

Vorwort

Die Statistiksoftware R gibt es schon seit 1992, aber erst seit kurzem erfährt sie die Aufmerksamkeit von Sozialwissenschaftlern, die sie eigentlich verdient hat. Auch ich bin erst seit relativ kurzer Zeit dabei: Dieses Buch entstand, während ich mich in R eingearbeitet habe. Ich hoffe, dass dadurch einige Fallen, in die ich beim Einarbeiten getappt bin, für die Leserinnen und Leser vermieden werden.

Ich möchte mich an dieser Stelle herzlich bei einer Reihe von Personen bedanken, die mir wesentlich bei der Entdeckung von R und dem Schreiben dieses Buchs geholfen haben.

Henriette Hunold, Tanja Kutscher und Martin Schultze haben sämtliche Kapitel in ihren Rohversionen gelesen und getestet und mich auf einige Fehler und didaktische Probleme aufmerksam gemacht. Walter Schreiber hat mir viele wertvolle Hinweise auf Websites, Pakete und Funktionen gegeben, die ich sonst womöglich nicht entdeckt hätte. Er hat außerdem das ganze Manuskript in einer einzigen Nacht durchgelesen, was ich ihm besonders hoch anrechne. Christian Geiser danke ich für seine Kommentare bezüglich meiner statistischen Erläuterungen. Ute Greveluhmann hat einige stilistische Besonderheiten und Tippfehler aufgedeckt.

Svenja Wahl und Grit Möller vom Beltz Verlag danke ich für ihre ständige Ansprechbarkeit, ihre inhaltliche und technische Unterstützung während des gesamten Schreibprozesses und für ihre Rückmeldungen zum Manuskript. Vielen Dank auch an Michael Eid, Mario Gollwitzer und Manfred Schmitt für die Vorversionen ihres genialen Lehrbuchs. Der Aufbau dieses Buchs ist weitgehend an ihrem Lehrbuch orientiert.

Mein Dank und meine Anerkennung gehen auch an die ganze R Community und besonders an diejenigen, die mir so schnell auf meine Mails geantwortet haben: Stephane Champely, John Fox, Kjetil Halvorsen, Rob Kabacoff, Ken Kelley, Mike Lawrence, William Revelle und David Winsemius.

So sehr mir die Arbeit an diesem Buch Spaß gemacht hat, so glücklich bin ich, dass es immer auch ein Leben ohne R gab. Danke, Wilhelm.

Für alle verbleibenden Unklarheiten und Fehler bin ich natürlich allein verantwortlich. Bitte kontaktieren Sie mich, wenn Sie Fehler entdecken sollten oder Vorschläge zur Verbesserung des Buches machen möchten.

Berlin, im November 2009

Maike Luhmann

1 Einleitung

Seit einiger Zeit macht die Statistiksoftware R auch außerhalb eingeweihter Statistiker-Kreise von sich reden. Aussagen wie „R macht so tolle Graphiken“ und „R kann sogar Strukturgleichungsmodelle“ hört man immer häufiger. Allerdings halten sich einige ungerechtfertigte Gerüchte hartnäckig, zum Beispiel „In R kann man seine Daten nicht sehen“ oder „R ist nur was für Programmierer“.

Diese und andere Vorurteile sollen mit diesem Buch aus der Welt geschafft werden. R ist viel leichter und schneller zu erlernen, als häufig geglaubt wird. Das spricht sich mittlerweile herum. R wird zunehmend auch von ganz normalen Anwendern benutzt. Jedes Jahr gibt es zudem mehr Fachbereiche, an denen nicht mehr kommerzielle Statistikprogramme, sondern R gelehrt wird. Mit Ihrer Entscheidung, R eine Chance zu geben, liegen Sie daher voll im Trend. Ich wünsche Ihnen dabei viel Erfolg und hoffentlich auch etwas Spaß!

Ach ja: In R kann man seine Daten sehen. Und Programmieren muss man nicht unbedingt können.

