

In diesem Kapitel betrachten wir ein kategorielles Merkmal mit $K \geq 2$ potenziellen Werten x_1, x_2, \dots, x_K . Die entsprechenden Stichprobenwerte bezeichnen wir mit X_1, X_2, \dots, X_n . Diese betrachten wir als stochastisch unabhängige Zufallsvariablen, wobei

$$\mathbb{P}(X_i = x_k) = p_k \quad \text{für } 1 \leq k \leq K$$

mit gewissen Parametern $p_1, p_2, \dots, p_K \geq 0$. Insbesondere ist $\sum_{k=1}^K p_k = 1$.

Beispiele

- Betrachten wir in Beispiel 1.8 die Variable „Rauchen“ mit den möglichen Ausprägungen $x_1 = \text{„nie“}$, $x_2 = \text{„gelegentlich“}$ und $x_3 = \text{„regelmäßig“}$. Wenn wir die 263 Befragten als rein zufällige Stichprobe aus der Grundgesamtheit aller Schweizerinnen und Schweizer im Alter von ca. 18–30 Jahren betrachten, können wir das obige Modell unterstellen. Dabei sind p_1, p_2, p_3 die relativen Anteile der nicht, gelegentlich bzw. regelmäßig rauchenden Personen in der Grundgesamtheit.
- Wir bleiben bei Beispiel 1.8, betrachten aber nun die Variable „Zufallsziffer“. Nun sei p_k die Wahrscheinlichkeit, dass eine rein zufällig aus der Population gewählte Person bei dieser Frage die Ziffer $k - 1 \in \{0, 1, \dots, 9\}$ angeben würde.
- Im Vorfeld einer Parlamentswahl werden n Wahlberechtigte rein zufällig befragt, welche der aufgestellten Parteien x_1, x_2, \dots, x_K sie wählen würden. Wenn die Zahl der Befragten deutlich kleiner ist als die Gesamtzahl der Wahlberechtigten, kann man obiges Modell unterstellen, wobei p_k der momentane relative Wähleranteil für Partei x_k ist.
- Ein technisches Gerät kann unter gewissen Standardbedingungen einwandfrei funktionieren (x_1), oder es tritt eines von $K - 1$ möglichen Problemen auf (x_2, \dots, x_K). Nun werden n gleichartige Geräte unter den besagten Bedingungen getestet. Dann ist p_k die Wahrscheinlichkeit, dass bei einem einzelnen Gerät Ausgang x_k beobachtet wird.

2.1 Punktschätzer und grafische Darstellungen

Für jede der K möglichen Ausprägungen berechnen wir ihre absolute Häufigkeit

$$H_k := \#\{i \leq n : X_i = x_k\}$$

sowie ihre relative Häufigkeit

$$\hat{p}_k := \frac{H_k}{n}$$

in der Stichprobe. Wie die Notation bereits andeutet, kann man \hat{p}_k als Punktschätzer für p_k deuten. Für diese Größen H_k und \hat{p}_k gilt:

Lemma 2.1 (Multinomialverteilung) Das Tupel $\mathbf{H} = (H_k)_{k=1}^K$ ist multinomialverteilt mit Parametern n und $\mathbf{p} = (p_k)_{k=1}^K$. Das heißt, für beliebige Tupel $\mathbf{h} = (h_k)_{k=1}^K \in \mathbb{N}_0^K$ ist

$$\mathbb{P}(\mathbf{H} = \mathbf{h}) = f_{n,\mathbf{p}}(\mathbf{h}) := \binom{n}{h_1, h_2, \dots, h_K} \prod_{k=1}^K p_k^{h_k}$$

mit dem Multinomialkoeffizienten

$$\binom{n}{h_1, h_2, \dots, h_K} := \begin{cases} \frac{n!}{h_1! h_2! \dots h_K!} & \text{falls } h_1 + h_2 + \dots + h_K = n, \\ 0 & \text{sonst.} \end{cases}$$

Diese Verteilung bezeichnen wir nachfolgend mit $\text{Mult}(n, \mathbf{p})$.

