Eine kategoriale Variable

6

ÜBERBLICK

Kategoriale Daten entstehen durch die Klassifikation von Beobachtungen in Kategorien. Meistens wird die interessierende Information dadurch gewonnen, dass ausgezählt wird, wie oft bestimmte Kategorien vorkommen. Dadurch können Häufigkeitsverteilungen und Prozentsätze berechnet werden. Dieses Kapitel beschäftigt sich damit, wie man prüfen kann, ob beobachtete Häufigkeiten mit bestimmten Annahmen übereinstimmen. Es werden die Grundlagen statistischen Testens besprochen. Außerdem wird behandelt, wie man für einen Prozentsatz, den man aus einer Stichprobe gewonnen hat, Bereiche ermittelt, von denen man annehmen kann, dass sie mit einer bestimmten Sicherheit den wirklichen Prozentsatz, wie er in der Population vorkommt, enthalten. Dieses Kapitel legt die Basis zum Verständnis der Inferenz- oder schließenden Statistik.

LERNZIELE

Nach Durcharbeiten dieses Kapitels haben Sie Folgendes erreicht:

- Sie wissen, was absolute und relative sowie beobachtete und erwartete Häufigkeiten sind, und können diese in R berechnen, tabellieren und in Form von Balken- und Kreisdiagrammen grafisch darstellen.
- Sie wissen, was eine Null- und Alternativhypothese ist und was Testen von Hypothesen bedeutet. Sie können ein- und zweiseitige Hypothesen unterscheiden und formulieren.
- Sie sind in der Lage, ein Signifikanzniveau festzulegen und einen p-Wert beim Testen einer Hypothese zu beurteilen.
- Sie können in R einen Chi-Quadrat-Test sowie einen Binomial-Test bei verschiedenen Problemstellungen berechnen. Dabei sind Sie in der Lage, das Ergebnis technisch und inhaltlich zu interpretieren.
- Sie kennen die Bedeutung der Begriffe Schwankungsbreite und Konfidenzintervall und können solche in R erstellen und grafisch darstellen.

6.1 Einleitung

Kategoriale Information erhält man, wenn etwas (ein Merkmal oder Charakteristikum), das man an verschiedenen Personen, Unternehmen, Pflanzen etc. (also Beobachtungseinheiten) registriert, in eine von mehreren Kategorien fällt. Dieser Prozess kann unterschiedlich verlaufen. Die Kategorien können schon von vorneherein feststehen und man braucht seine Beobachtungen nur mehr entsprechend zuzuordnen. Oder aber man sammelt die Information zunächst in freier Form (z. B. offene Fragen in einem Fragebogen), um anschließend nach einem Kategorisierungsschema (das man eventuell erst entwickeln muss) die Beobachtungen zuzuteilen. Diese beiden Vorgänge könnte man auch als KLASSIFIKATION bezeichnen.

Etwas anders verläuft der Prozess der AGGREGATION. Der Begriff kommt aus dem Lateinischen und bedeutet Anhäufung oder Vereinigung. Hier werden schon vorhandene Daten in einfachere, zusammenfassendere Strukturen transformiert. Zum Beispiel kann man die Körpergröße von Personen, die man in cm gemessen hat, zu drei Kategorien aggregieren, nämlich klein, mittel und groß. Oder, bei zugrunde liegender kategorialer Information, kann man z. B. die Berufe Tischler, Maurer etc. in

die Berufskategorie Handwerker zusammenfassen. Das Resultat einer Aggregation ist also oft eine Reihe von Kategorien einer Variable. Im Extremfall, besonders bei metrischen Daten, kann aber auch eine einzelne Maßzahl das Ergebnis sein, wenn z. B. die durchschnittliche Geburtenrate für ein Land bestimmt wird. Der ▶ Exkurs 6.1 illustriert am Beispiel von Berufen in Deutschland solche Vorgänge.

Wenn man nun kategoriale Daten erhoben oder mittels Klassifikation bzw. Aggregation gewonnen hat, geht es darum, diese Information in geeigneter Weise aufzubereiten, um die enthaltene Information erfassen, untersuchen und weitervermitteln zu können.

Der Vorgang hierbei ist das Auszählen, d. h., man zählt ab, wie oft Kategorie 1, Kategorie 2 etc. insgesamt vorkommen. Das Resultat dieser Auszählung nennt man *Häufigkeit* oder *absolute Häufigkeit*.

(Absolute) Häufigkeit:
die Zahl, die zustande kommt, wenn man abzählt, wie oft eine bestimmte Kategorie in Daten vorkommt

Eng verbunden mit dem Begriff absolute Häufigkeit ist die

(Absolute) Häufigkeitsverteilung:
eine Zusammenstellung (tabellarisch oder grafisch), wie oft einzelne Kategorien einer Variable in einer Stichprobe (oder auch in der Population) vorkommen

Bezieht man die absolute Häufigkeit auf die Gesamtanzahl von Beobachtungen, dann erhält man *relative Häufigkeiten* oder *Anteile*.

Relative Häufigkeit oder **Anteil:**

$$\textit{relative Häufigkeit einer Kategorie} = \frac{\textit{absolute Häufigkeit}}{\textit{Gesamtanzahl von Beobachtungen}}$$

Anteile werden oft als Prozentsätze (also als Anteile von 100) angegeben, d. h.

Prozent:

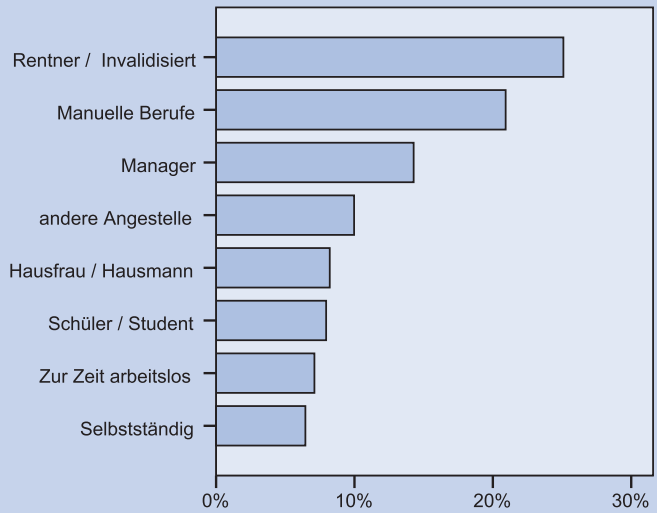
$$\textit{Prozentsatz einer Kategorie} = \textit{relative Häufigkeit} \times 100$$

Während Angaben von Anteilen in Prozentsätzen leichter zu lesen sind und auch allgemein üblich verwendet werden, hat die Darstellung mittels relativer Häufigkeiten den Vorteil, dass dadurch deren enge Verwandtschaft mit Wahrscheinlichkeiten ausgedrückt wird. Relative Häufigkeiten liefern (wenn zusätzlich die Gesamtzahl von Beobachtungen angegeben wird) die bedeutsamste Information, die man aus Daten gewinnen kann. Dies gilt nicht nur für kategoriale Daten. Auf diese Punkte werden wir später noch eingehen.

Exkurs 6.1 Wie kommt kategoriale Information zustande?

Beispiel: Berufe in Deutschland

Die simple Frage *Was arbeiten die Menschen eigentlich so?* ist gar nicht so einfach zu beantworten. Mögliche Quellen, wo man die entsprechende Information herbekommen könnte, sind die nationalen statistischen Behörden, für Deutschland das Statistische Bundesamt (www.destatis.de), für Österreich Statistik Austria (www.statistik.at), für die Schweiz das Bundesamt für Statistik (www.bfs.admin.ch) oder Eurostat als europäische Institution (ec.europa.eu/eurostat). Man könnte von einer dieser Quellen für Deutschland (2005) etwa folgende Grafik finden:



Die Grafik gibt einen guten Überblick, wie viel Prozent der Deutschen welcher Art von Tätigkeit nachgehen. Dieser Überblick ist allerdings aus zwei Gründen recht ungenau. Wenn wir wissen wollten, wie viele Personen tatsächlich selbstständig sind, müssten wir zunächst wissen, wie viele Personen insgesamt zur Erstellung der Grafik herangezogen worden waren (Kinder scheinen jedenfalls nicht dabei gewesen zu sein). Und zweitens kann man die angegebenen Prozentsätze nur ungefähr ablesen. Versuchen wir es dennoch. An anderer Stelle findet man auf der Webseite des deutschen statistischen Bundesamts, dass es im Jahr 2005 ca. 82 438 000 Einwohner in Deutschland gab, davon 11 649 800 Kinder und Jugendliche unter 16 Jahren. Für unsere Berechnung müssten wir als Gesamtzahl 70 788 200 verwenden. Laut Grafik sind etwa 7 % selbstständig, insgesamt gab es also 2005 knapp 5 Millionen Selbstständige.

Wir sehen hier zwei verschiedene Arten der Angabe von Zahlen im Zusammenhang mit kategorialer Information, nämlich (absolute) Häufigkeiten, d. h., wie oft ist insgesamt eine Kategorie vorgekommen, und Anteile (hier Prozentsätze), d. h., in welchem (relativen) Ausmaß kam eine bestimmte Kategorie im Verhältnis zu allen anderen Kategorien vor. Das eine Mal bezieht sich die

Information nur auf eine bestimmte Kategorie (ohne die anderen zu berücksichtigen), das andere Mal erhält man indirekt auch Information über weitere Kategorien. Auf diese Unterscheidung werden wir noch näher eingehen.

Der vorherigen Grafik kann man entnehmen, dass ca. 10 % als *andere Angestellte* bezeichnet werden. Was ist darunter zu verstehen? Aus den bisherigen Angaben ist dies nicht ohne Weiteres ersichtlich. Wir können annehmen, dass unter diesem Begriff eine größere Zahl von weniger häufig auftretenden Kategorien zusammengefasst wird. Solche Zusammenfassungen dienen der Übersichtlichkeit und der einfacheren Kommunizierbarkeit von wichtigen Informationsbestandteilen. Allerdings geht dabei natürlich ein Teil der ursprünglich vorhandenen Information verloren.

Tatsächlich wurde bei der Datenerhebung genauer gefragt.

