Ausprägung a_j übereinstimmen; wir wollen sie mit $h(a_j)$ bezeichnen. Der Anteilswert $f(a_j) = \frac{1}{n} \cdot h(a_j)$ heißt **relative Häufigkeit** von a_j. Es gelten natürlich die Beziehungen

$$\sum_{j=1}^{k} h(a_j) = n \quad \text{und} \quad \sum_{j=1}^{k} f(a_j) = 1 \ .$$

Für das Beispiel 3.1 erhält man die in folgender Tabelle angegebenen Häufigkeiten:

Ausprägung a_j	1	2	3	4	5	6	7
absolute Häufigkeit $h(a_j)$	1	1	2	12	17	9	8
relative Häufigkeit $f(a_j)$	$\frac{1}{50}$	$\frac{1}{50}$	$\frac{2}{50}$	$\frac{12}{50}$	$\frac{17}{50}$	$\frac{9}{50}$	$\frac{8}{50}$

Unter einer **Häufigkeitsverteilung** versteht man allgemein die Zuordnung von (absoluten oder relativen oder auch kumulierten) Häufigkeiten zu den Merkmalsausprägungen $a_1, \ldots, a_k$. Als

Darstellungsform für die absolute oder relative Häufigkeitsverteilung wählt man üblicherweise eine **Häufigkeitstabelle** (wie die vorstehende), ein **Stabdiagramm** (oder Säulendiagramm), ein **Kreissektorendiagramm** oder im Falle klassierter Daten ein **Histogramm**. Das Histogramm wird zusammen mit einigen Hinweisen für die Klassenbildung im nächsten Abschnitt behandelt.

Ein **Stabdiagramm** entsteht aus den in einem Koordinatensystem eingetragenen k Punkten $(a_j, h(a_j))$ dadurch, dass jeweils das Lot auf die Abszisse (d. h. die Merkmalsachse) eingezeichnet wird. Für das Beispiel 3.1 ergibt sich das in Abbildung 3.1 wiedergegebene Stabdiagramm. Selbstverständlich können an Stelle der absoluten auch die relativen Häufigkeiten verwendet werden. Wählt man die Stäbe etwas dicker, so spricht man von einem **Säulendiagramm**.

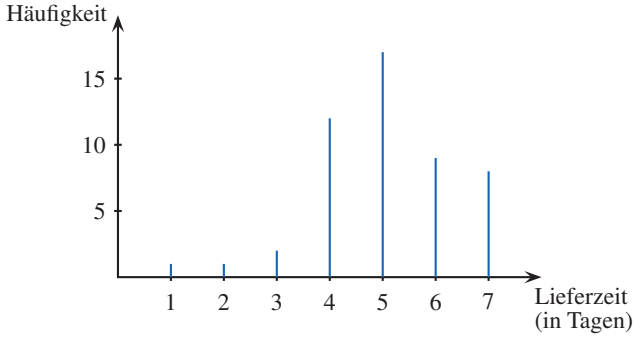

Abb. 3.1: Stabdiagramm für die Lieferzeiten aus Beispiel 3.1

Bei einem **Kreissektorendiagramm** werden die Häufigkeiten als Kreissektoren dargestellt, wobei die Sektorflächen (und damit auch die Winkel) proportional zu den Häufigkeiten gewählt werden müssen. Für das Beispiel 3.1 erhält man die Darstellung in Abbildung 3.2. In populären Darstellungen wird das Kreissektorendiagramm gern benutzt. Für nominalskalierte Merkmale besitzt es darüber hinaus den Vorteil, den bei Verwendung des Stabdiagramms möglicherweise suggerierten Eindruck zu vermeiden, dass eine Ordinal- oder gar eine Kardinalskala vorliege.

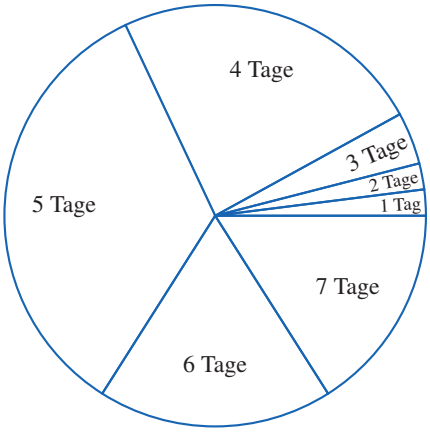

Abb. 3.2: Kreissektorendiagramm für die Lieferzeiten aus Beispiel 3.1

Aufgabe 3.2: Die Befragung von $n = 40$ Arbeitnehmern einer Textilfabrik nach dem für den Weg zur Arbeitsstätte benutzten Verkehrsmittel ergab folgende Urliste:

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

Dabei bedeuten 1: öffentliches Verkehrsmittel; 2: Pkw; 3: Motorrad; 4: Fahrrad; 5: zu Fuß. Erstellen Sie hieraus eine Häufigkeitstabelle sowie ein Kreissektorendiagramm. ■

3.1.2 Histogramm, Hinweise zur Klassenbildung

Wird ein stetiges oder quasistetiges Merkmal erhoben, so sind die in der Urliste vorkommenden Zahlen i. Allg. alle voneinander verschieden. Die Aufstellung von Häufigkeitstabellen oder Stabdiagrammen wäre dann eine brotlose Kunst; sie bringt – wie in Beispiel 3.3 illustriert – i. Allg. keinen nennenswerten Gewinn an Übersichtlichkeit.