Für $k = 1, 2, \dots, K$ ist H_k nach $\text{Bin}(n, p_k)$ verteilt, und die Schätzer \hat{p}_k erfüllen die Gleichungen

$$\begin{aligned} \mathbb{E}(\hat{p}_k) &= p_k, \\ \text{Var}(\hat{p}_k) &= \frac{p_k(1 - p_k)}{n} \leq \frac{1}{4n}, \\ \text{Cov}(\hat{p}_k, \hat{p}_l) &= \frac{-p_k p_l}{n} \quad \text{für } l \neq k. \end{aligned}$$

Dieses Lemma zeigt, dass \hat{p}_k ein unverzerrter Schätzer für p_k ist, dessen Fehler von der Größenordnung $O(n^{-1/2})$ ist. Genauer gesagt, ist

$$\mathbb{E}|\hat{p}_k - p_k| \leq \text{Std}(\hat{p}_k) \leq \frac{1}{2\sqrt{n}}.$$

Beweis von Lemma 2.1 Schreiben wir $\mathbf{H} = \mathbf{H}(\mathbf{X})$ mit dem Beobachtungsvektor $\mathbf{X} = (X_i)_{i=1}^n$ und $\mathcal{X} := \{x_1, x_2, \dots, x_K\}$, dann ist $\mathbb{P}(\mathbf{H} = \mathbf{h})$ gleich

$$\begin{aligned} \sum_{\tilde{\mathbf{x}} \in \mathcal{X}^n : \mathbf{H}(\tilde{\mathbf{x}}) = \mathbf{h}} \mathbb{P}(\mathbf{X} = \tilde{\mathbf{x}}) &= \sum_{\tilde{\mathbf{x}} \in \mathcal{X}^n : \mathbf{H}(\tilde{\mathbf{x}}) = \mathbf{h}} \prod_{i=1}^n p_{\tilde{x}_i} \\ &= \#\{\tilde{\mathbf{x}} \in \mathcal{X}^n : \mathbf{H}(\tilde{\mathbf{x}}) = \mathbf{h}\} \prod_{k=1}^K p_k^{h_k}. \end{aligned}$$

Die Frage ist nun, wie viele Tupel $\tilde{\mathbf{x}} \in \mathcal{X}^n$ mit $\mathbf{H}(\tilde{\mathbf{x}}) = \mathbf{h}$ existieren. Man kann auf $\binom{n}{h_1}$ Arten festlegen, an welchen Positionen der Wert x_1 steht. Danach gibt es $\binom{n-h_1}{h_2}$ Möglichkeiten, x_2 zu setzen, dann noch $\binom{n-h_1-h_2}{h_3}$ Möglichkeiten für x_3 und so weiter. Insgesamt erhalten wir

$$\binom{n}{h_1} \binom{n-h_1}{h_2} \binom{n-h_1-h_2}{h_3} \cdots \binom{n-h_1-\cdots-h_{K-1}}{h_K}$$

Möglichkeiten, und man kann leicht nachrechnen, dass dieses Produkt identisch ist mit dem Multinomialkoeffizienten $\binom{n}{h_1, \dots, h_K}$.

Analog kann man zeigen, dass $H_k \sim \text{Bin}(n, p_k)$. Nun schreiben wir

$$\hat{p}_k = n^{-1} \sum_{i=1}^n 1_{[X_i = x_k]}.$$

Dabei verwenden wir für eine beliebige Aussage A die Schreibweise

$$1_{[A]} := \begin{cases} 1, & \text{falls } A \text{ zutrifft,} \\ 0 & \text{sonst.} \end{cases}$$

Hieraus ergibt sich, dass $\mathbb{E}(\hat{p}_k) = n^{-1} \sum_{i=1}^n \mathbb{P}(X_i = x_k) = p_k$. Ferner folgt aus der stochastischen Unabhängigkeit der Zufallsvariablen X_i , dass

$$\begin{aligned} \text{Cov}(\hat{p}_k, \hat{p}_l) &= n^{-2} \sum_{i=1}^n \text{Cov}(1_{[X_i = x_k]}, 1_{[X_i = x_l]}) \\ &= n^{-1} (1_{[k=l]} p_k - p_k p_l). \end{aligned}$$