1. Hausfrau/Hausmann und verantwortlich für den Haushaltseinkauf und den Haushalt (ohne anderweitige Beschäftigung)
2. Schüler/Student
3. zurzeit arbeitslos
4. Rentner/Pensionär/Frührentner/Invalidisiert
5. Landwirt
6. Fischer
7. Freie Berufe (Rechtsanwalt, Arzt, Steuerberater, Architekt usw.)
8. Ladenbesitzer, Handwerker usw.
9. Selbstständige Unternehmer, Fabrikbesitzer (Alleininhaber, Teilhaber)
10. Freie Berufe im Angestelltenverhältnis (angestellte Ärzte, Anwälte, Steuerberater, Architekten usw.)
11. Leitende Angestellte/Beamte, Direktor oder Vorstandsmitglied
12. Mittlere Angestellte/Beamte (Bereichsleiter, Abteilungsleiter, Gruppenleiter, Lehrer, Technischer Leiter)
13. Sonstige Büroangestellte/Beamte
14. Angestellte/Beamte ohne Bürotätigkeit mit Schwerpunkt Reisetätigkeit (Vertreter, Fahrer)
15. Angestellte/Beamte ohne Bürotätigkeit z. B. im Dienstleistungsbetrieb (Krankenschwester, Bedienung in Restaurant, Polizist, Feuerwehrmann)
16. Meister, Vorarbeiter, Aufsichtstätigkeit
17. Facharbeiter
18. Sonstige Arbeiter

In der Grafik auf Seite 150 wurden die Kategorien 13 und 14 unter andere Angestellte zusammengefasst. Die mit etwa 15 % relativ stark vertretene Gruppe der Manager umfasste die Kategorien 10 bis 12. Doch selbst diese feinere Gruppierung (wie sie in der Eurobarometerumfrage verwendet wird) kann in noch detailliertere Kategorien aufgeschlüsselt werden. So verwendet die deutsche Bundesagentur für Arbeit die vom deutschen statistischen Bundesamt entwickelte Berufssystematik, die auf der untersten Ebene 29 000 Berufsbezeichnungen definiert.

Anhand dieses Beispiels können wir einige Grundüberlegungen anstellen. Die Beobachtungseinheiten sind hier einzelne Personen. Die Variable, die beobachtet (bzw. aufgezeichnet) wurde, könnte man mit „berufliche Tätigkeit“ bezeichnen. Die Ausprägung dieser Variable ist kategorial, d. h., es wurde jede einzelne Person gefragt, in welche der vorher festgelegten Kategorien sie fällt. Die Daten könnten folgendermaßen aussehen.

Person	Berufstätigkeit
1	Mittlere Angestellte/Beamte
2	Rentner/Invalidisiert
3	Hausfrau/Hausmann
4	Leitende Angestellte/Beamte
5	Mittlere Angestellte/Beamte
6	Sonstige Arbeiter
7	Rentner/Invalidisiert
⋮	⋮

Diese fiktive Liste enthält Einzelinformationen über die Art der Beschäftigung. Solch eine Liste nennt man Rohdatenliste und so ähnlich könnte auch eine Computerdatei aussehen, mittels derer die vorangegangene Grafik erstellt wurde. In dieser Art der Datendarstellung ist natürlich die meiste Information zu finden, zumal ja der Beruf jeder einzelnen Person verzeichnet ist. Allerdings kann man sich leicht vorstellen, wie lang diese Liste ist. Man muss daher die Daten reduzieren bzw. zusammenfassen, um die darin enthaltene Information untersuchen und weitervermitteln zu können.

6.2 Kommen alle Kategorien gleich häufig vor?

Fallbeispiel 1: Der „blaue Montag“

Datenfile: `kstand.dat`

In einer repräsentativen Studie (frei nach einem Bericht des österreichischen Wirtschaftsforschungsinstituts, 2008) wurde erhoben, an welchem Wochentag der letzte Krankenstand bei 300 Beschäftigten begann. Die folgende Häufigkeitstabelle zeigt, wie viele Personen an den einzelnen Wochentagen in Krankenstand gingen.

Mo	Di	Mi	Do	Fr	Sa	So
96	60	51	45	30	9	9

Man sieht, dass die meisten Krankenstände an einem Montag begannen, gefolgt von Dienstag und Mittwoch. Samstag und Sonntag sind die Häufigkeiten deutlich geringer. Es stellt sich die Frage, ob tatsächlich eine größere Gefahr besteht, am Wochenbeginn zu erkranken, oder ob diese Häufungen nur zufällig sind.

Gibt es Wochentage, an denen man eher erkrankt, oder ist es an allen Wochentagen gleich riskant zu erkranken?

Bevor wir uns überlegen, wie man diese Frage beantworten kann, wollen wir einige Prinzipien besprechen, die bei der Analyse von Daten bzw. Fragestellungen grundsätzlich beachtet werden sollten.

Im Beispiel des Krankenstandsbeginns an einzelnen Wochentagen ist die Grundstruktur der Daten relativ einfach, es gibt (nur) eine Variable, `WOCHENTAG`, die die Werte *Montag* bis *Sonntag* annehmen kann. Trotzdem sind die Rohdaten sehr unübersichtlich. Nach Einlesen der Datei `kstand.dat` in den Data Frame `kstand` ermöglichen wir den direkten Zugriff auf die Variablennamen in diesem Data Frame mittels `attach()` und sehen wir uns die ersten 100 Fälle (von 300) an.

R

```
> kstand <- read.table("kstand.dat", header = TRUE)
> attach(kstand)
> WOCHENTAG[1:100]
> kstand <- read.table("Rdata/kstand.dat", header = TRUE)
> attach(kstand)
> WOCHENTAG[1:100]
```

```
  [1] MO DO DO DO DI MO DI MO MO DI DO DI MO MO MO MI MO DI
 [19] MO SO FR MI MI DI MO SO FR FR DO MI MI DO MO FR DI FR
 [37] MO DO MI MO MO SO DO MO MI MO DI MO FR DI MO MI DI DI
 [55] SA SA MO MO DI SA SO FR FR DI DO MO DO MO MO MI MO DO
 [73] MO MO DO MI MI MO DO FR MI FR MO DI DI FR FR MI MI DI
 [91] MO MO MO DI MO MO MI MO DO MI
Levels: DI DO FR MI MO SA SO
```

Hier haben wir zwar die maximal zur Verfügung stehende Information, allerdings ist diese so nicht kommunizierbar. Wir benötigen also Methoden, wie wir die Inhalte kompakt darstellen können, ohne zu viel Information zu verlieren. Dies kann anhand von Grafiken oder numerischen Zusammenfassungen (z. B. Tabellen) geschehen. Die entsprechenden Methoden werden unter dem Begriff Datenbeschreibung oder deskriptive Statistik zusammengefasst und bestehen, je nach Datentyp, aus spezifischen Verfahren. Numerische und grafische Beschreibung der Daten sollten immer der Ausgangspunkt bei der Analyse von Daten sein.

6.2.1 Numerische Beschreibung

Bei einfacher kategorialer Information zählen wir aus, wie häufig jede Kategorie auftritt, und erhalten absolute Häufigkeiten. Dividieren wir diese noch zusätzlich durch die Gesamtanzahl an Beobachtungen, ergibt das die relativen Häufigkeiten. Wir können beide in Tabellenform darstellen und erhalten eine übersichtliche numerische Beschreibung des Datenmaterials, aus der schon einige Aspekte zur Beantwortung der Fragestellung ablesbar sind.

In R würden wir folgendermaßen vorgehen:

Eine einfache Häufigkeitstabelle erhalten wir mittels

R

```
> table(WOCHENTAG)
```

```
WOCHENTAG
DI DO FR MI MO SA SO
60 45 30 51 96  9  9
```

Die relativen Häufigkeiten errechnet man, indem die Anzahl der Beobachtungen in jeder Kategorie durch die Gesamtanzahl der Beobachtungen dividiert wird (die Gesamtanzahl entspricht natürlich der Länge des Datenvektors):

R

```
> table(WOCHENTAG)/length(WOCHENTAG)
```

```
WOCHENTAG
  DI   DO   FR   MI   MO   SA   SO
0.20 0.15 0.10 0.17 0.32 0.03 0.03
```

Alternativ geht es auch mittels

R

```
> prop.table(table(WOCHENTAG))
```

```
WOCHENTAG
  DI   DO   FR   MI   MO   SA   SO
0.20 0.15 0.10 0.17 0.32 0.03 0.03
```

Prozentwerte erhalten wir, indem wir noch mit 100 multiplizieren.

R

```
> prop.table(table(WOCHENTAG)) * 100
```

```
WOCHENTAG
DI DO FR MI MO SA SO
20 15 10 17 32  3  3
```

Wenn wir alles in einer Tabelle darstellen wollen, dann würden wir noch ein wenig Kosmetik betreiben.

Zunächst wollen wir die Wochentage, die beim Einlesen standardmäßig alphabetisch nach ihren Namen sortiert sind, in der richtigen Reihenfolge darstellen. Dazu müssen wir die `levels` des Faktors `WOCHENTAG` in die richtige Reihenfolge bringen. Dies geschieht, indem wir mit `factor()` einen neuen Faktor `Wochentag` mit den entsprechenden Eigenschaften definieren (wir können den Variablennamen `Wochentag` verwenden, weil R ja Groß- und Kleinschreibung unterscheidet).

R

```
> Wochentag <- factor(WOCHENTAG, levels = c("MO",
+     "DI", "MI", "DO", "FR", "SA", "SO"))
```

Dann legen wir die Vektoren für die absoluten und relativen Häufigkeiten bzw. für die Prozentwerte (in `absH`, `relH` und `proz`) an.

R

```
> absH <- table(Wochentag)
> relH <- table(Wochentag)/length(Wochentag)
```

Schließlich wollen wir nicht alle Kommastellen anzeigen und runden daher `relH` auf zwei Kommastellen bzw. `proz` auf ganzzahlig. Für die tabellarische Darstellung hängen wir zuletzt die entsprechenden Vektoren mit `cbind()` zusammen.

R

```
> proz <- relH * 100
> relH <- round(relH, digits = 3)
> proz <- round(proz, digits = 0)
> cbind(absH, relH, proz)
```

```
   absH relH proz
MO   96 0.32   32
DI   60 0.20   20
MI   51 0.17   17
DO   45 0.15   15
FR   30 0.10   10
SA    9 0.03    3
SO    9 0.03    3
```

Die ersten beiden Spalten zeigen die absoluten und relativen Häufigkeiten, die dritte die Prozentwerte für die Anzahl begonnener Krankenstände an den einzelnen Wochentagen. Man sieht, dass an Montagen und Dienstagen eine Häufung auftritt, während sich die Zahlen im weiteren Wochenverlauf verringern.

