7 Kategoriale Daten

Kategoriale Daten erhält man durch Klassifikation von auftretenden Beobachtungen in verschiedene Kategorien. Der Definition 6.1.3 folgend, sind dies also Daten, die nominalskaliert sind. Im Zusammenhang mit kategorialen Daten können mehrere Fragestellungen auftauchen, die man unter statistischen Gesichtspunkten untersuchen kann. Dabei unterscheidet man die beiden Szenarien, dass nur eine Variable untersucht wird (Abschnitt 7.1) oder dass zwei kategoriale Variablen von Interesse sind (Abschnitt 7.2).

7.1 Eine kategoriale Variable

Besitzt die untersuchte Variable mehr als zwei Kategorien als mögliche Merkmalsausprägungen, kann der χ^2-Anpassungstest überprüfen, ob die einzelnen Kategorien alle gleich häufig vorkommen bzw. ob die einzelnen Kategorien in einer vorgegebenen Häufigkeit auftreten (Abschnitt 7.1.1). Für den Sonderfall, dass die vorliegende Variable *dichotom* ist, d.h. nur zwei mögliche Ausprägungen besitzt, kann man die gleiche Fragestellung mit dem Binomialtest untersuchen (siehe Abschnitt 7.1.2).

7.1.1 Der χ^2-Anpassungstest

Betrachten wir zur Motivation dieses Tests den Datensatz `würfel.csv`, bei dem ein Würfel 50 mal geworfen und bei jedem Wurf das Ergebnis in der Variable `augenzahl` aufgezeichnet wurde. Zuerst wollen wir den Datensatz mit `read.csv2()` (s. Abschnitt 3.2.2) in den Workspace laden und uns die ersten zehn Beobachtungen des Datensatzes anzeigen lassen. Wir gehen in diesem, wie in den folgenden Fällen immer davon aus, dass die Datensätze[1] im Ordner **C:\R\Rohdaten** gespeichert sind.

```
> würfel <- read.csv2("C:/R/Rohdaten/würfel.csv")
> würfel[1:10, ]
 [1] 2 3 6 1 6 1 2 3 1 1
```

Einen schnellen Überblick, wie oft welche Augenzahl geworfen wurde, bekommt man mit der Funktion `table()`:

```
> table(würfel$augenzahl)

 1  2  3  4  5  6
16  8 10  8  4  4
```

[1] Die Datensätze sind online zu finden unter
`http://www.rrzn.uni-hannover.de/buch.html?&no_cache=1&titel=statistik_r`

Von den 50 Würfen waren also sechzehn eine 1, acht Würfe eine 2 usw. Bei einem *fairen*[2] Würfel würde man erwarten, dass die absoluten Häufigkeiten der Augenzahlen in etwa gleich groß sind. Im vorliegenden Fall wurden die Fünf und die Sechs aber deutlich weniger oft geworfen als die anderen Zahlen. Der Verdacht liegt hier nahe, dass der Würfel „gezinkt“ ist, d.h. dass gewisse Zahlen überdurchschnittlich oft geworfen werden. Zu untersuchen ist, ob diese Abweichungen noch zufällig sind, oder ob hier schon eine Regelmäßigkeit zugrunde liegt und der Würfel nicht fair ist. Neben der deskriptiven Darstellungsmöglichkeit mit `table()` können wir zudem ein Balkdendiagramm erstellen. Hierzu ist uns die Funktion `barplot()` aus Abschnitt 5.1.2 behilflich:

```
> barplot(table(würfel$augenzahl), ylab = "Anzahl", xlab = "Augenzahl")
```

Das Balkendiagramm ist in Abbildung 7.1 zu sehen.

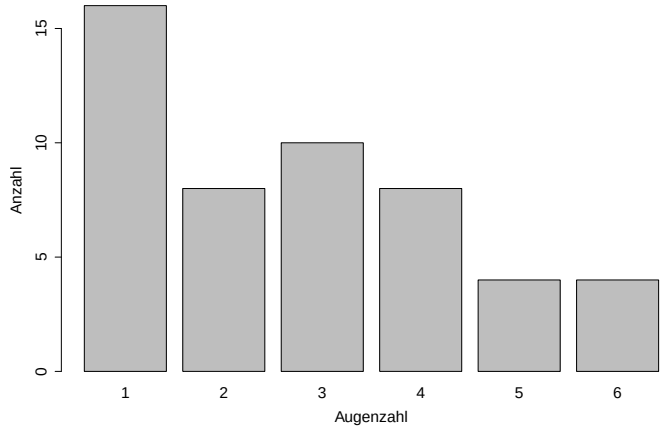

Abbildung 7.1. Balkendiagramm mit den Häufigkeiten der Augenzahlen von `würfel`.

