Einleitung

Willkommen bei *R für Dummies*, dem Buch, das die Lernkurve für die Statistik- und Programmiersprache R zu einem aufregenden Erlebnis ähnlich der Nordkurve macht.

Wir garantieren keineswegs, dass Sie nach der Lektüre dieses Buchs ein R-Guru sind, die folgenden Dinge werden Sie allerdings auf jeden Fall beherrschen:

- ✔ Sie führen Datenanalysen mit verschiedenen leistungsstarken Werkzeugen durch.
- ✔ Sie benutzen R für statistische Analysen sowie andere Daten verarbeitende Aufgaben.
- ✔ Sie lernen die Schönheit vektorbasierter Operationen im Vergleich zur Verwendung von Schleifen schätzen.
- ✔ Sie verstehen die Bedeutung folgender Codezeile:

```
knowledge <- apply(theory, 1, sum)
```

- ✔ Sie wissen, wie Sie Code von Mitgliedern der R-Gemeinde und anderen Entwicklern finden, herunterladen und einsetzen können.
- ✔ Sie wissen, wo Sie zusätzliche Hilfe und Ressourcen erhalten, um Ihre Fähigkeiten noch weiter auszubauen.
- ✔ Sie erzeugen wundervolle Grafiken und Visualisierungen Ihrer Daten und Ergebnisse.

Über dieses Buch

R für Dummies ist eine Einführung in die statistische Programmiersprache R. Wir beginnen zunächst mit der Benutzeroberfläche und arbeiten uns dann von ziemlich einfachen Konzepten der Sprache vor bis zu schon recht anspruchsvollen Themen der Datenverarbeitung und -analyse.

Jeder Schritt wird von einfach durchzuführenden Beispielen begleitet. Dieses Buch enthält zahlreiche Codeausschnitte, einige Baukastenfunktionen, die Sie später weiterverwenden können, sowie komplette Analyseskripte. All das hat im Wesentlichen einen Zweck: Sie sollen es selbst ausprobieren.

Wir versuchen erst gar nicht, eine technische Beschreibung zu unternehmen, wie R selbst programmiert wurde. Davon abgesehen sollen sich unsere Ausführungen, wie etwas funktioniert, in etwa die Waage halten mit den Gründen, warum das so funktioniert. R ist keine durchschnittliche Skriptsprache und hat einige Eigenschaften, die auf den ersten Blick überraschen mögen. Anstatt Ihnen einfach nur zu erzählen, wie Sie mit R sprechen sollen, glauben wir schon, dass es wichtig ist, Ihnen zu verraten, wie R Ihre Eingaben liest und interpretiert. Nachdem Sie dieses Buch gelesen haben, sollten Sie in der Lage sein, Daten in der von Ihnen gewünschten Form zu verarbeiten und Funktionen zu verwenden, die wir in diesem Buch nicht vorgestellt haben (die wir vorgestellt haben, möglichst auch).

Dieses Buch ist als Referenz gedacht. Sie müssen es nicht von Anfang bis Ende lesen. Stattdessen können Sie einfach das Inhaltsverzeichnis und das Stichwortverzeichnis nutzen, um die Informationen, die Sie brauchen, zu finden. In jedem Kapitel verweisen wir auf andere Kapitel, in denen Sie weitere Informationen finden.

Änderungen der zweiten Auflage

Seit der Publikation der ersten Auflage hat sich R kontinuierlich weiterentwickelt und verbessert. Um die Korrektheit des Buchs zu gewährleisten, haben wir den Code entsprechend der letzten Version von R (Version 3.4.1) angepasst. Basierend auf dem Feedback von Lesern, Studenten und Kollegen konnten wir einige Abschnitte überarbeiten und so Fragen klären und Ungenauigkeiten beheben. Beispielsweise haben wir den Code dahin gehend geändert, dass wir nun hochgestellte Gänsefüße statt Hochkommata innerhalb von Text-Strings verwenden. Auch bezeichnen wir jetzt die Basiseinheiten von Listen als Komponenten statt Elementen.

Änderungen der dritten Auflage

Die dritte Auflage des Buchs basiert auf der R-Version 4.0.4, die im Februar 2021 veröffentlicht wurde. Falls sich geänderte und neue Features auf die Themen und Beispiele auswirken, die im Buch vorgestellt werden, dann wurden Text und Code entsprechend angepasst.

Das aktuelle `rfordummies`-Paket enthält die Codebeispiele aus dem Buch. Sie können alles Weitere über das Paket in Anhang B erfahren.

