Lehr- und Handbücher der Statistik

Herausgegeben von Universitätsprofessor Dr. Rainer Schlittgen

Bisher erschienene Werke:

Böhning, Allgemeine Epidemiologie
Caspary · Wichmann, Lineare Modelle
Chatterjee · Price (Übers. Lorenzen), Praxis der Regressionsanalyse, 2. Auflage
Degen · Lorscheid, Statistik-Lehrbuch, 2. Auflage
Degen · Lorscheid, Statistik-Aufgabensammlung, 4. Auflage
Hartung, Modellkatalog Varianzanalyse
Harvey (Übers. Untiedt), Ökonometrische Analyse von Zeitreihen, 2. Auflage
Harvey (Übers. Untiedt), Zeitreihenmodelle, 2. Auflage
Heiler · Michels, Deskriptive und Explorative Datenanalyse
Kockelkorn, Lineare statistische Methoden
Miller (Übers. Schlittgen), Grundlagen der Angewandten Statistik
Naeve, Stochastik für Informatik
Oerthel · Tuschl, Statistische Datenanalyse mit dem Programmpaket SAS
Pflaumer · Heine · Hartung, Statistik für Wirtschaft- und Sozialwissenschaften: Deskriptive Statistik, 2. Auflage
Pflaumer · Heine · Hartung, Statistik für Wirtschafts- und Sozialwissenschaften: Induktive Statistik
Pokropp, Lineare Regression und Varianzanalyse
Rasch · Herrendörfer u. a., Verfahrensbibliothek, Band I und Band 2
Riedwyl · Ambühl, Statistische Auswertungen mit Regressionsprogrammen
Rinne, Wirtschafts- und Bevölkerungsstatistik, 2. Auflage
Rinne, Statistische Analyse multivariater Daten – Einführung
Rüger, Induktive Statistik, 3. Auflage
Rüger, Test- und Schätztheorie, Band I
Rüger, Test- und Schätztheorie, Band II: Statistische Tests
Schendera, Datenmanagement und Datenanalyse mit dem SAS-System
Schlittgen, Statistik, 10. Auflage
Schlittgen, Statistik-Trainer
Schlittgen, Statistische Inferenz
Schlittgen, GAUSS für statistische Berechnungen
Schlittgen, Angewandte Zeitreihenanalyse
Schlittgen, Statistische Auswertungen mit R
Schlittgen · Streitberg, Zeitreihenanalyse, 9. Auflage
Schürger, Wahrscheinlichkeitstheorie
Tutz, Die Analyse kategorialer Daten

Fachgebiet Biometrie

Herausgegeben von Dr. Rolf Lorenz

Bisher erschienene Werke:

Bock, Bestimmung des Stichprobenumfangs
Brunner · Langer, Nichtparametrische Analyse longitudinaler Daten

Statistische Auswertungen

Standardmethoden und Alternativen mit ihrer Durchführung in R

Von
Universitätsprofessor
Dr. Rainer Schlittgen

R. Oldenbourg Verlag München Wien

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.

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

Gedruckt auf säure- und chlorfreiem Papier
Gesamtherstellung: Druckhaus „Thomas Müntzer“ GmbH, Bad Langensalza

ISBN 3-486-57616-X

Vorbemerkung

Es gibt zahlreiche Bücher, die sich mit statistischen Auswertungsmethoden, speziell der Varianz- und der Regressionsanalyse beschäftigen. Die Bücher unterstellen unterschiedlich viel Vorwissen und haben jeweils eigene Ausrichtungen. Das gilt auch für den vorliegenden Text. Die ersten beiden Kapitel geben eine recht knappe Einführung in die wesentlichen Aspekte der Datenanalyse und statistischen Schlussweisen. In den folgenden Kapiteln wird sukzessive von einfacheren Fragestellungen zu komplexeren fortgeschritten; Gliederungsmerkmal sind die Anzahlen von Stichproben: eine, zwei, mehrere Stichproben und dann komplizierter strukturierte Stichproben. Im Laufe dieses Fortschreitens werden die Methoden vorgestellt. Daher reichen die beiden ersten Kapitel als Einstieg sicherlich aus. Allerdings wäre es ideal, wenn die Nutzerin bzw. der Nutzer bereits über Vorkenntnisse in Statistik verfügen würde. Dann könnte sie/er das erste Kapitel als Wiederholung lesen und das Hauptgewicht auf die anwendungsorientierten Aspekte legen.

Gerade bei der Varianz- und der Regressionsanalyse bildet die Normalverteilungsannahme den Standardzugang zu statistischen Auswertungen. Das kommt in dem "geflügelten Wort", dessen Ursprung nicht mehr ganz klar ist, zum Ausdruck:

> Die Mathematiker glauben an die Normalverteilung, weil sie sie für ein Naturgesetz halten.
> Die Naturwissenschaftler glauben an die Normalverteilung, weil sie sie für eine mathematisch bewiesene Tatsache halten.

Die Bedeutung der Normalverteilungstheorie beruht sicher auch darauf, dass für die Normalverteilung die ausgereifteste Theorie vorliegt, vor allem bei Problemen, die komplizierter strukturiert sind. Zudem sind Lösungsansätze für Probleme mit nichtnormalverteilten Stichproben vielfach aus Übertragungen der Normalverteilungstheorie entwickelt worden. Dabei ist die Unterstellung der Normalverteilung erst einmal der 'normale' Zugang. Wenn die Randbedingungen es zulassen, wird man ihn wählen. Ob er adäquat ist, sollte bei jeder konkreten Anwendung überprüft werden. Im Bedarfsfall sind auf schwächeren Voraussetzungen beruhende Methoden zu wählen. Gerade in jüngster Zeit sind hier im Bereich der nichtparametrischen Datenanalyse Fortschritte zu verzeichnen, die für die Varianzanalyse einen geschlossenen Auswertungs- und Interpretationsrahmen eröffnet.

Diese Vorstellung hat zu folgender Strukturierung der Kapitel geführt:

> Methoden, die auf der Normalverteilung basieren;
> Konsequenzen von Verletzungen der Annahmen;
> Alternativen.

Ein weiteres Bestimmungskriterium für diesen Text ist nicht zuletzt die Erkenntnis, dass nur das aktive Arbeiten dazu befähigt, Statistik zu lernen. Dies wird auch unterstützt durch die lernpsychologische Forschung, nach der Untersuchungen[1] gezeigt haben, dass der Mensch

[1] IZHD (1998,1999), zitiert nach: PD Dr. Nicole J. Saam: Einführung in die Statistik.

- 10 % von dem behält, was er nur liest,
- 20 % von dem, was er nur hört,
- 30 % von dem, was er nur beobachtet,
- 50 % von dem, was er hört und sieht,
- 70 % von dem, was er selber sagt,
- 90 % von dem, was er selber tut.

Die zahlreichen, auf realen Daten basierenden Beispiele können dementsprechend zwar als Muster dienen, sie können eigenständige Durchführung von Auswertungen nicht ersetzen. Da hierfür ein statistisches Auswertungspaket unerlässlich ist, wurde in der Vorstellung und Diskussion der Methoden auch ihre Umsetzung in die Statistik-Umgebung R integriert. R ist eine Umgebung, die über zahlreiche Funktionen zur statistischen Auswertung und zur grafischen Darstellung verfügt. Sie kann kostenfrei aus dem Internet heruntergeladen werden. Zugleich ist R selbst eine Programmiersprache. Der Standardset an Funktionen kann daher selbst weiterentwickelt werden. Es gibt eine große Anwendergemeinde, aus deren Reihe zahlreiche Ergänzungen geliefert werden. Das macht R zusätzlich attraktiv. Zudem ist R ein Dialekt der Programmiersprache S, dessen kommerzielle Variante S-Plus ebenfalls weit verbreitet ist. Zu S-Plus gibt es eine umfangreiche Literatur, die auch für R weitgehend ohne Änderungen verwendet werden kann.

Auf der anderen Seite ist R nicht ganz einfach zu erlernen. Dagegen gibt es zweierlei Therapien. Die eine, hier gewählte, verfolgt das Verabreichen in kleinen Dosen. Dazu werden im Text kleinere Auswertungen beispielhaft präsentiert. Eine Liste der Auswertungsprogramme findet sich am Ende des Buches. Im letzten Kapitel wird zudem eine Einführung in R gegeben. Nach dem Durcharbeiten des ganzen Buches sollte dann also der Einstieg geschafft sein. Um das eigene Arbeiten zu unterstützen sind die Datensätze sowie die Auswertungsprogramme auf meiner Homepage abgelegt. Der Link dahin ist:

http://www.rrz.uni-hamburg.de/IfStOek/indexSta.htm .

Um den zweiten Therapieansatz wenigstens zu benennen: Er besteht in der Schaffung einer Oberfläche, die es erlaubt, einfacher mit R umzugehen. Dies wurde im Rahmen des durch das BMBF geförderten Projektes 'Neue Statistik' realisiert. Zu dem Ergebnis, dem Statistik-Labor, gelangt man über die URL

http://www.statistiklabor.de .

Viel verdankt dieser Text dem Buch von Miller (1996). Dass er überhaupt geschrieben wurde, hängt damit zusammen, dass Millers Buch 1986 abgeschlossen wurde und schon daher viele neue Entwicklungen nicht mehr enthält. Zudem gibt es dort keine oder nur geringe Hintergrundinformationen zu den Methoden. Diese werden hier in hoffentlich adäquatem Umfang berücksichtigt. Denn ohne eine gute Basis kann kaum eine vernünftige Anwendung erreicht werden.

Wie üblich haben im Hintergrund zahlreiche Heinzelpersonen das Ihre zum Entstehen dieses Buches beigetragen. Ihnen allen möchte ich Dank sagen. Zuerst den Studierenden, die in mehreren Kursen die Erprobung des Textes aushalten mussten. Über den Einsatz in der Lehre ist speziell die endgültige Struktur gefunden worden. Dann haben Jesco Helde und Patrick Paulat gründlich Korrektur gelesen und Verbesserungsvorschläge gemacht. Ein ganz herzlicher Dank geht an Jörg Kaufmann. Er hat mir sein Manuskript 'Multiple Vergleiche bei ungleichen Stichprobenumfängen' zur Verfügung gestellt. Der Text hat davon wesentlich profitiert.
Weiterhin wäre das Buch nicht möglich gewesen ohne die Arbeit des LaTeX- und des R-Teams. Beide Teams arbeiten ehrenamtlich; ihre Arbeit kann nicht hoch genug eingeschätzt werden. Ihnen gilt mein besonderer Dank.

