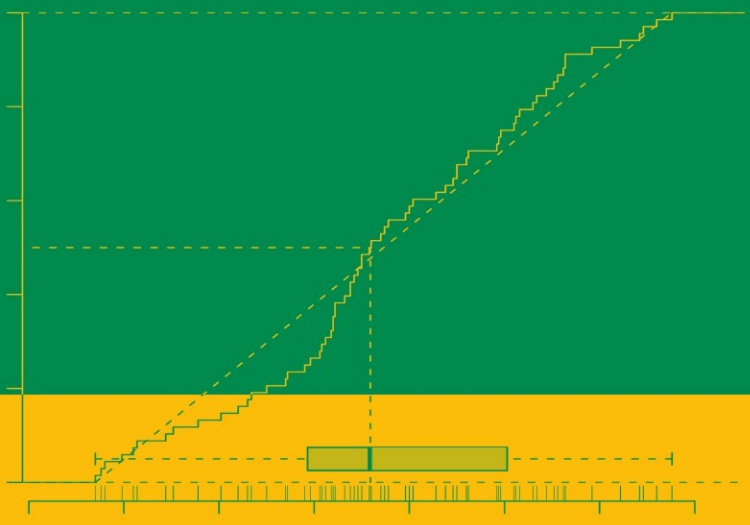
Kohn · Öztürk

Statistik für Ökonomen



Springer

Springer-Lehrbuch

Für weitere Bände:
http://www.springer.com/series/1183

Wolfgang Kohn · Riza Öztürk

Statistik für Ökonomen

Datenanalyse mit R und SPSS

Springer

Prof. Dr. Wolfgang Kohn
FH Bielefeld
FB Wirtschaft
Universitätsstr. 25
33615 Bielefeld
Deutschland
wolfgang.kohn@fh-bielefeld.de

Prof. Dr. Riza Öztürk
FH Bielefeld
FB Wirtschaft
Universitätsstr. 25
33615 Bielefeld
Deutschland
riza.oeztuerk@fh-bielefeld.de

ISSN 0937-7433
ISBN 978-3-642-14584-1 e-ISBN 978-3-642-14585-8
DOI 10.1007/978-3-642-14585-8
Springer Heidelberg Dordrecht London New York

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

Einbandentwurf: WMXDesign GmbH, Heidelberg

Gedruckt auf säurefreiem Papier

Springer ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.com)

Für

Coco Rindt,
Hannah, Sophia und Leonhard,
meine Eltern, Schwiegereltern und Freunde

meine Familie Güzel, Gülbahar und Serpil Öztürk
meinen Freund Beyhan Polat

Vorwort

Die Statistik ist ein enorm umfangreiches Wissensgebiet. Jedes Wissenschaftsgebiet benötigt zum Teil eigene statistische Verfahren und es sind auch Kenntnisse aus den Fachgebieten erforderlich, um die Anwendung und die Ergebnisse der Verfahren zu verstehen. So werden z. B. bei einer Datenerhebung per Fragebogen oder Interview u. a. psychologische Kenntnisse benötigt. Für die Datenverarbeitung und Berechnung der statistischen Verfahren werden Mathematik- und Informatikkenntnisse benötigt.

Die Interpretation der Ergebnisse setzt wiederum sehr gute Fachkenntnisse der entsprechenden Wissensgebiete voraus. Von daher ist es schwierig, in einer Darstellung alle Aspekte gleichermaßen zu berücksichtigen. Wir haben uns bemüht, die Eigenschaften und Zusammenhänge wesentlicher statistischer Verfahren aufzuzeigen und ihre Anwendung durch zahlreiche Beispiele und Abbildungen verständlich zu machen.

Das Buch wendet sich an all diejenigen, die sich im Rahmen eines ökonomisch bezogenen Studiums (wie BWL, Wirtschaftsinformatik, Wirtschaftspsychologie oder Wirtschaftsrecht) sich mit Statistik befassen. Das Buch hat zum Ziel, alle wesentlichen Schritte einer deskriptiven und induktiven statistischen Analyse zu erklären. Dies geschieht mit vielen Beispielen, Übungen und Programmanweisungen. Die Analyse metrischer Daten steht dabei im Vordergrund. Dies liegt daran, dass dieser Datentyp im Zentrum der klassischen Statistik steht.

Eine statistische Analyse wird heute mit dem Computer als Analysewerkzeug durchgeführt; allein schon wegen der Möglichkeit, Grafiken zu erzeugen. Die Anwendung der Statistikprogramme R und SPSS wird für alle wichtigen Analyseschritte erklärt. Beide Programme sind grundsätzlich verschieden in ihrer Anwendung. R wird durch Skriptbefehle gesteuert, d. h. es werden Befehlsfolgen eingegeben, die sich auch zu einem Programm ausweiten lassen. Hingegen werden in SPSS die Befehle durch Menüs eingegeben. Im vorliegenden Buch wird überwiegend R verwendet, da es sich besser zur Beschreibung einzelner Analyseschritte einsetzen lässt.

Die Daten können unter

`http://www.fh-bielefeld.de/fb5/kohn`

und

`http://www.fh-bielefeld.de/fb5/oeztuerk`

heruntergeladen werden. Sie sind dort als Textdateien und im SPSS-Format hinterlegt. Ferner werden auf der Webseite auch, wenn nötig, Korrekturen hinterlegt.

Das Buch wäre in dieser Form nicht ohne die hervorragenden Programmerweiterungen relax (R Editor for Literate Analysis and LaTeX) von Prof. Dr. Peter Wolf und sweave von Prof. Dr. Friedrich Leisch [11] für R möglich gewesen. Wir bedanken uns bei Dr. Dirk Martinke und Serpil Öztürk sowie insbesondere bei Coco Rindt für die vielen guten Hinweise und wertvollen Korrekturen, die zum Gelingen des Buches beitrugen.