1.1 Warum R?

Es gibt viele Gründe, die für R sprechen. Hier sind die wichtigsten:

- ▶ **R kann mehr.** Die Software enthält viele Funktionen, die man bei kommerziellen Statistik-Programmen wie zum Beispiel SPSS vergeblich sucht.
- ▶ **R ist schnell.** Das Programm wird ständig weiterentwickelt, sodass viele neue statistische Methoden zuerst in R und erst Jahre später in anderen Programmen implementiert sind.
- ▶ **R ist ansprechbar.** Die Programmierer der einzelnen Pakete sind über E-Mail erreichbar und reagieren meiner Erfahrung nach innerhalb weniger Tage. So können Fragen schnell geklärt und eventuelle Programmierfehler behoben werden.
- ▶ **R schläft nicht.** R-Nutzer gibt es auf der ganzen Welt, und viele von ihnen sind in der Mailingliste eingetragen. So kann man seine Fragen rund um die Uhr klären.
- ▶ **R ist kostenlos!**

1.2 Für wen ist dieses Buch?

Dieses Buch richtet sich sowohl an Einsteiger, die zum ersten Mal mit einer Statistiksoftware arbeiten, als auch an Umsteiger, die R als eine Alternative zu anderen Statistikprogrammen ausprobieren möchten. Besondere Computer- oder gar Programmierkenntnisse brauchen Sie nicht. Da dieses Buch aber keine Statistik-Einführung ist, sollten Sie die Grundlagen der hier behandelten statistischen Tests kennen. Dafür empfehle ich das Lehrbuch von Eid, Gollwitzer und Schmitt (2010), an dem auch die Gliederung dieses Buchs orientiert ist.

Ich bin selbst Psychologin und habe daher Datenbeispiele aus psychologischen Studien gewählt. Darüber hinaus behandle ich einige Funktionen und statistische Verfahren, die besonders für Psychologinnen und Psychologen von Interesse sind, z. B. die Berechnung von Skalenwerten (Kap. 7) oder Verfahren für die Testkonstruktion wie Itemanalysen und Faktorenanalysen (Kap. 18). Grundsätzlich richtet sich das Buch jedoch an Interessierte aus allen Fachrichtungen der Sozialwissenschaften.

Leider ist in diesem Einführungsbuch kein Platz, um komplexere Verfahren wie Strukturgleichungsmodelle oder Modelle der Item Response Theorie detailliert zu besprechen. Sie werden jedoch in der Lage sein, sich in die Befehle und Pakete, die Sie für solche Verfahren brauchen, selbst einzuarbeiten. Weitere Hinweise dazu finden Sie im Anhang B: Pakete.

1.3 Wie benutzt man dieses Buch?

Den Umgang mit einer Statistiksoftware lernt man am besten durch Ausprobieren. Daher empfehle ich, bei der Lektüre dieses Buchs immer einen Computer mit R dabeizuhaben. So können Sie die beschriebenen Funktionen direkt ausprobieren. Die Datensätze dafür können Sie von der Website zu diesem Buch herunterladen: <http://www.beltz.de/r-fuer-einsteiger>. Dort finden Sie auch zahlreiche Zusatzmaterialien.

Hinweise für Einsteiger

Ich habe mich bemüht, dieses Buch so zu schreiben, dass auch Einsteiger ohne jegliche Erfahrung mit Statistik-Software mit R zurechtkommen können. In diesem Buch stelle ich die statistischen Verfahren in der Reihenfolge vor, in der sie typischerweise gelehrt werden. Wenn Sie dieses Buch also in Begleitung zu einer Statistik-Vorlesung in Psychologie oder anderen sozialwissenschaftlichen Fächern lesen, können Sie die Kapitel einfach von vorne nach hinten durcharbeiten.

Hinweise für Umsteiger

Als Umsteiger haben Sie bereits Erfahrung mit anderen kommerziellen Statistik-Programmen und interessieren sich möglicherweise nur für bestimmte Funktionen

in R. In diesem Fall brauchen Sie natürlich nicht das Buch von vorne bis hinten durcharbeiten. Beginnen Sie stattdessen mit Kapitel 20. Dieses Kapitel ist ein Crash-Kurs für Umsteiger und dient als Wegweiser durch das Buch. In diesem Kapitel werden auch die wesentlichen Unterschiede zwischen R und SPSS vorgestellt.