Rainer Schlittgen

Inhaltsverzeichnis

Kapitel 1

Empirische und theoretische Verteilungen

1 Explorative Datenanalyse

1.1 Daten

Ein Datensatz besteht aus den Angaben zu einer oder mehreren Merkmalen oder (statistischen) Variablen, die durch wiederholte Beobachtung gewonnen wurden. Bei univariaten Daten, d. h. wenn nur eine einzelne Variable beobachtet wird, reicht es, die beobachteten Werte hintereinander aufzuführen.

Beispiel 1.1 (Displayangebot)
Ein Supermarkt untersucht die Auswirkung eines Displayangebotes auf den Umsatz von Produkten, siehe Büning u. a. (1981). Für $n = 80$ Produkte wurde während eines festen Angebotszeitraumes jeweils der Quotient

$$y_v = \frac{\text{Umsatz im Displayangebot}}{\text{Umsatz im regulären Regalangebot}} \quad (v = 1, \ldots, 80)$$

betrachtet. Die Daten sind:

1.20	0.89	0.88	0.80	1.36	0.99	1.05	0.83	1.38	1.10	1.08	0.88
1.40	1.14	1.15	1.00	1.41	1.25	1.20	1.05	1.46	1.33	1.28	1.07
1.55	1.36	1.36	1.09	1.86	1.39	1.43	1.15	1.89	1.61	1.50	1.22
1.99	1.65	1.55	1.32	2.24	1.80	1.63	1.34	2.33	1.80	1.75	1.34
2.40	1.80	1.77	1.35	2.55	1.89	1.78	1.36	5.06	1.90	1.94	1.39
7.87	2.06	2.10	1.40	10.48	2.23	2.19	1.46	2.37	5.32	1.46	3.65
1.53	1.54	1.60	1.63	1.70	1.80	4.50	6.18				

■

Besteht jede Beobachtung aus mehreren Werten, so werden die Daten als Datenmatrix $\mathbf{Y}$ organisiert. Deren Spalten sind durch die Variablen $Y_1, \ldots, Y_p$ gegeben, und

die Zeilen durch die Beobachtungswiederholungen, Objekte bzw. Subjekte $o_1, \ldots, o_n$. Die Beobachtungen zu einer Variablen werden also stets untereinander geschrieben.

$$\begin{array}{c} \\ o_1 \\ o_2 \\ \vdots \\ o_v \\ \vdots \\ o_n \end{array} \begin{array}{c} \begin{array}{cccccc} Y_1 & Y_2 & \ldots & Y_j & \ldots & Y_p \end{array} \\ \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_{v1} & y_{v2} & \cdots & y_{vj} & \cdots & y_{vp} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nj} & \cdots & y_{np} \end{pmatrix} \end{array}$$

Von einem Datensatz wird gesprochen, wenn zusätzlich zu den Werten der Variablen noch die Angaben zu den Variablen vorliegen und u.U. auch Bezeichnungen der Beobachtungen. Letztere werden oft als *Labels* bezeichnet.

Beispiel 1.2 (Firmenbewertung)
Im November 2001 veröffentlichte die Zeitschrift 'DMEuro' folgende Ergebnisse aus einer europäischen Studie, in der Firmen nach verschiedenen Kennziffern bewertet wurden. Hier stellen die Firmen die Beobachtungswiederholungen dar und die Kennziffern die Variablen bzw. Merkmale; als Labels wären die Firmennamen selbst vorstellbar.

Land : Hauptsitz der Firma
Inter : Internationalität (gemessen als außereuropäischer Umsatzanteil in Prozent)
Finanz : Finanzkraft (gemessen in Eigenkapitalquote in Prozent)
Inno : Innovationsfähigkeit (gemessen in durchschnittlichen Patenten der vergangenen Jahre)
Marke : Markenstärke (Kennzahl gemäß Fortune-Ranking)
Gesamt : Gesamtbewertung in der Studie

In der folgenden Form sind die Daten als Datei im ASCII-Format abgelegt:

Land	Inter	Finanz	Inno	Marke	Gesamt
D	45	30.2	3678	6.23	100
GB	64	51.0	90	7.63	96
NL	55	55.4	1048	5.85	95
CH	67	44.1	121	7.64	93
GB	54	49.2	228	7.70	91
D	65	20.4	1239	6.48	83
D	44	33.8	926	6.96	80
F	36	36.5	563	6.94	72
F	46	34.6	37	6.50	72
NL	58	14.6	320	6.82	69
D	40	24.5	182	5.72	61
E	35	29.6	9	5.74	60
D	31	13.5	750	6.40	58
D	42	13.1	540	5.62	57
I	23	36.7	55	5.45	57

F	33	18.8	234	5.23	52
D	22	14.8	11	5.66	44
I	27	13.5	122	4.35	41
F	9	21.6	197	4.97	40
D	25	26.4	280	0.00	31

■

R-Code 1.1 (Eingeben und -lesen, Anzeigen und Speichern von Daten)

```
y<-c(1.20, 0.89,0.88,0.80,1.36,0.99,1.05,0.83,1.38,1.10,1.08,0.88,
1.40,1.14,1.15,1.00,1.41,1.25,1.20,1.05,1.46,1.33,1.28,1.07,1.55,
1.36,1.36,1.09,1.86,1.39,1.43,1.15,1.89,1.61,1.50,1.22,1.99,1.65,
1.55,1.32,2.24,1.80,1.63,1.34,2.33,1.80,1.75,1.34,2.40,1.80,1.77,
1.35,2.55,1.89,1.78,1.36,5.06,1.90,1.94,1.39,7.87,2.06,2.10,1.40,
10.48,2.23,2.19,1.46,2.37,5.32,1.46,3.65,1.53,1.54,1.60,1.63,1.70,
1.80,4.50,6.18)
y <- scan("c:\\daten\\display.txt")
save(y,file="c:/daten/display.RData")
load("c:/daten/display.RData")
firmen <- read.table("c:/daten/firmen2001.dat",header=TRUE)
```

Bei den ersten Zeilen des Codes geht es um einen univariaten Datensatz. In der ersten Zeile werden die Daten mittels der Funktion `c` direkt eingegeben.

Dann wird unterstellt, dass die Displaydaten in der ASCII-Datei 'display.txt' gespeichert sind. Die einzelnen Werte sind durch Leerzeichen bzw. Zeilenumbrüche voneinander getrennt. Durch `scan` werden sie hintereinander eingelesen. Zu beachten ist, dass der Backslash (\) innerhalb der Anführungsstriche doppelt vorkommen muss. Stattdessen kann bei der Pfadangabe auch ein einzelner Schrägstrich (/) eingegeben werden. (Dies hat seinen Hintergrund im UNIX-Betriebssystem; es funktioniert aber auch unter Windows.)

Der eingelesene Datenvektor wird mit `<-` (zusammengesetzt aus < und -) der Variablen `y` zugewiesen. Der Inhalt von `y` wird durch Aufruf des Namens bzw. durch `print(y)` angefordert. angegeben. Wie viele der Daten bei der Ausgabe jeweils auf eine Zeile geschrieben werden, ist dynamisch und hängt von der Breite des R-Konsolenfensters ab.

Die Daten können mit der Variablenbezeichnung als R-Datensatz gespeichert werden. Die Dateierweiterung `RData` ist obligatorisch; wird kein absoluter Pfad angegeben, so wird die Datei im aktuellen Arbeitsverzeichnis angelegt.

In einer späteren Sitzung erhält man dann die Variable `y` mit den darin gespeicherten Werten durch den `load`-Befehl zurück.

Mit den letzten beiden Zeilen wird ein Datensatz eingelesen und angezeigt. Die Daten sind in einer ASCII-Datei gespeichert. Die Werte sind spaltenweise angeordnet und durch Leerzeichen voneinander getrennt.

Das Einlesen erfolgt durch den Aufruf von `read.table`. Es wird `header=TRUE` gesetzt, weil in der ersten Zeile der Datei die Variablennamen angegeben sind. Andernfalls müsste `header=FALSE` angegeben werden.

Die Anzeige, die auch einfach mit dem Aufruf `firmen` erfolgen kann, hat dann die in Tabelle des Beispiels 1.2 angegebene Gestalt, nur dass vorne eine Spalte mit fortlaufender Nummerierung hinzugefügt wird.

1.2 Darstellung der Daten

Bei der Datenaufbereitung erlauben Grafiken, einen ersten Gesamtüberblick über die Daten zu erhalten.

Die einfachste Darstellungsweise für univariate Daten bilden *Stabdiagramme.* Hier wird über jedem beobachteten Wert ein senkrechte Linie gezeichnet, deren Höhe die Häufigkeit repräsentiert, mit der dieser Wert im Datensatz auftritt. Meist werden Stabdiagramme für diskrete Variablen gezeichnet, d.h. wenn es nur wenige unterschiedliche Beobachtungswerte gibt. Aber auch bei so genannten stetigen Variablen, bei denen im Prinzip jeder Wert aus einem Intervall möglich ist, ist diese Darstellungsform sinnvoll.

Bei stetigen Variablen können die als Dezimalzahlen vorliegenden Beobachtungen zunächst halbgrafisch in Form eines *Stem-and-Leaf-Diagramms* dargestellt werden. Dazu wird die jeweils führende Ziffer eines Beobachtungswertes (der Stamm) links von einem senkrechten Strich aufgetragen und die nachfolgende zweite rechts davon (ein Blatt). Weitere Ziffern werden vernachlässigt. Die 'Blätter' werden zeilenweise aufsteigend geordnet. Nicht vorhandene nachfolgende Ziffern dürfen nicht zu einer Verkürzung des Stammes führen; ggf. entsteht rechts einfach eine Lücke.

Modifikationen sind leicht möglich. Eine davon ist die Aufteilung der führenden Ziffer auf zwei bzw. fünf oder zehn Zeilen. Diese möglichen Aufteilungen ergeben sich daraus, dass für jede Zeile stets die gleiche Anzahl von zweiten Ziffern möglich sein muss, um eine systematische Verzerrung der Darstellung auszuschließen.