Inhaltsverzeichnis

Teil I Einführung

1 Kleine Einführung in R 3
- 1.1 Installieren und Starten von R 3
- 1.2 R-Befehle ausführen 3
- 1.3 R-Workspace speichern 4
- 1.4 R-History sichern 4
- 1.5 Verwenden des R-Skripteditors 5

2 Kurzbeschreibung von SPSS 7
- 2.1 Der SPSS-Dateneditor 7
- 2.2 Statistische Analysen mit SPSS 9

3 Die Daten 11

4 Grundlagen 15
- 4.1 Grundbegriffe der Statistik 15
- 4.2 Datenerhebung und Erhebungsarten 16
- 4.3 Messbarkeitseigenschaften 17
- 4.4 Übungen 18
- 4.5 Fazit 18

Teil II Deskriptive Statistik

5 Eine erste Grafik 21

6 Häufigkeitsfunktion 25
- 6.1 Absolute Häufigkeit 25
- 6.2 Relative Häufigkeit 29
- 6.3 Übungen 31
- 6.4 Fazit 31

7 Mittelwert ... 33
7.1 Arithmetisches Mittel ... 33
7.2 Getrimmter arithmetischer Mittelwert ... 34
7.3 Gleitendes arithmetisches Mittel ... 37
7.4 Übungen ... 38
7.5 Fazit ... 38

8 Median und Quantile ... 39
8.1 Median ... 39
8.2 Quantile ... 42
8.3 Anwendung für die Quantile: Value at Risk (VaR) ... 44
8.4 Übungen ... 45
8.5 Fazit ... 46

9 Grafische Darstellungen einer Verteilung ... 47
9.1 Boxplot ... 47
9.2 Empirische Verteilungsfunktion ... 50
9.3 Histogramm ... 52
9.4 Übungen ... 56
9.5 Fazit ... 57

10 Varianz, Standardabweichung und Variationskoeffizient ... 59
10.1 Stichprobenvarianz ... 59
10.2 Standardabweichung ... 60
10.3 Variationskoeffizient ... 61
10.4 Übungen ... 63
10.5 Fazit ... 63

11 Lorenzkurve und Gini-Koeffizient ... 65
11.1 Lorenzkurve ... 65
11.2 Gini-Koeffizient ... 69
11.3 Übungen ... 70
11.4 Fazit ... 70

12 Wachstumsraten, Renditeberechnungen und geometrisches Mittel ... 71
12.1 Diskrete Renditeberechnung und geometrisches Mittel ... 71
12.2 Stetige Renditeberechnung ... 73
12.3 Übungen ... 76
12.4 Fazit ... 76

13 Indexzahlen und der DAX ... 77
13.1 Preisindex der Lebenshaltung nach Laspeyres ... 77
13.2 Basiseffekt bei Indexzahlen ... 79
13.3 Der DAX ... 80

13.4 Übungen 81
13.5 Fazit 82

14 Bilanz 1 83

Teil III Regression

15 Grafische Darstellung von zwei Merkmalen 87
15.1 QQ-Plot 87
15.2 Streuungsdiagramm 88
15.3 Übungen 90
15.4 Fazit 90

16 Kovarianz und Korrelationskoeffizient 91
16.1 Kovarianz 91
16.2 Korrelationskoeffizient 92
16.3 Übungen 95
16.4 Fazit 95

17 Lineare Regression 97
17.1 Modellbildung 97
17.2 Methode der Kleinsten Quadrate 98
17.3 Regressionsergebnis 100
17.4 Prognose 102
17.5 Lineare Regression bei nichtlinearen Zusammenhängen 103
17.6 Multiple Regression 105
17.7 Übungen 106
17.8 Fazit 106

18 Güte der Regression 107
18.1 Residuenplot 107
18.2 Streuungszerlegung und Bestimmtheitsmaß 109
18.3 Übungen 113
18.4 Fazit 113

19 Bilanz 2 115

Teil IV Wahrscheinlichkeitsrechnung

20 Grundzüge der diskreten Wahrscheinlichkeitsrechnung 119
20.1 Zufallsexperimente 119
20.2 Ereignisoperationen 120

20.3 Zufallsstichproben . . . 122
20.4 Wahrscheinlichkeitsberechnungen . . . 123
20.5 Kolmogorovsche Axiome . . . 130
20.6 Bedingte Wahrscheinlichkeit und Satz von Bayes . . . 132
20.7 Übungen . . . 139
20.8 Fazit . . . 141

21 Wahrscheinlichkeitsverteilungen . . . 143
21.1 Diskrete Zufallsvariablen . . . 143
21.2 Stetige Zufallsvariable . . . 144
21.3 Erwartungswert . . . 147
21.4 Varianz . . . 149
21.5 Kovarianz . . . 150
21.6 Übungen . . . 152
21.7 Fazit . . . 152

22 Normalverteilung . . . 153
22.1 Entstehung der Normalverteilung . . . 153
22.2 Von der Normalverteilung zur Standardnormalverteilung . . . 156
22.3 Berechnung von Wahrscheinlichkeiten normalverteilter Zufallsvariablen . . . 157
22.4 Berechnung von Quantilen aus der Normalverteilung . . . 160
22.5 Anwendung auf die parametrische Schätzung des *Value at Risk* für die BMW Aktie . . . 163
22.6 Übungen . . . 166
22.7 Fazit . . . 166

23 Weitere Wahrscheinlichkeitsverteilungen . . . 167
23.1 Binomialverteilung . . . 167
23.2 Hypergeometrische Verteilung . . . 173
23.3 Geometrische Verteilung . . . 177
23.4 Poissonverteilung . . . 179
23.5 Exponentialverteilung . . . 184
23.6 Übungen . . . 189
23.7 Fazit . . . 189

24 Bilanz 3 . . . 191

Teil V Schätzen und Testen

25 Schätzen . . . 195
25.1 Schätzen des Erwartungswerts . . . 196
25.2 Schätzen der Varianz . . . 197

25.3 Schätzen der Varianz des Stichprobenmittels . . . 200
25.4 Übungen . . . 200
25.5 Fazit . . . 201

26 Stichproben und deren Verteilungen . . . 203
26.1 Verteilung des Stichprobenmittels in einer normalverteilten Stichprobe . . . 203
26.2 Schwaches Gesetz der großen Zahlen . . . 204
26.3 Hauptsatz der Statistik . . . 205
26.4 Zentraler Grenzwertsatz . . . 207
26.5 Hauptsatz der Stichprobentheorie . . . 209
26.6 Übungen . . . 211
26.7 Fazit . . . 213