Diese Tabelle ist sehr spartanisch gehalten und würde in dieser Form auch nicht in einer Publikation oder Präsentation verwendet werden (▶ Exkurs 6.2: Einige Prinzipien zur Erstellung guter Tabellen). Für eine erste Analyse reicht es aber und man kann ja den R-Output leicht in Textverarbeitungsprogramme (wie z. B. Word) exportieren, um dann das Aussehen einer Tabelle zu modifizieren. Dies wird in Abschnitt 5.3 beschrieben.

Exkurs 6.2 Einige Prinzipien zur Erstellung guter Tabellen

Will man kategoriale Information in Form von Zahlenmaterial geeignet präsentieren, so wird man den R-Output nicht direkt verwenden, sondern noch ein wenig überarbeiten. Folgende Punkte sollte man beachten:

- Eine Tabelle sollte für sich allein stehen können, d. h., alles Wichtige sollte ohne weitere Erklärung verständlich sein.
- Angabe eines geeigneten Titels und der Datenquelle
- Benennung der Kategorien
- Bei ungeordneten Kategorien: diese gegebenenfalls nach ihrer Häufigkeit ordnen
- Angabe der Gesamtanzahl von Beobachtungen
- Zahlen runden (besonders Nachkommastellen nur angeben, wenn sie für das Verständnis der Größenordnung wichtig sind)
- Sehr große Tabelle eventuell in Teiltabellen zerlegen
- Keine vertikalen Linien zeichnen

Die folgende Tabelle soll diese Prinzipien illustrieren (die Daten stammen aus einer Kriminalstatistik des Staates New Jersey, aus dem Jahr 2005).

Verteilung begangener Morde nach Wochentagen in New Jersey, 2005[a]

	Sonntag	Montag	Dienstag	Mittwoch	Donnerstag	Freitag	Samstag	Gesamt
Häufigkeit	53	42	51	45	36	37	65	329
Prozent	16	13	16	14	11	11	20	

a. Quelle: http://www.njsp.com

6.2.2 Grafische Beschreibung

Der Spruch *Ein Bild sagt mehr als tausend Worte* klingt sehr abgedroschen, trotzdem stimmt er. Gerade dann, wenn Information klar gemacht und vermittelt werden soll, ist eine gute grafische Darstellung besonders wichtig. Bei kategorialen Daten haben sich einige Methoden bewährt, die hier beschrieben werden sollen.

Balkendiagramm (Bar Chart)

Diese Darstellungsweise ist wohl die wichtigste grafische Methode bei kategorialen Daten. Bei einem Balkendiagramm werden die Kategorien auf der horizontalen Achse und die absoluten oder relativen Häufigkeiten auf der vertikalen Achse aufgetragen, wobei die Höhe der einzelnen Balken den Häufigkeiten in den einzelnen Kategorien entspricht. Um zu veranschaulichen, dass die darzustellende Variable kategorial ist, lässt man einen kleinen Leerraum zwischen den Balken (man macht dies im Gegensatz zu einem sogenannten Histogramm, das wir später besprechen werden). Manchmal findet man auch die beiden Achsen vertauscht vor (▶ Exkurs 6.1), an der Information, die man aus solch einer Grafik ablesen kann, ändert sich dadurch aber nichts.

Für unser Beispiel des Beginntags von Krankenständen erzeugt man ein Balkendiagramm in R folgendermaßen:

R

```
> barplot(absH)
```

Am Balkendiagramm in ▶ Abbildung 6.1 sieht man (vielleicht noch besser als in der Häufigkeitstabelle), dass Montag und Dienstag die meisten Krankenstände beginnen, am Samstag und Sonntag die wenigsten.

Kreisdiagramm (Pie Chart)

Häufig begegnet man auch einer zweiten Form der grafischen Beschreibung von einfacher kategorialer Information, den Kreisdiagrammen (manchmal auch Torten- oder Kuchendiagramme genannt). Hierbei werden die Daten in Kreisform dargestellt. Die Kategorien bilden Kreissegmente, wobei deren Fläche proportional zu den relativen Häufigkeiten bzw. Prozenten des Auftretens der Kategorien dargestellt werden.

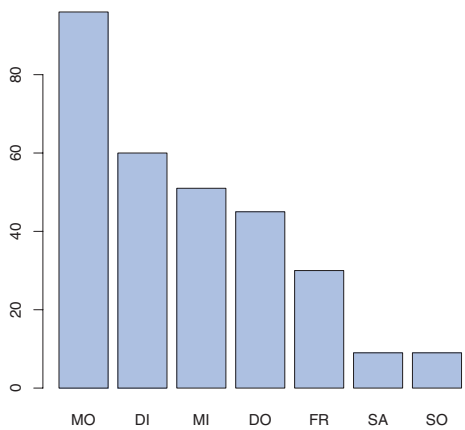


Abbildung 6.1: Balkendiagramm für Krankenstandsbeginn nach Wochentagen

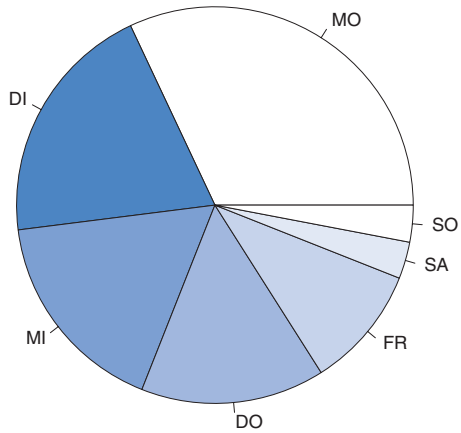


Abbildung 6.2: Kreisdiagramm für die Anzahl von begonnenen Krankenständen nach Wochentagen

Obwohl sie relativ beliebt sind, haben Kreisdiagramme doch eine Reihe von Nachteilen. Standardmäßig gibt R nur das Kreisdiagramm und die Werte für die Kategorien aus.

Für die Daten aus unserem Beispiel erzeugt man in R ein Kreisdiagramm (▶ Abbildung 6.2) mit

R

```
> pie(absH)
```

Man kann nicht (wie beim Balkendiagramm) die tatsächlichen Häufigkeiten ablesen und vergleichsweise sind Unterschiede nur schwer zu erkennen, außer sie sind sehr groß. Man sollte zumindest Häufigkeiten oder Prozente hinzufügen, um zu einer geeigneteren Darstellung (▶ Abbildung 6.3) zu kommen. Selbstdefinierte Beschriftungen werden im `pie()`-Befehl über die Option `labels=` spezifiziert. Hier haben wir die neuen Beschriftungen mittels der (sehr mächtigen) Funktion `paste()` (auf die wir hier nicht im Detail eingehen wollen) zusammengebastelt und das Resultat in `lab` gespeichert. (In der Zusammenfassung der R-Befehle am Ende des Kapitels gibt es ein paar Beispiele zu `paste()`, für eine ausführliche Beschreibung sei auf die Hilfeseite, erreichbar über `?paste`, verwiesen.) Außerdem fordern wir hier noch mittels der Option `clockwise = TRUE` an, dass die Kategorien im Uhrzeigersinn ausgegeben werden. Die adaptierte Grafik erhält man mit

R

```
> lab <- paste(names(absH), "\n(", proz, "%)\n", sep="")
> pie(absH, labels = lab, clockwise = TRUE)
```

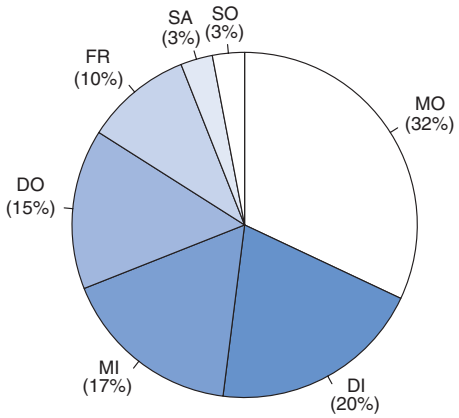


Abbildung 6.3: Kreisdiagramm mit Beschriftung

Im Vergleich zu Kreisdiagrammen sind Balkendiagramme (► Abbildung 6.1) lesbarer und informativer, weil Längen besser wahrgenommen werden können als Winkel. Zu erwähnen ist noch, dass die Beschriftungen, wie wir sie für das Kreisdiagramm in ► Abbildung 6.3 angebracht haben, in analoger Weise auch für Balkendiagramme möglich sind. Notwendig sind sie dort aber nicht, weil durch die Achsenbeschriftungen die relevante Information schon mitgeliefert wird.

6.2.3 Statistische Analyse der Problemstellung

Nachdem wir nun besprochen haben, wie die Information, die in den Daten steckt, numerisch und grafisch beschrieben werden kann, wollen wir uns der Beantwortung der eigentlichen Problemstellung zuwenden. Die Frage ist, ob es bestimmte Wochentage gibt, an denen Krankenstände häufiger beginnen als an anderen, oder ob sich der Beginn über die Wochentage gleich verteilt.

Um diese Frage mittels statistischer Methoden beantworten zu können, benötigen wir das Konzept der ERWARTETEN RELATIVEN HÄUFIGKEITEN und der ERWARTETEN (ABSOLUTEN) HÄUFIGKEITEN. *Beobachtete Häufigkeiten* (absolut oder relativ) entstehen durch Abzählen, *erwartete* Häufigkeiten sind solche, die man unter bestimmten Annahmen erwarten würde. In ► Exkurs 6.3 wird dieses Konzept anhand der Idee des *fairen Würfels* erläutert.