Unter der Annahme, dass alle 6 Würfelseiten gleichwahrscheinlich sind, beträgt die *erwartete relative Häufigkeit* genau 1/6, also etwa 16,7%. Die tatsächlichen relativen Häufigkeiten berechnet man mit der Funktion `prop.table()`:

```
> prop.table(table(würfel$augenzahl))

   1    2    3    4    5    6
0.32 0.16 0.20 0.16 0.08 0.08
```

Da diese relativen Häufigkeiten bei einem fairen Würfel alle nahezu gleich sein sollten, verstärken diese stark voneinander abweichenden Zahlen unseren Verdacht, dass der Würfel nicht fair ist. Die zu testende Nullhypothese lautet hier:

H_0 : Die erwarteten relativen Häufigkeiten sind für alle Kategorien gleich.

[2] Bei einem *fairen* Würfel oder *Laplace-Würfel* ist die Wahrscheinlichkeit für jede Augenzahl gleich groß, nämlich 1/6.

-Commander

Balkendiagramm und Häufigkeitstabelle können auch über den R-Commander erstellt werden. Dafür muss die verwendete Variable allerdings eine Faktorvariable sein (vgl. Abschnitt 2.1.3), die Variable `augenzahl` ist aber eine numerische Variable. Wir müssen also zuerst die Augenzahl in eine Variable vom Typ *factor* konvertieren. Dies erreicht man unter **Datenmanagement → Variablen bearbeiten → Konvertiere numerische Variablen in Faktoren ...**. Es öffnet sich das Dialogfeld aus Abbildung 7.2, in dem die einzige Variable `augenzahl` schon automatisch ausgewählt ist. Im Feld rechts unten geben wir `augenzahl.faktor` als Namen der neuen Variablen ein und wählen die Einstellung *Verwende Ziffern*. Letztere hat zur Folge, dass als Label der einzelnen Faktorstufen die ursprünglichen Ziffern verwendet werden. Bei der Einstellung *Verwende Etiketten* wäre nach Klick auf „OK“ ein weiteres Eingabefeld erschienen, in denen man die Labels der Faktorstufen selbst hätte festlegen können.

Für ein Balkendiagramm wie in Abbildung 7.1 geht man auf **Grafiken → Balkendiagramm ...** und fügt in die Felder neben *Label für die X-Achse* bzw. *Label für die Y-Achse* die Beschriftungen „Augenzahl“ bzw. „Anzahl“ ein. Optional kann man in das Feld *Graph title* noch eine Diagrammüberschrift bestimmen, was wir aber überspringen, indem wir den Eintrag aus dem Feld löschen und auf „OK“ gehen.

Mit der neu erstellten Variablen können wir uns außerdem unter **Statistik → Deskriptive Statistik → Häufigkeitsverteilung** Tabellen mit den absoluten und den relativen Häufigkeiten anzeigen lassen. Im sich öffnenden Dialogfeld ist die Variable `augenzahl.faktor` schon aktiviert, wir gehen nur noch auf „OK“ und sehen die beiden Tabllen im Ausgabefenster.

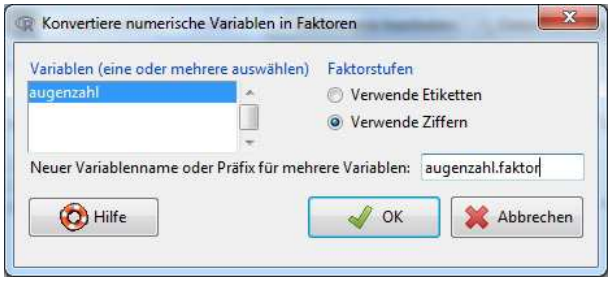

Abbildung 7.2. Dialogfeld zum Konvertieren der Variable `augenzahl` aus dem Datensatz `würfel` in eine Faktorvariable.

7.1.1 Bemerkung

Die obige Nullhypothese kann für dieses Stichprobendesign auch allgemeiner formuliert werden, wenn die relativen Häufigkeiten bestimmten Vorgaben entsprechen sollen. Beispielsweise liege eine Stichprobe aus einer Gesamtpopulation vor, wobei man von den Untersuchungseinheiten den Schulabschluss kennt. Man kann nun überprüfen, ob die Stichprobe hinsichtlich des Schulabschlusses „repräsentativ“ aus der Grundgesamtheit ausgewählt wurde. Hierbei vergleicht man die relativen Häufigkeiten der Stichprobe mit den Anteilen der Gesamtpopulation für den jeweiligen Schulabschluss, die als bekannt vorausgesetzt werden. Die Nullhypothese muss hierbei derart umformuliert werden, dass die relative Häufigkeit jeder Kategorie einer vorgegebenen erwarteten Häufigkeit entspricht. Für den durchzuführenden χ^2-Anpassungstest ergibt sich aber