27 Konfidenzintervalle für normalverteilte Stichproben . . . 215
27.1 Konfidenzintervall für μ_X bei bekannter Varianz . . . 215
27.2 Konfidenzintervall für μ_X bei unbekannter Varianz . . . 217
27.3 Approximatives Konfidenzintervall für μ_X . . . 219
27.4 Approximatives Konfidenzintervall für den Anteilswert θ . . . 220
27.5 Konfidenzintervall für die Varianz σ_X^2 . . . 221
27.6 Berechnung der Stichprobengröße . . . 222
27.7 Übungen . . . 223
27.8 Fazit . . . 223

28 Parametrische Tests für normalverteilte Stichproben . . . 225
28.1 Klassische Testtheorie . . . 225
28.2 Testentscheidung . . . 230
28.3 Gauss-Test für μ_X . . . 234
28.4 t-Test für μ_X . . . 235
28.5 t-Test für die Regressionskoeffizienten der Einfachregression . . . 239
28.6 Test für den Anteilswert θ . . . 243
28.7 Test auf Mittelwertdifferenz in großen Stichproben . . . 244
28.8 Test auf Mittelwertdifferenz in kleinen Stichproben . . . 246
28.9 Test auf Differenz zweier Anteilswerte . . . 247
28.10 Test auf Gleichheit zweier Varianzen . . . 248
28.11 Gütefunktion eines Tests . . . 250
28.12 Übungen . . . 254
28.13 Fazit . . . 256

29 Einfaktorielle Varianzanalyse . . . 257
29.1 Modell . . . 257
29.2 Test auf Gleichheit der Mittelwerte . . . 262
29.3 Test der Einzeleffekte . . . 263

29.4 Übungen . . . 266
29.5 Fazit . . . 266

30 Analyse kategorialer Daten . . . 267
30.1 Kontingenztabelle . . . 268
30.2 Randverteilungen . . . 270
30.3 Bedingte Verteilungen . . . 271
30.4 Logistische Regression . . . 276
30.5 Quadratische und normierte Kontingenz . . . 279
30.6 Unabhängigkeitstest . . . 281
30.7 Übungen . . . 282
30.8 Fazit . . . 283

31 Bilanz 4 . . . 285

A Glossar . . . 287

B Lösungen zu ausgewählten Übungen . . . 289

C Tabellen . . . 315

Literaturverzeichnis . . . 327

Sachverzeichnis . . . 329

Teil I
Einführung

Kapitel 1
Kleine Einführung in R

1.1 Installieren und Starten von R

Das R-Programm können Sie kostenlos für verschiedene Betriebssysteme von der R-Projektseite `www.r-project.org` herunterladen. Die Installation von R unter Windows erfolgt mit der Ausführung der Datei `R-x.x.x-win32.exe`. Die Versionsnummer `x.x.x` ändert sich regelmäßig. Ein Menü führt durch die Installation. Das R-Programm wird durch das R-Symbol gestartet. Es handelt sich dabei um die GUI (graphical user interface) Version. Es existiert auch eine Terminalversion.

Zusätzlich zum Basisprogramm können eine Vielzahl von Paketen dazu installiert werden, die den Funktionsumfang für spezielle statistischen Analyseverfahren erweitern. Der Menüeintrag steht unter `Pakete -> installieren`. Wir verwenden im Text das Paket `zoo`. Es stellt Funktionen für indexgeordnete (z. B. zeitlich geordnete) Daten zur Verfügung.

Die einzelnen Befehle werden im Text erläutert. Mehr Informationen zur Installation, Einrichtung und Verwendung des Programms finden Sie auf der oben genannten Internetseite und z.B. bei [3] oder [15].

1.2 R-Befehle ausführen

Im GUI Fenster des R-Programms können Befehle direkt eingegeben werden. Zum Beispiel werden mit dem Befehl

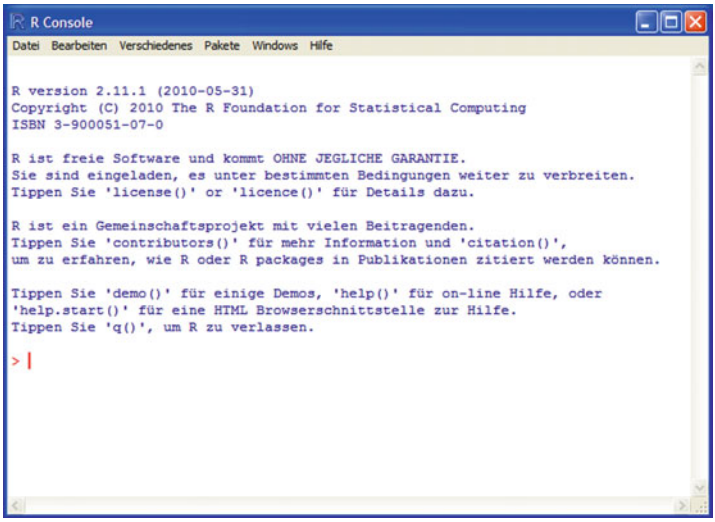
```
> x <- c(33.6, 33.98, 33.48, 33.17, 33.75, 33.61)
```

6 Werte der Variablen x zugewiesen. Der Dezimalpunkt kennzeichnet den Übergang zu den Dezimalstellen; das Komma trennt die Zahlen. Mit der `Enter`-Taste wird der Befehl ausgeführt. Der Inhalt der Variablen kann durch Eingabe von x und `Enter` ausgegeben werden.

W. Kohn, R. Öztürk, *Statistik für Ökonomen*, Springer-Lehrbuch,
DOI 10.1007/978-3-642-14585-8_1,

```
> x

[1] 33.60 33.98 33.48 33.17 33.75 33.61
```



Zu jedem Befehl kann mit `?Befehl` (z. B. `?c`) ein Hilfetext aufgerufen werden. In R stehen unter `Hilfe` Handbücher zur Verfügung, die die wichtigsten Befehle erläutern.

1.3 R-Workspace speichern

Mit der Speicherung des Workspace (`Datei -> Sichere Workspace`) werden alle Variablen und definierten Funktionen mit ihren Inhalten, die im R-Fenster eingegeben werden, gespeichert. Die erzeugte Datei kann mit dem Befehl `Datei -> Lade Workspace` bei einem erneuten Start von R eingelesen werden. Die Befehlsfolgen werden jedoch nicht gesichert. Dazu steht der folgende Befehl zur Verfügung.