1.4 Weiterentwicklungen und Aktualität des Buchs

R wird ständig weiterentwickelt. Täglich werden vorhandene Pakete aktualisiert oder neue Pakete zur Verfügung gestellt. Wenn Sie regelmäßig mit R arbeiten, werden Sie vermutlich immer wieder neue Pakete entdecken, die Ihnen die statistischen Analysen noch leichter machen. Für die Erstellung dieses Buchs habe ich mich bemüht, immer die aktuellsten Versionen zu verwenden (Stand November 2009). Wenn Sie trotzdem veraltete Funktionen oder gar Fehler finden, lassen Sie mich dies bitte wissen.

1.5 Verwendete Schriftarten

Um die Übersichtlichkeit des Texts zu erhöhen, werden in diesem Buch verschiedene Schriftarten verwendet (s. Tab. 1.1).

Tabelle 1.1 Verwendung von Schriftarten im Buch

| Text | Beispiel |
|----------------------|-------------------|
| Menübefehle | DATEI → SPEICHERN |
| Namen von Paketen | Hmisc-Paket |
| Namen von Funktionen | mean-Funktion |
| Befehle | > c(1, 2, 3) |

2 Installation

R ist als kostenlose Software im Internet verfügbar. Alles was wir brauchen, um loszulegen, ist also ein Internetzugang. Um R zu installieren, muss man Schreibrechte für das Laufwerk haben, auf dem R installiert werden soll. Auch für die Arbeit mit R sollte man Schreibrechte für das Laufwerk haben, da man so weitere Pakete herunterladen und installieren kann (zur Rolle von Paketen s. Abschn. 2.3).

2.1 Download

Die Internetseite www.cran.r-project.org ist der Ausgangspunkt für die Arbeit mit R. Die Abkürzung CRAN steht für Comprehensive R Archive Network. Hier werden sowohl die Software als auch verschiedene Zusatzdateien (so genannte Pakete, s. Abschn. 2.3) und Dokumente zur Verfügung gestellt. Direkt auf der Startseite dieser Website befindet sich die Option *Download and Install R*. Hier kann man R für die Betriebssysteme Linux, MacOS X und Windows herunterladen.

Zum Download der **Windows-Version** gelangt man über den Link *Windows*. Auf der neuen Seite wählen wir den Link *base*. Wir gelangen nun auf eine Seite mit dem Titel *R-2.X for Windows*, wo wir die Setup-Datei herunterladen und abspeichern können. (Das *X* steht für die aktuelle Versionsnummer. Auf der CRAN-Seite wird immer nur die neueste R-Version angeboten.)

Zum Download der **Mac-Version** gelangt man über den Link *MacOS X*. Wir gelangen nun auf die Seite *R for MacOS X*, wo wir die Datei *R-2.X.dmg* auswählen und herunterladen.

2.2 Installation

Die Installationsschritte unterscheiden sich etwas zwischen Windows und MacOS. Daher besprechen wir die Installation für die beiden Betriebssysteme getrennt.

2.2.1 Installation unter Windows

Wir starten die Installation von R, indem wir die heruntergeladene Setup-Datei öffnen. Als erstes müssen wir die Sprache für den Setup auswählen. Nun öffnet sich der Setup-Assistent. Dieser Assistent führt uns Schritt für Schritt durch die Installation. Wenn wir einen Schritt abgeschlossen haben, klicken wir auf **WEITER**. Wenn wir einen Schritt rückgängig machen möchten, klicken wir auf **ZURÜCK** und ändern die

Einstellungen. Die Installation kann jederzeit durch die Option `ABBRECHEN` beendet werden. Folgende Schritte werden bei der Installation durchlaufen:

- (1) Der Lizenztext wird angezeigt und sollte durchgelesen werden.
- (2) `ZIELORDNER AUSWÄHLEN`. Wir wählen den Ordner, in dem das Programm installiert wird. Standardmäßig wird der Ordner `C:\Programme\R\R-2.X` vorgeschlagen.
- (3) `AUSWAHL VON KOMPONENTEN`. Hier wählen wir die Dateien aus, die installiert werden sollen. Für die meisten Leser wird die Option `Benutzerinstallation` geeignet sein.
- (4) `STARTOPTIONEN ANPASSEN`. Wählen wir die Option `JA`, können wir im Folgenden bestimmte Einstellungen von R bei der Installation festlegen. Dazu zählen der Anzeigemodus (Darstellung der einzelnen Fenster), der Hilfestil (Hilfedateien als PDF oder im Windows-Format) sowie der Internetzugang. Wählen wir `NEIN`, werden die einzelnen Schritte übersprungen und die Standardeinstellungen für R übernommen. Wir wählen diese Option.
- (5) `STARTMENÜORDNER AUSWÄHLEN`. Hier kann man den Namen im Startmenü ändern und entscheiden, ob ein Ordner im Startmenü erstellt werden soll.
- (6) `ZUSÄTZLICHE AUFGABEN AUSWÄHLEN`. Hier können wir entscheiden, ob zusätzliche Symbole auf dem Desktop oder in der Schnellstartleiste angelegt werden sollen. Außerdem legen wir hier die Art der Verknüpfung von R mit Windows fest (Speicherung der Versionsnummer, R mit `.rdata`-Dateien verknüpfen). Für die meisten Leser sind die Standardeinstellungen ausreichend.
- (7) Wir haben es geschafft: Das Programm wird installiert!

2.2.2 Installation auf MacOS

Um R auf dem Mac zu installieren, öffnen wir die Setup-Datei `R-2.X.dmg`, die wir von der CRAN-Seite heruntergeladen haben (s. Abschn. 2.1). Dadurch öffnet sich der Setup-Assistent. Dieser Assistent führt uns Schritt für Schritt durch die Installation. Bei der Installation werden die folgenden Schritte durchlaufen:

- (1) Wichtige Informationen lesen und zustimmen.
- (2) Software-Lizenzvereinbarungen lesen und zustimmen.
- (3) Standardinstallation: Hier kann der Ort der Installation verändert werden.
- (4) Durchführung der Installation.
- (5) Fertig!

An dieser Stelle noch ein allgemeiner Hinweis: Dieses Lehrbuch wurde mit Windows erstellt. Die Benutzeroberfläche sieht auf dem Mac manchmal etwas anders aus. Grundsätzlich lassen sich aber alle Funktionen an denselben Stellen wiederfinden. Alle Pakete lassen sich auch auf MacOS verwenden.

10 Bivariate deskriptive Statistiken

In diesem Kapitel behandeln wir geeignete Tabellen und statistische Maße für zwei Variablen. Für die Beschreibung von zwei nominal- oder ordinalskalierten Variablen werden häufig Kontingenztabelle (auch als Kreuztabelle bekannt) erstellt (Abschn. 10.1). Um die Beziehung zwischen zwei Variablen zu quantifizieren, stehen verschiedene Zusammenhangsmaße zur Verfügung. Das bekannteste Zusammenhangsmaß ist die Produkt-Moment-Korrelation (Abschn. 10.2). Dieses Zusammenhangsmaß darf jedoch nur verwendet werden, wenn beide Variablen metrisch skaliert sind. Für nicht-metrische Variablen gibt es alternative Zusammenhangsmaße (Abschn. 10.3).

Die Beispiele in diesem Kapitel beziehen sich auf den Datensatz `erstis.rda` (s. Anhang A: Datensätze). Um die Befehle zu verkürzen, haben wir diesen Datensatz mit der `attach`-Funktion aktiviert (s. Abschn. 7.1.1).

