

Statistik in Theorie und Praxis

Mit Anwendungen in R

Bearbeitet von
Michael Falk, Johannes Hain, Frank Marohn, Hans Fischer, René Michel

1. Auflage 2014. Taschenbuch. xv, 507 S. Paperback

ISBN 978 3 642 55252 6

Format (B x L): 16,8 x 24 cm

Gewicht: 868 g

Weitere Fachgebiete > EDV, Informatik > Datenbanken, Informationssicherheit,
Geschäftssoftware > Mathematische & Statistische Software

Zu Inhaltsverzeichnis

schnell und portofrei erhältlich bei

The logo for beck-shop.de features the text 'beck-shop.de' in a bold, red, sans-serif font. Above the 'i' in 'shop' are three red dots of increasing size. Below the main text, the words 'DIE FACHBUCHHANDLUNG' are written in a smaller, red, all-caps, sans-serif font.

beck-shop.de
DIE FACHBUCHHANDLUNG

Die Online-Fachbuchhandlung beck-shop.de ist spezialisiert auf Fachbücher, insbesondere Recht, Steuern und Wirtschaft. Im Sortiment finden Sie alle Medien (Bücher, Zeitschriften, CDs, eBooks, etc.) aller Verlage. Ergänzt wird das Programm durch Services wie Neuerscheinungsdienst oder Zusammenstellungen von Büchern zu Sonderpreisen. Der Shop führt mehr als 8 Millionen Produkte.

Kapitel 2

Schätzen in Binomialmodellen

2.1 Motivation und Einführung

In diesem Kapitel stehen Zufallsexperimente mit *binärem* Ausgang im Zentrum, d.h. die Zufallsvariable kann nur zwei mögliche Ausprägungen annehmen. Probleme dieser Art finden beispielsweise in Industriebetrieben Anwendung, wie folgendes Beispiel zeigen soll.

Beispiel 2.1 *Ein Kaffeeunternehmen bezieht von einem Hersteller Einmalzuckerpackungen. Der Hersteller dieser Zuckerpackungen verspricht dem Kaffeeunternehmen, dass das Gewicht der Zuckerpackungen in einer Lieferung mehr als 7 Gramm beträgt. Der Anteil der Packungen, die diese Bedingung nicht erfüllen, soll dabei höchstens 1 % betragen. Da die Bedingung aus Zeitgründen nicht bei allen Zuckerpackungen überprüft werden kann, zieht das Kaffeeunternehmen eine Stichprobe von 540 Zuckerpackungen und überprüft an diesen die Vorgabe des Herstellers. Die zugehörigen Daten findet man im Datensatz `zuckerpackung` entnommen in leicht abgewandelter Form aus [11].*

Wir wollen im nächsten Schritt mit dem R-Commander den Anteil der Zuckerpackungen der Stichprobe bestimmen, die gegen die Vorgabe des Herstellers verstoßen.



Programmbeispiel 2.1 Wir gehen davon aus, dass `zuckerpackung` bereits der aktive Datensatz ist. Betrachtet man den Datensatz, beispielsweise indem man auf das Feld `Datenmatrix betrachten` geht, so erkennt man, dass im Datensatz nur eine Variable enthalten ist, nämlich das Gewicht der insgesamt 540 Zuckerpackungen. Wir müssen uns also in einem ersten Schritt eine neue Variable erzeugen, die anzeigt, ob eine Zuckerpackung gegen die Herstellervorgabe verstößt oder nicht. Dazu kodieren wir die Variable `gewicht` in eine neue Variable um, siehe Abschnitt 20.3.2 für genauere Erklärungen.