keine Änderung. □

Um die Nullhypothese zu untersuchen, betrachten wir nicht die erwarteten relativen Häufigkeiten, sondern die *erwarteten (absoluten) Häufigkeiten*. Unter der Annahme, dass alle Kategorien gleichwahrscheinlich sind, erhält man diese, indem man den Gesamtstichprobenumfang mit der erwarteten relativen Häufigkeit multipliziert. Für unser Beispiel mit dem Würfelwurf ergibt sich eine erwartete Häufigkeit von $50 \cdot (1/6) = 8.33$. Die Differenz aus tatsächlichen und erwarteten Häufigkeiten sind die *Residuen*. Je höher die Residuen, desto stärker ist die Abweichung zwischen tatsächlicher und erwarteter Häufigkeit, wobei die Abweichung hier postiv oder negativ sein kann. Diese Residuen können wir bei den Würfel-Daten einfach mit

```
> round(table(würfel$augenzahl) - (50 * (1 / 6)), 2)

    1     2     3     4     5     6
 7.67 -0.33  1.67 -0.33 -4.33 -4.33
```

berechnen, wobei wir mit der Funktion `round()` auf zwei Nachkommastellen runden. Unter der Annahme, dass H_0 gilt, sollten die Residuen nahe 0 sein. Je größer die absoluten Werte der Residuen, desto eher wird die Nullhypothese in Zweifel gezogen. Das ist das Prinzip des χ^2-*Anpassungstests*.

7.1.2 Hintergrund

Bei J möglichen Kategorien einer kategorialen Zufallsvariablen, ist die Teststatistik des χ^2-Anpassungstests definiert als

$$X^2 = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j},$$

*wobei O_j die tatsächliche Häufigkeit (O steht für **O**bserved) und E_j die erwartete Häufigkeit (E steht für **E**xpected) der Kategorie j ist. Unter der Nullhypothese ist X^2 approximativ χ^2-verteilt mit $J - 1$ Freiheitsgraden.*

7.1.3 Beispiel

Für den χ^2-Anpassungstest steht in R die Funktion `chisq.test()` bereit, die wir am Beispiel des Datensatzes `würfel` vorstellen wollen:

```
> chisq.test(table(würfel$augenzahl), p = rep(1 / 6, 6))

        Chi-squared test for given probabilities

data:  table(würfel$augenzahl)
X-squared = 11.92, df = 5, p-value = 0.0359
```

Im ersten Argument der Funktion steht das zu testende Objekt, hier also der Vektor der tatsächlichen Häufigkeiten, der mittels `table()` erzeugt wird. Mit dem Argument `p` geben wir den Vektor der erwarteten relativen Häufigkeiten an (siehe Abschnitt 2.1.2 für Details zur Funktion

`rep()`). Wie schon in Bemerkung 7.1.1 erwähnt, muss der Vektor für `p` nicht nur aus gleichen Einträgen bestehen.

Der p-Wert für diesen Test beträgt 0.0359, weshalb wir die Nullhypothese, dass die relativen Häufigkeiten in allen Kategorien gleich sind, auf dem 5%-Signifikanzniveau verwerfen können. Unser anfänglicher Verdacht, dass der Würfel gezinkt ist, hat sich also bestätigt.

Anders als in kommerziellen Softwarepakteten wie SPSS oder SAS, werden die Testergebnisse in R nicht in Form einer vorformatierten Tabelle ausgegeben. Dies ist für Quereinsteiger des Programms, die weiter verwendbare Ausgabetabellen von anderen Programmen gewohnt sind, womöglich etwas störend. Ein Vorteil der Testausgaben von R ist, dass diese oft noch sehr viele weitere „versteckte“ Informationen enthalten. Die Ergebnisse der Testprozeduren werden meist in Form einer Liste mit vielen weiteren Elementen ausgegeben, die man bei Bedarf aufrufen kann. Die Ausgabe der Funktion `chisq.test()` ist ebenfalls eine Liste, was man etwa mit der Funktion `str()` (s. Abschnitt 4.1.1) nachprüfen kann. Um beispielsweise nur den p-Wert für den obigen χ^2-Test aufzurufen, wählt man das Listenobjekt `p.value` aus, indem man die Liste wie in Abschnitt 2.1.6 erläutert, indiziert:

```
> chisq.test(table(würfel$augenzahl), p = rep(1 / 6, 6))$p.value
[1] 0.03590056
```