10.1 Kontingenztabelle

Kontingenz- bzw. Kreuztabelle sind eine besondere Form von Häufigkeitstabellen. In Kontingenztabelle werden alle Ausprägungen *mehrerer* Variablen miteinander kombiniert und für jede mögliche Kombination die Häufigkeit angegeben. Im folgenden Beispiel möchten wir wissen, wie viele Frauen und Männer jeweils in Berlin oder außerhalb Berlins gewohnt haben. Wir haben zwei Variablen (`geschl` und `berlin`) mit jeweils zwei Ausprägungen (weiblich und männlich bzw. ja und nein). Ähnlich wie bei den eindimensionalen Häufigkeitstabellen unterscheidet man auch hier zwischen absoluten und relativen Häufigkeiten sowie Prozentwerten.

10.1.1 Absolute Häufigkeiten

Eindimensionale Häufigkeitstabellen lassen sich in R mit dem `table`-Befehl anfordern (s. Abschn. 9.1). Diesen Befehl verwenden wir auch für Kontingenztabelle. Dazu werden einfach die beiden Variablen, die wir kombinieren möchten, als Argumente in der Funktion aufgenommen. Die Ausprägungen der ersten Variablen werden in den Zeilen, die Ausprägungen der zweiten Variablen in den Spalten dargestellt:

```
> table(berlin, geschl)
      geschl
berlin weiblich männlich
ja      94      49
nein    16      5
```

Darüber hinaus kann man die Zeilen- und Spaltensummen sowie die Gesamthäufigkeit anfordern. Am schnellsten geht das mit dem `addmargins`-Befehl:

```
> addmargins(table(berlin, geschl))
      geschl
berlin weiblich männlich Sum
ja      94      49 143
nein    16      5  21
Sum     110     54 164
```

Alternativ kann man die Zeilen- und Spaltensummen auch separat anfordern. Für die Zeilensumme bietet sich der `rowSums`-Befehl an. Der entsprechende Befehl für die Spaltensumme ist der `colSums`-Befehl. (`col` steht für *column*, der englische Begriff für Spalte):

```
> rowSums(table(berlin, geschl))
ja nein
143  21

> colSums(table(berlin, geschl))
weiblich männlich
110      54
```

10.1.2 Relative Häufigkeiten und Prozentwerte

Im vorherigen Kapitel wurde gezeigt, wie man relative Häufigkeiten und Prozentwerte anfordern kann und wie man diese Werte rundet. Diese Funktionen lassen sich genauso auf Kontingenztabellen anwenden. Bei Kontingenztabellen kommt aber noch eine Besonderheit dazu: Wir haben die Auswahl zwischen verschiedenen relativen Häufigkeiten, je nachdem, welchen Wert man als Referenz nimmt. Man hat die Wahl zwischen

- ▶ Gesamthäufigkeit
- ▶ Zeilensumme
- ▶ Spaltensumme

Relative Häufigkeiten in Bezug zur Gesamthäufigkeit

Diese relativen Häufigkeiten in Bezug zur Gesamthäufigkeit erhält man, wenn man den `prop.table`-Befehl wie gewohnt auf die Tabelle anwendet:

```
> prop.table(table(berlin, geschl))
      geschl
berlin weiblich männlich
ja      0.57317073 0.29878049
nein    0.09756098 0.03048780
```

Wenn man nun noch den `addmargins`-Befehl auf diesen Befehl anwendet, erhält man zusätzlich die relativen Zeilen- und Spaltenhäufigkeiten sowie die relative Gesamthäufigkeit, die natürlich den Wert 1 haben sollte:

```
> addmargins(prop.table(table(berlin, geschl)))
      geschl
berlin weiblich männlich      Sum
ja      0.57317073 0.29878049 0.87195122
nein    0.09756098 0.03048780 0.12804878
Sum     0.67073171 0.32926829 1.00000000
```

Multipliziert man diesen Ausdruck mit 100, erhält man die entsprechenden Prozentwerte. Um die Tabelle übersichtlicher zu machen, kann man die Werte mit dem `round`-Befehl runden.