Beispiel 3.3: In einem Kunststoff verarbeitenden Betrieb findet eine Vielzahl gleichartiger Ventile Verwendung. Diese sind einem besonders hohen Verschleiß ausgesetzt, was häufig zum Maschinenstillstand führt. Im Zuge der Planung einer kostenminimalen präventiven Instandhaltungsstrategie wurden folgende Ventillebensdauern (in Betriebsstunden) erhoben:

110	520	490	30	120	290	370	305	415	170	280	70	540	460	260
345	150	220	435	425	470	350	130	380	230	320	360	240	330	580

Das zugehörige Stabdiagramm zeigt Abbildung 3.3. ■

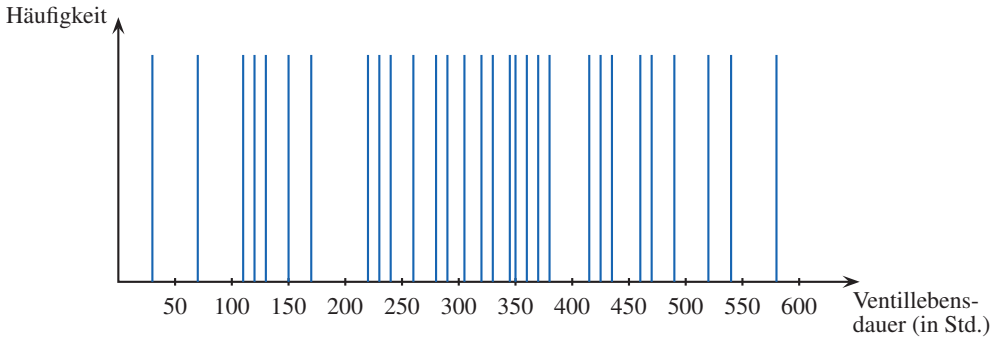

Abb. 3.3: Stabdiagramm für die Ventillebensdauern aus Beispiel 3.3

In derartigen Fällen bietet sich der Übergang zu klassierten Daten an. Wir teilen dazu die Merkmalsachse in Klassen ein und registrieren jeweils, welche **Klassenhäufigkeit** oder, wie man auch sagt, welche **Besetzungszahl** den einzelnen Klassen zukommt.

Wenn wir die j-te Klasse mit dem Symbol a_j und die Klassenhäufigkeit mit $h(a_j)$ bezeichnen, können wir den Begriff der Häufigkeitstabelle unverändert übernehmen; sie wird allerdings – wie in Abbildung 2.2 – zumeist vertikal angeordnet. Als grafische Darstellung solcher klassierten Häufigkeitsverteilungen verwendet man das **Histogramm**. Es besteht aus Rechtecken, die über den einzelnen Klassen so errichtet werden, dass die Rechtecksfläche proportional zur jeweiligen Klassenhäufigkeit ist. Führen wir in Beispiel 3.3, von 0 ab beginnend, 6 Klassen der Breite 100 ein, so erhalten wir die klassierte Häufigkeitstabelle und das

Histogramm von Abbildung 3.4. Nach Möglichkeit sollten die Klassengrenzen äquidistant gewählt werden; die Höhen der Histogramm-Rechtecke sind dann proportional zur Klassenhäufigkeit. **Unbeschränkte Randklassen** wie etwa die oberste Klasse in Abbildung 2.2 sollten nach Möglichkeit vermieden werden. Falls sie unvermeidbar sind, kann nur die Häufigkeitstabelle, jedoch nicht das Histogramm erstellt werden. Für die Mindestanzahl der zu bildenden Klassen gibt DIN 55302, Blatt 1 u. a. die Empfehlung, 10 Klassen bei $n \approx 100$, 13 Klassen bei $n \approx 1\,000$ und 16 Klassen bei $n \approx 10\,000$ vorzusehen.

