Statistik und ihre Anwendungen

Grundlagen der Datenanalyse mit R

Eine anwendungsorientierte Einführung

von
Daniel Wollschläger

1. Auflage

Springer 2012

Verlag C.H. Beck im Internet:
www.beck.de
ISBN 978 3 642 25799 5

Vorwort

Das vorliegende Buch liefert eine an human- und sozialwissenschaftlichen Anwendungen orientierte Einführung in die Datenauswertung mit R. R ist eine freie Umgebung zur umfassenden statistischen Analyse und grafischen Darstellung von Datensätzen, die befehlsorientiert arbeitet. Die Motivation für dieses Buch entstand aus dem Eindruck, dass sich R zwar unter statistischen Experten großer Beliebtheit erfreut, Anwender statistischer Verfahren aus Gebieten der empirischen Datenanalyse dagegen das Potential von R noch nicht gleichermaßen nutzen. Dieser Text soll daher jenen den Einstieg in R erleichtern, die in erster Linie grundlegende Auswertungsverfahren anwenden möchten, nicht aber über technische Vorkenntnisse mit befehlsgesteuerten Programmen verfügen.

Die hier getroffene Auswahl an statistischen Verfahren orientiert sich an den Anforderungen der Psychologie, sollte damit aber auch die wichtigsten Auswertungsmethoden anderer Human- und Sozialwissenschaften abdecken. Das Buch stellt die Umsetzung grafischer und deskriptiver Datenauswertung, nonparametrischer Verfahren, univariater linearer Modelle und multivariater Methoden vor. Dabei liegt der Fokus auf der Umsetzung der Verfahren mit R, nicht aber auf der Vermittlung statistischer Grundlagen. Von diesen wird hier angenommen, dass die Leserschaft mit ihnen vertraut ist. Auf Literatur zu den behandelten Verfahren wird zu Beginn der Abschnitte jeweils hingewiesen.

Kapitel 1 bis 3 dienen der Einführung in den Umgang mit R, in die zur Datenanalyse notwendigen Grundkonzepte sowie in die Syntax der Befehlssteuerung. Inhaltlich werden in Kap. 2 Methoden zur deskriptiven Datenauswertung behandelt, Kap. 3 befasst sich mit der Organisation von Datensätzen. Das sich an Kap. 4 zur Verwaltung von Befehlen und Daten anschließende Kap. 5 stellt Hilfsmittel für die inferenzstatistischen Methoden bereit. Diese werden in Kap. 6 (Regression), 7 (t-Tests und Varianzanalysen), 8 (klassische nonparametrische Tests), 9 (Resampling-Verfahren) und 10 (ausgewählte multivariate Methoden) behandelt. Das Buch schließt mit Kap. 11 zum Erstellen von Diagrammen und einem Ausblick auf den Einsatz von R als Programmiersprache in Kap. 12. Diese Reihenfolge der Themen ist bei der Lektüre keinesfalls zwingend einzuhalten. Da statistische Analysen praktisch meist gemeinsam mit der Datenorganisation und grafischen Illustra-

tion durchzuführen sind, empfiehlt es sich vielmehr, bereits zu Beginn auch Kap. 4 und 11 selektiv parallel zu lesen.

Um die Ergebnisse der R-eigenen Auswertungsfunktionen besser nachvollziehbar zu machen, wird ihre Anwendung an vielen Stellen durch manuelle Kontrollrechnungen begleitet. Die eigene Umsetzung soll zudem zeigen, wie auch Testverfahren, für die zunächst keine vorbereitete Funktion vorhanden ist, mit elementaren Mitteln prinzipiell selbst umgesetzt werden können. In den meisten Beispielen wird davon ausgegangen, dass die vorliegenden Daten bereits geprüft sind und eine hohe Datenqualität vorliegt: Fragen der Einheitlichkeit etwa hinsichtlich der Codierung von Datum und Uhrzeit, potenziell unvollständige Datensätze, fehlerhaft eingegebene oder unplausible Daten sowie doppelte Werte oder Ausreißer sollen ausgeschlossen sein. Besondere Aufmerksamkeit wird jedoch in einem eigenen Abschnitt dem Thema fehlender Werte geschenkt.