1. Gehe auf **Datenmanagement** → **Variablen bearbeiten** → **Rekodiere Variablen ...**
2. Zuerst bestimmt man den Namen der neuen Variablen. Dazu gibt man im Feld *Neuer Variablenname* ... den Namen `verstoss` ein (vgl. Abb. 2.2).
3. Wir möchten allen Beobachtungen, die gegen die Vorgabe verstoßen, den Wert `ja` in der neuen Variablen eintragen und allen anderen Beobachtungen den Wert `nein`. Im Feld *Eingabe der Rekodierungsanweisung* geben wir dazu folgendes ein (siehe Abb. 2.2):

```
lo:6.9 = "ja"
else    = "nein"
```

Die erste Zeile umfasst den Zahlenbereich von der kleinsten Beobachtung im Datensatz (`lo` steht für *lowest*) bis zum Wert 6.9. Das sind die Zuckerpackungen, die leichter als die Vorgabe sind und somit den Wert „ja“ zugewiesen bekommen. Da im Datensatz nur eine Nachkommastelle erfasst wird, genügt die Berücksichtigung nur einer Nachkommastelle. Alle anderen Beobachtungen erfüllen die Vorgabe und bekommen daher den Eintrag „nein“. Alternativ hätte man auch im ersten Teil der zweiten Zeile den Befehl `7:hi` eingeben können. Damit wären alle Beobachtungen von 7 bis zum größten Wert im Datensatz angesprochen worden (`hi` für *highest*).

Nachdem die kodierte Variable `verstoss` erzeugt wurde, kann man mit dieser eine Häufigkeitstabelle mit absoluten Häufigkeiten und Prozentwerten erstellen:

1. Gehe auf **Statistik** → **Deskriptive Statistik** → **Häufigkeitsverteilung**.
2. Im erscheinenden Dialogfeld braucht man nur noch auf zu gehen, da nur die Variable `verstoss` zur Auswahl möglich ist.



Im Ausgabefenster sehen wir das Ergebnis in Form von zwei Tabellen:

```
> .Table # counts for verstoss

ja nein
7  533

> round(100*.Table/sum(.Table), 2) # [...]

ja nein
1.3 98.7
```

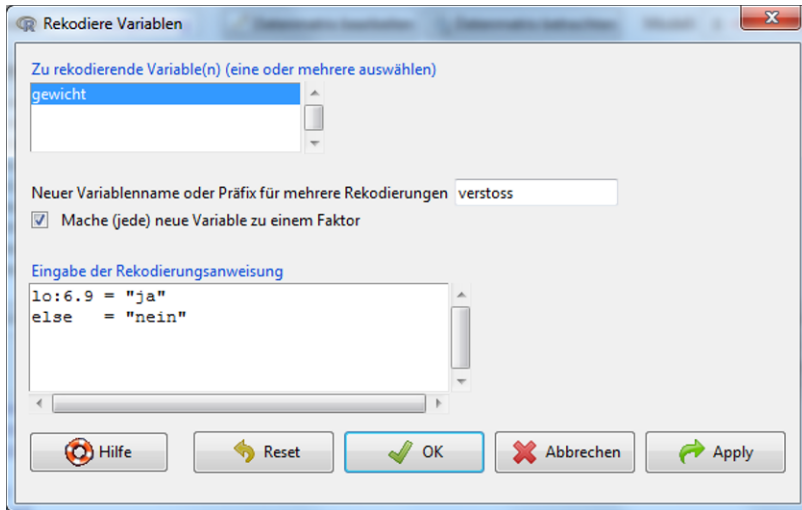


Abb. 2.2 Dialogfeld zur Umkodierung der Variable `gewicht` im Datensatz `zuckerpackungen`.

Die erste Tabelle gibt die absoluten Häufigkeiten aus, d.h. insgesamt verstoßen 7 Zuckerpackungen gegen die Vorgaben des Herstellers. Dies entspricht einem Anteil von etwa 1.3 %, was man dem zweiten Teil der Ausgabe entnimmt. Der Anteil der Zuckerpackungen, die nicht die Vorgabe des Herstellers erfüllen, beträgt also etwa 1.3 %, was etwas höher ist als die vorgegebene Fehlerquote von 1 %. Aus der Sicht des Zuckerunternehmens drängen sich nun folgende Fragen auf:

- Ist die Fehlerquote von 1.3 % noch vertretbar, oder lässt sich aufgrund dieses Werts schon darauf schließen, dass die tatsächliche Fehlerquote wirklich höher ist als die vorgegebenen 1 %?
- Aufgrund der Zufälligkeit der Stichprobe kann die Fehlerquote natürlich leicht von der Vorgabe von 1 % abweichen. In welchen Bereich aber darf sich die aus der Stichprobe ermittelte Fehlerquote über der Vorgabe befinden, ohne dass die eigene Angabe angezweifelt werden kann?