Klasse	Klassenhäufigkeit
0 bis unter 100	2
100 bis unter 200	5
200 bis unter 300	6
300 bis unter 400	8
400 bis unter 500	6
500 bis unter 600	3

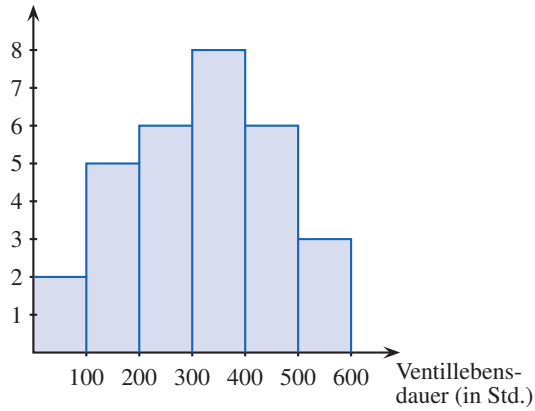

Abb. 3.4: Klassierte Häufigkeiten und Histogramm für die Ventillebensdauern aus Beispiel 3.3

Als weitere grafische Veranschaulichung klassierter Häufigkeitsverteilungen wird gelegentlich das **Häufigkeitspolygon** benutzt. Es besteht aus dem Streckenzug, der die Mitten aller oberen Rechteckseiten des Histogramms verbindet[1] (vgl. Lösung der nachfolgenden Aufgabe 3.4).

Aufgabe 3.4: Bei einer Stichprobe von $n = 20$ Studenten einer Wirtschaftsfakultät wurden folgende jährliche Ausgaben (in €) für Urlaubszwecke ermittelt:

1 000	580	520	350	620	800	120	600	550	420
470	200	560	480	1 000	600	1 150	800	250	650

Erstellen Sie hieraus unter Verwendung der Klassengrenzen 0, 300, 500, 700 und 1 200 das Histogramm sowie das Häufigkeitspolygon. ■

3.1.3 Kumulierte Häufigkeitsverteilung, empirische Verteilungsfunktion

Wie viele Beobachtungsdaten überschreiten eine gewisse Grenze? Wie viel Prozent erreichen maximal eine gewisse Grenze? Die Beantwortung derartiger Fragen erfordert die Kumulierung von Einzelhäufigkeiten. Aus diesem Grunde ist die Bildung von (absoluten oder relativen) kumulierten Häufigkeitsverteilungen zweckmäßig. Die Funktion

$$H(x) = \sum_{a_j \leqq x} h(a_j) \tag{1}$$

[1] Die Gleichheit der Gesamtfläche unter dem Histogramm und dem Häufigkeitspolygon ist allerdings nur dann gewährleistet, wenn die Klassengrenzen äquidistant sind und das Häufigkeitspolygon zu den beiden Fußpunkten auf der Merkmalsachse verlängert wird, die eine halbe Klassenbreite links der untersten bzw. rechts der obersten Klassengrenze liegen.

heißt **absolute kumulierte Häufigkeitsverteilung** des Merkmals X. Sie registriert für jedes vorgegebene x die Anzahl derjenigen Beobachtungswerte, die kleiner oder gleich x sind.[2] Analog ergibt sich die **relative kumulierte Häufigkeitsverteilung**:

$$F(x) = \sum_{a_j \leqq x} f(a_j) \; .$$

$F(x)$ wird häufig als **empirische Verteilungsfunktion** bezeichnet. Es gilt natürlich die Beziehung

$$F(x) = \tfrac{1}{n} \cdot H(x) \; .$$

Sowohl bei $H(x)$ als auch bei $F(x)$ handelt es sich um monoton wachsende Treppenfunktionen; sie springen an den realisierten Ausprägungen $a_1, a_2, \ldots, a_k$ gerade um die absolute bzw. relative Häufigkeit. Abbildung 3.5 verdeutlicht den typischen Verlauf.

$$H(x) = \begin{cases} 0 & \text{für} \quad x < 1 \\ 1 & \text{für} \quad 1 \leqq x < 2 \\ 2 & \text{für} \quad 2 \leqq x < 3 \\ 4 & \text{für} \quad 3 \leqq x < 4 \\ 16 & \text{für} \quad 4 \leqq x < 5 \\ 33 & \text{für} \quad 5 \leqq x < 6 \\ 42 & \text{für} \quad 6 \leqq x < 7 \\ 50 & \text{für} \quad 7 \leqq x \end{cases}$$

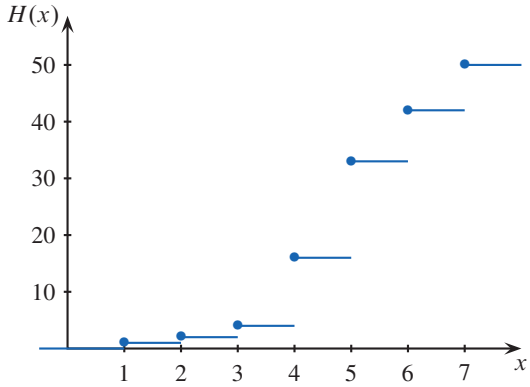

Abb. 3.5: Absolute kumulierte Häufigkeitsverteilung für die Lieferzeiten aus Beispiel 3.1

Aufgabe 3.5: Entnehmen Sie der Funktion $H(x)$ aus Abbildung 3.5, in wie vielen Fällen die Lieferzeit höchstens 5 Tage bzw. in wie viel Prozent der Fälle sie mehr als 3 Tage betragen hat. ■