Im Falle von $k = l$ ergibt sich die Formel $\text{Var}(\hat{p}_k) = n^{-1} p_k (1 - p_k)$, und $p_k (1 - p_k)$ ist gleich $1/4 - (p_k - 1/2)^2 \leq 1/4$. \square

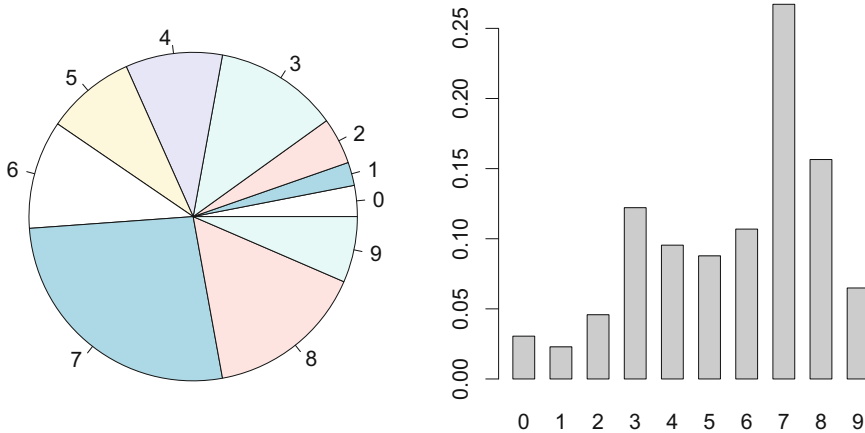


Abb. 2.1 Kuchen- und Stabdiagramm des Merkmals „Zufallsziffer“ in Beispiel 2.2

Grafische Darstellung Die absoluten oder relativen Häufigkeiten H_k bzw. \hat{p}_k kann man durch ein *Balkendiagramm* (*Stabdiagramm*, *bar chart*) oder ein *Kuchendiagramm* (*pie chart*) grafisch darstellen.

Für das Balkendiagramm werden die Ausprägungen x_k horizontal aufgelistet, und vertikal zeichnet man zu jedem x_k einen Balken mit Höhe H_k bzw. \hat{p}_k .

Für das Kuchendiagramm wird eine Kreisscheibe in K Sektoren („Kuchenstücke“) unterteilt. Jeder Sektor entspricht einer Ausprägung x_k , und seine Fläche ist proportional zu H_k bzw. \hat{p}_k .

Beispiel 2.2 („Zufallsziffern“)

Bei der Befragung in Beispiel 1.8 gaben $n = 262$ Studierende eine Zufallsziffer an. Die resultierenden absoluten und relativen Häufigkeiten waren:

x_j	0	1	2	3	4	5	6	7	8	9
H_j	8	6	12	32	25	23	28	70	41	17
\hat{p}_j	0,0305	0,0229	0,0458	0,1221	0,0954	0,0878	0,1069	0,2672	0,1565	0,0649

Abbildung 2.1 zeigt das entsprechende Stab- und Kuchendiagramm. Obwohl Kuchendiagramme sehr populär sind, lassen sich Stabdiagramme in der Regel leichter erfassen und interpretieren.

2.2 Konfidenzschranken für einen Binomialparameter

Nun konzentrieren wir uns auf eine Ausprägung x_k und betrachten nur die entsprechenden Größen $p = p_k$, $H = H_k$ und $\hat{p} = \hat{p}_k$. Wie schon gesagt wurde, ist H binomialverteilt mit Parametern n und p . An dieser Stelle empfehlen wir die Aufgaben 1 und 2.

Exakte Konfidenzschranken für p

Wir verwenden unser Kochrezept aus Kap. 1, diesmal mit den Verteilungsfunktionen $F_{n,p}$, $p \in [0, 1]$. Das heißt, $F_{n,p}(x) = \mathbb{P}_p(H \leq x) = \sum_{k=0}^x \binom{n}{k} p^k (1-p)^{n-k}$ für $x = 0, 1, \dots, n$. Zunächst müssen wir klären, inwiefern $F_{n,p}(x)$ in p monoton ist:

Lemma 2.3 Für beliebige $x \in \{0, 1, \dots, n-1\}$ ist

$$p \mapsto F_{n,p}(x)$$

stetig und streng monoton fallend auf $[0, 1]$ mit Randwerten $F_{n,0}(x) = 1$ und $F_{n,1}(x) = 0$. Genauer gesagt ist

$$F_{n,p}(x) = n \binom{n-1}{x} \int_p^1 u^x (1-u)^{n-1-x} du.$$

Die konkrete Integraldarstellung von $F_{n,p}(x)$ wird in Bemerkung 3.3 verwendet.