Anders als der Hersteller der Zuckerpackungen, wird das Kaffeeunternehmen die Frage umgekehrt angehen. Es wird sich fragen, wie klein die Fehlerquote mindestens sein muss, um nachweisen zu können, dass der Zuckerlieferant tatsächlich nicht gegen seine Angaben verstößt. Dem Kaffeeproduzenten geht es also darum die Einhaltung der Vorgaben zu zeigen, nicht den Verstoß der Vorgaben. Wir werden aber im Folgenden nur die Fragen des Zuckerherstellers untersuchen und verweisen für die Überprüfung der Sicht des Kaffeeunternehmens auf Aufgabe 17.

Eng verbunden mit der Frage nach einem Anteilswert ist die Frage nach einer Trefferwahrscheinlichkeit. Typischerweise treten Trefferwahrscheinlichkeiten in idealen Zufallsexperimenten auf, also in Zufallsexperimenten, die unter gleichen Bedingungen prinzipiell beliebig oft wiederholbar sind.

Beispiel 2.2 (Bekannte Trefferwahrscheinlichkeit) Eine Münze wird als fair bezeichnet, wenn bei einem Wurf die Wahrscheinlichkeit für beide Seiten gleich ist. Um zu überprüfen, ob eine vorliegende Münze fair ist, wird diese n mal geworfen und jedes Mal das Ergebnis notiert (Kopf oder Zahl).

- Angenommen die Münze wird $n = 20$ geworfen und 13 mal erscheint Kopf, d.h. der Anteil von Kopf beträgt $13/20 = 0.65$ und ist somit höher als der erwartete Anteil von 0.5: Ist die Münze fair?
- Sei nun $n = 100$, d.h. die Münze wird 100 mal geworfen und landet 65 mal mit Kopf nach oben. Der Anteil beträgt nun ebenfalls 0.65: Ist die Entscheidung ob die Münze fair ist nun eine andere wie im Fall für $n = 20$ oder nicht?

Wir werden in Kapitel 13 vorstellen, wie man die Frage, ob eine Münze fair ist, im Rahmen eines Schülerprojektes beantworten kann.

Beispiel 2.3 (Unbekannte Trefferwahrscheinlichkeit) Beim Wurf eines Reißnagels gibt es die beiden Möglichkeiten „Spitze zeigt nach oben“ oder „Spitze zeigt schräg nach unten“. Gibt es eine „Präferenz des Zufalls“? Was lässt sich über die unbekannte Trefferwahrscheinlichkeit sagen, wenn 200 Würfe auf eine ebene Fläche ergaben, dass 80 mal „Spitze zeigt nach oben“ eingetreten ist?

Allen drei Beispielen ist gemeinsam, dass sie zunächst einmal als ein Treffer/Niete-Modell angesehen werden können. In Beispiel 2.1 bedeutet Treffer „Gewicht der Zuckerpackung liegt unterhalb von 7 Gramm“, in Beispiel 2.2 „Münze zeigt Kopf“ und in Beispiel 2.3 etwa „Spitze zeigt nach oben“.

Für die Beantwortung der Fragen ist (nur) die Anzahl der Treffer von Bedeutung. Diese ist binomialverteilt, vorausgesetzt, das Treffer/Niete-Modell wird durch eine Bernoulli-Kette beschrieben (dazu mehr im nachfolgenden Abschnitt). In den Beispielen 2.2 und 2.3 ist diese Annahme wegen der Unabhängigkeit der Versuchsdurchführungen auch vernünftig. Ebenso ist in Beispiel 2.1 die Modellierung durch eine Binomialverteilung („Ziehen mit Zurücklegen“) sinnvoll, auch wenn die Gewinnung der Stichprobe strenggenommen durch „Ziehen ohne Zurücklegen“ erfolgte (hypergeometrische Verteilung), was wir im nächsten Abschnitt sehen werden.