Im Fall klassierter Daten interpretiert man als kumulierte Häufigkeitsverteilung $H(x)$ bzw. $F(x)$ denjenigen monoton wachsenden Polygonzug, der durch die über den Klassengrenzen eingetragenen absoluten bzw. relativen kumulierten Klassenhäufigkeiten verläuft. Dieser Interpretation liegt (wie beim Histogramm) die Vorstellung zu Grunde, dass innerhalb der Klassen eine gleichmäßige Verteilung eines stetigen oder quasistetigen Merkmals vorliegt. Das Histogramm aus Abbildung 3.4 führt nach dieser Konstruktionsvorschrift zu der in Abbildung 3.6 dargestellten kumulierten Häufigkeitsverteilung $H(x)$.

3.2 Lageparameter

Eine wesentlich radikalere Komprimierung der Ausgangsdaten als beim Übergang zur Häufigkeitsverteilung ist dann gegeben, wenn die gesamte Urliste durch eine einzige Zahl, einen

[2] Korrekter, aber optisch unschöner, müsste (1) lauten:

$$H(x) = \sum_{j \,|\, a_j \leqq x} h(a_j) \; .$$

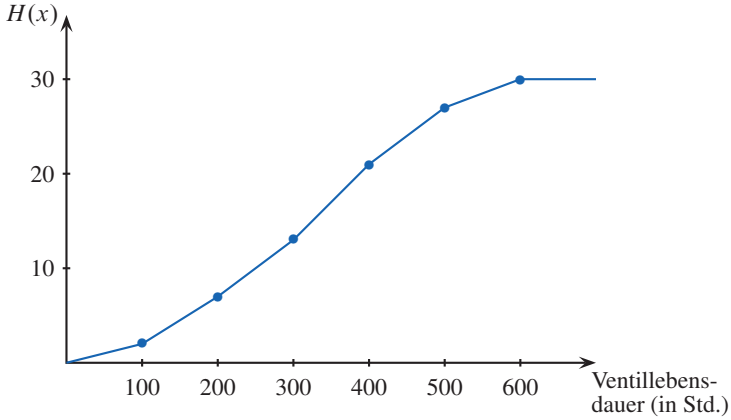

Abb. 3.6: Kumulierte Häufigkeitsverteilung, die sich für die klassierte Häufigkeitsverteilung von Abbildung 3.4 bei Zugrundelegung einer gleichmäßigen Verteilung innerhalb der Klassen ergibt

so genannten **Lageparameter**, charakterisiert werden soll. An einen Lageparameter wird die Forderung gestellt, dass er „möglichst gut“ beschreibt, **wo** das gesamte Datenmaterial auf der Merkmalsachse lokalisiert ist. Im Folgenden werden die wichtigsten Lageparameter besprochen.

3.2.1 Modalwert, Median, arithmetisches und geometrisches Mittel

Diejenigen Ausprägungen, die die größte Häufigkeit aufweisen, werden als **Modalwerte** x_{Mod} bezeichnet. Eingipfelige Häufigkeitsverteilungen besitzen genau einen Modalwert; so hat beispielsweise die Häufigkeitsverteilung aus Abbildung 3.1 den Modalwert 5. Der Modalwert ist der wichtigste Lageparameter für nominalskalierte Merkmale. Man sollte die Angabe eines Modalwertes allerdings generell auf den Fall unimodaler, d. h. eingipfeliger Verteilungen beschränken. Andere Bezeichnungen für den Modalwert sind **häufigster Wert**, **dichtester Wert** oder auch **Modus**.

Der **Median** x_{Med} ist durch die Eigenschaft definiert, dass mindestens 50 % aller Merkmalswerte kleiner oder gleich x_{Med} und mindestens 50 % aller Merkmalswerte größer oder gleich x_{Med} sind. Wenn die n Merkmalswerte bereits der Größe nach gemäß $x_1 \leqq x_2 \leqq \cdots \leqq x_n$ geordnet sind, so liegt der Median „in der Mitte“ dieser Zahlen. Genauer gilt: Im Falle einer ungeraden Beobachtungsanzahl n ist

$$x_{\text{Med}} = x_{\frac{n+1}{2}} \ . \tag{2a}$$

Im Falle eines geraden n erfüllt jeder Wert im Intervall $[x_{\frac{n}{2}}; x_{\frac{n}{2}+1}]$ obige Bedingung. Im letzteren Fall wird der Median meistens als Intervallmitte, d. h. als

$$x_{\text{Med}} = \tfrac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) \tag{2b}$$

festgesetzt. Für die Daten aus Abbildung 3.1 ist der Median gleich 5; er stimmt hier also – was nicht generell zutrifft – mit dem Modalwert überein. Der Median – häufig auch als **Zentralwert** bezeichnet – ist der wichtigste Lageparameter für ordinalskalierte Merkmale.