Exkurs 6.3 Der faire Würfel

Ein Würfel wird dann als „fair" bezeichnet, wenn alle sechs Seiten gleich häufig auftreten, oder mit anderen Worten, wenn die Wahrscheinlichkeit für jede der möglichen Seiten 1 bis 6 gleich groß ist. Die Erfahrung zeigt uns, dass wir nicht damit rechnen können, dass bei sechsmaligem Würfeln als Resultat alle sechs Augenzahlen je einmal vorkommen. Allerdings, wenn wir nicht nur sechs Mal, sondern viel öfter würfeln, wie das auch bei Gesellschaftsspielen der Fall ist, dann erwarten wir schon, dass alle Seiten am Ende ungefähr gleich oft aufgetreten sind. Würden wir

immer längere Serien würfeln, also 10 000 Mal, 100 000 Mal etc., dann sollten die Prozentsätze des Auftretens der einzelnen Augenzahlen immer ähnlicher werden und wir könnten sagen, dass der Würfel fair ist. Das setzt aber voraus, dass der Würfel verschiedene Voraussetzungen erfüllt, nämlich vollkommen gleichmäßige Beschaffenheit des Materials, aus dem er hergestellt ist, exakt gleiche Kantenlängen (nicht nur auf den hundertstel Millimeter genau) etc. Solch einen Würfel gibt es natürlich nicht, aber man kann einen herstellen, bei dem diese Abweichungen nur sehr klein und daher vernachlässigbar sind. Bei manchen Würfeln kann einem aber schon der Verdacht kommen, dass er nicht fair ist. Wie könnten wir das nun prüfen? Die Antwort ist Ausprobieren. Nehmen wir einmal an, wir hätten 30 Mal gewürfelt und uns die Häufigkeiten für die einzelnen Seiten wie folgt notiert:

gewürfelte Augenzahl	1	2	3	4	5	6
beobachtete Häufigkeit	3	8	4	5	6	4

Unter der Annahme, dass der Würfel fair ist, hätten wir eigentlich erwartet, dass jede Seite gleich häufig, nämlich zu 1/6 bzw. zu 16.7 % aufgetreten wäre. Diese erwarteten Anteile nennt man auch ERWARTETE RELATIVE HÄUFIGKEITEN. Multipliziert man sie mit der Gesamtanzahl der Würfe, hier mit 30, so erhält man die ERWARTETEN (ABSOLUTEN) HÄUFIGKEITEN. Demnach hätten wir je 5 Mal die Augenzahlen 1 bis 6 erwartet und nicht 3, 8, 4, ... Diese Abweichungen schreiben wir dem Zufall zu. Stutzig hätte uns womöglich gemacht, wenn z. B. die Seite mit 6 dreißig Mal und die anderen Seiten kein einziges Mal vorgekommen wären. Aber ab wann sollte uns die Sache verdächtig vorkommen? Diese Frage wollen wir versuchen zu beantworten.

Für unsere Fragestellung wollen wir die Annahme prüfen, ob der Krankenstandsbeginn sich gleichmäßig über die Wochentage verteilt. Unter dieser Annahme müssten von den insgesamt 300 Beginntagen 300/7 = 43.9 an jedem Wochentag auftreten, wir würden also ca. 43 an jedem Tag *erwarten*.

In R können wir eine entsprechende Tabelle erzeugen.

R

```
> erwH <- rep(300/7, 7)
> residuum <- absH - erwH
> erwH <- round(erwH, digits = 1)
> residuum <- round(residuum, digits = 1)
> cbind(absH, erwH, residuum)
```

```
   absH erwH residuum
MO   96 42.9     53.1
DI   60 42.9     17.1
MI   51 42.9      8.1
DO   45 42.9      2.1
FR   30 42.9    -12.9
SA    9 42.9    -33.9
SO    9 42.9    -33.9
```

Die Tabelle hat drei Spalten. In der ersten (`absH`) finden wir die tatsächlich beobachteten Häufigkeiten für die einzelnen Wochentage, in der nächsten (`erwH`) die erwarteten Häufigkeiten und in der letzten (`residuum`) die Differenz der ersten beiden Spalten.

Wir sehen, dass die Differenzen zwischen −33 und 53 liegen. Wie beim fairen Würfel können wir uns nun fragen, ob diese Abweichungen zufällig zustande gekommen sind oder ob mehr dahinter steckt. Je größer die Abweichungen insgesamt sind, umso weniger werden wir an Zufall glauben. Sind sie groß genug, dann werden wir die Frage verneinen, dass die Beginntage der Krankenstände gleichmäßig verteilt sind. Aber wo ist die Grenze zwischen *rein zufällig* und nicht mehr *durch reinen Zufall* erklärbar? Hier kann uns die Statistik weiterhelfen.

Exkurs 6.4 Statistische Tests – Hypothesen – p-Wert Was ist noch Zufall und was nicht mehr?

Am Beginn steht immer eine Fragestellung, z. B. ob ein Würfel fair ist. Die Methode, die man verwendet, um solch eine Frage zu beantworten, nennt man STATISTISCHEN TEST. Es gibt zwei mögliche Ergebnisse: Wir glauben entweder

- der Würfel ist fair oder
- der Würfel ist nicht fair.

Diese beiden Statements nennt man STATISTISCHE HYPOTHESEN. Die sogenannte NULLHYPOTHESE (auch mit H_0 bezeichnet) lautet:

- der Würfel ist fair oder
  H_0: alle Augenzahlen beim Würfeln sind gleich wahrscheinlich

Das logische Gegenteil ist die ALTERNATIVHYPOTHESE (oder H_A), die uns helfen soll, solche Annahmen zu überprüfen. Sie lautet:

- der Würfel ist nicht fair oder
  H_A: zumindest eine Augenzahl ist wahrscheinlicher als die anderen

Ein STATISTISCHER TEST hilft uns, eine Entscheidung zugunsten der Null- oder der Alternativhypothese zu treffen. Man kann dabei eine Wahrscheinlichkeit angeben, ob beobachtete Daten für die Nullhypothese sprechen. Diese Wahrscheinlichkeit bezeichnet man als P-WERT. Die Entscheidungsregel lautet:

- Ist der p-Wert klein, glaubt man nicht an die Nullhypothese, man glaubt eher, dass die Alternativhypothese zutrifft – der Würfel ist nicht fair.
- Ist der p-Wert groß, glaubt man an die Nullhypothese – der Würfel ist fair.

Im ersten Fall sagt man, *die Nullhypothese wird* (zugunsten der Alternativhypothese) *verworfen*, im zweiten Fall sagt man, *die Nullhypothese wird beibehalten*.

Wann ist ein p-Wert groß bzw. klein?

Dies ist keine statistische Frage, sondern ein Frage der subjektiven Sicherheit. Üblicherweise verwendet man einen Vergleichswert von 0.05 oder 0.01. Diese Festlegung wird SIGNIFIKANZNIVEAU (auch α) genannt. Ein Wert von 0.05 besagt, dass man sich in einem von 20 Fällen irrt und

dass man die Nullhypothese verwirft, obwohl sie eigentlich zutreffend ist. Das heißt, wenn man sehr viele Versuche mit einem fairen Würfel durchführen würde, dann käme man in 5 % zufällig zu dem Ergebnis, dass der Würfel nicht fair ist, obwohl das eigentlich nicht stimmt (Abschnitt 6.9).

Wir registrieren Differenzen zwischen den beobachteten Häufigkeiten und den Häufigkeiten, die wir erwarten würden, wenn die Nullhypothese gültig wäre. Aus diesen Differenzen können wir uns eine einzelne Maßzahl überlegen, die solche Abweichungen beschreibt. Sie soll groß sein, wenn die Abweichungen groß sind, und umgekehrt. Es gäbe natürlich verschiedene Möglichkeiten, solch eine Maßzahl zu konstruieren, die wichtigste stammt aber von Karl Pearson. Wir wollen sie einem allgemeinen Sprachgebrauch nach Pearson X^2 nennen.

Berechnung von X^2 bei eindimensionalen Häufigkeitsverteilungen

$$X^2 = \sum_{j=1}^{J} \frac{(o_j - e_j)^2}{e_j} \tag{6.1}$$

- o_j ... beobachtete Häufigkeit für die Kategorie j (o steht für *observed*)
- e_j ... erwartete Häufigkeit für die Kategorie j (e steht für *expected*)
- J ... Gesamtanzahl der Kategorien

Der X^2-Wert gibt uns Auskunft über die Größe der Abweichungen zwischen erwarteten und beobachteten Häufigkeiten. Anhand der Formel 6.1, in die ja die Differenz $o_j - e_j$ eingeht, sieht man, dass das X^2 mit der Größe der Abweichungen steigt. Für unser Beispiel (wie wir gleich sehen werden) ist dieser Wert $X^2 = 131.76$. Nun sollte uns dieser Wert einen Anhaltspunkt darüber geben, ob die Abweichungen zwischen den erwarteten und den beobachteten Häufigkeiten auffällig sind (man sagt auch *statistisch signifikant* bzw. *bedeutsam* in dem Sinn, dass man eher nicht daran glaubt, dass sich der Krankenstandsbeginn gleichmäßig über die Wochentage verteilt).

In R können wir die Berechnung folgendermaßen durchführen:

R

```
> chisq.test(table(Wochentag))
```

```
        Chi-squared test for given probabilities

data:  table(Wochentag)
X-squared = 132, df = 6, p-value < 2.2e-16
```

Im Output finden wir drei Zahlen. Die erste, `X-squared`, gibt uns den Wert 131.76. Das ist das vorher besprochene Pearson X^2 aus Formel 6.1. Der zweite Wert, `df`, bedeutet FREIHEITSGRADE (aus dem Englischen *degrees of freedom*) und ist die Anzahl der Kategorien minus 1 ($df = J - 1$). Die dritte Zahl, `p-value`, ist hier die wichtigste. Sie gibt eine Wahrscheinlichkeit an und liegt daher zwischen 0 und 1. Sie gibt uns (sehr vereinfacht gesagt) an, wie plausibel unsere Annahme ist, dass die Beginntage von

Krankenständen gleichmäßig auf alle Wochentage verteilt sind. Diese Zahl ist hier extrem klein[1], d. h., die Nullhypothese ist sehr unplausibel. Eine Interpretation des Ergebnisses könnte so aussehen.

Fallbeispiel 1: Der „blaue Montag“: Interpretation

Die Nullhypothese, dass der Beginn von Krankenständen an allen Wochentagen gleich häufig vorkommt, musste auf Grund eines Chi-Quadrat-Tests verworfen werden ($X^2 = 131.76$, $df = 6$, $p < 0.001$). Die Daten weisen darauf hin, dass der Beginn eines Krankenstands an verschiedenen Wochentagen unterschiedlich häufig auftritt. Die meisten Krankenstände beginnen an Montagen, gefolgt von Dienstag und Mittwoch. Am Wochenende sind die Häufigkeiten sehr gering.

Die letzte Ausgabe und die Schlussfolgerungen, die wir daraus gezogen haben, bedürfen einiger zusätzlicher Erläuterungen. (In Abschnitt 6.9 am Ende dieses Kapitels werden die theoretischen Grundlagen ausführlicher beschrieben.)

Zunächst ist Ihnen vielleicht der Unterschied aufgefallen, dass wir die beiden Begriffe X^2 (X-Quadrat) und χ^2 (Chi-Quadrat) verwendet haben. Der Grund ist, dass man zwischen X^2, einer Zahl, die aus beobachteten Daten errechnet wurde, und Chi-Quadrat, einer theoretischen Größe, unterscheiden sollte (eine ausführlichere Erklärung gibt der erwähnte Abschnitt 6.9).