Untersuchungen aus dem Bereich der Wahrnehmungspsychologie haben zu der folgenden Empfehlung für die Anzahl der Zeilen geführt. Dabei ist zu berücksichtigen, dass eine eventuell nicht besetzte Zeile nicht einfach weggelassen werden darf:

$$\text{Anzahl der Zeilen im Stem-and-Leaf-Diagramm} \approx 10 \cdot \log_{10}(n). \tag{1.1}$$

Beispiel 1.3 (Displayangebot - Fortsetzung)
Die Displaydaten des Beispiels 1.1 führen zu dem in der Abbildung 1.1 wiedergegebenen, mit dem Programm R erzeugten Stem-and-Leaf-Diagramm.
Das dargestellte Stem-and-Leaf-Diagramm hat 11 Zeilen; die Faustregel empfiehlt $10 \cdot \ln(80) \approx 19$ Zeilen. So wäre es günstiger, die Zeilen jeweils noch einmal aufzuteilen, so dass die zweiten Ziffern 0 bis 4 in der ersten und die Ziffern 5 bis 9 in der zweiten der beiden zu einer gemeinsamen ersten Ziffer gehörenden Zeilen zu stehen kämen. ∎

Bei stetigen Variablen sind *Histogramme* die Standardform der Darstellung. Das Histogramm geht von einer vorgegebenen Klasseneinteilung $y_0^* < y_1^* < \ldots < y_m^*$ aus und ist so definiert, dass der Flächeninhalt über einer Klasse proportional zur relativen Häufigkeit der Beobachtungen in dieser Klasse ist. Es nimmt für eine Stichprobe $y_1, \ldots, y_n$ die folgende Form an:

```
The decimal point is at the |
 0 | 88999
 1 | 00111111111222333333444444444444555555666666678888888899999
 2 | 01122233445
 3 | 6
 4 | 5
 5 | 13
 6 |
 7 | 9
 8 |
 9 |
10 | 5
```

Abbildung 1.1: Stem-and-Leaf-Diagramm

$$\tilde{p}(y) = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} I_{(y_{k-1}^*, y_k^*]}(y_i)\frac{1}{y_k^* - y_{k-1}^*} & \text{für } y_{k-1}^* < y \leq y_k^* \\ 0 & \text{sonst}. \end{cases} \tag{1.2}$$

Dabei erfaßt die Indikatorfunktion $I_{(y_{k-1}^*, y_k^*]}(y_i)$, ob y_i im Intervall $(y_{k-1}^*, y_k^*]$ liegt. Sie ist eins, wenn dies gilt und null sonst. Allgemein:

$$I_A(y) = \begin{cases} 1 & y \in A \\ 0 & \text{sonst} \end{cases} \tag{1.3}$$

Die Klassen werden oft gleich breit gewählt. Dann gilt für die Anzahl der Klassen die Richtlinie für die Anzahl der Zeilen eines Stem-and-Leaf-Diagramms entsprechend:

$$\text{Anzahl der Klassen} \approx 10 \cdot \log_{10}(n). \tag{1.4}$$

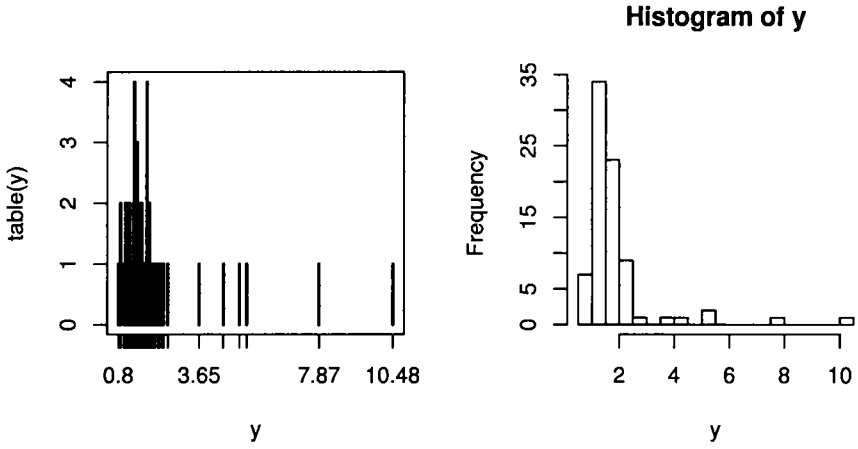

Abbildung 1.2: Displayangebot

Beispiel 1.4 (Displayangebot - Fortsetzung)
Für die Displayangebots-Daten, siehe das Beispiel 1.1, ergeben sich die in der Abbildung 1.2 dargestellten Diagramme. Das Stabdiagramm macht zumindest deutlich, dass die Beobachtungen im unteren Bereich stark konzentriert sind. Es gibt zudem nicht viele gleiche Werte. Größere Werte kommen nur vereinzelt vor. ■

R-Code 1.2 (Einfache Diagramme für univariate Daten)

```
stem(y)              # Stem-and-Leaf-Diagramm
par(mfrow=c(1,2))    # Halbieren des Grafik-Fensters
plot(table(y))       # Stabdiagramm
hist(y,breaks=19)    # Histogramm
```

Stem-and-Leaf-Diagramm, Stabdiagramm und Histogramm lassen sich in R sehr leicht erstellen. Die Displaydaten sind dabei in der Variablen y gespeichert. `stem` zählt nicht zu den eigentlichen Grafikfunktionen; die Ausgabe erfolgt auch im Konsolen- und nicht im Grafikfenster.
Für das Stabdiagramm wird mit dem dritten Aufruf zunächst mit `table(y)` eine Häufigkeitstabelle erstellt. Die grafische Darstellung mittels `plot` ergibt das Gewünschte. Beim Histogramm wird über `breaks=19` die Anzahl der Klassen festgelegt. Auch die explizite Angabe der Klassengrenzen wäre möglich. Die Grafik ist in der Abbildung 1.2 wiedergegeben.
Das # dient als Kommentarzeichen; alles, was auf einer Zeile hinter dem #-Zeichen steht, wird ignoriert.

Das Histogramm ist unstetig und hängt sehr stark von der Wahl der Klassengrenzen ab. Als naheliegende Verbesserung bietet sich daher an, die Klassen über den Bereich der Beobachtungen 'gleiten' zu lassen. Das führt bei einer festen Klassenbreite h zu

$$\tilde{p}(y) = \frac{1}{h \cdot n} \sum_{i=1}^{n} I_{(y-h/2,y+h/2]}(y_i) = \frac{1}{h \cdot n} \sum_{i=1}^{n} I_{(-1/2,1/2]}\left(\frac{y - y_i}{h}\right).$$

Das Resultat ist i.d.R. von sehr unruhiger Gestalt. Eine Verbesserung im Sinne eines glatteren Funktionsverlaufes erhält man durch die Ersetzung der Indikatorfunktion durch eine stetige Funktion, einen sogenannten *Kern* $K(u)$:

$$\hat{p}(y) = \frac{1}{h \cdot n} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right). \tag{1.5}$$

Dabei muss $K(u)$ die gleichen Eigenschaften wie $I_{(-1/2,1/2]}(u)$ haben, nämlich $K(u) \geq 0$ und $\int_{-\infty}^{\infty} K(u)\,dy = 1$. Diese beiden Eigenschaften zeichnen gerade Dichtefunktionen aus. Für $K(u)$ wird daher oft die Dichte der Standardnormalverteilung genommen. Eine andere Möglichkeit ist der Epanechnikov-Kern. Das Resultat wird als *Kerndichteschätzung* bezeichnet. Die Wahl der Bandbreite h ist ein ähnlich kritischer Punkt wie die Klassenbreite beim Histogramm. Hier gibt Silverman (1986, S.43ff) eine gute Diskussion.

Beispiel 1.5 (Displayangebot - Fortsetzung)
In der Abbildung 1.3 ist der Normalverteilungskern mit der Kerndichteschätzung der Displaydaten aus dem Beispiel 1.1 dargestellt. Die Bandbreite wurde automatisch bestimmt. ■

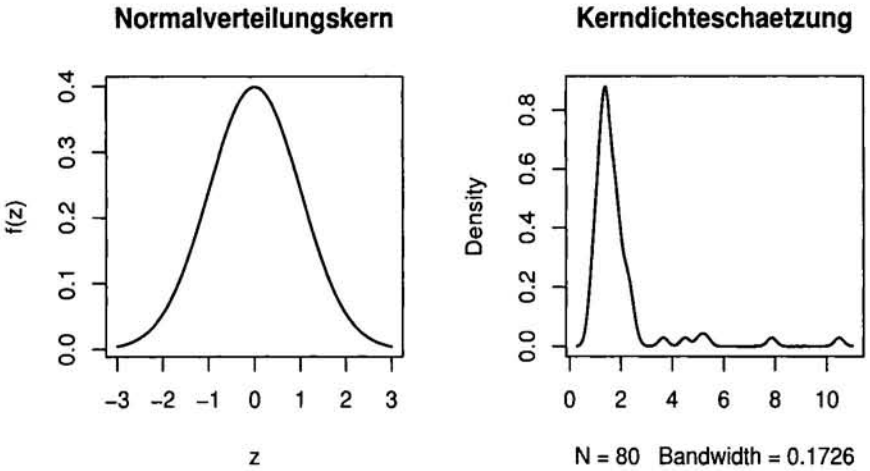

Abbildung 1.3: Displayangebot

R-Code 1.3 (Kerndichteschätzung)

```
d<-density(y,kernel="gaussian")
z<-seq(-3,3,.1)
f<-dnorm(z,mean=0,sd=1)
par(mfrow=c(1,2))
plot(z,f,type="l",lwd=1.5,ylab="f(z)",main="Normalverteilungskern")
plot(d,lwd=1.5,main="Kerndichteschaetzung")
```

Die Displaydaten sind in der Variablen y gespeichert. In der ersten Zeile wird die Kerndichteschätzung durchgeführt. Die Bandbreite wird nach einem geeigneten Kriterium automatisch bestimmt. Für die Darstellung werden die Punkte auf der Abszisse in der zweiten Zeile erzeugt und der Variablen `z` zugewiesen. Die Funktion `dnorm` bestimmt dann die zugehörigen Werte der Normalverteilungsdichte mit Erwartungswert 0 und Standardabweichung 1. Der Parameter `type="l"` verlangt Verbindungslinien, `main` legt die Überschrift fest. Die Anführungszeichen machen den Text als String oder Zeichenkette kenntlich. Das Resultat ist die Abbildung 1.3.