Der bekannteste Lageparameter – im Alltag meist als **Durchschnittswert** bezeichnet – ist das **arithmetische Mittel** $\bar{x}$, das bei Verwendung der Urliste bzw. der Häufigkeitsverteilung folgendermaßen berechnet wird:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{bzw.} \quad \bar{x} = \sum_{j=1}^{k} a_j\, f(a_j) \ . \tag{3}$$

Es ist auf kardinalskalierte Merkmale zugeschnitten und besitzt die Form eines speziellen gewogenen Mittels[3] der Merkmalswerte; jedem Merkmalswert wird das Gewicht $\frac{1}{n}$ bzw. der Merkmalsausprägung a_j das Gewicht $f(a_j)$ zugeordnet.

Bestehen die Merkmalswerte z. B. aus Wachstumsfaktoren oder Aufzinsungsfaktoren, die über sukzessive Zeitperioden hinweg beobachtet wurden, so ist – vgl. Aufgabe 3.9 – an Stelle des arithmetischen Mittels das **geometrische Mittel**

$$x_{\text{Geom}} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

zu verwenden. Denn nur dieses hat die Eigenschaft, über n Perioden hinweg denselben Effekt zu erzielen wie die sukzessive Multiplikation mit $x_1, x_2, \ldots, x_n$.

3.2.2 Eigenschaften der Lageparameter und Vergleich

Den bisherigen Ausführungen ist folgende Empfehlung zu entnehmen:

Skalierung	nominal	ordinal	kardinal
zu verwendender Lageparameter	Modalwert	Median	arithm. Mittel[4]

Einige relativierende und ergänzende Bemerkungen sollten jedoch hinzugefügt werden:

- Im Gegensatz zum Modalwert stimmen das arithmetische Mittel sowie der nach (2b) berechnete Median u. U. mit keinem der Beobachtungswerte überein. Der altbekannte Vorwurf, es gebe den „mittleren Bundesbürger“ oder den „mittleren Haushalt“ überhaupt nicht, kann in gewissen Fällen berechtigt sein.
- In der Tat kann der Modalwert auch bei ordinal- oder kardinalskalierten Merkmalen den sinnvolleren Lageparameter darstellen. Als Standardbeispiel sei die Erhebung von Schuhgrößen, Konfektionsgrößen u. Ä. aufgeführt, bei der die Kenntnis des Modalwertes eine fundiertere Produktionsentscheidung abzuleiten gestattet als die Kenntnis des Medians oder des arithmetischen Mittels.
- Im Unterschied zum arithmetischen Mittel, das auf Ausreißer (oder Fehler) in den Daten sehr empfindlich reagiert, erweist sich der Median gegenüber Ausreißern als robust, vgl. Aufgabe 3.7.
- Bei klassierten Daten ist der exakte Wert von $\bar{x}$ nicht zu ermitteln. Als **Ersatzgröße** wird das gewogene Mittel der Klassenmitten verwendet, wobei die relativen Klassenhäufigkeiten als Gewichte fungieren.

[3] Allgemein kann man für jeden Satz von nichtnegativen Gewichten $g_1, \ldots, g_n$ mit $g_1 + \cdots + g_n = 1$ ein **gewichtetes** oder **gewogenes arithmetisches Mittel** der Zahlen $y_1, \ldots, y_n$ definieren gemäß $g_1 y_1 + g_2 y_2 + \cdots + g_n y_n$.

[4] Bei Wachstumsfaktoren ist jedoch – wie erwähnt – das geometrische Mittel zu verwenden.

- **Transformiert** man die Beobachtungsdaten x_i gemäß

$$y_i = a + bx_i$$

linear, so transformieren sich die jeweiligen arithmetischen Mittel $\bar{x}$ und $\bar{y}$ offensichtlich nach demselben Gesetz

$$\bar{y} = a + b\bar{x} \; ; \tag{4}$$

nach dem Muster der Aufgabe 3.8 kann (4) zur Rechenvereinfachung ausgenutzt werden.

- Im Fall r disjunkter (d. h. paarweise elementfremder) statistischer Massen $M_1, M_2, \ldots, M_r$, deren jeweilige arithmetische Mittel mit $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_r$ bezeichnet seien, berechnet sich das Gesamtmittel (d. h. das arithmetische Mittel für die Gesamtmasse $M = M_1 \cup M_2 \cup \cdots \cup M_r$) als gewogenes Mittel

$$\boxed{\bar{x}_{\text{Ges}} = \frac{1}{n} \sum_{j=1}^{r} n_j \bar{x}_j} \; , \tag{5}$$

wobei n_j die Anzahl der Elemente von M_j und $n = n_1 + \cdots + n_r$ die Anzahl der Elemente der Gesamtmasse M bezeichnen.