1.4 R-History sichern

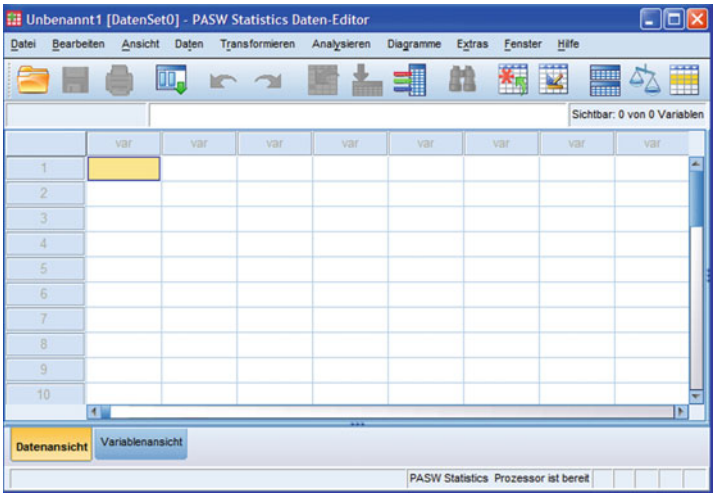
Mit der Sicherung der History (`Datei -> Speichere History`) werden die im R-Fenster eingegebenen Befehle als Datei gesichert. Mit dem Befehl `Datei -> Lade History` können die Befehle wieder in R eingelesen werden. Die Datei kann mit jedem Editor (auch mit dem R-Skripteditor) bearbeitet werden.

1.5 Verwenden des R-Skripteditors

Unter dem Menüeintrag `Datei -> Neues Skript` kann eine Datei erstellt werden, in die die R-Befehle eingetragen werden kann. Mit dem Befehl `Strg-R` können die Befehlszeilen zur Ausführung übergeben werden. Die Datei kann mit dem Befehl `Datei -> Öffne Skript` wieder geöffnet und bearbeitet werden.

Kapitel 2
Kurzbeschreibung von SPSS

SPSS ist eine kommerzielle Statistiksoftware, für die an vielen Hochschulen eine Campuslizenz existiert. Die Installation erfolgt entweder über eine CD/DVD oder als Netzwerkinstallation durch die Hochschulrechenzentren. Das Basismodul der Software ermöglicht eine Datenverwaltung und umfangreiche statistische und grafische Analysen, die vor allem für die Sozialwissenschaften ausgelegt sind. Wir verwenden die SPSS Version 18.0.2.



SPSS wird in der Regel menügesteuert. Die Menüaufrufe sind im Text angegeben. Das Programm verfügt aber auch über eine eigene Befehlssyntax, auf die in diesem Buch jedoch nicht eingegangen wird.

2.1 Der SPSS-Dateneditor

Nach dem Programmstart erscheint der so genannte SPSS-Dateneditor auf dem Bildschirm. Es ist eine Tabelle, in der Daten eingegeben werden können.

W. Kohn, R. Öztürk, *Statistik für Ökonomen*, Springer-Lehrbuch,
DOI 10.1007/978-3-642-14585-8_2, © Springer-Verlag Berlin Heidelberg 2010

Eine neue Datendatei wird über das Menü `Datei > Neu > Daten` angelegt. Eine bereits existierende SPSS Datendatei wird über den Menüpunkt `Datei > Öffnen > Daten` geöffnet. Über das gleiche Menü kann auch ein vorhandener Datensatz mit einem anderen Dateiformat (z. B. Excel) importiert werden. Es kann immer nur eine Datendatei zur Zeit im Dateneditor bearbeitet werden. SPSS speichert die Daten mit der Endung `sav`. Der Dateneditor stellt zwei Ansichten der Daten bereit.

Die Datenansicht zeigt die Tabelle mit den Daten. In den Zeilen stehen die Beobachtungen; in den Spalten stehen die Variablen. Zwischen den Zellen kann man mit der Maus oder mit den Pfeiltasten auf der Tastatur wechseln. Es besteht keine Möglichkeit in der Datentabelle selbst Berechnungen durchzuführen oder Formeln einzugeben. Der Wert in den Zellen kann durch Anklicken der Zelle oder Drücken der `F2`-Taste auf der Tastatur verändert werden.

In die Variablenansicht wechselt man durch Auswahl des Registers `Variablenansicht` unten links im Programmfenster. Die Variablenansicht enthält Informationen zu den Variablen. Unter `Name` wird der Name einer Variable eingegeben. Dabei ist zu beachten, dass Variablennamen in SPSS maximal 8 Zeichen lang sein, keine Sonderzeichen enthalten dürfen und mit einem Buchstaben beginnen müssen. In den weiteren Spalten werden weitere Details zu den Daten festgelegt. In vielen Fällen sind Voreinstellungen von SPSS ausreichend. Folgende Datentypen stehen unter `Typ` zur Verfügung:

- Numerische Variablen sind Variablen, deren Werte Zahlen sind. Unter `Dezimalstellen` kann angegeben werden, wie viele Dezimalstellen gespeichert und im Dateneditor angezeigt werden.
- Es können verschiedene Datumsformate ausgewählt werden. Bei zweistelligen Jahresangaben wird das Jahrhundert durch die Einstellung in den Optionen bestimmt. Dieses kann im Menü `Bearbeiten > Optionen` im Register `Daten` verändert werden.
- String-Variablen (Zeichenketten) können sowohl Buchstaben als auch Ziffern enthalten. Sie können nicht für Berechnungen verwendet werden.
- Punkt-Variablen verwenden einen Punkt zur Tausender-Trennung und ein Komma als Dezimaltrennzeichen.
- Komma-Variablen verwenden ein Komma zur Tausender-Trennung und einen Punkt als Dezimaltrennzeichen.
- In der wissenschaftlichen Notation werden die Werte in der Exponentialdarstellung (einem `E` und einer Zehnerpotenz mit Vorzeichen) angezeigt.
- Spezielle Währung