Relative Häufigkeiten in Bezug zur Zeilensumme

Um die Häufigkeiten in Bezug auf die Zeilensumme anzufordern, wird der `prop.table`-Befehl um das Argument `,1` erweitert. Wenn man diese relativen Häufigkeiten innerhalb einer Zeile addiert, erhält man den Wert 1:

```
> prop.table(table(berlin, geschl),1)
      geschl
berlin weiblich männlich
ja      0.6573427 0.3426573
nein    0.7619048 0.2380952
```

Anstelle der relativen Häufigkeiten kann man auch direkt die Prozentwerte anfordern. Dafür steht der `rowPercent`-Befehl im `Rcmdr`-Paket zur Verfügung. Damit diese Funktion verwendet werden kann, muss zunächst das `abind`-Paket geladen werden.

```
> rowPercents(table(berlin, geschl))
      geschl
berlin weiblich männlich Total Count
ja      65.7    34.3    100    143
nein    76.2    23.8    100    21
```

In der Ausgabe werden für jede Zeile die Prozentwerte angegeben. In der Spalte `Total` können wir die Summe der Prozentwerte für jede Zeile ablesen. Wenig überraschend sind diese Summen in jeder Zeile genau 100. In der Spalte `Count` werden zusätzlich die absoluten Zeilenhäufigkeiten angegeben.

Tipp

Der `rowPercents`-Befehl hat eine eingebaute Rundungsfunktion. Dazu ergänzt man die gewünschte Anzahl der Dezimalstellen als zusätzliches Argument in dem Befehl. In dem folgenden Beispiel runden wir auf zwei Nachkommastellen:

```
> rowPercents(table(berlin, geschl), 2)
      geschl
berlin weiblich männlich Total Count
ja      65.73    34.27    100    143
nein    76.19    23.81    100    21
```

Relative Häufigkeiten in Bezug zur Spaltensumme

Um die Häufigkeiten in Bezug auf die Spaltensumme anzufordern, wird der `prop.table`-Befehl um das Argument `, 2` erweitert. Wenn man diese relativen Häufigkeiten innerhalb einer Spalte addiert, erhält man den Wert 1:

```
> prop.table(table(berlin, geschl), 2)
      geschl
berlin weiblich männlich
ja      0.8545455 0.9074074
nein    0.1454545 0.0925926
```

Der `colPercents`-Befehl ist das Pendant zum `rowPercents`-Befehl. Mit diesem Befehl kann man direkt die Spaltenprozentwerte anfordern und erhält darüber hinaus die absoluten Häufigkeiten für die einzelnen Spalten:

```
> colPercents(table(berlin, geschl))
      geschl
berlin weiblich männlich
ja      85.5    90.7
nein    14.5     9.3
Total   100.0   100.0
Count   110.0    54.0
```

Tipp

Obwohl Kontingenztabelle meist für nominal- oder ordinalskalierte Variablen erstellt werden, kann man auch metrische Variablen verwenden. Die dabei entstehenden Tabellen können unter Umständen aber sehr groß ausfallen. Probieren Sie dies einmal mit der Kombination der Variablen `alter` und `abi` aus!

10.1.3 Mehrdimensionale Kontingenztabelle

Man kann auch drei oder mehr Variablen in Kontingenztabelle miteinander kombinieren. Dafür erweitert man den Befehl um den entsprechenden Variablennamen. R teilt die eigentlich dreidimensionale Tabelle in mehrere zweidimensionale Kontingenztabelle auf. Das heißt konkret, dass für jede Ausprägung der Variablen, die als letztes aufgeführt ist, eine eigene Kontingenztabelle erstellt wird, in der die beiden anderen Variablen kombiniert werden.

In dem folgenden Beispiel wurden die drei Variablen `berlin`, `wohnort.alt` und `geschl` eingegeben. Die Variable `geschl` wird in dem Befehl als letzte aufgeführt, daher erhalten wir in der Ausgabe zwei Kontingenztabelle, eine für die Frauen und eine für die Männer:

```
> table(berlin, wohnort.alt, geschl)
, , geschl = weiblich

      wohnort.alt
berlin alte BL neue BL Berlin Ausland
ja      19     9    52     9
nein     4     6     2     3

, , geschl = männlich

      wohnort.alt
berlin alte BL neue BL Berlin Ausland
ja       7     6    28     4
nein     1     3     1     0
```