Beweis von Lemma 2.3 Die Funktion $p \mapsto F_{n,p}(x)$ ist ein Polynom und somit stetig und differenzierbar. Dass $F_{n,0}(x) = 1$ und $F_{n,1}(x) = 0$, ergibt sich einfach durch Einsetzen. Außerdem kann man mit elementaren Rechnungen zeigen, dass

$$\frac{d}{dp} F_{n,p}(x) = -n \binom{n-1}{x} p^x (1-p)^{n-1-x} < 0 \quad \text{für } 0 < p < 1.$$

Dies beweist die strikte Monotonie von $p \mapsto F_{n,p}(x)$, und

$$F_{n,p}(x) = F_{n,p}(x) - F_{n,1}(x) = n \binom{n-1}{x} \int_p^1 u^x (1-u)^{n-1-x} du. \quad \square$$

Abbildung 2.2 illustriert die Monotonieaussage von Lemma 2.3. Diese Monotonieeigenschaft impliziert die drei folgenden Verfahren:

(i) Mit einer Sicherheit von $1 - \alpha$ ist $F_{n,p}(H) > \alpha$. Letztere Ungleichung ist gleichbedeutend mit

$$p \begin{cases} < b_\alpha(H), & \text{falls } H < n, \\ \leq 1, & \text{falls } H = n. \end{cases}$$

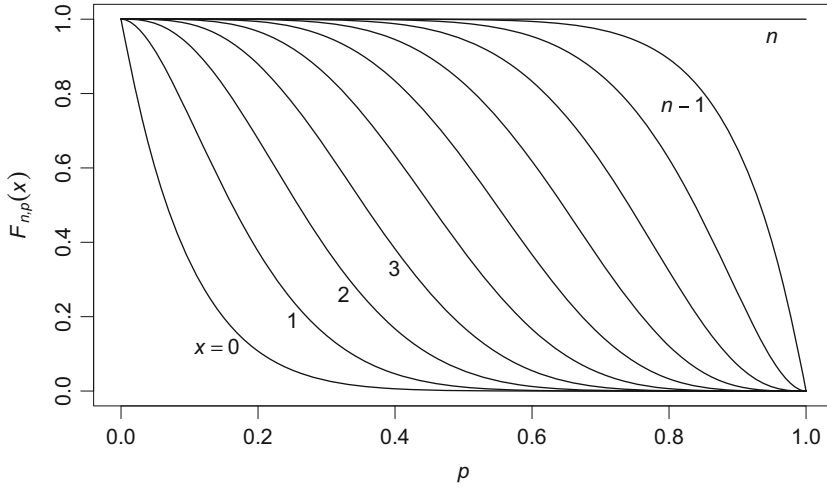


Abb. 2.2 Die Funktionen $p \mapsto F_{n,p}(x)$ für $n = 10$ und $x = 0, 1, \dots, n$

Dabei setzen wir

$$b_\alpha(h) := \begin{cases} \text{eind. Lösung } p \text{ von } F_{n,p}(h) = \alpha & \text{für } h = 0, 1, \dots, n-1, \\ 1 & \text{für } h = n. \end{cases}$$

Somit erhalten wir eine *obere* $(1 - \alpha)$ -Konfidenzschranke $b_\alpha(H)$ für p . Das heißt, wir können garantieren, dass

$$\mathbb{P}_p(p \leq b_\alpha(H)) \geq 1 - \alpha \quad \text{für beliebige } p \in [0, 1].$$

(ii) Mit einer Sicherheit von $1 - \alpha$ ist $F_{n,p}(H - 1) < 1 - \alpha$, was gleichbedeutend mit folgender Ungleichung ist:

$$p \begin{cases} \geq 0, & \text{falls } H = 0, \\ > a_\alpha(H), & \text{falls } H > 0. \end{cases}$$