In der Variablenansicht können noch weitere Variableneigenschaften festgelegt werden. Unter `Spaltenformat` wird angegeben, wie viele Zeichen eine Variable hat (z. B. 8 Ziffern). Da die kurzen Variablennamen in SPSS nicht sehr aussagekräftig sind, ist es möglich, für jede Variable ein Variablenlabel (bis zu 120 Buchstaben) zu vergeben. Ob im Dateneditor die Variablennamen oder die Variablenlabels angezeigt werden, wird über den Menüpunkt `Bearbeiten > Optionen` im Register `Allgemein` festgelegt werden. Auch für die codierten Ausprägungen kategorialer

Variablen können `Wertelabels` vergeben werden. Hier muss man auf den kleinen Pfeil klicken, damit sich die entsprechende Dialogbox öffnet. Über das Menü `Ansicht > Wertelabels` wird bestimmt, ob im Dateneditor die Originalwerte oder die Wertelabels angezeigt werden. In der Datenansicht kann weiterhin die `Ausrichtung` der Werte in der Spalte festgelegt werden, sowie das `Messniveau`. Außerdem können einzelne Merkmalsausprägungen als fehlende Werte definiert werden (z. B. 99 = `keine Angabe`).

2.2 Statistische Analysen mit SPSS

Die Analysemethoden werden im Menüpunkt `Analysieren` ausgewählt. Bei manchen Analysen ist eine `Transformation` der Daten notwendig. Die Ausgabe der Analysen erfolgt in einem neuen Programmfenster, dem Ausgabefenster. In diesem Fenster stehen ebenfalls die Menüpunkte zur Datenanalyse zur Verfügung. Die einzelnen Anweisungsschritte werden im Text angegeben.

Kapitel 3
Die Daten

Ohne Daten ist in den meisten Fällen eine statistische Analyse nicht möglich. Zur anschaulichen und realistischen Beschreibung der hier eingesetzten statistischen Verfahren werden die Schlusskurse der BMW Aktie (Börsenplatz Frankfurt) vom 09.08. bis 16.11.04 als Grundlage verwendet. Sie stammen von der Internetseite des Handelsblattes.

Warum diese Daten? Es sind „echte" Daten, also kein erfundener Datensatz und liefern daher eine realistische statistische Analyse. Ferner sind Aktienkurse sicherlich für Ökonomen interessanter als Körpermaße und sie eignen sich für viele statistische Methoden, die wir hier vorstellen. Daher wird dieser Datensatz, da wo es sinnvoll ist, auf eine statistische Methode angewendet.

Bevor die Daten in den Programmen verfügbar sind, müssen sie von einer Datei eingelesen oder per Hand eingegeben werden. Die Datendateien können verschiedene Formate besitzen. Wir haben für beide Programme Datendateien vorbereitet.[1]

An Hand der folgenden Werte werden wir viele statistische Analysetechniken erklären und interpretieren.

R-Anweisung
Schauen Sie sich zuerst mit einem Editor (in Windows z. B. Notepad) die Datei `bmw.txt` an. Aus der Datei wollen wir die Schlusskurse der BMW Aktie in R übernehmen. Mit dem Befehl `read.table()` können so genannte ASCII Dateien in das R Programm eingelesen werden. In den Anführungszeichen wird der Dateiname angegeben. Mit `sep=";"` wird die Trennung (Separation) zwischen den Zahlen eingestellt, die im vorliegenden Fall eine Semikolon ist. Die Option `header=T` bewirkt, dass eine Kopfzeile mit den Variablennamen in der Datendatei vorhanden ist und erst in der zweiten Zeile der Datei der Datensatz beginnt. Wichtig ist, dass vorher über `Datei > Verzeichnis wechseln` auch der Pfad zum entsprechenden Verzeichnis (Ordner) eingestellt wird.

[1] Die Daten können von der Internetadresse, die im Vorwort genannt ist, heruntergeladen werden.

W. Kohn, R. Öztürk, *Statistik für Ökonomen*, Springer-Lehrbuch,
DOI 10.1007/978-3-642-14585-8_3,

```
> bmw <- read.table("bmw.txt", sep = ";", header = TRUE)
> s.bmw <- bmw$Schluss
> s.bmw <- s.bmw[length(s.bmw):1]
> s.bmw
```

Die letzte Befehlszeile kehrt die Reihenfolge der eingelesenen Werte um, damit der älteste Wert zuerst und der neuste Wert zuletzt ausgegeben wird. Mit dem `read.table()` Befehl wird ein so genannter *data frame* erzeugt. Dies ist eine Tabelle der eingelesenen Werte.

```
 [1] 33.60 33.98 33.48 33.17 33.75 33.61 33.61 33.45 33.25 33.87
[11] 34.14 34.19 34.47 34.55 34.31 33.89 34.25 34.46 34.58 34.76
[21] 34.82 34.96 34.48 34.56 35.30 35.38 34.72 35.21 35.23 35.11
[31] 34.78 34.25 33.60 33.79 33.59 33.36 33.71 33.10 34.02 34.65
[41] 34.81 34.67 34.82 34.30 34.23 33.53 33.75 33.69 33.85 33.66
[51] 34.00 33.56 33.54 33.35 32.38 32.72 33.69 33.15 33.73 33.80
[61] 33.01 32.40 32.35 32.57 32.55 32.89 32.76 32.49
```

Mit dem Befehl `names()` werden die Variablennamen der Datentabelle angezeigt. Die Variablennamen in der Kopfzeile der Datentabelle können auch mit dem Befehl `attach()` direkt für den Aufruf in R verfügbar gemacht werden. Dies ermöglicht einen direkten Zugriff auf die Variablennamen in der Datentabelle. So muss dann nur noch `Schluss` statt `bmw$Schluss` eingegeben werden.

```
> names(bmw)
> attach(bmw)
```

Die Datentabelle kann mit dem `Bearbeiten -> Dateneditor` angesehen und bearbeitet werden. Soll eine neue Datentabelle erstellt werden, so muss zuerst mit `y <-data.frame()` ein leere Datentabelle erzeugt werden. Als Name für die Datentabelle wurde hier y gewählt.