Bedeutsam ist weiterhin die grafische Darstellung der *empirischen Verteilungsfunktion* $\hat{F}(y)$. Sie gibt die relative Häufigkeit der Beobachtungen, die kleiner oder gleich

y sind, als Funktion von y an:

$$\hat{F}(y) = h(Y \leq y) = \frac{1}{n} \sum_{v=1}^{n} I_{(-\infty,y]}(y_v). \tag{1.6}$$

$\hat{F}(y)$ hat Sprungstellen bei den beobachteten Werten $y_1, \ldots, y_n$. Die geordneten Werte $y_{1:n} \leq y_{2:n} \leq \cdots \leq y_{n:n}$ sind gerade gleich den empirischen *p-Quantilen* für $p = 1/n, 2/n, \ldots, n/n$. Allgemein sind p-Quantile definiert durch:

$$y_p = \hat{F}^{-1}(p) = \inf\{y | \hat{F}(y) \geq p\}. \tag{1.7}$$

Die bis jetzt betrachteten Grafik-Typen stellen jeweils den kompletten Datensatz dar. *Box-and-Whisker-Plots*, auch kurz als Box-Plots bezeichnet, reduzieren diese Information, vermitteln dabei aber noch einen guten Eindruck von der Struktur der Beobachtungen. Sie basieren auf den extremsten Werten $y_{1:n}, y_{n:n}$, den Quantilen $y_{0.25}, y_{0.75}$ (auch unteres und oberes *Quartil* genannt) sowie dem *Median*[1]

$$\tilde{y} = \frac{1}{2}(y_{\lfloor (n+1)/2 \rfloor :n} + y_{\lfloor (n+2)/2 \rfloor :n}). \tag{1.8}$$

Beispiel 1.6 (Displayangebot - Fortsetzung)
Für die Displaydaten aus dem Beispiel 1.1 sind empirische Verteilungsfunktion und Box-and-Whisker-Plot in der Abbildung 1.4 zusammengefasst. Beim Box-Plot werden die extremsten Werte als einzelne Punkte dargestellt.

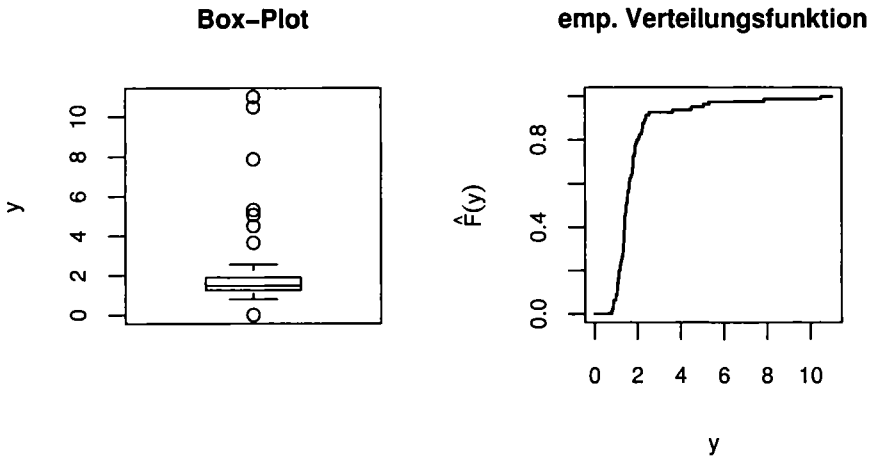

Abbildung 1.4: Displayangebot

■

[1]Hierbei bezeichnet $\lfloor y \rfloor$ die *Gaußsche Klammer*, d.h. die größte ganze Zahl $\leq y$.

R-Code 1.4 (Box-Plot und empirische Verteilungsfunktion)

```
par(mfrow=c(1,2))
boxplot(y,ylab="y",main="Box-Plot")
y<-sort(y)
y<-c(0,y,11)
F<-c(1:80)/80
F<-c(0,F,1)
plot(y,F,type="s",lwd=1.6,ylab=expression(hat(F)(y)),
main="emp. Verteilungsfunktion")
```

Für die Erstellung der empirischen Verteilungsfunktion werden die Daten erst einmal sortiert. Dann werden sie mittels `y<-c(0,y,11)` um etwas extremere Werte als das Minimum und das Maximum ergänzt, damit die Darstellung der Treppenfunktion an den Rändern nicht beim Minimum und Maximum endet. Natürlich muss auch `F` entsprechend ergänzt werden. Der Hauptpunkt beim Zeichnen der Treppenfunktion ist dann die Option `type="s"` in der `plot`-Funktion.
Wenn übrigens eine Zeile nicht mit einer vollständigen Anweisung abgeschickt wird, so erscheint auf der Folgezeile ein Pluszeichen. Dann kann der Befehl vervollständigt werden.

1.3 Maßzahlen der Lage und Streuung

Neben der grafischen Darstellung interessiert man sich für eine summarische Beschreibung eines Datensatzes. Hier sind die wichtigsten Charakteristika das Niveau und die Ausbreitung. Die üblichen Maßzahlen zur Beschreibung von Lage und Streuung der Werte einer Variablen Y sind das *arithmetische Mittel* $\bar{y}$ und die *Varianz* $s^2(y)$ bzw. die *Standardabweichung* $s(y)$:

$$\bar{y} = \frac{1}{n}\sum_{v=1}^{n} y_v, \qquad s^2(y) = \frac{1}{n-1}\sum_{v=1}^{n}(y_v - \bar{y})^2, \qquad s(y) = \sqrt{s^2(y)}. \tag{1.9}$$

Bei Datenanalysen möchte man oft mit Maßzahlen der Lage und Streuung arbeiten, die speziell von einzelnen extremen Werten, sogenannten *Ausreißern*, nicht zu stark beeinflusst werden. Maßzahlen, bei denen auch extreme Beobachtungen nur einen geringen Einfluss besitzen, werden als resistent (oder robust) bezeichnet. Zu solchen *resistenten Maßzahlen* gehören der Median $\tilde{y}$, das *getrimmte Mittel* $\bar{y}_\alpha$, der *Quartilsabstand* s_Q und der *MAD*, der *Median der absoluten Abweichungen vom Median.* Die drei letztgenannten Maßzahlen sind:

$$\bar{y}_\alpha = \frac{1}{n - 2[\alpha n]} \sum_{v=[\alpha n]}^{[(1-\alpha)n]} y_{v:n} \tag{1.10}$$

$$s_Q = y_{0.75} - y_{0.25} \tag{1.11}$$

$$\text{MAD} = \text{Median}\{|y_v - \tilde{y}| : \ v = 1, \ldots, n\} \tag{1.12}$$

Der Median und der MAD haben beide einen hohen *Bruchpunkt*; d.h. sie können einen großen Anteil von extremen Werten vertragen, bevor die Schätzungen vollkommen unbrauchbar werden. Dies wird in der Definition 2.10 genauer gefasst.

Zusätzlich wurde als sehr resistente Maßzahl für die Streuung noch die *kürzeste Länge des α-Teils* $s_{s(\alpha)}$ (s im Subskript steht für 'shortest') vorgeschlagen und als zugehörige Maßzahl für die Lage der Mittelpunkt $m_{s(\alpha)}$ der beiden geordneten Werte, welche den kürzesten α-Teil festlegen. Die Bezeichnung dieser Maßzahl entspricht gerade ihrer Definition:

$$s_{s(\alpha)} = \min\{y_{(v+w):n} - y_{v:n} | 1 \leq v \leq v + w \leq n, (w+1)/n \geq \alpha\} \tag{1.13a}$$

und für die dadurch bestimmten Indizes v und $v + w$:

$$m_{s(\alpha)} = \frac{1}{2}(y_{u:n} + y_{(u+w):n}). \tag{1.13b}$$

$s_{s(0.5)}$ wird auch als *kürzeste Hälfte* bezeichnet. Geeignete multivariate Verallgemeinerungen haben eine große Bedeutung bei der explorativen Datenanalyse erlangt.

R-Code 1.5 (Univariate Maßzahlen)

```
mean(y)                                # arithmetisches Mittel
var(y)                                 # Varianz
sd(y)                                  # Standardabweichung
median(y)                              # Median
mean(y,trim=0.1)                       # 10% getrimmtes Mittel
quantile(y,0.75)-quantile(y,0.25)      # Quartilsabstand
mad(y)                                 # MAD
y<-sort(y); n<-length(y); d<-matrix(y,n/2,2)
d1<-d[,2]-d[,1]
i<-c(1:(n/2)); i<-min(i[d1==min(d1)])
msh<-(d[i,1]+d[i,2])/2                 # Mitte der kürzesten Hälfte
ssh<-(1+15/(n-1))*(d[i,2]-d[i,1])      # kürzeste Hälfte mit Faktor
```

Zu beachten ist bei der in R implementierten Quantilsfunktion, dass generell linear interpoliert wird.
Für die mit der kürzesten Hälfte zusammenhängenden Maßzahlen ist ein etwas größerer Aufwand nötig. Zuerst werden die Beobachtungen aufsteigend sortiert, mit `length(y)` wird die Anzahl der Beobachtungen angefordert. R-Befehle werden durch einen Zeilenwechsel oder wie hier durch ein Semikolon getrennt. Der Befehl `matrix` erstellt aus dem Vektor `y` eine $(n/2, 2)$-Matrix. Da die Matrix spaltenweise aufgebaut wird, stehen in der ersten Spalte die Beobachtungen 1

bis 40 und in der zweiten die Beobachtungen 41 bis 80 des geordneten Datensatzes. So sind jeweils zwei Werte nebeneinander angeordnet, die gerade $n/2$ Beobachtungen voneinander entfernt sind. Die Zeile mit der kleinsten Differenz wird in der dritten Zeile bestimmt. Durch `i[d1==min(d1)]` werden die Komponenten des Zeilenindex `i` ausgewählt, die den kleinsten Abstand voneinander haben. Da dies mehrere sein können, wird noch einmal ein eindeutiger Index durch den zweiten `min`-Befehl angefordert. Die Maßzahlen ergeben sich dann in den beiden folgenden Zeilen vier und fünf. Das Gitter-Symbol # dient zum Einfügen von Kommentaren; ab diesem Zeichen bis zum Zeilenende wird bei der Ausführung von R alles ignoriert.