R und RStudio

R für Dummies kann mit jedem Betriebssystem verwendet werden, auf dem R läuft. Ob Sie macOS, Linux oder Windows nutzen, mit diesem Buch bringen wir Sie auf den Weg mit R.

R ist mehr eine Programmiersprache als ein Anwendungsprogramm. Wenn Sie R herunterladen, laden Sie auch eine Konsolenanwendung herunter, die zu Ihrem Betriebssystem passt. Erwarten Sie jedoch keine Wunder: Die Funktionalität ist sehr eingeschränkt und variiert zwischen den einzelnen Betriebssystemen.

RStudio ist eine Plattform-übergreifende Anwendung, auch benannt als Integrated Development Environment (IDE), mit einigen sehr netten Eigenschaften in Bezug auf R. Dieses Buch setzt zwar keine spezielle Konsolenanwendung voraus. Da RStudio jedoch eine über alle Betriebssysteme einheitliche Anwendung ist, glauben wir, dass Sie es schnell zum Laufen bekommen. Daher nutzen wir lieber RStudio als den uneinheitlichen Editor, um die Konzepte im Buch vorzustellen.

Konventionen in diesem Buch

Codeausschnitte erscheinen wie in dem folgenden Beispiel, indem wir eine Million Würfe zweier sechsseitiger Würfel simulieren:

```
> set.seed(42)
> throws <- 1e6
> dice <- sapply(1:2,
+ function(x) sample(1:6, throws, replace = TRUE)
+ )
> table(rowSums(dice))

     2      3      4      5      6      7      8
 28007  55443  83382 110359 138801 167130 138808
     9     10     11     12
110920  83389  55816  27945
```

Jede Zeile des R-Codes in diesem Beispiel beginnt mit einem der folgenden Symbole:

- **>:** Das Anweisungssymbol >. Es ist nicht Teil des Codes. Geben Sie es nicht ein, wenn Sie den Code selbst ausprobieren.
- **+:** Das Fortsetzungssymbol +. Es zeigt an, dass diese Zeile eigentlich noch zur vorhergehenden gehört. Genau genommen müssen Sie gar keine Zeilenumbrüche in Ihren Code einbauen. Wir tun dies jedoch häufig, um die Lesbarkeit des Codes zu verbessern. Darüber hinaus ist es hilfreich, damit er auf die Buchseiten passt.

Die Zeilen, die weder mit dem einen noch mit dem anderen Symbol beginnen, sind Ausgaben von R. In unserem Beispiel erhalten Sie die Gesamtzahl an Würfen, in denen die Summe der Augen 2, 3, ..., 12 betrug. Zum Beispiel war die Summe der Augen in 28.007 von einer Million Würfen gleich zwei.

Sie können diese Codeausschnitte kopieren und in R ausführen. Achten Sie darauf, sie genau abzuschreiben. Es gibt nur drei Ausnahmen:

- Geben Sie nicht das Kommandosymbol > ein.
- Geben Sie nicht das Fortsetzungssymbol + ein.
- Tabulatoren oder Leerzeichen können Sie beliebig im Code verteilen, solange es nicht innerhalb von Schlüsselwörtern ist. Mit dem Zeilenvorschub sollten Sie etwas vorsichtiger umgehen.

Wenn R eine Eingabe von Ihnen erwartet, zeigt es das mit dem Symbol > ganz links in der Zeile, etwa so:

```
> print("Hallo Welt!")
```

Wenn Sie diese Anweisung in die Konsole eingeben und ↵ drücken, antwortet R mit:

```
[1] "Hallo Welt"
```

Aus Bequemlichkeit werden sowohl die Eingabe als auch die Ausgabe in einem Block angezeigt:

```
> print("Hallo Welt!")
[1] "Hallo Welt!"
```

Schließlich sei noch auf die Schriftart von R-Code im Buch hingewiesen. Die meisten Wörter in R sind von englischen Wörtern abgeleitet. Zwar ist dies bei einem deutschen Text nicht so verwirrend wie bei einem englischen. Dennoch setzen wir R-Funktionen, Argumente und Schlüsselwörter in `Monofont`. Funktionen werden immer zusammen mit nachgestellten geschlossenen Klammern dargestellt – zum Beispiel `plot()`. Funktionsargumente geben wir grundsätzlich nicht an und weichen davon nur in wichtigen Fällen ab.

Manchmal geht es um Menübefehle, zum Beispiel Datei | Sichern (File | Save). Dies bedeutet lediglich, dass Sie gebeten werden, das Menü Datei (File) zu öffnen und anschließend die Option Sichern (Save) zu wählen.