- Obige Lageparameter besitzen je eine **Optimalitätseigenschaft**, die sich in Form der Minimierung eines (realen oder fiktiven) Schadens formulieren lässt. Sucht man denjenigen Lageparameter λ, der die Quadratsumme

$$\sum_{i=1}^{n} (x_i - \lambda)^2 \tag{6}$$

minimiert, so ergibt sich $\lambda = \bar{x}$ als Optimalwert. Sucht man dagegen denjenigen Lageparameter λ, der die Betragssumme

$$\sum_{i=1}^{n} |x_i - \lambda|$$

minimiert, so ergibt sich $\lambda = x_{\text{Med}}$ als Optimalwert. Sucht man denjenigen Lageparameter λ, der die Summe

$$\sum_{i=1}^{n} s(x_i, \lambda) \quad \text{mit} \quad s(x_i, \lambda) = \begin{cases} 0, & \text{falls } x_i = \lambda \\ 1, & \text{falls } x_i \neq \lambda \end{cases}$$

minimiert, so ergibt sich schließlich $\lambda = x_{\text{Mod}}$.

Diese Optimalitätseigenschaften besitzen auch außerhalb der Statistik, nämlich z. B. bei Standortproblemen (vgl. Aufgabe 3.10), einige Anwendungsmöglichkeiten.

Aufgabe 3.6: Bestätigen Sie, dass $\bar{x}$ den Ausdruck (6) minimiert. Verifizieren Sie dabei, dass

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \tag{7}$$

gilt, d. h. dass die Summe aller Abweichungen von $\bar{x}$ gleich null ist. ■

Aufgabe 3.7:

a) Berechnen Sie für die Ventillebensdauern aus Beispiel 3.3 sowohl $\bar{x}$ als auch den Median. Berechnen Sie zum Vergleich den aus dem Histogramm der Abbildung 3.4 resultierenden Ersatzwert für $\bar{x}$.

b) Durch einen Schreibfehler sei als letzter Beobachtungswert in Beispiel 3.3 statt 580 der Wert 5 800 eingetragen worden. Wie wirkt sich dieser Fehler auf $\bar{x}$ bzw. x_{Med} aus? ■

Aufgabe 3.8: Die Verbraucher-Beratungsstelle einer Kleinstadt stellte an einem Stichtag die Preise für Normal-Benzin bei den 20 örtlichen Tankstellen fest. Es ergaben sich die bereits der Größe nach geordneten Beobachtungswerte:

91,4	91,4	91,9	91,9	91,9	91,9	91,9	92,9	93,9	93,9
95,9	95,9	95,9	96,9	97,9	97,9	97,9	98,9	98,9	98,9

a) Bestimmen Sie aus diesen Messergebnissen den Modus, den Median und – nach geeigneter linearer Transformation der Merkmalswerte – das arithmetische Mittel.

b) Um sich ein genaueres Bild von der Lage auf dem Benzinmarkt in dieser Region zu machen, wurden die Preise für Normal-Benzin bei weiteren 12 Tankstellen in der unmittelbaren Umgebung der Stadt festgestellt. Es ergab sich (bei diesen 12 Tankstellen) ein mittlerer Preis von 96,5. Berechnen Sie das arithmetische Mittel aller 32 Benzinpreise. ■

Aufgabe 3.9: Im Jahre 1995 betrug der bundesdeutsche Verbrauch an Primärenergie 486,9 Mio. t Steinkohleeinheiten.[5] Für die Jahre bis 1999 „wuchs“ der Verbrauch um die folgenden (gerundeten) Faktoren:

1996 gegenüber 1995 um 1,03
1997 gegenüber 1996 um 0,99
1998 gegenüber 1997 um 0,99
1999 gegenüber 1998 um 0,98

Bestimmen Sie den durchschnittlichen Wachstumsfaktor und vergleichen Sie diesen mit $\bar{x}$. ■

Aufgabe 3.10: Ein Zweigwerk ist in Form einer Werkstraße angelegt. Abbildung 3.7 verdeutlicht die Situation und gibt die Anzahl der Beschäftigten in der Pforte sowie den drei Maschinenhallen an. Die 1 000 Beschäftigten wurden bisher mit werkseigenen Bussen zur Kantine des Hauptwerks transportiert. Nun soll eine eigene Kantine an dem Punkt der Werkstraße errichtet werden, für den die Summe der Wegstrecken aller 1 000 Beschäftigten minimal ist. Bestimmen Sie diesen Standort. ■

	Pforte	Maschinenhalle 1	Maschinenhalle 2	Maschinenhalle 3
Anzahl der Beschäftigten →	3	200	300	497
Abstand von der Pforte →	0	100	600	900

Abb. 3.7: Daten zur Standortplanung der Kantine aus Aufgabe 3.10

[5] Zitiert nach: Institut der Deutschen Wirtschaft (2001): Deutschland in Zahlen, S. 91.

3.3 Streuungsparameter

Aus der alleinigen Angabe eines Lageparameters kann natürlich noch nicht entnommen werden, ob die Beobachtungswerte im Wesentlichen in der Nähe dieses Lageparameters oder aber weiter entfernt davon liegen.