Diese wird oft mit dem Faktor $f = (1 + \frac{15}{n-1})$ versehen, um $s'_{s(0.5)} = f \cdot s_{s(0.5)}$ im Fall der Normalverteilung mit der Standardabweichung vergleichbar zu machen, siehe Grübel (1988).

Beispiel 1.7 (Display-Angebot - Fortsetzung)
Bei den bereits mehrfach betrachteten Display-Daten erhält man die folgenden Werte für die angesprochenen Maßzahlen:

$\bar{y}$	$\tilde{y}$	$\bar{y}_{0.1}$	$m_{s(0.5)}$	$s^2(y)$	$s(y)$	s_Q	MAD	$s'_{s(0.5)}$
1.87	1.48	1.57	1.34	2.07	1.44	0.617	0.474	0.69

Die Maßzahlen der Lage unterscheiden sich nicht unwesentlich. Alle robusten Lagemaße sind kleiner als das arithmetische Mittel. Das resultiert aus den einzelnen großen Werten. Die Maßzahlen der Streuung sind prinzipiell untereinander nicht vergleichbar, da sie auf unterschiedlichen Konzepten beruhen. ■

In der Praxis werden oft einfach das arithmetische Mittel und der Median berechnet. Unterscheiden sich die beiden nicht sehr, so ist man zufrieden und arbeitet mit dem arithmetischen Mittel weiter.

1.4 Bivariate Daten

Die grafische Darstellung bivariater Daten erfolgt in der Regel mittels *Streudiagrammen*, bei denen die Wertepaare als Punkte in einem Koordinatensystem dargestellt werden. Selten findet man grafische Darstellungen von zweidimensionalen Histogrammen oder bivariaten Kerndichteschätzungen.

Als eine Verallgemeinerung der Box-Plots können *konvexe Hüllen* angesehen werden. Eine konvexe Hülle ist ein geschlossener Streckenzug, der sich aus linearen Verbindungen von Randpunkten des Streudiagramms ergibt. Die Randpunkte werden dabei so bestimmt, dass alle Punkte des Streudiagramms sowie ihre paar weisen linearen Verbindungen in diesem Streckenzug liegen. Die äußere Hülle entspricht den Endpunkten der Whiskers beim Box- und Whiskers-Plot. Nun lässt man die auf ihr liegenden Punkte weg und konstruiert für die restlichen Daten erneut eine konvexe Hülle. Dieses Vorgehen wird fortgesetzt, bis gerade noch (i.d.R. etwas mehr als) 50%

der Daten übrig bleiben; die zugehörige konvexe Hülle ist dann ein Analogon zur Box.

Beispiel 1.8 (Gebrauchtwagen)
Die Angebote an gebrauchten 5erBMW's einer Ausgabe der Zeitung 'Zweite Hand', siehe Thadewald (1998), enthalten u.a. den Kilometerstand sowie die Preisvorstellung des Anbieters. In dem Streudiagramm 1.5 sind die Preise (in 1000 DM) gegen die km (in 1000) eingezeichnet. Die konvexen Hüllen machen deutlich, dass es einige 'extreme Angebote' gibt.

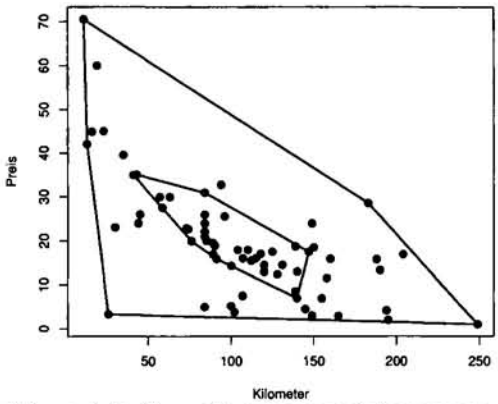

Abbildung 1.5: Streudiagramm mit konvexen Hüllen

■

R-Code 1.6 (Streudiagramm mit konvexen Hüllen)

```
y<-read.table("c:/daten/bmw5.txt")
plot(y,type="p",xlab="Kilometer",ylab="Preis",pch=16,cex=1.2)
n<-nrow(y)
c1<-chull(y)
lines(rbind(y[c1,],y[c1[1],]),lwd=2)
while (nrow(y) > n/2) {
c1<-chull(y) ;
if ((nrow(y)-length(c1))<n/2){break}
y <- y[-c1,] }
c1<-chull(y)
lines(rbind(y[c1,],y[c1[1],]),lwd=2)
```

Der Code zur Erstellung eines Streudiagrammes ist einfach. Zunächst also der einfache Teil. Zunächst werden die 67 Datenpaare in einen Datensatz gelesen. Beim Plot-Befehl gibt `pch=16` an, dass ausgefüllte Punkte geplottet werden. `cex` macht sie etwas größer, als es die Standardeinstellung vorsieht.

Für die Bestimmung der Punkte, die die Eckpunkte der konvexen Hüllen ausmachen, wird mit `nrow` zuerst die Anzahl der Zeilen des Datensatzes bestimmt. Der Befehl `chull(y)` ergibt die Indizes der Zeilen des Datensatzes, wo die zugehörigen Datenpunkte die Ecken der konvexen Hülle darstellen. Diese äußere konvexe Hülle wird dann dargestellt. `lines` fügt die Linien dem bereits erstellten Plot hinzu. Damit ein geschlossener Linienzug entsteht, muss der erste Punkt noch einmal unten angehängt werden. Dies geschieht mit `rbind`.
Das Ansprechen der Zeilen ist ein Zugriff auf den ersten Index eines zweidimensionalen Gebildes. Das wird durch die Angabe vor dem Komma in der eckigen Klammer ausgedrückt. Die erste Spalte würde entsprechend mit `y[,1]` ausgewählt. Das negative Vorzeichen der Indizes führt dazu, dass die entsprechenden Zeilen entfernt werden. Über eine Schleife, siehe S. 308, werden nun so lange die äußeren Datenpunkte entfernt, dass die restlichen immer noch mindestens die Hälfte der Beobachtungen ausmachen. Damit dies nicht unterschritten wird, wird die Schleife im letzten Durchgang mit `break` verlassen. Abschließend wird die innere konvexe Hülle dargestellt.

Die klassischen Maßzahlen zur Beschreibung der Lage bivariater Daten (x_v, y_v), $v = 1, \ldots, n$, ist einfach der *Schwerpunkt*, d.h. der Vektor $(\bar{x}, \bar{y})$ der beiden arithmetischen Mittel. Der Vektor der beiden Mediane, $(\tilde{x}, \tilde{y})$, ist der sogenannte *Medianpunkt*. Er ist ein resistentes Lagemaß.

Zur Erfassung der Streuung bivariater Daten nimmt man zunächst wieder die univariaten Maßzahlen. Diese berücksichtigen aber nicht das gemeinsame Streuungsverhalten. Die Maßzahl dafür ist die *Kovarianz* $s(x, y)$ bzw. die standardisierte Version, der *Korrelationskoeffizient* $r(x, y)$:

$$s(x, y) = \frac{1}{n-1} \sum_{v=1}^{n} (x_v - \bar{x})(y_v - \bar{y}), \tag{1.14a}$$

$$r(x, y) = \frac{s(x, y)}{\sqrt{s^2(x) s^2(y)}}. \tag{1.14b}$$

Natürlich gilt $s(x, y) = s(y, x)$. Für $s^2(x)$ schreibt man wegen $s^2(x) = \frac{1}{n-1} \sum_{v=1}^{n} (x_v - \bar{x})(x_v - \bar{x})$ auch $s(x, x)$.
Varianzen und Kovarianz werden in der *Kovarianzmatrix* zusammengefasst:

$$\mathbf{S} = \begin{pmatrix} s(x, x) & s(x, y) \\ s(x, y) & s(y, y) \end{pmatrix}. \tag{1.15}$$

Analog ist die Korrelationsmatrix definiert.

Der Korrelationskoeffizient erfüllt die Ungleichung

$$-1 \leq r(x, y) \leq 1, \tag{1.16}$$

wobei das Gleichheitszeichen genau dann gilt, wenn alle Punkte auf einer Geraden liegen. Die Korrelation erfasst also den linearen Zusammenhang.

Als Faustregel für die Interpretation des Korrelationskoeffizienten verwendet man häufig die folgende Einteilung:

$\lvert r(x,y)\rvert$		Interpretation
0	:	keine Korrelation
0-0.5	:	schwache oder niedrige Korrelation
0.5-0.8	:	mittlere Korrelation
0.8-1	:	starke oder hohe Korrelation
1	:	perfekte Korrelation

Beispiel 1.9 (Gebrauchtwagen - Fortsetzung)
Für die Kilometerzahlen der gebrauchten 5erBMW ergibt sich ein Durchschnitt von 105.13; der durchschnittlich geforderte Preis beträgt 19.82. Weiter erhält man:

	Kovarianzmatrix			Korrelationsmatrix	
	Km	DM		Km	DM
Km	2727.40	-465.38	Km	1.00	-0.68
DM	-465.38	173.18	DM	-0.68	1.00

Preis (DM) und gefahrene Kilometer (Km) sind negativ korreliert; mehr gefahrene Kilometer bedeuten einen niedrigeren Verkaufswert. Allerdings ist die Korrelation nur als mittelstark einzustufen. Es gibt ja auch andere Bestimmungsfaktoren für den Wiederverkaufswert; dazu zählt etwa die Ausstattung. ■

Wie die anderen klassischen Maßzahlen zur Lage und Streuung sind auch $s(x,y)$ und $r(x,y)$ sehr empfindlich in Bezug auf Ausreißer. Daher verwendet man in der Praxis oft den *Rangkorrelationskoeffizienten* r_{Spear} von Spearman, um die Stärke des Zusammenhanges zu erfassen. Er ist der übliche Korrelationskoeffizient, berechnet für die Rangwerte der Beobachtungen. *Rangwerte* sind dabei die "Platznummern" der aktuellen Beobachtungen. Sofern gleiche Beobachtungswerte vorliegen, wird über die Rangwerte gemittelt.