Dabei setzen wir

$$a_\alpha(h) := \begin{cases} 0 & \text{für } h = 0, \\ \text{eind. Lösung } p \text{ von } F_{n,p}(h-1) = 1 - \alpha & \text{für } h = 1, 2, \dots, n. \end{cases}$$

Dies liefert eine *untere* $(1 - \alpha)$ -Konfidenzschranke $a_\alpha(H)$ für p , das heißt,

$$\mathbb{P}_p(p \geq a_\alpha(H)) \geq 1 - \alpha \quad \text{für beliebige } p \in [0, 1].$$

(iii) Wenn man den unbekannten Parameter p sowohl nach unten als auch nach oben abschätzen will, kann man das $(1 - \alpha)$ -Vertrauensintervall $[a_{\alpha/2}(H), b_{\alpha/2}(H)]$ für p verwenden. Dies ist die Methode von C. Clopper und Egon S. Pearson¹ [4]. Andere Methoden liefern tendenziell etwas kleinere Konfidenzintervalle, lassen sich aber schwieriger berechnen und begründen.

► **Bemerkung** Die Gleichung $F_{n,p}(x) = \gamma$ lässt sich für $x = 0$ und $x = n - 1$ explizit lösen. Ansonsten benötigt man numerische Verfahren, beispielsweise Bisektionsalgorithmen; siehe Aufgabe 3.

Beispiel (Qualitätskontrolle)

Der Hersteller eines bestimmten Geräts ist davon überzeugt, dass die Wahrscheinlichkeit p für den Ausfall eines solchen Gerätes unter bestimmten Bedingungen nahezu gleich null ist. Um dies zu untermauern, unterzieht er n solche Geräte einem Belastungstest und ermittelt die Zahl H von Ausfällen. Aus seiner Sicht wäre die Berechnung einer oberen Vertrauensschranke $b_\alpha(H)$ sinnvoll.

Angenommen, er beobachtet $H = 0$ Ausfälle. Dann ist $\hat{p} = 0$, und die obere Vertrauensschranke $b_\alpha(0)$ ist die Lösung p der Gleichung $F_{n,p}(0) = (1 - p)^n = \alpha$. Der Hersteller kann also mit einer Sicherheit von $1 - \alpha$ davon ausgehen, dass p kleiner ist als

$$b_\alpha(0) = 1 - \alpha^{1/n}.$$

Im Falle von $n = 50$ Geräten und $\alpha = 0,05$ ergibt sich beispielsweise die obere 95%-Vertrauensschranke $b_{0,05}(0) \approx 0,0582$.

Angenommen, der Hersteller testet $n = 50$ Geräte, und genau eines davon fällt aus. Dann ist $\hat{p} = 0,02$, und die obere Vertrauensschranke $b_{0,05}(1)$ ist die eindeutige Lösung p der Gleichung $(1 - p)^{50} + 50p(1 - p)^{49} = 0,05$. Durch geschicktes Ausprobieren kann man zeigen, dass $0,0913 \leq b_{0,05}(1) \leq 0,0914$.

Beispiel 2.4 (Meinungsumfrage)

Die Mitglieder einer Interessenvereinigung möchten ihre Stadtregierung davon überzeugen, dass die Mehrheit der Bürgerinnen und Bürger für die Beibehaltung einer bestimmten Straßenbahnlinie ist. Hierzu werden $n = 100$ Bürgerinnen und Bürger befragt, von denen sich $H = 67$ Personen für die Beibehaltung aussprechen. Dies liefert den Schätzwert $\hat{p} = 0,67$ für den unbekannten relativen Anteil p von Befürwortenden. Um die Unsicherheit bei dieser Schätzung zu berücksichtigen, ist aus Sicht der Interessenvereinigung eine untere Vertrauensschranke $a_\alpha(67)$ sinnvoll. Diese ist die Lösung p der Gleichung $F_{n,p}(66) = 1 - \alpha$. Speziell für $\alpha = 0,05$ ergeben numerische Berechnungen, dass $0,5845 \leq a_{0,05}(67) \leq 0,5846$; siehe auch Abb. 2.3. Man kann also mit einer Sicherheit von 95 % davon ausgehen, dass der relative Anteil p größer ist als 0,5845.