Prinzipiell kann man obige Fragestellungen mit Hilfe eines Binomialtests beantworten. Bezeichnen H_0 die Nullhypothese, H_1 die Alternativhypothese und $p \in (0, 1)$ den (unbekannten) Anteilswert bzw. die unbekannte Trefferwahrscheinlichkeit, so lautet in Beispiel 2.1 aus Sicht des Kaffeeunternehmens das Testproblem $H_0 : p \leq 0.01$, $H_1 : p > 0.01$. In den Beispielen 2.2 und 2.3 würde man das zweiseitige Testproblem $H_0 : p = 0.5$, $H_1 : p \neq 0.5$ betrachten. Wie man hier genau vorgeht, wird im Abschnitt 2.7.1 besprochen.

Neben Tests gehören Schätzer zu den grundlegenden Verfahren der schließenden Statistik. Dabei sind die Intervallschätzungen bzw. die sogenannten Konfidenzintervalle, gerade im Hinblick auf die Praxis, von großer Bedeutung. Mehr dazu in Abschnitt 2.4. Wir werden sehen, dass Konfidenzintervalle mehr Information liefern als einzelne Tests. Insbesondere lassen sich mittels Konfidenzintervallen Testentscheidungen sofort treffen. Darüber hinaus erweist sich ein Hypothesentest bei der Frage nach der Trefferwahrscheinlichkeit in Beispiel 2.3 als wenig hilfreich. Viel

interessanter wäre dagegen eine Aussage der Form „Mit 99 %iger Sicherheit enthält das Intervall (0.37, 0.45) die unbekannte Trefferwahrscheinlichkeit“.

2.2 Schätzen eines Anteilswertes

Bevor wir uns den eigentlichen Fragestellungen nähern, müssen noch einige neue Begriffe eingeführt werden. Die Menge aller Beobachtungsobjekte, über die man bei einer statistischen Untersuchung eine Aussage treffen möchte, wird **Grundgesamtheit** oder **Population** genannt. Eine ausgewählte Teilmenge der Grundgesamtheit bezeichnet man als **Stichprobe**. Eine **Wahrscheinlichkeitsstichprobe** ist eine Stichprobe, die nach einem festgelegten stochastischen Modell gezogen wird. Meistens ist es so, dass eine (Wahrscheinlichkeits-)Stichprobe durch unabhängige und identisch verteilte Zufallsvariable (engl: independent and identically distributed (iid) random variables) X_1, \dots, X_n beschrieben wird. Man denke z. B. an das *Ziehen mit Zurücklegen* (Binomialmodell,) siehe Bemerkung 2.5. Eine nicht unabhängige Stichprobe erhält man z. B. durch das *Ziehen ohne Zurücklegen* (hypergeometrisches Modell).

Beispiel 2.4 *Je nach statistischer Fragestellung unterscheiden sich die Grundgesamtheiten, was an folgenden Beispielen verdeutlicht werden soll:*

- *Angenommen man möchte die durchschnittliche Körpergröße der männlichen Bevölkerung in Deutschland bestimmen. In diesem Fall bestünde die Grundgesamtheit logischerweise aus allen volljährigen männlichen Bürgern. Da man aus offensichtlichen Gründen nicht die Größe von allen Personen der Grundgesamtheit messen kann, empfiehlt sich die Ziehung einer zufälligen Stichprobe.*
- *Die Grundgesamtheit für das Problem mit den Zuckerpackungen, das in Abschnitt 2.1 genauer vorgestellt wurde, sind alle Zuckerpackungen der Lieferung an den Kaffeehersteller. Der Datensatz `zuckerpackungen` ist eine Stichprobe aus dieser Grundgesamtheit.*

Im Folgenden sei eine Grundgesamtheit mit N Elementen und eine Stichprobe vom Umfang n gegeben, die aus einer rein zufälligen Auswahl von Elementen der Grundgesamtheit besteht. Darüber hinaus soll diese **dichotom** sein, d.h. die Grundgesamtheit zerfällt in zwei Teilmengen: Den Elementen (Untersuchungseinheiten), die eine bestimmte Eigenschaft (E) besitzen, und den Elementen, die E nicht besitzen. Gefragt ist nach dem **Anteilswert**

$$p := \frac{\text{Anzahl der Elemente mit Eigenschaft E}}{\text{Anzahl aller Elemente der Grundgesamtheit}} \quad (2.1)$$

Das statistische Problem besteht darin, den Parameter p aufgrund einer Stichprobe zu schätzen.