Was Sie nicht lesen müssen

Sie können dieses Buch so lesen, wie es für Sie am besten passt. Wenn Sie jedoch unter Zeitdruck stehen oder weniger an Details interessiert sind, können Sie problemlos alle rein technischen Informationen auslassen (zum betreffenden Symbol siehe weiter hinten in dieser Einleitung). Darüber hinaus können Sie auch alle grau hinterlegten Kästen auslassen. Zwar enthalten sie interessante Informationen, jedoch nichts, was für das Verständnis des Themas benötigt wird. Wenn Sie aber doch alles lesen, seien Sie gnädig wegen so mancher (absichtlicher) Wiederholung.

Törichte Annahmen über den Leser

Folgende Annahmen treffen wir über Sie als Leser sowie Ihren Computer:

- ✔ **Sie kennen sich mit Ihrem Computer bestens aus.** Sie wissen, wie man Software herunterlädt und installiert. Sie haben Zugang zum Internet und wissen, wie man dort Informationen findet.
- ✔ **Sie sind kein Programmierer.** Wenn Sie doch Programmierer sind und weitere oder andere Sprachen gewohnt sind, schauen Sie sich gern die rein technischen Informationen an (zum betreffenden Symbol siehe weiter hinten in dieser Einleitung). Dort erfahren Sie mehr dazu, wo R genauso oder anders tickt als andere Sprachen.
- ✔ **Sie sind kein Statistiker, aber Sie verstehen die Grundlagen der Statistik.** *R für Dummies* ist kein Buch über Statistik, obgleich wir Ihnen zeigen, wie man mit R einfache statistische Analysen durchführen kann. Wenn Sie mehr über Statistik erfahren wollen, empfehlen wir Ihnen *Statistik für Dummies* oder *Statistik mit R für Dummies* (beide erschienen im Wiley-VCH Verlag).

✔ **Sie wollen neue Dinge entdecken.** Sie mögen es, Probleme zu knacken, und haben keine Angst, mal etwas in der R-Konsole auszuprobieren.

Wie dieses Buch aufgebaut ist

R für Dummies gliedert sich in sechs Teile. Folgende Themen erwarten Sie in den einzelnen Teilen:

Teil I: Sind Sie beReit?

In diesem Teil lernen Sie R kennen und schreiben Ihr erstes Skript. Sie machen sich mit dem sehr nützlichen Vektorkonzept in R vertraut und führen Berechnungen simultan auf vielen Variablen aus. Sie lernen den R-Arbeitsbereich (englisch *workspace*) kennen, das heißt, wie Sie Variablen erzeugen, verändern und entfernen. Sie finden heraus, wie Sie Ihre Arbeit speichern und wie Sie Skriptdateien laden und verändern, die Sie in vorangegangenen Sitzungen erstellt haben. Darüber hinaus zeigen wir Ihnen ein paar Grundlagen in R, zum Beispiel wie Sie den Funktionsumfang erweitern, indem Sie Packages installieren.

Teil II: Arbeiten mit R

In diesem Teil füllen wir Sie ab mit den drei R: Reading (lesen), wRiting (schreiben) und aRithmetic (rechnen) – mit anderen Worten, wie Sie mit Text und Zahlen arbeiten, und nicht zu vergessen: mit Datumswerten. Hier lernen Sie auch die für das Leben mit R unerlässlichen Konzepte der Listen und Datensätze (*data frame*) kennen.

Teil III: Programmieren in R

R ist eine Programmiersprache. Daher ist es unerlässlich, dass Sie wissen, wie Sie Funktionen schreiben und durchblicken. In diesem Teil zeigen wir Ihnen genau das sowie wie Sie die Ablaufsteuerung mit `if` beeinflussen oder Code wiederholt mithilfe von Schleifen ausführen. Für den unwahrscheinlichen Fall bereiten wir Sie auch auf den Umgang mit Warnungen und Fehlermeldungen vor, die Sie möglicherweise mit Ihrem Code hervorrufen. Auch für die anschließende Fehlersuche machen wir Sie fit.

Teil IV: Daten zum Reden bringen

In diesem Teil stellen wir Ihnen die verschiedenen Datenstrukturen vor, die Sie in R verwenden können. Dazu gehören Listen und Datensätze (*data frame*). Sie erfahren, wie Sie Daten in R hinein- sowie herausbekommen (zum Beispiel indem Sie Dateien einlesen oder aus der Zwischenablage kopieren). Darüber hinaus sehen Sie, wie R mit anderen Anwendungen zusammenarbeiten kann, zum Beispiel mit Microsoft Excel.