SPSS-Anweisung

In SPSS starten Sie das Einlesen der ASCII Datei `bmw.txt` unter dem Menüpunkt `Datei > Textdatei einlesen`. Folgen Sie den Menüanweisungen. Die Datei besitzt kein vordefiniertes Format und in der ersten Zeile stehen die Variablennamen. Der Variablen `Datum` muss das Format `Datum` mit `tt.mm.jjjj` zugewiesen werden. Die verbleibenden Variablen sind vom `Kommatyp`, d. h. die Dezimalstelle wird als Punkt ausgewiesen, die Tausenderstelle wird durch ein Komma angezeigt. Speichern Sie die eingelesen Werte ab.

Es liegen 68 Werte vor. Die **Anzahl der Werte** wird im allgemeinen mit n bezeichnet. Warum so viele Werte? Statistische Analysen sollten sich niemals nur auf wenige Werte stützen, weil statistische Aussagen erst verlässlich werden, wenn viele Beobachtungen vorliegen. Daher wurde hier ein Zeitraum von etwa 100 Tagen gewählt. Die Wochenenden und Feiertage entfallen; danach verbleiben 68 Werte. Die Anzahl der Werte n ist eine wichtige statistische Angabe.

Mit den Aktienkursen lassen sich sofort viele, aber nicht alle statistischen Analyseschritte anwenden. Daher werden die BWM Aktienkurse künstlich in 3 Kategorien (niedrig, mittel, hoch) unterteilt, um eine andere Art von Daten zu erzeugen. Der Bereich von 32 bis 33 wird willkürlich als niedriger Kurs festgelegt. Der Bereich von 33 bis 34.5 als mittlerer Kurs und über 34.5 als hoher Kurs festgelegt. Diese Daten sind jetzt also nur noch in Kategorien verfügbar und sie werden als **ordinale Daten** bezeichnet. Das Kursniveau lässt sich nur als Kategorie angeben. Ein messbarer Abstand ist nun nicht mehr definiert (siehe Abschn. 4.3).

R-Anweisung
Mit den folgenden Anweisungen werden die Kurse in die drei Kategorien eingeteilt. Der Vektor `c('niedrig','mittel','hoch')` weist die Bezeichnungen der 3 Kategorien der Variablen `bmw.levels` zu. Der Befehl `cut` teilt die Kurse in die 3 Bereiche ein: von 32 bis einschließlich 33, von 33 bis einschließlich 34.5 und von 34.5 bis einschließlich 35.5 (siehe hierzu Seite 27f). Mit dem Befehl `factor` werden den Kurskategorien die Namen zugewiesen. Also ein Kurs von 33.6 wird als „mittel“ ausgewiesen.

```
> bmw.levels <- c("niedrig", "mittel", "hoch")
> bmw.cut <- cut(s.bmw, breaks = c(32, 33, 34.5, 35.5))
> bmw.factor <- factor(bmw.cut, labels = bmw.levels)
> bmw.factor

 [1] mittel  mittel  mittel  mittel  mittel  mittel  mittel
 [8] mittel  mittel  mittel  mittel  mittel  mittel  hoch
[15] mittel  mittel  mittel  mittel  hoch    hoch    hoch
[22] hoch    mittel  hoch    hoch    hoch    hoch    hoch
[29] hoch    hoch    hoch    mittel  mittel  mittel  mittel
[36] mittel  mittel  mittel  mittel  hoch    hoch    hoch
[43] hoch    mittel  mittel  mittel  mittel  mittel  mittel
[50] mittel  mittel  mittel  mittel  mittel  niedrig niedrig
[57] mittel  mittel  mittel  mittel  mittel  niedrig niedrig
[64] niedrig niedrig niedrig niedrig niedrig
Levels: niedrig mittel hoch
```

Kapitel 4
Grundlagen

4.1 Grundbegriffe der Statistik

Die Grundbegriffe der Statistik dienen dazu, das Vokabular und die Symbole festzulegen, um Beobachtungen zu beschreiben.

Die Analyse fängt mit der Beobachtung eines Objekts an. In unserem Fall sind es die Börsenkurse der BMW Aktie. Das einzelne Objekt, das beobachtet wird, wird **statistische Einheit** genannt und mit ω_i bezeichnet. Der Index i nummeriert die statistischen Einheiten von 1 bis n. Der Index i ist eine natürliche Zahl. Die Anzahl der statistischen Einheiten wird mit n bezeichnet. Es wird die Folge der statistischen Einheiten betrachtet.

$$\omega_1, \omega_2, \ldots, \omega_n$$

Die Folge ist Gegenstand einer statistischen Untersuchung. Die statistischen Einheiten müssen **sachlich**, **zeitlich** und **räumlich** abgegrenzt sein. Wird z. B. eine Aktie beobachtet, so muss eindeutig bezeichnet werden, welche Aktie beobachtet wurde, wann und wo (welcher Börsenplatz). Die so gewonnene Menge der statistischen Einheiten wird als **Stichprobe** mit Ω_n bezeichnet.

$$\Omega_n = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

Sie stammt aus einer übergeordneten Menge, die als **Grundgesamtheit** Ω bezeichnet wird. Sie ist die Menge aller statistischen Einheiten. In statistischen Untersuchungen werden in der Regel Stichproben untersucht. Besitzt man eine Stichprobe, also z. B. die Kurse einer Aktie, so sind meistens nicht alle Eigenschaften der statistischen Einheit (bei einer Aktie z. B. Kurs, Börse, Typ) interessant, sondern nur bestimmte, z. B. der Aktienkurs.

Eine Eigenschaft der statistischen Einheit wird als **Merkmal** $X(\omega_i)$ bezeichnet und häufig mit X abgekürzt. Die möglichen Werte, die das Merkmal annehmen kann, werden **Merkmalsausprägungen** genannt. Die Beobachtung einer Merkmalsausprägung wird als **Merkmalswert** oder **Beobachtungswert** mit x_i bezeichnet.