Um die Problemstellung mathematisch greifbarer zu machen, modelliert man das Schätzproblem mit Hilfe eines **Urnenmodells**. In diesem entspricht die Grundge-

samtheit gleichartigen, von 1 bis N nummerierten Kugeln. Außerdem entsprechen r rote Kugeln den r Elementen aus der Grundgesamtheit mit Eigenschaft E , d.h. der Zähler in (2.1) wird hier mit r bezeichnet. Demzufolge entsprechen $N - r$ schwarze Kugeln den Elementen aus der Grundgesamtheit, die die Eigenschaft E nicht besitzen. Obige Gleichung lässt sich dann umformulieren zu

$$p = \frac{r}{N}$$

Betrachtet man nun eine Zufallsstichprobe vom Umfang n als n -maliges **Ziehen ohne Zurücklegen** aus einer Urne mit N Kugeln, so lässt sich die zufällige Anzahl der gezogenen roten Kugeln durch ein **hypergeometrisches Modell** beschreiben. Die Wahrscheinlichkeit, dass die Stichprobe genau k rote Kugeln enthält beträgt in diesem Modell

$$H_{n,r,N-r}(\{k\}) := \frac{\binom{r}{k} \binom{N-r}{n-k}}{\binom{N}{n}}, \quad (2.2)$$

wobei

$$\binom{N}{n} := \frac{N!}{n!(N-n)!} \quad (2.3)$$

der **Binomialkoeffizient von N und n** ist. (Lies: „ N über n “.)

Die Schwierigkeit dabei ist aber, dass in der Praxis meist der Populationsumfang N und auch die Anzahl r der Elemente mit der Eigenschaft E unbekannt ist. In Beispiel 2.4 wird man nicht alle Zuckerpackungen aus der Lieferung zählen und schon gar nicht die, die die Vorgaben des Herstellers verletzen. Genausowenig weiß man im gleichen Beispiel nicht die Anzahl aller männlichen Bürger in Deutschland. Man weiß oft nur, dass N im Vergleich zum Stichprobenumfang n sehr groß ist.

Um diese Problematik zu umgehen, modifiziert man das Urnenmodell und betrachtet **Ziehen mit Zurücklegen**. Die Anzahl der gezogenen roten Kugeln lässt sich nun durch ein **Binomialmodell** beschreiben. Dabei deutet man den Anteil p in (2.1) als Wahrscheinlichkeit für das Auftreten der Eigenschaft E bei einem zufällig gewählten Element der Grundgesamtheit. Die Wahrscheinlichkeit, dass die Stichprobe genau k rote Kugeln enthält, beträgt

$$B_{n,p}(\{k\}) := \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \leq k \leq n, \quad (2.4)$$

wobei $r/N = p$ und $(N-r)/N = 1-p$. Das einfachere Binomialmodell in (2.4) ist eine gute Approximation des hypergeometrischen Modells in (2.2), wenn r und $N-r$ (und damit N) groß sind im Verhältnis zu n (und damit zu k), siehe Aufgabe 1.

Im Folgenden gehen wir also davon aus, dass ein Zufallsexperiment mit **binärem** Ausgang zugrunde liegt, d.h. das Zielergebnis entweder eintritt (Treffer) oder nicht (Niete). Wird dieses Treffer/Niete-Experiment n -mal unabhängig wiederholt, so spricht man von einer **Bernoulli-Kette der Länge n** . Haben sich nach Durchführung dieses Zufallsexperiments k Treffer ergeben, so stellt sich die Frage, was man

aufgrund dieser Information über die unbekannte Trefferwahrscheinlichkeit p aussagen kann. Modelliert man die *vor* Durchführung der Experimente *zufällige* Trefferzahl als Zufallsvariable S_n , so ist diese binomialverteilt mit den Parametern n und p , oder kurz ausgedrückt $S_n \sim B_{n,p}$.

$$P_p(S_n = k) = B_{n,p}(\{k\}) := \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \leq k \leq n, \quad (2.5)$$

Die Binomialverteilung $B_{n,p}$ ist ein Wahrscheinlichkeitsmaß auf $\{0, 1, \dots, n\}$, genauer auf der Potenzmenge von $\{0, 1, \dots, n\}$. Durch die Einzelwahrscheinlichkeiten $B_{n,p}(\{k\})$, $k = 0, \dots, n$, ist das Wahrscheinlichkeitsmaß eindeutig festgelegt gemäß $B_{n,p}(A) = \sum_{k \in A} B_{n,p}(\{k\})$, $A \subset \{0, 1, \dots, n\}$. In der folgenden Bemerkung wollen wir auf die stochastische Modellierung näher eingehen.

