

Auf einen Blick

1	Über dieses Buch	13
2	Einführung	19
3	R Base und RStudio: Installation und erste Schritte	25
4	Die Programmiersprache R	51
5	Grundlagen der statistischen Datenanalyse	73
6	Daten einlesen und für die Analyse vorbereiten	93
7	Daten analysieren mit einfacher Statistik	129
8	Umfassendes Praxisbeispiel	187
9	Abschluss	233

Inhalt

Materialien zum Buch	11
----------------------------	----

1 Über dieses Buch

1.1 Für wen ist dieses Buch? Für Sie?	13
1.2 Was sind die Ziele, was können Sie hier lernen?	13
1.3 Was Sie nicht lernen werden	15
1.3.1 Verwandte Programmiersprachen	15
1.3.2 Big Data	15
1.3.3 Datenbankabfragen	16
1.3.4 Andere Datenformen	16
1.3.5 Kompliziertere Rechenverfahren	16
1.4 Wie Sie mit diesem Buch arbeiten	17

2 Einführung

2.1 Statistik und Data Science im Vergleich	21
2.1.1 Statistik	21
2.1.2 Data Science	22
2.2 Was ist R, und warum sollten Sie das überhaupt lernen?	23

3 R Base und RStudio: Installation und erste Schritte

3.1 R Base	25
3.1.1 Für Windows	27
3.1.2 Für Linux	28
3.1.3 Für macOS	29

3.2	RStudio	29
3.2.1	Für Windows	31
3.2.2	Für Linux	31
3.2.3	Für (Mac) OS X	31
3.3	Wie sieht die Oberfläche aus, und was bedeuten die einzelnen Bereiche?	31
3.4	Die R-Konsole	33
3.4.1	Erste Berechnungen in der R-Konsole	33
3.4.2	Skriptverlauf und Fehleranalyse	34
3.5	Mein erstes Skript	35
3.5.1	Skriptvorlage erstellen	35
3.5.2	Schreiben und Ausführen des ersten Skriptes	38
3.5.3	Abspeichern des erstellten Skripts	39
3.5.4	Laden des gespeicherten Skripts	41
3.6	Hilfe!	42
3.6.1	R-eigene Hilfe	42
3.6.2	Hilfe aus der riesigen R-Community	45

4	Die Programmiersprache R	51
4.1	Objekte	51
4.2	Funktionen	52
4.3	Pakete (Packages)	53
4.4	Ein paar Vokabeln	61
4.4.1	Die wichtigsten Funktionen und Befehle für Anfänger	61
4.4.2	Cheat-Sheets	67
4.5	Kommentare	68
4.6	Groß- und Kleinschreibung und andere Syntaxregeln	68
4.6.1	Benennung	68
4.6.2	Klammern	69
4.6.3	Syntax	69
4.7	Computer sind dumm	71

5 Grundlagen der statistischen Datenanalyse 73

5.1	Fragestellung und Studiendesign	73
5.2	Von Daten und Datensätzen	74
5.2.1	Querschnittsdaten	75
5.2.2	Längsschnitt- oder Paneldaten	78
5.2.3	Datenqualität	82
5.2.4	Datentypen in R (!)	82
5.2.5	Statistik	83
5.3	Berechnung des Durchschnittswertes (Mittelwert, arithmetisches Mittel)	84
5.4	Wachstumsberechnung (Veränderungsberechnung)	84
5.5	Trend und lineare Regression	86
5.6	Beispieldatensatz »Zigarettenverbrauch«	88

6 Daten einlesen und für die Analyse vorbereiten 93

6.1	Daten aus Excel einlesen	93
6.2	Daten im .csv-Format einlesen	98
6.2.1	Einlesen von .csv-Dateien mit RStudio	98
6.2.2	Umwandeln in Excel und weiter mit Abschnitt 6.1	101
6.3	Umgang mit Datumsangaben	102
6.3.1	Datumsangaben als character-Strings	103
6.3.2	Datumsangaben aus Individualkomponenten	108
6.4	Daten vorbereiten	109
6.4.1	Excel	112
6.4.2	In R	114
6.5	Not available! – Der Umgang mit fehlenden Werten	122
6.5.1	NAs entdecken mit is.na()	122
6.5.2	Löschen von NAs	126

7 Daten analysieren mit einfacher Statistik 129

7.1	Beispiel 1: Zigarettenkonsum	129
7.1.1	Mittelwert	132
7.1.2	Wachstum	133
7.1.3	Trend/lineare Regression	136
7.2	Beispiel 2: 100 Jahre Wohlstandsentwicklungen in Indien	144
7.2.1	Zusammenstellung der benötigten Daten	144
7.2.2	Analyse	157
7.3	Visualisierung	166
7.3.1	Grundlagen und Funktionsweise von ggplot2	166
7.3.2	Einfaches Streudiagramm mit dem Datensatz Luftqualität	167
7.3.3	Barplot: Beispiel Zigarettentatensatz	170
7.3.4	Einfacher Lineplot: Beispiel Zigaretten- und Gapminder-Daten	179

8 Umfassendes Praxisbeispiel 187

8.1	Fragestellung: Was will ich wissen?	188
8.2	Datenbeschaffung	189
8.3	Daten laden und Überblick verschaffen	191
8.3.1	Daten laden	191
8.3.2	Überblick verschaffen	191
8.4	Daten vorbereiten und bereinigen	195
8.4.1	NAs entfernen	195
8.4.2	Unplausible Werte entfernen	196
8.4.3	Weitere Anpassungen	196
8.5	Verarbeitung der Daten im Dreiklang	196
8.6	Kommunikation	219
8.7	Dokumentation	219
8.8	Bonus: Folgeanalyse in der Zeitreihe: Krankheitstage	219

Anhang235

A Datenmaterial	235
A.1 Zigaretten 1	235
A.2 Zensus Länder	236
A.3 Luftqualität	237
A.4 Zigaretten 2	243
A.5 Indien 1990–2000	244
A.6 Personaldaten	245
A.7 Personaldaten – Erweitert	252
 Index	261