W. Kohn, R. Öztürk, *Statistik für Ökonomen*, Springer-Lehrbuch,
DOI 10.1007/978-3-642-14585-8_4,

Tabelle 4.1 Grundbegriffe

Begriff	Symbol	Erklärung	Beispiel
statistische Einheit	ω	Informationsträger	Personen, Institutionen, Objekte
Merkmal	X	interessierende Eigenschaft einer statistischer Einheit	Alter, Börsenwert, Gewicht, Körpergröße, Kosten, Beurteilung, Produktanzahl, Farbe
Merkmalsausprägung	x	Werte, Zustände, Kategorien, die ein Merkmal annehmen kann	Euro, kg, cm, Befinden, Noten, ja und nein, schwarz, weiß, blau, ...
Merkmalswert	x_i	der beobachtete Wert einer Merkmalsausprägung	31 Jahre, 1300 Euro, 23 cm, 73 kg, schwarz

Die BMW Aktie ist die statistische Einheit, die in einem bestimmten Zeitraum an einem bestimmten Börsenplatz beobachtet wird. Der Aktienkurs ist das Merkmal und die einzelnen Werte sind die Merkmalswerte. Jede Beobachtung ist eine Realisation des Merkmals X. Das Merkmal X ist eine Funktion der statistischen Einheit. In Tabelle 4.1 sind die eben erklärten Begriffe zusammengestellt.

4.2 Datenerhebung und Erhebungsarten

Grundlage jeder Statistik sind Daten. Hierbei muss zwischen der Datenerhebung und der Erhebungsart unterschieden werden.

Mit der Datenerhebung wird die Art und Weise der Datengewinnung beschrieben. Erfolgt die Datenerhebung durch eine direkte Befragung in Form von Fragebogen oder Interview, dann handelt es sich um eine **Primärstatistik**. Werden die Daten aus bestehenden Statistiken gewonnen, dann spricht man von einer **Sekundärstatistik**.

Hiervon zu unterscheiden ist die Erhebungsart. Die Erhebungsart beschreibt, wie die statistischen Objekte ausgewählt werden. Werden die Objekte mit einem Zufallsprinzip ausgewählt, so spricht man von einer Zufallsstichprobe. Diese können unterschiedlich ausgestaltet sein.

Um eine **Zufallsstichprobe** (siehe auch Abschn. 20.3) ziehen zu können, müssen die Elemente der Grundgesamtheit nummerierbar sein. Hierfür wird häufig das so genannte **Urnenmodell** verwendet. Es wird eine Urne angenommen, die n gleiche Kugeln enthält, die sich nur durch ein Merkmal z. B. eine Zahl oder Farbe unterscheiden. Es gibt mindestens zwei Arten von Kugeln in der Urne. Die Kugeln werden als gut durchmischt angenommen. Die Entnahme der Kugeln kann auf zwei verschiedene Arten erfolgen: **mit** und **ohne Zurücklegen**. Zieht man die **Stichprobe mit Zurücklegen**, dann liegt eine **einfache Zufallsstichprobe** vor. Die einzelnen Stichprobenzüge sind voneinander **unabhängig**. Werden die gezogenen Elemente nicht mehr in die Urne zurückgelegt, wie beim Ziehen der Lottozahlen, so handelt es sich auch um eine **Zufallsstichprobe ohne Zurücklegen**, bei der aber

die einzelnen Züge **abhängig** voneinander sind. Mit jedem Zug verändert sich die Zusammensetzung in der Urne.

Je größer die Urne ist, desto schwieriger ist eine gleich wahrscheinliche Auswahl der einzelnen Kugeln. Daher werden für größere Zufallsstichproben meistens Zufallszahlen erzeugt. Für viele Zufallsstichproben muss vorher festgelegt werden, aus wie vielen Elementen sie gezogen werden sollen. Dies setzt die Kenntnis der Grundgesamtheit voraus. Sie ist aber manchmal nicht oder nur mit großem Aufwand feststellbar.

Ein anderes Problem kann die Suche nach einer bestimmten statistischen Einheit sein, die durch den Zufallsprozess ausgewählt wurde. Ist diese statistische Einheit eine Person, so verursacht die Suche nach ihr datenschutzrechtlich Probleme. Daher wird in der Praxis häufig eine nicht zufällige Stichprobe gezogen. Diese sind einfacher und häufig kostengünstiger zu gewinnen.

Eine **nicht zufällige Stichprobe** liegt vor, wenn ein Stichprobenplan, z. B. bestimmte **Quoten**, eingehalten werden müssen. Man spricht dann häufig auch von **repräsentativen Stichproben**. Für die Wahrscheinlichkeitsrechnung und die schließende Statistik sind diese Stichproben der Theorie nach aber ungeeignet.

Bei allen Erhebungsarten sind die Rechtsvorschriften des **Datenschutzes** zu beachten. Die Datenerhebung und die Erhebungsart ist immer sehr sorgfältig zu planen, da sie die Grundlage jeder weiteren statistischen Analyse sind.

4.3 Messbarkeitseigenschaften

Merkmale weisen im Wesentlichen drei verschiedene Messbarkeitseigenschaften auf. Das einfachste Messniveau ist die Unterscheidung der Merkmalsausprägungen der Art nach. Es wird als **nominales Messniveau** bezeichnet. Das Geschlecht ist z. B. ein nominales Merkmal, da es nur der Art nach unterschieden werden kann. Die Merkmalsausprägungen werden als **Kategorie** bezeichnet.

Kann eine eindeutige Reihenfolge in den Merkmalsausprägungen erzeugt werden, so liegt ein **ordinales Messniveau** vor. Zensuren sind z. B. ein ordinales Merkmal, da es geordnet werden kann, jedoch ist ein Abstand zwischen den Zensuren nicht definiert. Auch die Zensuren stellen nur Kategorien dar.

Ist der Abstand zwischen den Merkmalsausprägungen definiert, so wird von einem **metrischen Messniveau** gesprochen. Zum Beispiel kann der Abstand zwischen zwei deutschen Aktienkursen in Euro gemessen werden. Manchmal wird das metrische Messniveau weiter unterteilt in Intervall-, Verhältnis- und Absolutskala, auf die wir hier nicht weiter eingehen wollen. In Tabelle 4.2 sind die verschiedenen Messeigenschaften zusammengefasst.

Im vorliegenden Text wird die statistische Analyse vorwiegend mit metrischen Werten erläutert. Für dieses Messniveau existieren die meisten Verfahren. Die Analyse kategorialer Daten (nominales und ordinales Messniveau zusammengefasst) wird an geeigneten Stellen in den Kapiteln immer wieder aufgegriffen. In Kap. 30 werden Verfahren zur Analyse kategorialer Merkmale beschrieben.