Bemerkung 2.5 Die Zufallsvariable S_n ist eine Zählvariable. Sie zählt die zufällige Anzahl der Treffer. Der Grundraum Ω , auf dem die Zufallsvariable S_n definiert ist bzw. der zugrundeliegende Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P_p)$, bleibt abstrakt. Dieser lässt sich natürlich explizit angeben. Kodiert man Treffer mit 1 und Nieten mit 0, so wählt man $\Omega = \{0, 1\}^n$, als σ -Algebra \mathcal{A} auf Ω die Potenzmenge von Ω und mit $\omega = (\omega_1, \dots, \omega_n)$ als Wahrscheinlichkeitsmaß

$$\begin{aligned} P_p(\{\omega\}) &:= \prod_{j=1}^n p^{\omega_j} (1-p)^{1-\omega_j} \\ &= p^{\sum_{j=1}^n \omega_j} \cdot (1-p)^{n-\sum_{j=1}^n \omega_j} \end{aligned}$$

Dies ist eine formale Beschreibung einer Bernoulli-Kette der Länge n mit Trefferwahrscheinlichkeit p . Die Zufallsvariable $S_n : \Omega \rightarrow \mathbb{R}$, $S_n(\omega) = \sum_{j=1}^n \omega_j$ besitzt unter P_p eine $B_{n,p}$ -Verteilung (Aufgabe 2):

$$P_p(\{\omega \in \Omega : S_n(\omega) = k\}) = P_p(S_n = k) = B_{n,p}(\{k\}), \quad k = 0, \dots, n$$

Die Zählvariable S_n lässt sich auch in der Form $S_n = \sum_{i=1}^n X_i$ schreiben. Dabei sind $X_i : \Omega \rightarrow \{0, 1\}$ die Koordinatenprojektionen: $X_i(\omega) = \omega_i$, $\omega = (\omega_1, \dots, \omega_n) \in \Omega$, $i = 1, \dots, n$. Diese sind unabhängig und identisch verteilt mit $X_i \sim B_{1,p}$ (Bernoulli-Verteilung), siehe Aufgabe 2. Die Zufallsvariablen X_1, \dots, X_n stellen dann eine (Wahrscheinlichkeits-)Stichprobe dar. Da es nur auf die zufällige Anzahl der Erfolge ankommt, beziehen sich die folgenden Ausführungen ausschließlich auf S_n .

In der Wahrscheinlichkeitsrechnung geht man von einem bekannten Binomialmodell aus (d.h. p ist bekannt) und studiert Verteilungseigenschaften von S_n , berechnet also die Wahrscheinlichkeit von Ereignissen wie z. B. $\{S_n = k\}$, $0 \leq k \leq n$. In der Statistik ist die Situation umgekehrt: Hier liegt eine konkrete Realisierung k von S_n vor und es wird versucht, aufgrund des eingetretenen Ereignisses $\{S_n = k\}$ auf den unbekannten Parameter p zu schließen. Zunächst gilt

$$P_p(S_n = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} > 0, \quad 0 < p < 1,$$

jeder Parameter $p \in (0, 1)$ ist also theoretisch möglich. Die zentrale Frage ist somit: Welcher Modellparameter p ist der Plausibelste?