Im Anschluss daran entdecken Sie, wie einfach es ist, fortgeschrittene Manipulationen an Ihren Daten vorzunehmen. Wir zeigen Ihnen, wie Sie eine Teilmenge Ihrer Daten auswählen

und wie Sie sie sortieren und ordnen. Wir erklären, wie Sie verschiedene Datensätze vereinigen können, wenn sie gemeinsame Spalten haben. Schließlich zeigen wir Ihnen eine sehr leistungsstarke generische Strategie, Daten zu teilen oder zu kombinieren und Funktionen auf Teilmengen Ihrer Daten anzuwenden. Nachdem Sie diese Strategie verstanden haben, können Sie sie immer wieder verwenden und anspruchsvolle Datenanalysen in wenigen Schritten durchführen.

Es juckt uns natürlich, Ihnen ein paar statistische Analysen zu zeigen. Schließlich ist Statistik die Domäne von R. Wir versprechen jedoch, es einfach zu halten. Nachdem Sie diesen Teil gelesen haben, werden Sie wissen, wie Sie Ihre Variablen und Daten mit R beschreiben und verdichten. Sie werden einige klassische Tests (zum Beispiel den t-Test) ausführen und Zufallszahlen für Simulationen erzeugen und verwenden.

Schließlich zeigen wir Ihnen ein paar Grundlagen, wie Sie lineare Modelle einsetzen können – zum Beispiel lineare Regression und Varianzanalyse (ANOVA). Darüber hinaus zeigen wir Ihnen, wie Sie R verwenden, um Vorhersagen auf Basis Ihrer Modelle zu treffen.

Teil V: Mit Grafiken arbeiten

Es heißt: »Ein Bild sagt mehr als tausend Worte.« Das ist sicherlich wahr, wenn es darum geht, Ihre Analysen mit anderen zu teilen. In diesem Teil entdecken Sie, wie Sie einfache und anspruchsvolle Grafiken einsetzen, um Ihre Daten zu veranschaulichen. Von Balken- und Liniendiagrammen angefangen geht es weiter bis hin zu *lattice*-Grafiken, mit denen Sie mehrdimensionale Daten in Scheiben schneiden und anschaulich machen können.

Teil VI: Der Top-Ten-Teil

In diesem Teil zeigen wir Ihnen, wie Sie Dinge in R erledigen, für die Sie wahrscheinlich bis heute Microsoft Excel verwendet haben – zum Beispiel Pivot- und Wertetabellen (englisch *lookup tables*). Darüber hinaus geben wir Ihnen zehn Tipps, wie Sie am besten mit Paketen (*package*) arbeiten, die nicht Teil des Basissystems sind.

Symbole, die in diesem Buch verwendet werden

Im Laufe der Lektüre dieses Buchs werden Sie über verschiedene Symbole stolpern. Diese Symbole kennzeichnen bestimmte Informationen.

Wenn Sie dieses Symbol sehen, können Sie sicher sein, dass sich hier ein Hinweis befindet, der Ihre Arbeit vereinfacht oder beschleunigt – oder beides.

Natürlich brauchen Sie das Buch nicht auswendig zu lernen. Wenn Sie jedoch dieses Symbol sehen, sollten Sie ernstlich in Erwägung ziehen, den zugehörigen

Hinweis im Gedächtnis zu behalten. Oft handelt es sich um ein Entwurfsmuster oder einen Ausdruck, dem Sie in mehr als einem Kapitel begegnen.

Wenn Sie dieses Symbol sehen, passen Sie auf! Es weist auf etwas hin, dass Sie – möglicherweise nach reiflicher Überlegung – eher nicht machen möchten. Obschon es sehr unwahrscheinlich ist, dass R ein richtiges Unglück verursacht, warnt Sie dieses Symbol vor Folgen, die zumindest zu Verwirrung führen können.

Rein technische Informationen, die Sie getrost überspringen können, sind mit diesem Symbol gekennzeichnet. Wir tun unser Bestes, die Informationen so interessant und relevant wie nur irgend möglich zu gestalten. Gleichzeitig nehmen Sie keinen Schaden, wenn Sie sich – zum Beispiel unter Zeitdruck – auf das absolut Notwendige konzentrieren wollen und großzügig über die so gekennzeichneten Textpassagen hinwegsehen.

Wie es weitergeht

Es gibt nur einen Weg, R zu lernen: es zu nutzen! In diesem Buch versuchen wir, Sie mit R bekannt zu machen, jedoch müssen Sie sich selbst an Ihren PC setzen und damit experimentieren. Tun Sie irgendetwas, damit das Buch offen neben dem Computer liegen bleibt, und greifen Sie in die Tasten!