Tabelle 4.2 Messbarkeitseigenschaften

Begriff	Erklärung	Beispiel
Nominalskala	Merkmalswerte beschreiben nur einen Zustand oder Status, keine Ordnung	Farbe, Geschlecht, Herkunft, Beruf
Ordinalskala	Merkmalswerte können geordnet werden, eine Reihenfolge kann hergestellt werden, z. B. etwas ist besser oder schlechter, aber kein Abstand messbar	Einstellungen, Befinden, Noten
Metrische Skala	Merkmalswerte können geordnet werden, man kann die genauen Abstände zwischen den Merkmalswerten messen	Geldeinheiten, Temparaturen, Stückzahlen, Längenmaß, Gewichtsmaß

4.4 Übungen

Übung 4.1 Warum sind statistische Einheiten abzugrenzen?

Übung 4.2 Was ist der Unterschied zwischen Merkmal und Merkmalswert?

Übung 4.3 Ist die repräsentative Stichprobe eine Zufallsstichprobe?

Übung 4.4 Um welches Messniveau handelt es sich bei den Aktienkursen?

Übung 4.5 Durch welche Informationen unterscheiden sich metrische, ordinale und nominale Merkmale?

4.5 Fazit

In diesem Kapitel wurden verschiedene Grundlagen für die Statistik erläutert. Die Grundbegriffe legen ein Vokabular für die Statistik fest. Die Datenerhebung und Erhebungsart zeigen verschiedene Aspekte zur Gewinnung von Werten auf. Eine Stichprobe (Datenerhebung und Erhebungsart) ist immer sehr sorgfältig zu planen. Denn jede weitere statistische Analyse greift auf diese Werte zurück. Der Datenschutz ist eng mit der Erhebung der Stichprobe verbunden und muss unbedingt beachtet werden. Die Messbarkeitseigenschaften von Merkmalen legen die späteren statistischen Analyseverfahren fest. Oft sucht man auch nach geeigneten Merkmalen, um bestimmte Verfahren anwenden zu können.

Teil II
Deskriptive Statistik

Kapitel 5
Eine erste Grafik

Die Datenanalyse beginnt in der Regel damit, dass die Werte in einer Grafik dargestellt werden. Im Fall von metrischen zeitlich geordneten Daten ist dies eine **Verlaufsgrafik**.

Die Kurse schwanken zwischen rd. 32 Euro und 36 Euro. Ist dies viel? Wo liegt das Zentrum? Wir werden auf diese Fragen in den folgenden Kapiteln eingehen.

R-Anweisung
Eine einfache Grafik wird mit

```
> plot(s.bmw, type = "b")
```

erzeugt. Mit der Option `type='b'` (b = both) werden die Datenpunkte mit Linien verbunden.

R-Anweisung für Experten
Für Abb. 5.1 wird eine Formatierung der x-Achse mit der folgenden Befehlsfolge notwendig:

```
> n.bmw <- length(s.bmw)
> datum <- Datum
> datum <- as.Date(datum[n.bmw:1], "%d.%m.%Y")
> plot(datum, s.bmw, type = "b", bty = "n", xaxt = "n")
> t <- seq(1, n.bmw, 10)
> axis(side = 1, at = datum[t], labels = format(datum[t],
+     format = "%d.%m.%y"))
```

Eine Alternative zu den obigen Formatanweisungen ist die Verwendung der `zoo` Bibliothek. Mit dieser Bibliothek können Variablen mit einem Datum als

W. Kohn, R. Öztürk, *Statistik für Ökonomen*, Springer-Lehrbuch,
DOI 10.1007/978-3-642-14585-8_5, © Springer-Verlag Berlin Heidelberg 2010

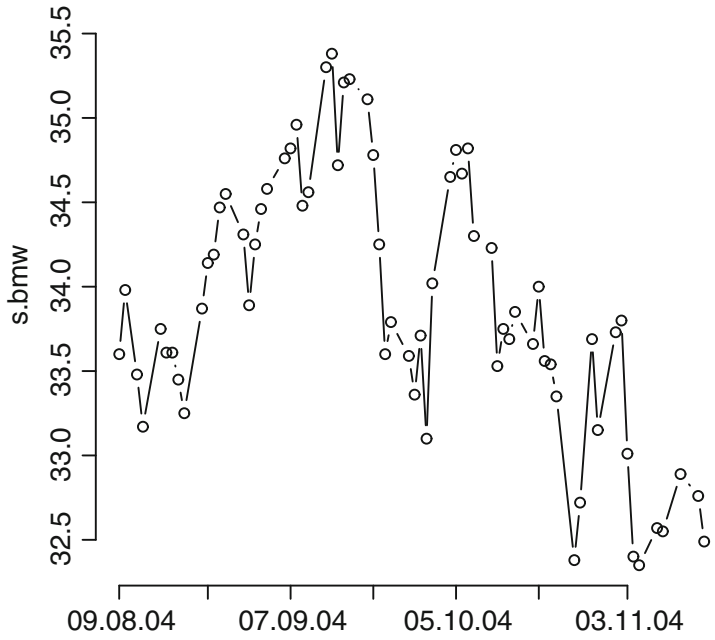


Abb. 5.1 Kursverlauf BMW Schlusskurse

Index erstellt werden. Bei der Grafik wird dann das Datum als Index auf der x-Achse abgetragen. Jedoch können einige R Funktionen diese Variablen nicht direkt verarbeiten. Sie müssen dann mit `as.numeric()` in eine numerische Variable transformiert werden.

```
> bmw <- read.table("bmw.txt", sep = ";", header = TRUE)
> z.bmw <- zoo(bmw$Schluss, as.Date(bmw$Datum, "%d.%m.%Y"))
> n.bmw <- length(bmw$Schluss)
> z.bmw <- z.bmw[n.bmw:1]
> plot(z.bmw)
```

Mit der Option `xaxp=c(`x_1, x_2, n`)` kann die Einteilung der x-Achse verändert werden. x_1 und x_2 geben Anfang und Ende der Einteilung an. n bestimmt die Unterteilung. `yaxp` ist der entsprechende Befehl für die y-Achse. Die ausführliche Beschreibung der Option ist mit `?par` unter `xaxp` auf den Hilfeseiten von R nachzulesen.