Beispiel 3.11: Das arithmetische Mittel der 50 Lieferzeiten aus Beispiel 3.1 ist $\bar{x} = \frac{252}{50} = 5{,}04$. Denselben Wert erhält man z. B. auch für die Häufigkeitsverteilung $h(5 \text{ Tage}) = 48$ und $h(6 \text{ Tage}) = 2$, die offensichtlich eine weitaus geringere Variabilität oder „Streuung" aufweist. ▪

Auf Grund des in Beispiel 3.11 illustrierten Phänomens erscheint es ratsam, die Angabe eines Lageparameters durch die zusätzliche Angabe eines Streuungsparameters zu flankieren.

3.3.1 Spannweite, durchschnittliche Abweichung, mittlere quadratische Abweichung, Standardabweichung, Variationskoeffizient

Diese Überschrift enthält die in der Praxis verwendeten Streuungsparameter. Sie setzen stets kardinalskalierte Beobachtungswerte $x_1, \ldots, x_n$ voraus und sind folgendermaßen definiert: Die **Spannweite**

$$SP = \max_i x_i - \min_i x_i \tag{8}$$

ist die Differenz zwischen dem größten und dem kleinsten Beobachtungswert. Die **durchschnittliche Abweichung** von einem Lageparameter λ

$$\bar{s} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \lambda| = \sum_{j=1}^{k} |a_j - \lambda| \cdot f(a_j) \tag{9}$$

ist das arithmetische Mittel der Abstände aller Beobachtungswerte von λ.[6] Die **mittlere quadratische Abweichung**

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{10}$$

ist das arithmetische Mittel der quadrierten Abstände aller Beobachtungswerte von $\bar{x}$. Unter Verwendung der Häufigkeitsverteilung berechnet man s^2 folgendermaßen:

$$s^2 = \sum_{j=1}^{k} (a_j - \bar{x})^2 \cdot f(a_j) \; .$$

Die **Standardabweichung** $s = \sqrt{s^2}$ ist die positive Wurzel aus der mittleren quadratischen Abweichung. Der (nur für positives $\bar{x}$ betrachtete) **Variationskoeffizient** $V = \frac{s}{\bar{x}}$ ist der Quotient aus Standardabweichung s und arithmetischem Mittel $\bar{x}$.

Die weitaus wichtigsten Streuungsparameter sind s und s^2. Sie werden im nachfolgenden Abschnitt detaillierter behandelt. Die Spannweite SP ist zwar am einfachsten zu berechnen,

[6] Wegen der in Abschnitt 3.2.2 angesprochenen Minimalitätseigenschaft des Medians wird $\bar{s}$ meist durch $\lambda = x_{\text{Med}}$ und gelegentlich durch $\lambda = \bar{x}$ konkretisiert.

jedoch – da sie nur auf den beiden extremen Beobachtungswerten beruht – sehr „ausreißerempfindlich“. Der Variationskoeffizient V ist maßstabsunabhängig und quantifiziert das, was man intuitiv unter „Variabilität“ versteht, vielfach besser als die anderen Streuungsparameter (vgl. Beispiel 3.12). Er wird häufig als Prozentwert angegeben.

Beispiel 3.12: Wie in Beispiel 3.1 werden wieder 50 Lieferzeiten betrachtet, allerdings für ein anderes Produkt und einen anderen Lieferanten. Die Lieferzeiten seien jeweils um 20 Tage länger als diejenigen aus Beispiel 3.1. Sie betragen also 24, 25, 24, 21, 25, 24, 23, 24, 25, 26 usw. und schwanken zwischen 21 und 27 Tagen. Die Schwankungen sind im Verhältnis zur Dauer der Lieferfrist kleiner geworden; die Variabilität ist – so wird man gefühlsmäßig urteilen – kleiner als in Beispiel 3.1. Die vier Streuungsparameter SP, $\bar{s}$, s^2 und s bleiben aber jeweils unverändert; lediglich der Variationskoeffizient V zeigt eine Verringerung der Variabilität an: Während er im Beispiel 3.1 bei $\frac{s}{5{,}04}$ liegt, beträgt er jetzt $\frac{s}{25{,}04}$. ■

Aufgabe 3.13: Berechnen Sie für die Häufigkeitsverteilung aus Beispiel 3.11 die fünf Streuungsparameter SP, $\bar{s}$, s^2, s und V. ■

3.3.2 Eigenschaften der mittleren quadratischen Abweichung und der Standardabweichung

- Werden die Beobachtungswerte x_i gemäß $y_i = a + bx_i$ **linear transformiert**, so transformieren sich die zugehörigen mittleren quadratischen Abweichungen s_x^2 bzw. s_y^2 folgendermaßen:

$$s_y^2 \overset{(10)}{=} \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \overset{(4)}{=} \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - a - b\bar{x})^2 = b^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \overset{(10)}{=} b^2 s_x^2 \ .$$