Aufgrund der Rangtransformation ist der Wert von $r_{Spear}(x,y)$ nicht mehr als Stärke des linearen Zusammenhanges interpretierbar. Der Rangkorrelationskoeffizient misst vielmehr die Stärke eines monotonen Zusammenhanges.

Beispiel 1.10
In der Tabelle sind fünf Beobachtungspaare angegeben.

Y_1	Y_2	$R(Y_1)$	$R(Y_2)$
3.0	1.5	2	1
4.4	5.0	3.5	3
2.8	9.3	1	5
5.1	5.0	5	3
4.4	5.0	3.5	3

In der Spalte $R(Y_1)$ sind die Rangwerte der Beobachtungen von Y_1 gelistet. Da der Wert 4.4 zweimal vorkommt, erhalten beide den Mittelwert der fortgezählten Rang-

zahlen zugeordnet. Bei drei gleichen, wie in der zweiten Spalte, bekommen alle drei den zugehörigen mittleren Rang zugewiesen. Aus den Werten der dritten und vierten Spalte werden nun Paare gebildet, für die der Korrelationskoeffizient berechnet wird. Man erhält $r_{Spear} = -0.23$. Der Korrelationskoeffizient der ursprünglichen Werte ist $r = -0.15$. ■

R-Code 1.7 (Bivariate Maßzahlen)

```
mean(y)                          # Vektor der Mittelwerte
var(y)                           # Kovarianzmatrix
cor(y)                           # Korrelationsmatrix
apply(y,2,median)                # Vektor der Mediane
cor(rank(y[,1]),rank(y[,2]))     # Rangkorrelationsmatrix
cor(apply(y,2,rank))             # Rangkorrelationsmatrix
```

Die Funktion `median(y)` ist nur für univariate Daten implementiert. Daher muss mit dem Befehl `apply` gearbeitet werden. Er dient zur Anwendung einer Funktion auf einzelne Spalten oder Zeilen, je nachdem ob das zweite Argument eine 2 oder eine 1 ist. Die Befehle auf den beiden letzten Zeilen sind gleichwertig; sie ergeben beide den Rangkorrelationskoeffizienten.

2 Zufallsvariablen und Verteilungen

2.1 Wahrscheinlichkeitsverteilungen

Der Anteil der Elemente der Grundgesamtheit, bei denen ein Merkmal Y einen bestimmten Wert y hat, ist die Wahrscheinlichkeit dafür, dass bei einer einfachen zufälligen Auswahl eines Elementes aus der Grundgesamtheit eines gezogen wird, bei dem Y gerade diesen Wert hat. Das führt auf die *Wahrscheinlichkeitsfunktion* $f(y) = \mathrm{P}(Y = y)$; diese ist nur für $y_1 \leq y_2 \leq \ldots \leq y_k$ von null verschieden, wenn die y_i die unterschiedlichen Realisationsmöglichkeiten von Y sind.

Entsprechend gibt die empirische Verteilungsfunktion $\hat{F}(y)$ des Merkmals Y in der Grundgesamtheit an der Stelle y die Wahrscheinlichkeit an, dass bei einer einfachen Zufallsauswahl ein Element aus der Grundgesamtheit mit $y_i \leq y$ gezogen wird. Folglich wird diese spezielle empirische Verteilungsfunktion hier zur *theoretischen Verteilungsfunktion*:

$$F(y) = \mathrm{P}(Y \leq y)\,.$$

Die Verteilungsfunktionen stellen die Informationen dar, welche der statistischen Betrachtung zugänglich sind. Die Verbindung der Werte von Y mit den konkreten 'Merkmalsträgern' oder Personen in einer Grundgesamtheit wird bei der statistischen Betrachtung außer acht gelassen. Nunmehr wird die Verteilungsfunktion $F(y)$ als 'Grundgesamtheit' angesehen.

Mit dieser Idealisierung können für $F(y)$ insbesondere auch stetige Funktionen verwendet werden. Dies ist etwa für Messungen aller Art sinnvoll. Hier unterstellt man gerne, dass die Messgenauigkeit im Prinzip beliebig hoch sei, um stetige Verteilungen zu rechtfertigen. Hier wird dies pragmatischer gesehen; als Modelle brauchen Verteilungfunktionen nur gute Approximationen der Realität darzustellen.

Stetige Verteilungsfunktionen lassen sich mittels Integralen angeben:

$$F(y) = \int_{-\infty}^{y} f(u)du. \tag{1.17}$$

Die zu integrierende Funktion wird als *Dichtefunktion* oder kurz *Dichte* bezeichnet. Die Angabe von $f(y)$ und $F(y)$ ist gleichwertig.

Im Idealfall ergeben sich Verteilungen aus einfachen Modellannahmen.

Beispiel 1.11 (Pareto-Verteilung)
Die *Pareto-Verteilung* ist ein Beispiel für eine stetige Verteilung, die aus einfachen Annahmen ableitbar ist. Paretos Einkommensgesetz, das er aus empirischen Untersuchungen gewonnen hat, besagt, dass die Verteilung des Einkommens von Personen mit einem Mindesteinkommen y beschrieben werden kann durch

$$N_y = Ay^{-\alpha}.$$

Hier ist N_y die Anzahl der Personen, deren Einkommen $\geq y$ ist, und A, α sind populationsspezifische Konstanten. Nun wird eine Einkommensgrenze k festgehalten und der Anteil der Einkommensbezieher mit einem Einkommen $\geq y$ an denen, die ein Einkommen von mindestens k haben, betrachtet:

$$\frac{N_y}{N_k} = \left(\frac{k}{y}\right)^{\alpha}.$$

Somit ist die zugehörige Verteilung durch die Verteilungsfunktion gegeben:

$$F(y) = \begin{cases} 0 & \text{für } y < k \\ 1 - \left(\frac{k}{y}\right)^{\alpha} & \text{für } y \geq k. \end{cases}$$

Die Dichte erhält man daraus als Ableitung von $F(y)$:

$$f(y) = F'(y) = \begin{cases} 0 & \text{für } y < k \\ \alpha \frac{k^{\alpha}}{y^{\alpha+1}} & \text{für } y \geq k. \end{cases}$$

■

Da der Zugang zu Verteilungsmodellen über einfache Modellannahmen eher selten gelingt, und nicht für jedes empirische Phänomen eine eigene Verteilung ermittelt

werden kann, greift man auf idealtypische Verteilungsfunktionen zurück. Das sind dann solche, die häufig empirische Phänomene hinreichend gut erfassen. Bevorzugt sind dabei Verteilungen, die von wenigen Parametern abhängen, so dass mit der Adjustierung der Parameter schon durch einen Verteilungstyp viele empirische Phänomene erfasst werden können. In der statistischen Praxis stellt sich dann die Aufgabe, eine der bekannten Verteilungen auszuwählen, die mit den Daten hinreichend gut übereinstimmt. Diese Verteilungen werden häufig nicht durch die Verteilungsfunktion $F(y)$ angegeben, sondern durch dazu äquivalente Beschreibungen, vor allem durch Wahrscheinlichkeitsfunktionen bzw. Dichtefunktionen.

Beispiel 1.12 (Normalverteilung)
Ein oft verwendeter Verteilungstyp ist die *Normalverteilung* mit der Dichte

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right). \tag{1.18}$$

Wenn ausgedrückt werden soll, dass Y eine Normalverteilung mit den Parametern μ und σ^2 besitzt, wird $Y \sim \mathcal{N}(\mu,\sigma^2)$ geschrieben. ■

R-Code 1.8 (Normalverteilung)

```
pnorm(y,0,1) # Verteilungsfunktion
dnorm(y,0,1) # Dichtefunktion
qnorm(p,0,1) # Quantile
rnorm(n,0,1) # Zufallszahlen
```

R verfügt über eine beträchtliche Anzahl von Verteilungen, siehe Seite 323. Es können jeweils die Dichte und die Verteilungsfunktion für einen Vektor vorgegebener Werte, Quantile für einen Vektor von Anteilen aufgerufen sowie Zufallszahlen generiert werden. Für jeden Verteilungstyp unterscheiden sich die einzelnen Funktionen nur durch den typischen ersten Buchstaben. y ist jeweils ein Vektor von möglichen Werten, d.h. hier von reellen Zahlen. Die Werte 0 und 1 sind Voreinstellungen für die Parameter μ und σ (nicht von σ^2!). Sie können geändert werden; dann sind Werte für beide anzugeben. Die Komponenten des Vektors p müssen die Bedingung $0 < p < 1$ erfüllen. n ist schließlich eine positive ganze Zahl, die die Anzahl der zu erzeugenden Zufallszahlen angibt.
Das Gitter-Symbol # dient zum Einfügen von Kommentaren; ab diesem Zeichen bis zum Zeilenende wird bei der Ausführung von R alles ignoriert.

Dass die Normalverteilung eine so große Rolle in der Anwendung spielt, hängt u.a. damit zusammen, dass über die Parameter μ und σ^2 die Verteilung verschoben und gestaucht bzw. gestreckt werden kann. Dies ergibt sich allgemein auf folgende Weise.

Y habe die Verteilung mit der Verteilungsfunktion $F(y)$ und der Dichte $f(y)$. Dann gilt für die linear transformierte Zufallsvariable $V = a + bY$, $b > 0$:

$$\begin{aligned} F_V(v) &= \mathrm{P}(V \leq v) = \mathrm{P}(a + bY \leq v) = \mathrm{P}\left(Y \leq \frac{v-a}{b}\right) \\ &= F\left(\frac{v-a}{b}\right). \end{aligned} \tag{1.19}$$

Damit ist die Dichte

$$f_V(v) = F_V'(v) = F'\left(\frac{v-a}{b}\right) = f\left(\frac{v-a}{b}\right)\frac{1}{b}. \tag{1.20}$$

Beispiel 1.13 (Lineartransformation bei Normalverteilung)
Startet man mit einer $\mathcal{N}(0,1)$-Verteilung,

$$f(z;0,1) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right),$$

so führt die Lineartransformation $Y = \mu + \sigma Z$ zu der in (1.18) angegebenen allgemeinen Form der Dichtefunktion. Alle Normalverteilungen lassen sich also aus der *Standardnormalverteilung* $\mathcal{N}(0,1)$ durch Lineartransformation gewinnen. Die Verteilungsfunktion der $\mathcal{N}(0,1)$-Verteilung wird wie üblich mit $\Phi(z)$ bezeichnet und die zugehörige Dichte mit $\phi(z)$. ∎

In dem R-Code 1.8 ist auch der Befehl zur Erzeugung von normalverteilten *Zufallszahlen* angegeben. Solche Zufallszahlen sind u.a. wichtig, um statistische Verfahren zu untersuchen. Damit können nämlich die Randbedingungen für die Verfahren vorgegeben werden.