Im Buch wird an verschiedenen Punkten Bezug zu anderer Software genommen. Die folgenden dabei verwendeten Namen sind durch eingetragenes Warenzeichen der jeweiligen Eigentümer geschützt: Eclipse, Excel, Java, Linux, MacOS, MySQL, ODBC, OpenGL, OpenOffice, Oracle, RStudio, S, S+, SAS, SPSS, SQLite, Stata, TIBCO, Trellis, Unix, Windows.

Änderungen in der zweiten Auflage

In der vorliegenden zweiten Auflage bezieht sich das Buch auf Version `2.14` von R. Gegenüber der vorangehenden Auflage wurde es stärker aufgabenorientiert strukturiert sowie an vielen Stellen überarbeitet und inhaltlich ergänzt. So geht es etwa ausführlicher auf die Verarbeitung von Zeichenketten und Datumsangaben ein (Abschn. 2.12, 2.13) und beinhaltet eine vertiefte Darstellung der Kreuzvalidierung und Diagnostik von Regressionsmodellen (Abschn. 6.5, 6.6). Die Auswertung varianzanalytischer Fragestellungen berücksichtigt jetzt die Schätzung von Effektstärken (Abschn. 7.2–7.7). Als Tests auf gleiche Variabilität werden zusätzlich jene nach Fligner-Killeen sowie nach Mood und Ansari-Bradley besprochen (7.1.3, 8.4). Das neue Kap. 9 führt in die Anwendung von Bootstrapping und Permutationstests ein. Bei multivariaten Verfahren ist die Diskriminanzanalyse ebenso hinzugekommen wie eine Behandlung des allgemeinen linearen Modells (Abschn. 10.8, 10.9). Schließlich geht der Text nun auf Möglichkeiten zur Darstellung von Bitmap-Grafiken ein (Abschn. 11.5.10) und beschreibt detaillierter, welche Möglichkeiten für Funktionsanalyse und Debugging R bietet (Abschn. 12.3). Die R-Beispielauswertungen sind ausführlicher kommentiert und abschnittsübergreifend konsistenter. Der überarbeitete Index wurde nach inhaltlichen Schlagworten, R-Funktionen und Zusatzpaketen getrennt.

Korrekturen, Ergänzungen und Anregungen sind herzlich willkommen. Die verwendeten Daten sowie alle Befehle des Buches und ggf. notwendige Berichtigungen können Sie unter dieser Adresse beziehen:
`http://www.uni-kiel.de/psychologie/dwoll/r/`

Danksagung

Mein besonderer Dank gilt den Personen, die an der Entstehung des Buches in frühen und späteren Phasen mitgewirkt haben: Abschn. 1.1 bis 1.2.3 entstanden auf der Grundlage eines Manuskripts von Dieter Heyer und Gisela Müller-Plath am Institut für Psychologie der Martin-Luther-Universität Halle-Wittenberg, denen ich für die Überlassung des Textes danken möchte. Zahlreiche Korrekturen und viele Verbesserungsvorschläge wurden dankenswerterweise von Wolfgang Ramos, Julian Etzel, Erwin Grüner, Johannes Andres, Sabrina Flindt und Susanne Wollschläger beigesteuert. Johannes Andres danke ich für seine ausführlichen Erläuterungen der statistischen Grundlagen. Die Entstehung des Buches wurde beständig durch die selbstlose Unterstützung von Heike Jores und Vincent van Houten begleitet. Niels Peter Thomas und Alice Blanck vom Springer Verlag möchte ich herzlich für die freundliche Kooperation und Begleitung der Veröffentlichung danken.

Zuvorderst ist aber dem Entwickler-Team von R sowie der zahlreichen Zusatzpakete Dank und Anerkennung dafür zu zollen, dass sie in freiwillig geleisteter Arbeit eine hervorragende Umgebung zur statistischen Datenauswertung geschaffen haben, deren mächtige Funktionalität hier nur zu einem Bruchteil vorgestellt werden kann.

Kiel, *Daniel Wollschläger*
November 2011 dwoll@psychologie.uni-kiel.de