  Aus der Beziehung

$$\boxed{s_y^2 = b^2 s_x^2} \tag{11}$$

  ergibt sich für die Standardabweichungen s_x bzw. s_y sofort der Zusammenhang

$$s_y = |b| \cdot s_x \ .$$

- Eine einfache Umrechnung der mittleren quadratischen Abweichung führt auf eine alternative Darstellungsform:

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

  Das Resultat

$$\boxed{s^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2} \tag{12}$$

  wird als **Verschiebungssatz** bezeichnet.[7]

[7] Der in der Physik bewanderte Leser wird sicher die Analogie zu dem gleichnamigen Satz der Mechanik erkennen; s^2 entspricht nämlich dem Trägheitsmoment bei Rotation um den Schwerpunkt ($= \bar{x}$) und $\frac{1}{n}\sum x_i^2$ dementsprechend dem Trägheitsmoment bei Rotation um den Nullpunkt.

- Die drei Streuungsparameter SP, $\bar{s}$ und s besitzen dieselbe Dimension (nämlich diejenige der Beobachtungswerte), so dass die Frage nach einem Größenvergleich sinnvoll ist. In der Tat lässt sich nachweisen, dass die Standardabweichung generell zwischen $\bar{s}$ (mit $\lambda = x_{\text{Med}}$ oder $\lambda = \bar{x}$ gebildet) und SP liegt:

$$\bar{s} \leqq s \leqq SP \; .$$

- Für r disjunkte statistische Massen $M_1, \ldots, M_r$, deren arithmetische Mittel bzw. mittlere quadratische Abweichungen mit $\bar{x}_1, \ldots, \bar{x}_r$ bzw. $s_1^2, \ldots, s_r^2$ bezeichnet seien, berechnet sich die mittlere quadratische Abweichung für die Gesamtmasse $M = M_1 \cup \cdots \cup M_r$ folgendermaßen:

$$s_{\text{Ges}}^2 = \frac{1}{n} \sum_{j=1}^{r} n_j s_j^2 + \frac{1}{n} \sum_{j=1}^{r} n_j (\bar{x}_j - \bar{x}_{\text{Ges}})^2 \; . \tag{13}$$

  Dabei bezeichnet n_j wie in Abschnitt 3.2.2 die Anzahl der Elemente von M_j und $n = n_1 + \cdots + n_r$ die Anzahl der Elemente der Gesamtmasse M; das Gesamtmittel $\bar{x}_{\text{Ges}}$ ist nach (5) zu berechnen. Der erste Ausdruck auf der rechten Seite von (13), also das gewogene Mittel der s_j^2, wird gelegentlich **interne mittlere quadratische Abweichung** oder kürzer **interne Varianz** genannt; der zweite Summand wird entsprechend als **externe mittlere quadratische Abweichung** bzw. als **externe Varianz** bezeichnet.

- Im Vorgriff auf ein wahrscheinlichkeitstheoretisches Resultat (vgl. Abschnitt 8.6.3) sei eine Interpretation der Standardabweichung s erwähnt, die auf zahlreiche kardinalskalierte Merkmale anwendbar ist: Im Intervall $[\bar{x} - s; \bar{x} + s]$ liegen etwa 68 % aller Beobachtungswerte, im Intervall $[\bar{x} - 2s; \bar{x} + 2s]$ liegen etwa 95 % aller Beobachtungswerte und im Intervall $[\bar{x} - 3s; \bar{x} + 3s]$ liegen praktisch alle Beobachtungswerte.

- Für stetige Aktienrenditen x_i wird s meist als (historische) **Volatilität** bezeichnet.

Aufgabe 3.14: Für die 20 Normal-Benzin-Preise aus Aufgabe 3.8 errechnet man eine mittlere quadratische Abweichung von 7,875. Für die 12 weiteren Preise aus Teil b) sei die mittlere quadratische Abweichung 2,875. Berechnen Sie die mittlere quadratische Abweichung aller 32 Preise. ■

Aufgabe 3.15: Fassen Sie die 50 Lieferzeiten, die der Häufigkeitsverteilung aus Beispiel 3.11 zu Grunde liegen, in zwei Gruppen M_1 und M_2 zusammen, wobei M_1 aus den 48 Lieferzeiten der Dauer 5 und M_2 aus den beiden Lieferzeiten der Dauer 6 besteht. Berechnen Sie hierfür die interne und die externe Varianz. Vergleichen Sie die externe Varianz mit der in Aufgabe 3.13 ermittelten mittleren quadratischen Abweichung. ■

3.4 Konzentrationsmaße

Für eine Wettbewerbswirtschaft spielen Konzentrationstendenzen sowie ihre Offenlegung bzw. Verhinderung offensichtlich eine bedeutsame Rolle. Ein befriedigendes quantitatives Maß der Konzentration könnte deshalb einen nützlichen Beitrag für die aktuelle Wirtschaftspolitik leisten und die Diskussion um das erreichte oder gewünschte Ausmaß der Konzentration versachlichen.