Gleichverteilte Zufallszahlen werden zwar mittels mathematischer Algorithmen erzeugt, verhalten sich aber regellos und jede Ziffer kommt etwa gleich häufig vor. Sie sind natürlich nicht genau zufällig; nach einer langen Periode wiederholen sie sich. Man spricht deshalb auch von Pseudozufallszahlen. Eine allgemeine Methode zur Erzeugung von nicht gleichverteilten Zufallszahlen basiert auf dem folgenden Satz.

Satz 1.14 (Wahrscheinlichkeitsintegraltransformation)
Sei Y eine Zufallsvariable mit einer stetigen Verteilungsfunktion $F(y)$, die für $0 < F(y) < 1$ streng monoton wachsend ist. Dann ist die Zufallsvariable $U = F(Y)$ gleichverteilt über dem Intervall $[0,1] : U \sim \mathcal{R}(0,1)$.
Umgekehrt hat $Y = F^{-1}(U)$ die Verteilungsfunktion $F(y)$ falls U über dem Intervall $(0,1)$ gleichverteilt ist.

Beweis: Es ist

$$\mathrm{P}(U \leq u) = \mathrm{P}(F(Y) \leq u) = \mathrm{P}(Y \leq F^{-1}(u)) = F(F^{-1}(u)) = u.$$

Damit ist der erste Teil gezeigt. Der zweite folgt analog. ∎

Beispiel 1.15 (Pareto-Verteilung)
Bei der Pareto-Verteilung ergibt sich für $0 < p < 1$:

$$p = F(y_p) = 1 - \left(\frac{k}{y_p}\right)^{\alpha} \Longrightarrow y_p = \frac{k}{(1-p)^{1/\alpha}}.$$

Daher sind Pareto-verteilte Zufallszahlen bei vorgegebenen Parametern k und α dadurch zu erhalten, dass $\mathcal{R}(0,1)$-Zufallszahlen u erzeugt und gemäß $k/(1-u)^{1/\alpha}$ transformiert werden. ∎

2.2 Gestaltparameter von Verteilungen

In der überwiegenden Zahl der Fälle ist man nicht an der gesamten Verteilung interessiert, sondern nur an Form- oder Gestaltparametern von Verteilungen. Dies sind allgemeine Maßzahlen, welche die Gestalt einer Verteilung charakterisieren, wie speziell die Lage, die Ausbreitung und die Schiefe. Diese Maßzahlen hängen i.d.R. mit den Verteilungsparametern zusammen. Auch darüber lassen sich die einzelne Vertreter aus den 'Familien' von Wahrscheinlichkeitsverteilungen festlegen. Sie werden daher oft etwas nachlässig ebenfalls als Parameter bezeichnet. Die gebräuchlichsten dieser Formparameter lassen sich über die *Momente von Verteilungen* definieren.

Allgemein werden die *Momente* definiert durch:

$$\mu_r = \begin{cases} \sum\limits_{y_i} y_i^r f(y_i) & \text{falls } Y \text{ diskret} \\ \int\limits_{-\infty}^{\infty} y^r f(y)dy & \text{falls } Y \text{ stetig}. \end{cases} \tag{1.21}$$

Das erste Moment ist der *Erwartungswert*: $\mathrm{E}(Y) = \mu_1$. Dafür wird einfach μ geschrieben.

Der Erwartungswert ist der am meisten verwendete theoretische Lageparameter.

Beispiel 1.16 (Erwartungswert der Poisson-Verteilung)
Y sei Poisson-verteilt, habe also die Wahrscheinlichkeitsfunktion

$$f(y) = e^{-\lambda} \cdot \frac{\lambda^y}{y!} \qquad (y = 0, 1, 2, \ldots).$$

Den Erwartungswert von Y erhält man leicht unter Ausnutzen der Eigenschaft einer Wahrscheinlichkeitsfunktion, dass die Summe über alle Realisationsmöglichkeiten eins ergibt:

$$\begin{aligned} \mathrm{E}(Y) &= \sum_{y=0}^{\infty} y \cdot e^{-\lambda}\frac{\lambda^y}{y!} = \lambda \sum_{y=1}^{\infty} e^{-\lambda}\frac{\lambda^{y-1}}{(y-1)!} = \lambda \sum_{z=0}^{\infty} e^{-\lambda}\frac{\lambda^z}{z!} \\ &= \lambda \cdot 1 = \lambda. \end{aligned}$$

∎

Der *Erwartungswert einer Funktion* $g(Y)$ einer Zufallsvariablen ist

$$\mathrm{E}(g(Y)) = \begin{cases} \sum\limits_{y_i} g(y_i)f(y_i) & \text{falls } Y \text{ diskret} \\ \int\limits_{-\infty}^{\infty} g(y)f(y)dy & \text{falls } Y \text{ stetig}\,. \end{cases} \tag{1.22}$$

Damit sind die Momente Erwartungswerte spezieller Funktionen der Zufallsvariablen.

Für eine wichtige, im Folgenden oft verwendete Eigenschaft des Erwartungswertes werden mehrdimensionale Dichten bzw. Wahrscheinlichkeitsfunktionen benötigt. Im zweidimensionalen Fall lassen sich diese noch anschaulich vorstellen. Bei der bivariaten Dichte $f_{\mathbf{y}}(y_1, y_2)$ des Zufallsvektors $\mathbf{y} = (Y_1, Y_2)'$ ist das durch $f_{\mathbf{y}}(y_1, y_2)$ begrenzte Volumen über einem Rechteck auf der (y_1, y_2)-Ebene gleich der Wahrscheinlichkeit, dass der Zufallsvektor einen Wert aus diesem Rechteck annimmt:

$$\mathrm{P}(a < Y_1 \leq b, c < Y_2 \leq c) = \int_a^b \int_c^d f_{\mathbf{y}}(y_1, y_2)dy_1dy_2.$$

Die *Randverteilungen* sind gegeben durch

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{\mathbf{y}}(y_1, y_2)\, dy_2\,, \quad f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f_{\mathbf{y}}(y_1, y_2)\, dy_1\,.$$

Y_1 und Y_2 sind unabhängig, wenn $f_{\mathbf{y}}(y_1, y_2) = f_{Y_1}(y_1)f_{Y_2}(y_2)$.

Die Erweiterung auf höhere Dimensionen ist - zumindest formal - nicht schwierig. Allgemein werden die Zufallsvariablen $Y_1, \ldots, Y_p$ als Komponenten eines *Zufallsvektors* betrachtet. Dafür wird geschrieben

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}, \quad \text{bzw.} \quad \mathbf{y} = (Y_1, \ldots, Y_p)'.$$

Wie üblich wird dann notationell nicht mehr zwischen dem Zufallsvektor $\mathbf{y} = (Y_1, \ldots, Y_p)'$ und dem Vektor der Realisationen $\mathbf{y} = (y_1, \ldots, y_p)'$ unterschieden. Worum es sich jeweils handelt, muss aus dem Kontext erschlossen werden.

Eine multivariate Dichte oder Wahrscheinlichkeitsfunktion $f_{\mathbf{y}}(\mathbf{y})$ eines Zufallsvektors ist eine p-dimensionale, reellwertige Funktion. Integrale bzw. Summen über geeignete Bereiche ergeben Wahrscheinlichkeiten. Die einzelnen Komponenten sind unabhängig, wenn die gemeinsame Dichte bzw. Wahrscheinlichkeitsfunktion das Produkt der univariaten Randdichten bzw. Wahrscheinlichkeitsfunktionen ist:

$$f_{\mathbf{y}}(\mathbf{y}) = f_{\mathbf{y}}(y_1, \ldots, y_p) = f_{Y_1}(y_1) \cdots f_{Y_p}(y_p). \tag{1.23}$$

Satz 1.17 (Erwartungswert einer Linearkombination)
Der Erwartungswert einer Linearkombination von Zufallsvariablen ist gleich der Linearkombination der Erwartungswerte:

$$\mathrm{E}\left(\sum_{i=1}^{k} a_i Y_i\right) = \sum_{i=1}^{k} a_i \mathrm{E}(Y_i). \tag{1.24}$$

Beweis: Es wird nur der Fall zweier Zufallsvariablen betrachtet. Seien also X und Y zwei stetige Zufallsvariablen mit der gemeinsamen Dichte $f(x,y)$ und den Erwartungswerten $\mathrm{E}(X)$ und $\mathrm{E}(Y)$. Dann gilt für $Z = aX + bY$:

$$\begin{aligned}
\mathrm{E}(aX+bY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (ax+by) f(x,y)\,dxdy \\
&= a\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\cdot f(x,y)\,dxdy + b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\cdot f(x,y)\,dxdy \\
&= a\int_{-\infty}^{\infty} x\int_{-\infty}^{\infty} f(x,y)\,dxdy + b\int_{-\infty}^{\infty} y\int_{-\infty}^{\infty} f(x,y)\,dxdy \\
&= a\int_{-\infty}^{\infty} x\cdot f_X(x)\,dx + b\int_{-\infty}^{\infty} y\cdot f_Y(y)\,dy \\
&= a\cdot \mathrm{E}(X) + b\cdot \mathrm{E}(Y).
\end{aligned}$$

■

Die Erwartungswerte $\mathrm{E}((Y-\mu)^r)$ werden als *r-te zentrale Momente* bezeichnet. Dafür wird auch geschrieben:

$$\mu_r' = \mathrm{E}((Y-\mu)^r). \tag{1.25}$$

Das zweite zentrale Moment ist die *Varianz* $\mathrm{Var}(Y) = \mathrm{E}((Y-\mu)^2)$. Damit, oder noch besser mit der *Standardabweichung* $\sqrt{\mathrm{Var}(Y)}$, wird die Streuung oder Ausbreitung einer Verteilung charakterisiert. In der Regel wird für die Varianz das Symbol σ^2 verwendet.