Der Begriff Freiheitsgrade wird uns noch öfter begegnen. Vereinfacht gesagt dienen Freiheitsgrade dazu, den p-Wert zu bestimmen. Freiheitsgrade werden auch im Abschnitt 6.9 näher erklärt.

Der dritte Begriff, `p-value` oder p-Wert, ist wie erwähnt der wichtigste Wert, wenn es darum geht, eine (statistische) Fragestellung zu beantworten bzw. die Gültigkeit einer Annahme zu evaluieren. Wir haben hier nicht den Wert `2.2e-16`, der in der Ausgabe stand, verwendet, sondern $p < 0.001$. Das ist eine Geschmacksfrage, aber der leichteren Lesbarkeit halber haben wir uns für die obige Variante entschieden. Der angegebene Wert heißt ja, dass in weniger als 1 von 1000 Fällen ein solches oder noch extremeres Ergebnis zufällig zu erwarten ist, und das deutet ja schon an, dass die Gültigkeit der Nullhypothese sehr unplausibel ist.

Exkurs 6.5 Wie interpretiert man ein Ergebnis?

Zweck einer Interpretation ist das Zusammenführen von technischen Ergebnissen eines statistischen Verfahrens mit den inhaltlichen Aspekten der Fragestellung, für die man die statistische Methode verwendet hat. Es gibt keine exakten, allgemeingültigen Regeln, wie man ein Ergebnis

1 R verwendet bei sehr kleinen oder sehr großen Zahlen (was sehr groß ist, hängt von den Programmvoreinstellungen ab) die Exponentialdarstellung. Dabei bedeutet z. B. `1.0e+3` $= 1 \cdot 10^3 = 1000$ oder wie in unserem Beispiel `2.2e-16` $= 2.2 \cdot 10^{-16} = 0.0000000000000000221$. Es „rutscht“ also das Dezimalzeichen im ersten Beispiel um 3 Stellen nach rechts und im zweiten um 16 Stellen nach links.

interpretiert. Die genaue Länge und Form hängt davon ab, für welchen Zweck man eine Interpretation erstellt. Manche wissenschaftliche Zeitschriften oder Institutionen geben gewisse Formvorlagen oder Richtlinien. Im Allgemeinen sollte eine Interpretation aus zwei Teilen bestehen, aus einem technischen und einem inhaltlichen.

- Technische Interpretation: Hier wird oft die Nullhypothese (eventuell auch Alternativhypothese) formuliert, welches statistische Verfahren (eventuell auch warum) angewendet wurde. Ebenso gibt man Kennzahlen des speziellen statistischen Verfahrens und den p-Wert sowie das Signifikanzniveau, das man verwendet hat, an. Es sollte auch erwähnt werden, ob die Nullhypothese verworfen oder beibehalten wird.
- Inhaltliche Interpretation: Was bedeutet das technische Ergebnis. Die Fragestellung wird beantwortet und es werden wichtige beschreibende Fakten (wie z. B. Prozentsätze) angegeben.

Je nach Fragstellung und Methode wird eine Interpretation anders aussehen. Beispiele können Sie den Kästen Interpretation entnehmen, die immer am Ende der Analyseabschnitte in diesem Buch angeführt werden.

Das Ergebnis des Fallbeispiels 1, wie wir es bisher analysiert haben, deutet darauf hin, dass zu Wochenbeginn sich mehr Leute krank schreiben lassen als während des Rests der Woche. Ist das also ein Beleg für den „blauen Montag“? Ganz so einfach wird es wohl nicht sein. Es wurde bisher nicht in Betracht gezogen, dass man auch am Wochenende erkranken kann, die Krankenstandsmeldung aber erst zu Wochenbeginn erfolgt. Dies soll im Folgenden berücksichtigt werden.

Fallbeispiel 1: Der „blaue Montag“ (Teil 2)

Wenn man am Wochende erkrankt, wird die Krankschreibung möglicherweise erst am Montag oder wegen eines Arztbesuchs erst am Dienstag erfolgen. Teilt man die Woche in zwei Hälften, von Samstag bis Dienstag und von Mittwoch bis Freitag, könnte man die These des „blauen Montags“ besser überprüfen.

Gibt es Unterschiede in der Anzahl der Krankenstandsmeldungen zwischen der ersten und der zweiten Wochenhälfte?

Diese Frage lässt sich auf Grund der bisherigen Überlegungen leicht beantworten, wir müssen nur die Daten (die erhobenen Wochentage) in zwei Kategorien einteilen und dann den Test anstatt für die einzelnen Wochentage nun für die Wochenabschnitte durchführen. In R würden wir folgendermaßen vorgehen:

Zunächst erzeugen wir eine neue Variable `Wochenhaelfte` mittels der Funktion `ifelse()`. Diese Funktion erlaubt die Erstellung einer neuen Variable mit zwei Werten, je nachdem, ob eine Bedingung erfüllt ist oder nicht. In unserem Beispiel wird für jeden Wert von `Wochentag` geprüft, ob er in der Menge der Werte SA bis DI enthalten ist (Operator `%in%`). Wenn ja, dann hat die neue Variable `Wochenhaelfte` den Wert `"SA-DI"`, wenn nein, dann `"MI-FR"`.

R

```
> Wochenhaelfte <- ifelse(Wochentag %in% c("SA",
+     "SO", "MO", "DI"), "SA-DI", "MI-FR")
```

Eine Darstellung der Häufigkeiten sowie den Chi-Quadrat-Test erhalten wir mittels

R

```
> table(Wochenhaelfte)
> ct <- chisq.test(table(Wochenhaelfte))
```

```
Wochenhaelfte
MI-FR SA-DI
  126   174

        Chi-squared test for given probabilities

data:  table(Wochenhaelfte)
X-squared = 7.68, df = 1, p-value = 0.005584
```

Fallbeispiel 1: Der „blaue Montag“: Interpretation (Teil 2)

Der Chi-Quadrat-Test ergab, dass die Nullhypothese, nach der Krankenstände in den beiden Wochenhälften gleich häufig beginnen, verworfen werden musste ($X^2 = 7.68$, $df = 1$, $p = 0.006$). Die Daten weisen darauf hin, dass der Beginn eines Krankenstands in den beiden Wochenhälften unterschiedlich häufig auftritt. In der Wochenhälfte von Samstag bis Dienstag gibt es mehr Krankschreibungen als von Mittwoch bis Freitag.

Nicht berücksichtigt wurde, dass die beiden Wochenhälften unterschiedlich viele Tage umfassen. Wie man so etwas in eine Analyse miteinbezieht, wird im nächsten Abschnitt behandelt.

Da wir die Variablen aus dem Data Frame `kstand`, auf die wir uns wir uns vorher mit `attach(kstand)` direkten Zugriff verschafft haben, nicht mehr benötigen, lösen wir diese Zugriffsmöglichkeit mit dem Befehl `detach()` wieder auf.

R

```
> detach(kstand)
```

Wenn man mittels `attach()` einen Data Frame einer R Session zuordnet, also dann direkt die Variablennamen verwenden kann, ohne den Data Frame angeben zu müssen, dann sollte man nicht vergessen, diese Zuordnung mittels `detach()` wieder aufzuheben. Der Grund dafür ist, dass eventuell zwei gleiche Variablennamen in zwei verschiedenen Data Frames verwendet werden und nur jener aus dem zuletzt zugeordneten Data Frame zur Verfügung steht. R gibt dann zwar eine Meldung aus, trotzdem kann das leicht zu Fehlern führen, die man besser vermeidet.

6.3 Entsprechen Häufigkeiten bestimmten Vorgaben?

Im Abschnitt 6.2 haben wir uns mit der Frage beschäftigt, ob beobachtete Häufigkeiten für Kategorien einer Variable mit der Hypothese in Einklang stehen, dass in Wirklichkeit alle Kategorien gleich wahrscheinlich sind. Wir wollen diese Problemstellung nun insofern erweitern, als die Anteile für Kategorien unterschiedlich spezifiziert sein können. Nehmen wir an, eine Variable hätte drei Kategorien. Wir könnten prüfen, ob alle Kategorien zu je 1/3 vorkommen. Nun wollen wir den Fall untersuchen, ob zum Beispiel in der ersten Kategorie 50 %, in der zweiten 35 % und in der dritten 15 % vorkommen. Die dabei gestellte statistische Frage ist, ob die Anteile von einzelnen Kategorien in einer Stichprobe den tatsächlichen Anteilen in der Population entsprechen. Wir wir sehen werden, ist die statistische Methode zur Beantwortung solcher Fragen sehr ähnlich zum ersten Abschnitt. Eine typische Anwendung, wie sie in der Praxis häufig vorkommt, wird in Fallbeispiel 2 dargestellt.

Fallbeispiel 2: Repräsentativität einer Stichprobe

Eine Meinungsforscherin hat eine Telefonumfrage an 200 zufällig ausgewählten Telefonteilnehmern in Österreich zum Thema einer erwünschten Gesetzesmaßnahme zur speziellen Förderung von Familien mit Kindern durchgeführt. Aus Angaben der Statistik Austria für 2007 weiß sie, dass die insgesamt etwa 2.31 Millionen Familien sich folgendermaßen aufteilten: 31.2 % Ehepaare ohne Kinder, 42.4 % Ehepaare mit Kindern, 7.3 % Lebensgemeinschaften ohne Kinder, 6.1 % Lebensgemeinschaften mit Kindern und 13 % alleinerziehende Elternteile. Die Meinungsforscherin interessierte, ob ihre Stichprobe repräsentativ bezüglich der Familienstruktur war, d. h., ob die Anteile der verschiedenen Arten von Familien in ihrer Telefonstichprobe mit den Anteilen aller österreichischen Familien übereinstimmen. In ihrer Stichprobe konnte sie folgende Häufigkeiten feststellen:

Familientyp	Häufigkeit	Prozent
Ehepaare ohne Kinder	42	21
Ehepaare mit Kindern	98	49
Lebensgemeinschaft ohne Kinder	6	3
Lebensgemeinschaft mit Kindern	20	10
Alleinerziehende Elternteile	34	17
Gesamt	200	

Ist die Stichprobe repräsentativ für die Population bezüglich der Familienstruktur?

Wie schon zuvor (Abschnitt 6.2) beginnen wir die Analyse mit einer Darstellung der Daten.

6.3.1 Numerische und grafische Beschreibung

Wir werden, wie im vorigen Abschnitt, wieder eine Häufigkeitstabelle erstellen. Da hierzu aber kein Datenfile zur Verfügung steht, müssen wir die Tabelle selbst erzeugen. Zunächst wollen wir die Variable für die beobachteten Häufigkeiten, `FbeobH`, generieren, wobei wir die Bezeichnungen für die einzelnen Kategorien gleich mitdefinieren können.