Verallgemeinerung Der erste Teil von Lemma 2.3 ist ein Spezialfall einer allgemeineren Aussage über Monotonieeigenschaften von Verteilungsfunktionen, die wir später noch verwenden werden:

¹ Karl Pearson (1857–1936) und Egon S. Pearson (1885–1980): Vater und Sohn, bedeutende britische Statistiker.

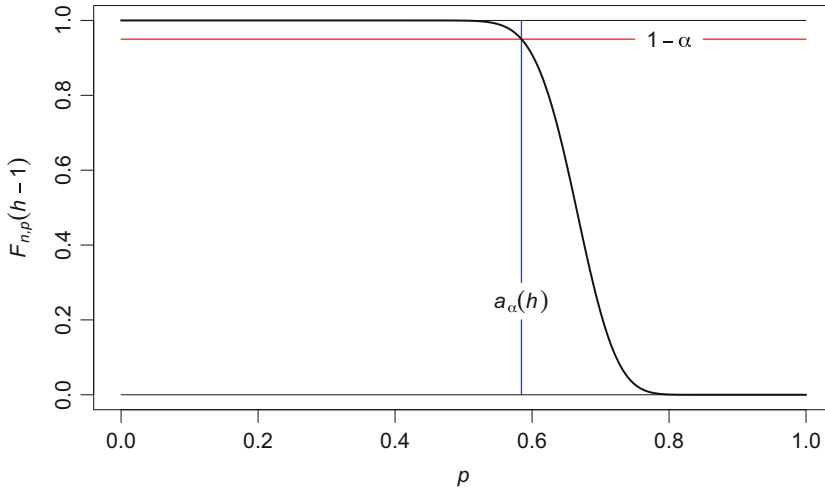


Abb. 2.3 Untere Konfidenzschranke $a_{0,05}(67)$ für p bei $n = 100$

Lemma 2.5 (Monotonieaussagen in Verteilungsfamilien) Gegeben seien nichtnegative Gewichte w_0, w_1, w_2, \dots derart, dass $0 < \sum_{k \geq 0} w_k \theta^k < \infty$ für beliebige $\theta > 0$. Nun definieren wir für einen beliebigen Parameter $\theta \in (0, \infty)$ Wahrscheinlichkeitsgewichte

$$f_\theta(x) := w_x \theta^x / \sum_{k \geq 0} w_k \theta^k, \quad x \in \mathbb{N}_0,$$

und eine Verteilungsfunktion F_θ mit

$$F_\theta(x) := \sum_{k=0}^x f_\theta(k), \quad x \in \mathbb{N}_0.$$

Im Falle von $\min\{k : w_k > 0\} \leq x < \sup\{k : w_k > 0\}$ ist $F_\theta(x)$ eine stetige und streng monoton fallende Funktion von $\theta > 0$, wobei $\lim_{\theta \rightarrow 0} F_\theta(x) = 1$ und $\lim_{\theta \rightarrow \infty} F_\theta(x) = 0$.

Beispiele

Hier folgen zwei Beispiele für solche Verteilungsfamilien:

- Poissonverteilungen $\text{Poiss}(\theta)$, $\theta > 0$: $w_k = 1/k!$;
- Binomialverteilungen $\text{Bin}(n, p)$, $0 < p < 1$: $\theta = p/(1-p)$ und $w_k = \binom{n}{k}$.

Im Zusammenhang mit „Chancenquotienten“ werden wir in Kap. 7 eine weitere Familie dieser Bauart kennenlernen.

Approximative Vertrauensschranken für p

In vielen Lehr- und Handbüchern werden noch approximative Vertrauensschranken propagiert, was für schnelle Vorauswertungen in Ordnung ist. Angesichts der heute verfügbaren Rechner ist aber die Berechnung exakter Vertrauensschranken kein Problem mehr. Wir beschreiben nun zwei Varianten von approximativen Schranken. Zuvor erinnern wir an die Definition der Normalverteilungen.

