

Jürgen Groß

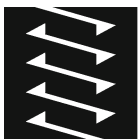
Grundlegende Statistik mit R

Jürgen Groß

Grundlegende Statistik mit R

Eine anwendungsorientierte Einführung
in die Verwendung der Statistik Software R

STUDIUM



VIEWEG+
TEUBNER

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

Dr. Jürgen Groß

E-Mail: gross@statistik.uni-dortmund.de

1. Auflage 2010

Alle Rechte vorbehalten

© Vieweg+Teubner Verlag | Springer Fachmedien Wiesbaden GmbH 2010

Lektorat: Ulrike Schmickler-Hirzebruch | Nastassja Vanselow

Vieweg+Teubner Verlag ist eine Marke von Springer Fachmedien.

Springer Fachmedien ist Teil der Fachverlagsgruppe Springer Science+Business Media.

www.viewegteubner.de



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

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Umschlaggestaltung: KünkelLopka Medienentwicklung, Heidelberg

Druck und buchbinderische Verarbeitung: MercedesDruck, Berlin

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.

Printed in Germany

ISBN 978-3-8348-1039-7

Vorwort

Das Buch zeigt, wie die statistische Aufbereitung und Auswertung von Daten mit Hilfe des unter der GPL frei verfügbaren Paketes R vorgenommen werden kann. Auf der Basis von aufeinander aufbauenden Lerneinheiten wird das notwendige Rüstzeug vermittelt, um auch ohne vorherige Programmierkenntnisse statistische Auswertungen durchführen zu können. Dabei werden eine Reihe statistischer Methoden beispielhaft angewendet.

Zielgruppe sind Studierende unterschiedlicher Fachrichtungen, die sich im Rahmen ihres Studiums mit der statistischen Aufbereitung und Auswertung von Daten beschäftigen.

- Es soll gezeigt werden, wie statistische Auswertungen auf möglichst unkomplizierte Art und Weise durchgeführt werden können. Daher werden nicht alle Möglichkeiten, die R bietet, in diesem Buch Erwähnung finden. Insbesondere der Aspekt der Programmierung, der beispielsweise von Ligges (2008) abgehandelt wird, wird auf das notwendige Maß eingeschränkt. Stattdessen ist es das Ziel, aufzuzeigen, wie bestimmte statistische Methoden praktisch angewendet werden können.
- Die Verfahren werden anhand von Beispiel-Datensätzen erläutert, die von R selbst zur Verfügung gestellt werden. Die Beispiele sind in den meisten Fällen in Form einer Aufgabe gestellt, deren mögliche Lösung unter Verwendung von R dann jeweils diskutiert wird.
- Das Buch hat den Charakter eines Kurses. Es kann als ein Begleiter und sinnvolle Ergänzung für Vorlesungen verwendet werden, die grundlegende statistische Methoden behandeln, wie sie z.B. in Fischer (2005), Schlittgen (2003) oder Fahrmeier et al. (2003) abgehandelt werden.

Die ersten drei Kapitel (Teil 1) des Buches bieten eine Einführung in den grundlegenden Umgang mit R. Kapitel 4 bis 18 (Teil 2) beschäftigen sich mit dem Umgang mit Daten, der Datenanalyse, sowie grundlegenden statistischen Verteilungsmodellen. Kapitel 19 bis 24 (Teil 3) beinhalten weiterführende Methoden der Varianz- und Regressionsanalyse, sowie den grundlegenden Umgang mit Methoden der Zeitreihenanalyse.

- Der größte Vorteil von R liegt natürlich in der freien Verfügbarkeit unter GPL. Nachdem man sich an die Bedienung etwas gewöhnt hat, ist es recht unkompliziert eigene oder fremde Datensätze zu laden und aufzubereiten.
- Ein weiterer Vorteil liegt darin, dass es eine große Gemeinschaft von Beitragenden gibt. Fehlerhafte Prozeduren werden schnell als solche erkannt und korrigiert. Zu-

dem gibt es zu nahezu jedem statistischen Themengebiet frei verfügbare Pakete, welche entsprechende Methoden bereits zur Verfügung stellen.

- Schließlich ist R auch eine vollständige komplexe Programmiersprache. Der fortgeschrittene Anwender kann damit eigene Methoden implementieren, oder auch bestehende Funktionen für eigene Zwecke geeignet ändern.

Einsteigern oder Wiedereinsteigern fällt es zunächst meist schwer, herauszufinden, welche Möglichkeiten von R zur Bearbeitung bestimmter statistischer Fragestellungen zur Verfügung gestellt werden. Hier ist es die Absicht und Hoffnung des Buches, sinnvolle Hilfestellungen zu geben und als nützliches Nachschlagewerk dienen zu können.

Mein Dank gilt den Autoren von R, siehe R Development Core Team (2009), sowie den zahlreichen Beitragenden. Zudem möchte ich mich bei all denjenigen Studierenden bedanken, deren Lernbereitschaft und motivierendes Interesse für die Anwendung statistischer Methoden mit Hilfe von R für mich eine Anregung dargestellt haben, dieses Buch zu schreiben. Schließlich danke ich Annette Möller, die durch intensives Korrekturlesen zu zahlreichen Verbesserungen beigetragen hat. Sämtliche verbleibenden Fehler und Irrtümer gehen zu Lasten des Autors.