R

```
> FbeobH <- c("Ehepaare ohne Kinder" = 42,
+   "Ehepaare mit Kindern" = 98,
+   "Lebensgemeinschaft ohne Kinder" = 6,
+   "Lebensgemeinschaft mit Kindern" = 20,
+   "Alleinerziehende Elternteile" = 34)
```

Die Funktion `c()` (combine) haben wir schon kennengelernt. Sie dient dazu, einen Vektor zu erzeugen. Wir können aber auch gleichzeitig den einzelnen Elementen dieses Vektors Namen zuordnen. In unserem Beispiel sind die Namen unter Anführungszeichen gesetzt. Das ist im Allgemeinen nicht notwendig, allerdings enthalten unsere Namen Leerzeichen, die R in diesem Fall falsch interpretieren würde. Daher müssen wir diese Namen mit Anführungszeichen versehen.

Als Nächstes erzeugen wir noch einen Vektor mit den dazugehörigen Prozentwerten

R

```
> FbeobP <- FbeobH/sum(FbeobH) * 100
```

Wenn wir eine Tabelle aus diesen beiden Variablen erzeugen wollen, bietet es sich an, eine Matrix zu erstellen (wir wollen sie `Familientyp` nennen) und den beiden Spalten noch einen Namen zu geben.

R

```
> Familientyp <- cbind(FbeobH, FbeobP)
> colnames(Familientyp) <- c("Häufigkeit", "Sample %")
> Familientyp
```

```
                              Häufigkeit Sample %
Ehepaare ohne Kinder                  42       21
Ehepaare mit Kindern                  98       49
Lebensgemeinschaft ohne Kinder         6        3
Lebensgemeinschaft mit Kindern        20       10
Alleinerziehende Elternteile          34       17
```

Diese Tabelle entspricht jener aus Fallbeispiel 2. Sie enthält zwar die gesamte Information zur Stichprobe, aber hinsichtlich der Fragestellung wäre es günstiger, auch noch die Zahlen aus der gesamten österreichischen Bevölkerung hinzuzufügen. Man bekäme dann gleich einen Eindruck, wie ähnlich oder unähnlich die Prozente aus der Population und der Stichprobe sind.

Wir fügen also die Prozentwerte der Population zu der Tabelle hinzu. Dazu definieren wir eine Variable `FpopP` und tragen die Werte ein, wie sie im Fallbeispiel 2 angegeben sind, und fügen sie zu der Matrix `Familientyp` mittels `cbind()` hinzu.

R

```
> FpopP <- c(31.2, 42.4, 7.3, 6.1, 13.0)
> Familientyp <- cbind(Familientyp, "Pop. %" = FpopP)
```

```
                              Häufigkeit Sample % Pop. %
Ehepaare ohne Kinder                  42       21   31.2
Ehepaare mit Kindern                  98       49   42.4
Lebensgemeinschaft ohne Kinder         6        3    7.3
Lebensgemeinschaft mit Kindern        20       10    6.1
Alleinerziehende Elternteile          34       17   13.0
```

Man erkennt, dass die Prozentsätze differieren. Es fällt auf, dass generell Familien mit Kindern stärker in der Stichprobe vertreten sind als jene ohne Kinder. Da die Umfrage zum Thema einer Gesetzesmaßnahme zur speziellen Förderung von Familien mit Kindern stattfand, könnte eventuell die Bereitschaft zur Beantwortung bei solchen Personen größer gewesen sein, die in einer Familie mit Kindern leben. Auch die Wahrscheinlichkeit, solche Personen eher zu Hause am Festnetz zu erreichen, könnte eine Rolle gespielt haben.

Darstellung verschiedener Variablen in einer Grafik

Zur grafischen Beschreibung der Daten aus Fallbeispiel 2 bieten sich GRUPPIERTE BALKENDIAGRAMME an. Sie sind eine Erweiterung der Balkendiagramme aus Abschnitt 6.2.2. Bei gruppierten Balkendiagrammen werden mehrere kategoriale Variablen gleichzeitig oder eine kategoriale Variable aufgeschlüsselt nach verschiedenen Gruppen dargestellt (im Detail gehen wir darauf in Kapitel 9 ein).

In R erhalten wir ein gruppiertes Balkendiagramm ganz einfach, indem wir zu dem vorher schon in Abschnitt 6.2.2 beschriebenen Befehl `barplot()` die Option `beside = TRUE` hinzufügen. Natürlich müssen wir auch die beiden Vektoren für Stichproben-Prozent und Populations-Prozent, die in der Matrix `Familientyp` in den Spalten 2 und 3 stehen, angeben. Schließlich erzeugen wir noch eine Legende mit der Option (legend) und skalieren die y-Achse mittels `ylim`, um Platz für die Legende zu schaffen.

R

```
> barplot(Familientyp[,2:3], beside = TRUE,
+     legend = rownames(Familientyp), ylim = c(0,70))
```

Das so hergestellte gruppierte Balkendiagramm findet sich in ▶ Abbildung 6.4. Für Publikationen, Berichte oder Präsentationen könnte man diese Grafik noch schöner machen, z. B. die Schriftgrößen ändern. Gerade was die Erzeugung hochwertiger Grafiken betrifft, ist R sehr mächtig. Eine detaillierte Darstellung aller Möglichkeiten würde den Rahmen sprengen (einige finden sich in Abschnitt 5.2).

▶ Abbildung 6.4 ist zum Vergleich der Prozentsätze aus der Stichprobe und der Population dennoch ganz gut geeignet. Man kann sehr schön die Unterschiede und

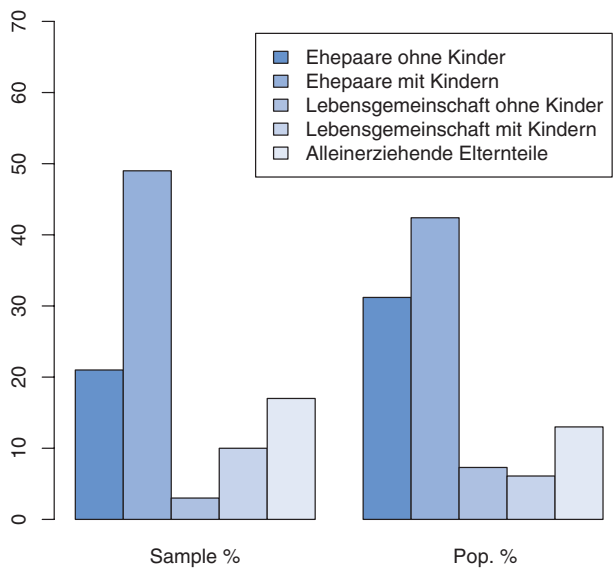


Abbildung 6.4: Gruppiertes Balkendiagramm für die Prozentwerte aus einer Stichprobe und der Population

deren Größenordnung erkennen. Es verstärkt sich der Eindruck, den wir schon aus der Tabelle auf Seite 168 gewonnen haben. In der Stichprobe sind jene Befragten überproportional vertreten, die in Familien mit Kindern leben. Gegenteiliges gilt für Ehepaare ohne Kinder, die in der Stichprobe unterrepräsentiert sind. Die Frage ist jetzt, sind diese Unterschiede bedeutsam oder nur auf Zufall zurückzuführen.

6.3.2 Statistische Analyse der Problemstellung

Ganz ähnliche Überlegungen, wie wir sie schon im Abschnitt 6.2.3 angestellt haben, treffen auch hier zu. Allerdings ist die Annahme jetzt nicht, dass alle Kategorien gleich häufig vorkommen (Nullhypothese), sondern dass sie in bestimmter Weise festgelegt sind. In unserem Beispiel, der Frage nach Repräsentativität der Stichprobe, sind die festgelegten Anteile jene der Familientypen in der Population. Wir können zwar die Formel (6.1) verwenden, aber wir müssen die erwarteten Häufigkeiten anders bestimmen.

Wir kennen die Anteile in der Population, nämlich 0.312 für Ehepaare ohne Kinder, 0.424 für Ehepaare mit Kindern etc. Wir können aus diesen jene Häufigkeiten bestimmen, die wir erwarten würden, wenn in der Stichprobe der 200 Personen die Stichprobenhäufigkeiten mit den Populationswerten übereinstimmen würden. Wir müssen dazu nur die Populationsanteile mit der Stichprobengröße 200 multiplizieren. Wenn wir eine bezüglich der Familienstruktur repräsentative Stichprobe hätten, dann müssten $200 \times 0.312 = 62.4$, also 62 Befragte, der Kategorie Ehepaar kinderlos angehören. Allgemein gilt

Berechnung von erwarteten Häufigkeiten

$$e_j = n\pi_j \tag{6.2}$$

- e_j ... erwartete Häufigkeit für die Kategorie j
- n ... Stichprobengröße
- π_j ... relative Häufigkeit in der Population (Wahrscheinlichkeit für Kategorie j)

Die weiteren Schritte sind analog zu Abschnitt 6.2.3. Wenn die Abweichungen zwischen beobachteten und erwarteten Häufigkeiten zu groß werden, verwerfen wir die Nullhypothese, dass die beobachteten Werte mit den erwarteten übereinstimmen. Die Abweichungen werden dann nicht mehr dem Zufall zugeschrieben, sondern man glaubt an einen systematischen Unterschied.

Die Berechnung des Tests in R erfolgt über

R

```
> chisq.test(Familientyp[,1],
+     p = Familientyp[,3]/sum(Familientyp[,3]))
```

Der Unterschied zur Spezifikation auf Seite 162 besteht darin, dass i) die Variable `Familientyp[,1]` schon tabellierte Werte, also Häufigkeiten, enthält und daher der `table()`-Befehl nicht mehr benötigt wird, und ii) dass die erwarteten relativen Häufigkeiten über die Option `p` angegeben werden.

```
        Chi-squared test for given probabilities

data:  Familientyp[, 1]
X-squared = 21.2, df = 4, p-value = 0.0002840
```

Wie wir am p-Wert sehen, ist die Plausibilität der Nullhypothese sehr klein. Daher verwerfen wir die Annahme, dass die Stichprobenanteile der einzelnen Familientyp-Kategorien jenen in der Population entsprechen.