Beispiel 1.18 (Varianz einer $\mathcal{R}(0,1)$-Verteilung)
Für eine über dem Intervall $(0,1)$ gleichverteilte Zufallvariable Y ist $\mathrm{E}(Y) = 1/2$ und

$$\mathrm{Var}(Y) = \int_0^1 \left(y-\frac{1}{2}\right)^2 dy = \frac{1}{3}\left(y-\frac{1}{2}\right)^3\Bigg|_0^1 = 2\cdot\frac{1}{3}\cdot\frac{1}{8} = \frac{1}{12}.$$

■

Beispiel 1.19 (Erwartungswert und Varianz der Normalverteilung)
Die Zufallsvariable Y sei $\mathcal{N}(\mu, \sigma^2)$-verteilt. Dann ist der Parameter μ gleich dem Erwartungswert

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} y f(y; \mu, \sigma^2)\, dy = \mu\,,$$

und σ^2 ist gleich der Varianz

$$\mathrm{Var}(Y) = \int_{-\infty}^{\infty} (y-\mu)^2 f(y; \mu, \sigma^2)\, dy = \sigma^2\,.$$

■

Dass die Varianz eine Maßzahl für die Ausbreitung oder die "Streuung" einer Verteilung ist, wird durch die *Ungleichung von Tschebyscheff* fundiert:

Satz 1.20 (Tschebyscheff-Ungleichung)
Sei Y eine Zufallsvariable mit Erwartungswert μ und Varianz σ^2. Dann gilt

$$\mathrm{P}(|Y-\mu| > \epsilon) \leq \frac{\mathrm{Var}(Y)}{\epsilon^2} \tag{1.26a}$$

bzw.

$$\mathrm{P}(|Y-\mu| \leq \epsilon) \geq 1 - \frac{\mathrm{Var}(Y)}{\epsilon^2}. \tag{1.26b}$$

Beweis: X habe die Dichte $f(x)$. Dann erhält man die Abschätzungen

$$\begin{aligned}
\mathrm{Var}(Y) &= \int_{-\infty}^{\infty} (y-\mu)^2 f(y) dy \geq \int_{(y-\mu)^2 > \epsilon^2} (y-\mu)^2 f(y) dy \\
&\geq \int_{(y-\mu)^2 > \epsilon^2} \epsilon^2 f(y) dy = \epsilon^2 \int_{|y-\mu| > \epsilon} f(y) dy \\
&= \epsilon^2 \cdot \mathrm{P}(|Y-\mu| > \epsilon).
\end{aligned}$$

Damit ergibt sich die Behauptung. ■

Die Interpretation der Tschebyscheff-Ungleichung ist klar: Die untere Schranke für die Wahrscheinlichkeit, dass Y einen Wert annimmt, der sich von $\mu = \mathrm{E}(Y)$ um nicht mehr als ϵ unterscheidet, hängt von der Varianz ab. Je größer die Varianz ist, desto größer muss ϵ gewählt werden, damit die Schranke gleich bleibt.

Die wichtigsten Eigenschaften der Varianz sind im folgenden Satz festgehalten.

Satz 1.21 (Eigenschaften der Varianz)
Für die Varianz einer Zufallsvariablen Y gilt:

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2; \tag{1.27a}$$

$$\mathrm{Var}(a + bY) = b^2\mathrm{Var}(Y); \tag{1.27b}$$

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{E}((X - \mu_X)(Y - \mu_Y)). \tag{1.27c}$$

Bei der letzten Beziehung ist X eine weitere Zufallsvariable und μ_X, μ_Y sind die zugehörigen Erwartungswerte.
Beweis: Die erste Beziehung ergibt sich aus:

$$\mathrm{Var}(Y) = \mathrm{E}(Y - \mu)^2 = \mathrm{E}(Y^2 - 2\mu Y + \mu^2) = \mathrm{E}(Y^2) - 2\mu^2 + \mu^2 = \mathrm{E}(Y^2) - \mu^2.$$

Die zweite ergibt sich ebenso unmittelbar:

$$\mathrm{Var}(a + bY) = \mathrm{E}(a + bY - (a + b\mu))^2 = \mathrm{E}(b^2(Y^2 - \mu)^2) = b^2\mathrm{E}(Y - \mu)^2 = b^2\mathrm{Var}(Y).$$

Die letzte ist lediglich etwas aufwändiger:

$$\begin{aligned}\mathrm{Var}(X + Y) &= \mathrm{E}(X + Y - (\mu_X + \mu_Y))^2 = \mathrm{E}((X - \mu_X) + (Y - \mu_Y))^2 \\ &= \mathrm{E}((X - \mu_X)^2 + 2(X - \mu_X)(Y - \mu_Y) + (Y - \mu_Y)^2) \\ &= \mathrm{E}(X - \mu_X)^2 + \mathrm{E}(Y - \mu_Y)^2 + 2\mathrm{E}((X - \mu_X)(Y - \mu_Y)).\end{aligned}$$

■

Die Zerlegung (1.27a) ist oft für die Bestimmung der Varianz vorteilhaft.

Beispiel 1.22 (Varianz der Poisson-Verteilung)
Ist Y Poisson-verteilt, so ist

$$\begin{aligned}\mathrm{E}(Y^2) &= \sum_{y=0}^{\infty} y^2 \cdot e^{-\lambda} \cdot \frac{\lambda^y}{y!} = \sum_{y=1}^{\infty} y \cdot e^{-\lambda} \cdot \frac{\lambda^y}{(y-1)!} \\ &= \sum_{y=1}^{\infty} (y-1) \cdot e^{-\lambda} \cdot \frac{\lambda^y}{(y-1)!} + \sum_{y=1}^{\infty} e^{-\lambda} \cdot \frac{\lambda^y}{(y-1)!} \\ &= \lambda^2 \sum_{z=0}^{\infty} e^{-\lambda} \cdot \frac{\lambda^z}{z!} + \lambda \\ &= \lambda^2 + \lambda.\end{aligned}$$

Damit folgt $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2 = \lambda$. ■

Mit dem dritten und vierten zentralen Moment, $\mathrm{E}((Y-\mu)^3)$ und $\mathrm{E}((Y-\mu)^4)$, werden zwei weitere Gestaltparameter der Verteilung definiert, die *Schiefe* und die *Wölbung*,

$$\frac{\mathrm{E}(Y - \mu)^3}{\sigma^3} \quad \text{und} \quad \frac{\mathrm{E}(Y - \mu)^4}{\sigma^4}. \tag{1.28}$$

Für *symmetrische Verteilungen* wie die Normalverteilung ist $\mathrm{E}(Y - \mu)^3/\sigma^3 = 0$. Bei einer Verteilung, deren rechte Flanke größere Wahrscheinlichkeitsmasse als die linke aufweist, ist die Maßzahl positiv. Solche Verteilungen heißen *rechtsschief*. Umgekehrt ist die Maßzahl der Schiefe negativ für *linksschiefe Verteilungen*, bei denen die linke Flanke größere Wahrscheinlichkeitsmasse als die rechte aufweist.

Beispiel 1.23 (Schiefe der Pareto-Verteilung)
Bei der Pareto-Verteilung mit einem Parameter $\alpha > r$ erhält man:

$$\mathrm{E}(Y^r) = \int_k^\infty y^r \alpha \frac{k^\alpha}{y^{\alpha+1}}\, dy.$$

Für $\alpha + 1 - r > 1$ folgt:

$$\begin{aligned}\mathrm{E}(Y^r) &= \alpha k^\alpha \frac{-1}{\alpha - r} \frac{1}{y^{\alpha-r}}\Bigg|_k^\infty \\ &= \frac{\alpha}{\alpha - r} k^r = k^r \left(1 - \frac{r}{\alpha}\right)^{-1}.\end{aligned}$$

Dies ergibt speziell:

$$\mathrm{E}(Y) = \frac{\alpha}{\alpha - 1} k,$$

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2 = \frac{\alpha}{\alpha - 2} k^2 - \left(\frac{\alpha}{\alpha - 1} k\right)^2 = \frac{\alpha k^2}{(\alpha - 2)(\alpha - 1)^2}$$

und

$$\mathrm{E}((Y - \mu)^3) = \mathrm{E}(Y^3 - 3\mu Y^2 + \mu Y - \mu^3) = \mathrm{E}(Y^3) - 3\mu \mathrm{E}(Y^2) + 3\mu^2 \mathrm{E}(Y) - \mu^3,$$

was zu der Schiefe führt:

$$\gamma_1(Y) = 2 \frac{\alpha + 1}{\alpha - 3} \sqrt{\frac{\alpha - 2}{\alpha}}.$$

Die Schiefe nimmt mit wachsendem α ab. So gilt beispielsweise:

α	3.5	4	5	7.5	10	15
$\gamma_1(Y)$	11.784	7.071	4.648	3.235	2.811	2.483

■

Die Wölbung ist ein eher vages Konzept. (Dies gilt allerdings auch für die Lage und Streuung, vgl. Mosteller & Tukey 1977). Mit $\mathrm{E}(Y - \mu)^4/\sigma^4$ wird eine von Lage und Streuung unabhängige Verteilung der Wahrscheinlichkeitsmasse von den Schultern der Verteilung in die Flanken und das Zentrum bzw. in die umgekehrte Richtung beschrieben, vgl. Balanda & MacGillivray (1988). Als Vergleichsbasis dient die Normalverteilung, für die $\mathrm{E}(Y - \mu)^4/\sigma^4 = 3$ gilt. Daher nimmt man an Stelle der Wölbung als Maßzahl oft den *Exzess*, die um den Wert 3 verringerte Wölbung. Der Exzess wird i.a. nur bei symmetrischen Verteilungen betrachtet. Zur Abkürzung werden die folgenden Bezeichnungen eingeführt:

$$\gamma_1 = \gamma_1(Y) = \frac{\mathrm{E}(Y - \mu)^3}{\sigma^3}, \quad \gamma_2 = \gamma_2(Y) = \frac{\mathrm{E}(Y - \mu)^4}{\sigma^4} - 3. \tag{1.29}$$