Oben haben wir die Residuen der einzelnen Augenzahlen als Differenz der beobachteten mit den erwarteten Häufigkeiten berechnet. Dies kann man auch mit den Ergebnissen der Funktion `chisq.test()` machen. Die dazu benötigten erwarteten absoluten Häufigkeiten erhält man durch Indizieren mit `expected`:

```
> table(würfel$augenzahl) -
+ chisq.test(table(würfel$augenzahl), p = rep(1 / 6, 6))$expected

        1          2          3          4          5          6
 7.6666667 -0.3333333  1.6666667 -0.3333333 -4.3333333 -4.3333333
```

Die Werte entsprechen denen aus unserer eigenen Rechnung weiter oben. ∗

Wie schon in Hintergrund 7.1.2 erwähnt, ist der χ^2-Anpassungstest nur ein *asymptotischer* Test. Dies bedeutet, dass man bei jeder Testentscheidung einen kleinen Fehler begeht. Ist der Stichprobenumfang aber groß genug, ist dieser Fehler so klein, dass wir ihn getrost vernachlässigen können. Laut Genschel & Becker (2005) muss für die erwartete Häufigkeit $E_j \geq 5, j = 1, \ldots, J$, gelten. Ist diese Voraussetzung bei einem durchgeführten Test verletzt, erscheint neben dem Testergebnis noch eine Warnmeldung in R (`Chi-Quadrat-Approximation kann inkorrekt sein`). Die Voraussetzungen des χ^2-Anpassungstests sind hier am Ende nun noch einmal zusammengefasst.

7.1.4 Voraussetzungen (χ^2-Anpassungstest)
Es liegt eine Stichprobe $x_1, \ldots, x_n$ bestehend aus Realisationen von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ vor, deren Wertebereich aus J Kategorien besteht. Für jede dieser Kategorien muss gelten, dass $E_j \geq 5, j = 1, \ldots, J$ ist, d.h. die erwartete Häufigkeit muss in jeder Kategorie mindestens den Wert 5 betragen.


-Commander

Für den χ^2-Anpassungstest gehen wir, wie im Beispiel oben, auf **Statistik** → **Deskriptive Statistik** → **Häufigkeitsverteilung** und setzen zusätzlich ein Häkchen bei *Chi-Quadrat-Anpassungstest*. Nach Klick auf „OK“ ist in der Ausgabe neben den Tabellen mit den absoluten und relativen Häufigkeiten auch das Ergebnis des Tests zu sehen.

7.1.2 Der Binomialtest

Beschränkt sich die Anzahl der Kategorien einer Stichprobe auf zwei (z.B. männlich/weiblich oder erkankt/nicht erkrankt), führt dies zum *Binomialtest*.

7.1.5 Voraussetzungen (Binomialtest)
Die Stichprobe $x_1, \ldots, x_n$ besteht aus Realisationen von unabhängigen Wiederholungen eines Zufallsexperiments mit *dichotomen* Ausgang, d.h. es gibt genau zwei Merkmalsausprägungen.

Häufig entscheidet man sich für einen der beiden Ausgänge des Zufallsexperiments und bezeichnet dieses Ereignis dann auch als „Treffer“ und die zugehörige Wahrscheinlichkeit als „Trefferwahrscheinlichkeit“ π. Allerdings ist π unbekannt. Die Nullhypothese, die der Binomialtest untersucht, lautet nun

$$H_0 : \pi = \pi_0,$$

wobei $\pi_0 \in (0, 1)$ ein hypothetischer Wert für die Trefferwahrscheinlichkeit ist.

7.1.6 Hintergrund
Unter der Nullhypothese ist jedes Element der Stichprobe eine Realisation eines Bernoulli-Experiments mit Trefferwahrscheinlichkeit π. Die Teststatistik beim Binomialtest ist daher die Gesamttrefferzahl

$$B = \sum_{i=1}^{n} X_i,$$

die unter H_0 binomialverteilt ist gemäß $B(n, \pi)$.

7.1.7 Beispiel
Zur Verdeutlichung des Binomialtests betrachten wir den Datensatz `zuckerpackung.csv`, entnommen aus Weiß (2006). Der Datensatz ist eine Stichprobe vom Umfang $n = 540$, entnommen aus einer Lieferung von Einmalzuckerpackungen an einen Kaffeehersteller. Der Lieferant verspricht dem Kaffeehersteller, dass das Gewicht der Zuckerpackungen zwischen 7.0 und 9.0 Gramm liegt und dass höchstens 2.5% der Zuckerpackungen diese Bedingung nicht einhalten. Die erste Variable des Datensatzes enthält das Gewicht der Zuckerpackung, die zweite Variable zeigt an, ob das Gewicht gegen die obige Bedingung verstößt (`1`) oder nicht (`0`). Lesen wir zuerst den Datensatz in R ein und verschaffen uns einen Überblick über die Daten.