Fallbeispiel 2: Repräsentativität: Interpretation

Der Chi-Quadrat-Test zur Überprüfung der Nullhypothese, dass die Häufigkeiten für die Familienarten in der Stichprobe jenen in der Population entsprechen, zeigt ein signifikantes Ergebnis ($X^2 = 21.2$, $df = 4$, $p < 0.001$). Demnach ist nicht davon auszugehen, dass die Stichprobe bezüglich der Familienarten repräsentativ ist. Familien mit Kindern sind stärker in der Stichprobe vertreten als in der Population.

Am Ende des vorigen Abschnitts (Abschnitt 6.2), als wir die Beginntage von Krankenständen untersuchten, hatten wir gemutmaßt, dass das signifikante Ergebnis beim Vergleich der beiden Wochenhälften möglicherweise darauf zurückzuführen ist, das die beiden Wochenhälften einmal 3 und einmal 4 Tage umfassten. Basierend auf den Überlegungen dieses Kapitels, können wir das berücksichtigen. Anstatt anzunehmen, dass die erwarteten Häufigkeiten der Krankenstandsmeldungen in den beiden Wochenhälften gleich sind, spezifizieren wir die erwarteten Anteile proportional zur Anzahl der Tage in den beiden Wochenhälften, also 3/7 für `"MI-FR"` und 4/7 für `"SA-DI"`. Wir erhalten den adaptierten Test mit

R

```
> chisq.test(table(Wochenhaelfte), p = c(3/7, 4/7))
```

```
        Chi-squared test for given probabilities

data:  table(Wochenhaelfte)
X-squared = 0.09, df = 1, p-value = 0.7642
```

Das Ergebnis hat sich dramatisch verändert. Der X^2-Wert ist nun sehr klein, was auf eine große Übereinstimmung von beobachteten und erwarteten Häufigkeiten hindeutet, und der p-Wert ist groß geworden. Insgesamt gibt es also wenig Evidenz dafür, dass Erkrankungen sich zu Wochenbeginn häufen (es spricht also nicht sehr viel für den berühmten „blauen Montag“).

Die Interpretation dieses Ergebnisses könnte folgendermaßen lauten: Der Chi-Quadrat-Test ergab, dass die Nullhypothese, nach der die Häufigkeiten des Beginns von Krankenständen proportional zur Anzahl der Tage in den beiden Wochenhälften (`"MI-FR"` und `"SA-DI"`) verteilt sind, beizubehalten ist ($X^2 = 0.09$, $df = 1$,

$p = 0.764$). Demnach gibt es in keiner der beiden Wochenhälften überproportional viele Krankschreibungen.

6.4 Hat ein Prozentsatz (Anteil) einen bestimmten Wert?

Bisher haben wir untersucht, wie man analysieren kann, ob Häufigkeiten bzw. Anteile mehrerer Kategorien bestimmten Vorgaben entsprechen. Jetzt wollen wir uns auf einzelne Kategorien konzentrieren.

Fallbeispiel 3: Glaube an paranormale Phänomene

Im Jahr 2005 führte das Gallup Institut eine telefonische Umfrage an 1008 repräsentativ ausgewählten Amerikanern zum Thema paranormale Phänomene durch. Zu insgesamt zehn solcher Phänomene sollten die Befragten angeben, ob sie an deren Existenz glaubten. Die folgende Tabelle gibt eine Übersicht über die Phänomene und die Prozentsätze derer, die an ihre Existenz glaubten.

Glaube an	Prozent
Außersinnliche Wahrnehmung	41
Häuser, in denen es spukt	37
Geister	32
Telepathie	31
Hellsehen	26
Astrologie	25
Totenbeschwörung	21
Hexen	21
Wiedergeburt	20
Spiritismus	9

Eine Analyse der Einzelergebnisse ergab, dass 73 % an zumindest eines dieser Phänomene glaubten. Aus einer ähnlichen Untersuchung aus dem Jahr 2001 war bekannt, dass damals 76 % an zumindest eines dieser Phänomene glaubten. Das Gallup Institut interpretiert dies als leichten Rückgang. (Quelle: http://www.gallup.com/poll/16915/three-four-americans-believe-paranormal.aspx)

War der Glaube an paranormale Phänomene bei Amerikanern zwischen 2001 und 2005 rückläufig?

Wenn man die Daten einer einzelnen Kategorie analysieren möchte, dann kann man im Wesentlichen die gleichen Methoden verwenden wie in den vorhergehenden Abschnitten 6.2.3 und 6.3.2. Man muss sich nur vor Augen halten, dass es für diese

einzelne Kategorie eigentlich zwei Möglichkeiten gibt, nämlich *trifft zu* bzw. *trifft nicht zu*. Wenn man also eine einzelne Kategorie untersucht, dann verhält sich diese wie eine „neue" Variable mit zwei Kategorien. Auf das Fallbeispiel 3 übertragen bedeutet dies *„Glaube an zumindest ein paranormales Phänomen"* und *„Glaube an keines"*.

Numerische und grafische Beschreibung

Zur numerischen Beschreibung wird es wohl genügen, eine Häufigkeitstabelle wie in Abschnitt 6.2.1 zu erstellen oder die Zahlen einfach anzugeben.

Als grafische Darstellung bietet sich wieder ein Balkendiagramm an. In Analogie zu Abschnitt 6.2.2 kann es folgendermaßen erstellt werden. Das Resultat findet sich in ▶ Abbildung 6.5.

R

```
> para <- c(ja = 0.73, nein = 1 - 0.73)
> barplot(para)
```

Etwas schöner wird die Grafik, wenn man einige Plotparameter ändert: Mittels `ylim = c(0, 1)` skaliert man die y-Achse auf das Intervall 0 bis 1, durch `width = c(0.5, 0.5)` kann man die Balken etwas schmäler machen. Das geht aber nur in Kombination mit einer Änderung der Achsenskalierung der horizontalen Achse, die wir mit `xlim = c(0, 1.5)` spezifizieren.

Mit dem modifizierten R-Befehl

R

```
> barplot(para, ylim = c(0, 1), width = c(0.5, 0.5),
+     xlim = c(0, 1.5))
```

erhalten wir die modifizierte Grafik (▶ Abbildung 6.6).

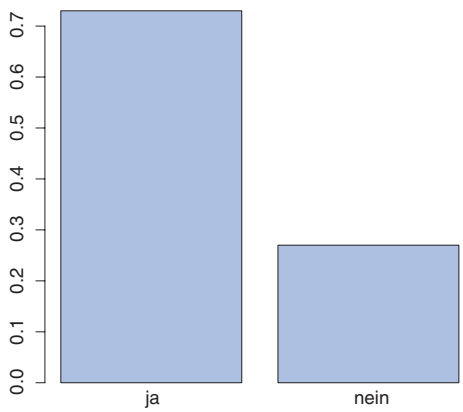


Abbildung 6.5: Balkendiagramm für Glaube an paranormale Phänomene

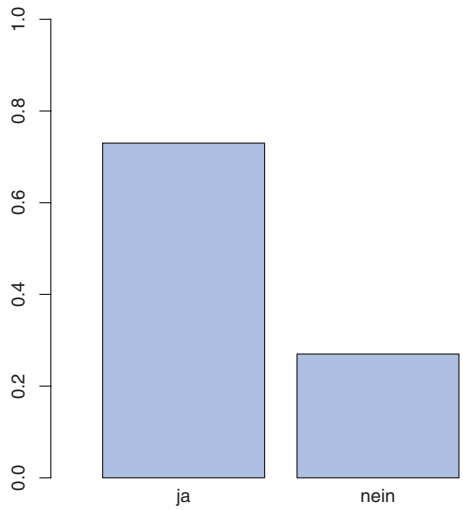


Abbildung 6.6: Modifiziertes Balkendiagramm für Glaube an paranormale Phänomene

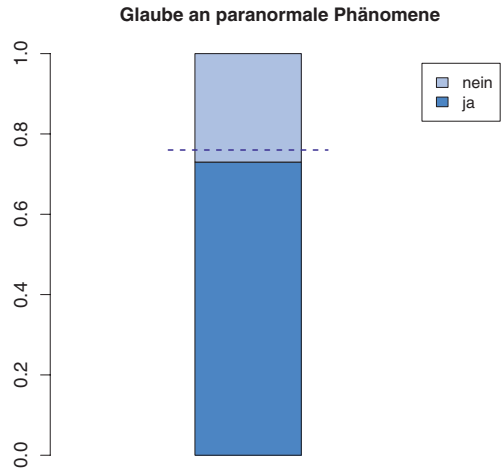


Abbildung 6.7: Gestapeltes Balkendiagramm mit Referenzlinie

Allerdings interessiert für die Fragestellung eigentlich nur die Kategorie `ja`. Dazu könnte man ein gestapeltes Balkendiagramm verwenden (diese werden in Abschnitt 7.1.2 genauer erklärt). Und man möchte vielleicht auch noch die 76 %-Rate aus dem Jahr 2001 miteinbeziehen. Das Resultat könnte so aussehen (▶ Abbildung 6.7): Die dazu notwendigen R-Befehle sind

R

```
> bb <- barplot(as.matrix(para), ylim = c(0, 1),
+     width = 0.2, xlim = c(-0.2, 0.6), legend = names(para),
+     main = "Glaube an paranormale Phänomene")
> lines(c(bb - 0.15, bb + 0.15), c(0.76, 0.76),
+     lty = "dashed", col = "blue", lwd = 1.5)
```

Neben den vorher schon beschriebenen Optionen `ylim`, `width` und `xlim` werden noch `legend` und `main` verwendet. Mit `legend = names(para)` kann man eine Legende für die Kategorien der gezeichneten Variable anfordern und mit `main` wird ein Titel ausgegeben. Die gestrichelte Linie zeichnen wir mit dem `lines()`-Befehl.

Die Grafik in ▶ Abbildung 6.7 ist keine Standard-R-Grafik und ihre Erstellung ist ein wenig komplizierter als das bisher Beschriebene. Wie wollen die Vorgehensweise dennoch beschreiben, um zu zeigen, wie man auch speziellere Grafiken herstellen kann. Das Folgende kann aber beim ersten Lesen übersprungen werden. Wenn man die Funktion `barplot()` und gleichzeitig eine Ausgabe anfordert (wie hier in `bb`), dann wird die Grafik (trotzdem) gezeichnet, aber in das Ausgabeobjekt (also hier `bb`) werden die x-Werte für die Mitte der Balken gespeichert. Das kann man sich zunutze machen, wenn man weitere Elemente zur Grafik hinzufügen möchte, aber die Koordinaten nicht weiß.