2.3 Punktschätzer

Eine plausible Wahl für den Modellparameter p ist die **relative Trefferhäufigkeit**, definiert als Quotient der Anzahl der Treffer des Experiments k und der Gesamtlänge des Experiments n , d.h.

$$\hat{p}(k) := \frac{k}{n}$$

Man spricht von einer **Punktschätzung**, da man sich auf die Angabe einer Zahl als Schätzwert beschränkt. Die Funktion \hat{p} heißt **Punktschätzer**. Ist k eine Realisierung der $B_{n,p}$ -verteilten Zufallsvariablen S_n , so ist die Schätzung $\hat{p}(k) = k/n$ eine Realisierung der Zufallsvariablen

$$\hat{p}(S_n) = \frac{S_n}{n}$$

Im Folgenden schreiben wir für die zufällige relative Trefferhäufigkeit auch R_n , also $S_n/n = R_n$. Wenn p der (wahre) zugrunde liegende Parameter ist, so beträgt wegen $S_n = nR_n$ die Wahrscheinlichkeit, dass $R_n = p$ ist,

$$P_p(R_n = p) = \begin{cases} 0, & \text{falls } np \notin \{0, 1, \dots, n\} \\ \binom{n}{np} p^{np} (1-p)^{n-np}, & \text{falls } np \in \{0, 1, \dots, n\} \end{cases} \quad (2.6)$$

Beispiel 2.6 Für $n = 20$ und $p = 0.5$ bzw. $p = 0.2$ möchten wir die Wahrscheinlichkeit für $R_n = p$ berechnen, also

$$P_{0.5}(R_n = 0.5) = P_{0.5}(S_n = 10) \quad \text{bzw.} \quad P_{0.2}(R_n = 0.2) = P_{0.2}(S_n = 4)$$

Dies ist die Wahrscheinlichkeit dafür, dass bei einem Experiment der Länge 20 mit einer Trefferwahrscheinlichkeit von 50 % bzw. 20 % genau 10 bzw. 4 Treffer erzielt werden. Dazu verwenden wir R als Rechenhilfe.



Programmbeispiel 2.3 Mit Hilfe des R-Commander kann man sich für gegebenes n und p die Wahrscheinlichkeiten für alle $k \in \{0, 1, \dots, n\}$ anzeigen lassen:

1. Gehe hierzu auf **Verteilungen** → **Diskrete Verteilungen** → **Binomial-Verteilung** → **Wahrscheinlichkeiten der Binomial-Verteilung** ...

2. Es öffnet sich ein neues Dialogfeld wie in Abb. 2.4. Im Feld *Binomial trials* gibt man den gewünschten Wert für n ein, in diesem Fall also die Zahl 20, und geht auf **OK**.
3. In der Ausgabe wird eine Liste angezeigt, bei der in der ersten Spalte der Wert von k und in der Spalte daneben die Wahrscheinlichkeit aus (2.6) angezeigt wird. Für $k = 10$ erhalten wir die Zahl $1.761971e - 01$. Diese Notation ist die übliche wissenschaftliche Darstellungsweise und muss interpretiert werden als $1.761971 \cdot 10^{-1} = 0.1761971$.
4. Für $p = 0.2$ kann man sich ebenfalls die Wahrscheinlichkeiten anzeigen lassen. Dazu muss man aber im Dialogfeld in Abb. 2.4 im Feld *Probability of success* die Standardeinstellung von 0.5 auf 0.2 ändern.

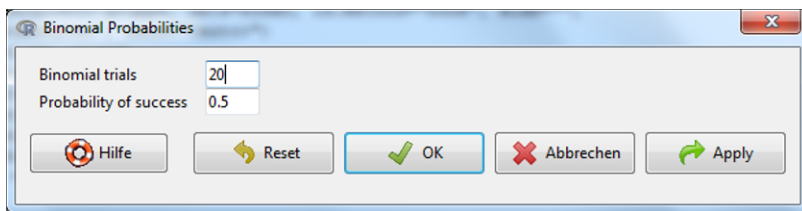


Abb. 2.4 Dialogfeld zur Berechnung von Wahrscheinlichkeiten für vorgegebenen Werte für n , p und k gemäß (2.6).