Dortmund, im Februar 2010

Jürgen Groß

Inhalt

Vorwort	v
1 Schnellstart	1
1.1 R installieren	1
1.2 Die R Konsole	2
1.3 Erste Schritte	2
1.4 Skripte	5
1.5 Weitere Schritte	7
1.6 R beenden	10
2 Zusätzliche Pakete	11
2.1 Paketnamen und Hilfeseiten	11
2.2 Hilfen zu Paketen	12
2.3 Laden installierter Pakete	12
2.4 Weitere Pakete installieren	13
3 Attribute von R-Objekten	15
3.1 Das Klassenattribut	15
3.2 Umgang mit Attributen	16
4 Umgang mit Vektoren	19
4.1 Statistische Variablen	19
4.2 Operatoren	20
4.3 Funktionen	21
4.4 R-Objekte	22
4.5 Folgen von Zahlen	23
4.6 Indizierung	24

4.7	Logische Vektoren	25
4.8	Benannte Vektorelemente	28
4.9	Datentypen	28
4.10	Qualitative Variablen	29
4.11	Das zyklische Auffüllen	32
4.12	Runden	33
5	Umgang mit Datensätzen	35
5.1	Die Datenmatrix	35
5.2	Beispieldatensätze in R	36
5.3	Indizierung	37
5.4	Fehlende Werte	41
5.5	Gruppierende Variablen und Teildatensätze	43
5.6	Datensätze, Matrizen, Listen	45
5.7	Datensätze einbinden	47
5.8	Datensätze sortieren	48
6	Datensätze einlesen	49
6.1	Textformat	49
6.2	Fremde Dateiformate	51
7	Empirische Kenngrößen	53
7.1	Minimum und Maximum	53
7.2	Mittelwert und Median	54
7.3	Empirische Quantile	56
7.4	Kenngrößen Zusammenfassung	59
7.5	Streuung	60
7.6	Kenngröße pro Gruppe	61
8	Empirische Verteilungen	65
8.1	Klassenbildung und Häufigkeiten	65
8.2	Balken- und Stabdiagramm	66
8.3	Histogramm	67
8.4	Boxplot	70

8.5	Weitere grafische Darstellungen	72
9	Umgang mit Grafiken	77
9.1	Bildschirmfenster	77
9.2	Grafiken erstellen	78
9.3	Weitere Grafik-Einrichtungen	84
10	Theoretische Verteilungen	85
10.1	Zufallsvariablen	85
10.2	Dichten	87
10.3	Komplexe Grafiken	90
10.4	Verteilungsfunktion	93
10.5	Quantilfunktion	96
11	Pseudozufallszahlen	99
11.1	Reproduzierbarkeit	99
11.2	Ziehen von Stichproben aus Mengen	100
11.3	Ziehen von Stichproben aus Verteilungen	101
11.4	Simulationen	102
12	Das Ein-Stichproben Verteilungsmodell	107
12.1	Verteilungs-Quantil Diagramm	107
12.2	Normal-Quantil Diagramm	110
12.3	Maximum-Likelihood Anpassung	111
12.4	Kern-Dichteschätzer	117
13	Zwei-Stichproben Verteilungsmodelle	119
13.1	Verteilungen zweier Variablen	119
13.2	Streudiagramme	122
13.3	Korrelationskoeffizienten	127
13.4	Die bivariate Normalverteilung	130
14	Kontingenztafeln	133
14.1	Verbundene und unverbundene Stichproben	133

14.2	Häufigkeitsverteilung	134
14.3	Grafische Darstellungen	137
14.4	Mehrdimensionale Felder	139
15	Statistische Tests	141
15.1	Kritische Werte	141
15.2	P-Werte	143
15.3	Der Binomialtest	143
15.4	Statistische Hypothesentests	144
16	Ein- und Zwei-Stichprobentests	149
16.1	Der t -Test	149
16.2	Der Wilcoxon-Test	155
16.3	Test auf gleiche Varianzen	158
17	Tests auf Zusammenhang	159
17.1	Test auf Korrelation	159
17.2	Der Chi-Quadrat Unabhängigkeitstest	162
17.3	Zusammenhänge in 2×2 Kontingenztafeln	166
18	Anpassungstests	171
18.1	Der Chi-Quadrat Anpassungstest	171
18.2	Der Kolmogorov-Smirnov Anpassungstest	174
18.3	Tests auf Normalverteilung	175
19	Einfachklassifikation	179
19.1	Das Modell der Einfachklassifikation	179
19.2	Der F -Test	181
19.3	Parameterschätzer	184
19.4	Ungleiche Gruppen Varianzen	185
19.5	Nicht Normalverteilung	187
19.6	Der paarweise t -Test	187
20	Lineare Einfachregression	191
20.1	Das Modell	191

20.2	Modell Kenngrößen	192
20.3	Die Kenngrößen Zusammenfassung	198
20.4	Linearität	199
20.5	Normalität	200
20.6	Modell ohne Interzept	201
20.7	Prognosen	202
20.8	Grafische Darstellungen	203
21	Multiple Regression	205
21.1	Das multiple Regressionsmodell	205
21.2	Modell Kenngrößen	206
21.3	Anpassung	209
21.4	Diagnostische Diagramme	212
21.5	Schrittweise Regression	213
21.6	Eingebettete Modelle	218
21.7	Qualitative Einflussgrößen	219
22	Logistische Regression	223
22.1	Generalisierte lineare Modelle	223
22.2	Das Logit Modell	224
22.3	Das Logit Modell für Tabellen	231
23	Zeitreihen	235
23.1	Datenstrukturen	235
23.2	Grafische Darstellungen	239
23.3	Glätten	239
23.4	Differenzen- und Lag-Operator	242
23.5	Empirische Autokorrelationsfunktion	242
23.6	Das Periodogramm	245
23.7	Einfache Zeitreihen Modelle	250
24	ARIMA Modelle	251
24.1	Modellbeschreibung	251

24.2 Modellbildung	253
24.3 Modelldiagnose	257
24.4 Saisonale ARIMA Modelle	258
Literaturverzeichnis	261
Index	263