In unserem Beispiel benötigen wir das für die gestrichelte Linie, die den Wert der Nullhypothese beschreibt. Mit `lines()` können wir eine Linie zeichnen. Die ersten Werte Spezifikationen in `lines()` geben die x-Werte und die y-Werte jener Punkte an, die mit einer Linie verbunden werden sollen, also x_1, x_2 und y_1, y_2. Die y-Werte sind klar, sie beschreiben den Wert der Nullhypothese, nämlich 0.76. Sie werden mit dem Vektor `c(0.76, 0.76)` definiert. Die x-Werte erhalten wir so: `bb` ist der Mittelpunkt des Balkens, die Breite haben wir mit `width = 0.2` festgelegt. Wenn wir von `bb` ausgehen und die halbe Balkenbreite (für einmal nach links und einmal nach rechts) dazu zählen, dann erhalten wir die x-Werte für eine Linie, die genau innerhalb des Balkens liegt. Wir geben noch einen kleinen Wert (0.05) dazu, damit die Linie über den Balken hinausragt. Der Vektor für die x-Werte wird also mit `c(bb - 0.15, bb + 0.15)` spezifiziert. Schließlich soll die Linie gestrichelt (`lty = 'dashed'`, `lty` steht für *line type*) und blau (`col = 'blue'`, `col` steht für *colour*) sein und außerdem möchten wir sie ein wenig dicker (`lwd = 1.5`, `lwd` steht für *line width*, 1.5 heißt 1.5 Mal so dick wie die Standarddicke).

Man sieht, dass die Zustimmungsrate zur Existenz paranormaler Phänomene sich nicht sehr verändert hat. Die entsprechende statistische Methode im nächsten Abschnitt wird uns darüber Aufschluss geben, ob diese Rate tatsächlich niedriger ist oder nicht.

6.4.1 Statistische Analyse der Problemstellung

Die Frage, die wir untersuchen wollen, lautet: War der Glaube an paranormale Phänomene bei Amerikanern zwischen 2001 und 2005 rückläufig? Übersetzt in eine statistische Fragestellung, könnte man das auch so formulieren: Ist die relative Häu-

figkeit des Glaubens an die Existenz mindestens eines Phänomens im Jahr 2005 (in Wirklichkeit) niedriger als die entsprechende relative Häufigkeit aus dem Jahr 2001.

Als statistische Hypothesen formuliert:

- Nullhypothese H_0: $\pi = 0.76$
  Die relative Häufigkeit der Zustimmung 2005, π, entspricht der relativen Häufigkeit der Zustimmung 2001 ($\pi_0 = 0.76$).
- Alternativhypothese H_A: $\pi < 0.76$
  Die relative Häufigkeit der Zustimmung 2005, π, ist kleiner als die relative Häufigkeit der Zustimmung 2001 ($\pi_0 = 0.76$).

Allgemein würde man schreiben:

Hypothesen beim Test eines Anteils

H_0: $\pi = \pi_0$ (Nullhypothese)
H_A: $\pi < \pi_0$ (Alternativhypothese)

- π ... die unbekannte tatsächliche Häufigkeit der Zustimmung 2001. (Da wir sie nicht kennen, aber etwas über sie wissen wollen, verwenden wir für sie bei der Berechnung r, die beobachtete relative Häufigkeit aus der Stichprobe.)
- π_0 ... der Wert, den wir kennen (oder den wir festlegen) und gegen den wir prüfen wollen. In unserem Beispiel ist er 0.76.

Die statistische Problemstellung ist von der Idee her gleich, wie wir sie schon in Abschnitt 6.2.3 kennengelernt haben:

- Wir gehen davon aus, dass in Wirklichkeit die relative Häufigkeit der Zustimmung 2005 jener aus dem Jahr 2001 entspricht, dass also tatsächlich H_0 gilt.
- Wir sammeln Daten aus einer Stichprobe (für das Jahr 2005), berechnen die relative Häufigkeit r (sie ist die beste Information, die wir für das unbekannte π haben) und vergleichen sie mit jener Zahl, die wir kennen, nämlich 0.76. Allgemein können wir π_0 statt 0.76 einsetzen.
- Da wir ja eine Zufallsstichprobe verwenden, wird r nicht genauso groß wie π sein, sondern ein wenig abweichen. Wenn die Abweichung aber zu groß wird, dann werden wir nicht glauben, dass H_0 gilt, sondern eher H_A.
- Die Prüfmethode zur Entscheidung, welche der beiden Hypothesen zutrifft, nennen wir statistischen Test.

Im Unterschied zum Chi-Quadrat-Test aus Abschnitt 6.2.3 und 6.3.2, wo wir absolute Häufigkeiten (beobachtete und erwartete) verglichen haben, beschäftigen wir uns jetzt mit Anteilen oder relativen Häufigkeiten (auch beobachtete, nämlich r, und erwartete, nämlich π_0). Der Test, den wir jetzt verwenden werden, heißt Ein-Stichproben-Test für Anteile bzw. Binomial-Test. Beide Tests prüfen die gleiche Nullhypothese, unterscheiden sich aber in den Details ihrer Berechnung (und bezüglich ihrer mathematisch-statistischen Eigenschaften). Doch dazu noch später.

In R führen wir den Ein-Stichproben-Test für Anteile folgendermaßen durch: Die beobachtete Häufigkeit ist

R

```
> beobH <- round(0.73 * 1008)
> beobH
```

```
[1] 736
```

Aus den Angaben von Gallup (siehe Fallbeispiel 3) wissen wir nur, dass 73% von $n = 1008$ Befragten zumindest an ein paranormales Phänomen glauben. Wir multiplizieren also die relative Häufigkeit 0.73 mit 1 008, der Stichprobengröße, um die beobachtete Häufigkeit zu erhalten. Zusätzlich runden wir noch auf eine ganze Zahl, also `round(0.73*1008)`.

Für unsere Berechnung müssen wir in der Funktion `prop.test` folgende fünf Argumente spezifizieren:

R

```
> prop.test(beobH, 1008, p = 0.76, alternative = "less",
+     correct = FALSE)
```

- Das erste Argument muss die beobachtete Häufigkeit des Werts (der Kategorie) sein, die wir überprüfen wollen.
- Das zweite Argument ist die Stichprobengröße.
- Der dritte zu spezifizierende Wert ist der Wert, gegen den wir prüfen wollen, also jener der Nullhypothese bzw. π_0. Hier: `p=0.76`
- Dann geben wir an, ob es sich um einen zweiseitigen oder einseitigen Test handelt (die Begriffe *zweiseitig* bzw. *einseitig* werden in ▶ Exkurs 6.6 besprochen). Je nach Fragestellung spezifizieren wir für die Option `alternative` entweder `"less"` oder `"greater"` bzw. `"two.sided"` (die Voreinstellung, die verwendet wird, wenn man nichts angibt). Hier verwenden wir `"less"`, weil wir ja prüfen wollen, ob sich die Zustimmung verringert hat.
- Schließlich setzen wir noch `correct = FALSE`. Diese Option ist standardmäßig auf `TRUE` gesetzt und würde bei der Berechnung eine Kontinuitätskorrektur verwenden. Eine solche ist dann sinnvoll, wenn die zu prüfende Variable (hier Glaube an mindestens ein paranormales Phänomen) eigentlich metrisch ist und nur durch Gruppierung der Werte kategorial gemacht wurde. Das ist hier aber nicht der Fall.

Wir erhalten

```
        1-sample proportions test without continuity correction

data:  beobH out of 1008, null probability 0.76
X-squared = 4.92, df = 1, p-value = 0.01326
alternative hypothesis: true p is less than 0.76
95 percent confidence interval:
 0.000 0.753
sample estimates:
```

```
   p
0.73
```

Neben den Angaben, welche Variablen bzw. Werte wir in der Funktion angegeben haben, finden wir den Wert, den wir für die H_0 spezifiziert haben (`null probability 0.76`), sowie neben X^2 und df den p-Wert. Dieser ist 0.01457 und damit wesentlich kleiner als 0.05. Wir verwerfen daher die Nullhypothese. Zur weiteren Information gibt R auch noch die Spezifikation für die Alternativhypothese (`alternative hypothesis: true p is less than 0.76`) sowie die relative Häufigkeit aus der Stichprobe (`sample estimates`) aus. Die weitere Ausgabe (`95 percent confidence interval`) bezieht sich auf eine andere inferenzstatistische Methode, die wir im (nächsten) Abschnitt 6.5 behandeln werden.

Fallbeispiel 3: paranormale Phänomene: Interpretation

Der Ein-Stichproben-Test zur Überprüfung der Nullhypothese, dass der Anteil der Amerikaner, die an die Existenz von paranormalen Phänomenen glauben, kleiner als 76 % ist, erbrachte ein signifikantes Ergebnis ($p = 0.015$). Demnach besteht Evidenz dafür, dass sich der Anteil jener, die an übersinnliche Phänomene glauben, zwischen 2001 und 2005 (wenn auch nicht stark, aber dennoch nachweislich) verringert hat.

Es gibt, wie oben erwähnt, einen zweiten Test, den man zur Beantwortung der formulierten Nullhypothese verwenden kann, den Binomial-Test. Während die Funktion `prop.test()` die Berechnung über die χ^2-Verteilung durchführt, verwendet der Binomialtest die sogenannte Binomialverteilung zur Berechnung der Wahrscheinlichkeit für das Zutreffen von Null- bzw. Alternativhypothese. Dieser Test berechnet die „exakten" Wahrscheinlichkeiten und ist bei kleinen Stichproben zu bevorzugen. Allerdings kann bei großen Stichproben der Rechenaufwand erheblich werden und dann ist es einfacher, Methoden zu verwenden, die den p-Wert möglichst genau approximieren. Je größer die Stichprobe ist, umso besser ist die Approximation (eine Faustregel besagt, dass der Wert sowohl von $n \cdot \pi_0$ als auch $n \cdot (1-\pi_0)$ größer als 10 sein soll). Die Approximationen beruhen auf der Grundidee, den p-Wert so zu berechnen, als ob man unendlich viele Stichproben gezogen hätte. Man nennt den p-Wert dann auch asymptotischen p-Wert. Die Grundideen hierfür werden im Abschnitt 6.9 erläutert.

In unserem Beispiel ist die Stichprobe relativ groß und daher ist das Ergebnis des (asymptotischen) Ein-Stichproben-Tests recht genau. Zur Illustration, wollen wir dennoch auch den (exakten) Binomial-Test anwenden. Die Spezifikation und auch der Output unterscheiden sich nicht wesentlich. Statt `prop.test()` verwendet man `binom.test()`.

R

```
> binom.test(beobH, 1008, p = 0.76, alternative = "less")
```