Zur Angabe einer Wahrscheinlichkeit für einen einzelnen Wert von k , verwenden wir die Funktion `dbinom()`, da dies mit dem R-Commander nicht möglich ist. Wir geben dazu im Skriptfenster folgenden Befehl ein:

```
dbinom(10, 20, 0.5)
dbinom(4, 20, 0.2)
```

Das erste Argument ist die Zahl der Treffer, das zweite die Angabe von n und das letzte Argument die Trefferwahrscheinlichkeit p . In der Ausgabe erscheint das Ergebnis:

```
> dbinom(10, 20, 0.5)
[1] 0.1761971

> dbinom(4, 20, 0.2)
[1] 0.2181994
```

Im übrigen ist die Funktion `dbinom()` identisch mit der in (2.6) definierten Funktion. Wird also im ersten Argument ein Wert für $k \notin \{0, 1, \dots, n\}$ angegeben, ist das Ergebnis 0 und es erscheint zudem eine Warnmeldung im Meldungsfenster.



Es gilt somit:

$$P_{0.5}(S_n = 10) \approx 17.6\% \quad \text{und} \quad P_{0.2}(S_n = 4) \approx 21.8\%$$

Da für die Punktschätzung $\hat{p}(k)$ im Allgemeinen

$$\hat{p}(k) \neq p$$

gilt, stellt sich natürlich die Frage nach der Genauigkeit der Schätzung. Um die Qualität einer Schätzung besser beurteilen zu können, werden Gütekriterien formuliert. Zwei wichtige Kriterien beziehen sich auf den **Erwartungswert** und die **Varianz** des Schätzers $R_n = \hat{p}(S_n)$. Nach den üblichen Rechenregeln für den Erwartungswert und die Varianz (siehe z. B. [5], [6]) gilt:

$$E_p(R_n) = \frac{1}{n} E_p(S_n) = \frac{1}{n} np = p \quad (2.7)$$

und

$$\text{Var}_p(R_n) = \frac{1}{n^2} \text{Var}_p(S_n) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n} \quad (2.8)$$

Bemerkung 2.7 (i) Gleichung (2.7) drückt aus, dass der Schätzer R_n **erwartungstreu** und damit in einem ganz bestimmten Sinne repräsentativ ist: Der Erwartungswert der Verteilung von $R_n = S_n/n$ ist gleich p , wenn die zugrunde liegende Binomialverteilung den (wahren) Parameter p besitzt. Der Schätzer R_n ergibt also „im Mittel“ gerade p . Im Fall der Erwartungstreue liegt also keine systematische Unter- bzw. Überschätzung von p vor. Anstelle von einem erwartungstreuen Schätzer spricht man auch von einem **unverzerrten** Schätzer.

(ii) Gleichung (2.8) besagt, dass die Varianz von R_n mit wachsendem Stichprobenumfang n abnimmt, ganz gleich, welches p tatsächlich zugrunde liegt. Eine Schätzung ist somit umso genauer, je größer n ist. Nach der Tschebyschow-Ungleichung (vgl. Lemma 2.13) ist der Schätzer R_n **konsistent** in dem Sinne, dass

$$\lim_{n \rightarrow \infty} P_p(\{|R_n - p| > \varepsilon\}) = 0$$

Man sagt, dass R_n **in Wahrscheinlichkeit** oder auch **stochastisch** gegen p konvergiert.



Programmbeispiel 2.5 Erzeugen wir uns mit dem R-Commander 100 binomialverteilte (Pseudo-)Zufallszahlen mit $n = 20$ und $p = 0.5$. Zur Generierung von Zufallszahlen siehe auch die Bemerkungen in Abschnitt 11.1.1.

1. Gehe im Menü auf **Verteilungen** → **Diskrete Verteilungen** → **Binomial-Verteilung** → **Zufallsstichprobe auf einer Binomial-Verteilung ...** Es öffnet sich das Dialogfeld wie in Abb. 2.6 links.
2. Zunächst geben wir in das Feld ganz oben einen Namen für das neue Objekt mit den Zufallszahlen ein, hier also `zufall.binomial`.
3. Der Wert für n wird im Feld mit der Beschriftung *Binomial trials* angegeben, die Trefferwahrscheinlichkeit p im Feld darunter (*Probability of success*). Wir

Statistik in Theorie und Praxis

Mit Anwendungen in R

Falk, M.; Hain, J.; Marohn, F.; Fischer, H.; Michel, R.

2014, XV, 507 S. 140 Abb., Softcover

ISBN: 978-3-642-55252-6