Definition (Normalverteilung)

Eine reellwertige Zufallsvariable X heißt *normalverteilt mit Erwartungswert $\mu \in \mathbb{R}$ und Standardabweichung $\sigma > 0$* , wenn sie nach der Dichtefunktion $\phi_{\mu,\sigma}$ verteilt ist; dabei ist

$$\phi_{\mu,\sigma}(x) := \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right) \quad \text{mit} \quad \phi(z) := (2\pi)^{-1/2} \exp(-z^2/2).$$

Damit gleichbedeutend ist die Aussage, dass $\mathbb{P}(X \leq x) = \Phi((x-\mu)/\sigma)$ für beliebige $x \in \mathbb{R}$, wobei

$$\Phi(x) := \int_{-\infty}^x \phi(z) dz.$$

Als Symbol für diese Verteilung verwenden wir $\mathcal{N}(\mu, \sigma^2)$. Im Spezialfall, dass $\mu = 0$ und $\sigma = 1$, nennen wir X *standardnormalverteilt*, und $\mathcal{N}(0, 1)$ ist die *Standardnormalverteilung*.

Dass X normalverteilt ist mit Erwartungswert μ und Standardabweichung $\sigma > 0$, ist gleichbedeutend damit, dass $Z := (X - \mu)/\sigma$ standardnormalverteilt ist. Mit anderen Worten: X lässt sich schreiben als $X = \mu + \sigma Z$ mit standardnormalverteiletem Z . Aus Aufgabe 5 ergibt sich dann, dass tatsächlich $\mathbb{E}(X) = \mu$ und $\text{Std}(X) = \sigma$.

Die Verteilungsfunktion $\Phi : \mathbb{R} \rightarrow (0, 1)$ der Standardnormalverteilung ist bijektiv mit Grenzwerten $\Phi(-\infty) = 0$ und $\Phi(\infty) = 1$. Ihre Umkehrfunktion bezeichnen wir mit Φ^{-1} . Aus der Symmetrie von $\mathcal{N}(0, 1)$ um 0 folgt, dass

$$\Phi(-x) = 1 - \Phi(x) \quad \text{für } x \in \mathbb{R}$$

sowie

$$\Phi^{-1}(\gamma) = -\Phi^{-1}(1 - \gamma) \quad \text{für } \gamma \in (0, 1).$$

Wilsons Methode Der Zentrale Grenzwertsatz (siehe Anhang, Abschn. A.3) beinhaltet, dass für beliebige Zahlen $-\infty \leq r < s \leq \infty$ gilt:

$$\mathbb{P}_p\left(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \in [r, s]\right) \rightarrow \Phi(s) - \Phi(r) \quad \text{wenn } np(1-p) \rightarrow \infty. \quad (2.1)$$

Für große Werte von $np(1-p) = \text{Var}(H)$ kann man also mit einer Sicherheit von ungefähr $1 - \alpha$ davon ausgehen, dass

$$\begin{aligned}\hat{p} &\leq p + c_{\alpha,n} \sqrt{p(1-p)} \quad \text{bzw.} \\ \hat{p} &\geq p - c_{\alpha,n} \sqrt{p(1-p)} \quad \text{bzw.} \\ |\hat{p} - p| &\leq c_{\alpha/2,n} \sqrt{p(1-p)}\end{aligned}$$

mit

$$c_{\alpha,n} := \Phi^{-1}(1 - \alpha) / \sqrt{n}.$$

Die vorangehenden Ungleichungen lassen sich nach p auflösen; siehe Aufgabe 6. Sie sind äquivalent zu

$$\begin{aligned}p &\geq \frac{\hat{p} + c^2/2 - c \sqrt{\hat{p}(1-\hat{p}) + c^2/4}}{1 + c^2} \quad \text{mit } c = c_{\alpha,n} \quad \text{bzw.} \\ p &\leq \frac{\hat{p} + c^2/2 + c \sqrt{\hat{p}(1-\hat{p}) + c^2/4}}{1 + c^2} \quad \text{mit } c = c_{\alpha,n} \quad \text{bzw.} \\ p &\in \left[\frac{\hat{p} + c^2/2 \pm c \sqrt{\hat{p}(1-\hat{p}) + c^2/4}}{1 + c^2} \right] \quad \text{mit } c = c_{\alpha/2,n} \quad (2.2)\end{aligned}$$

und liefern somit approximative $(1 - \alpha)$ -Konfidenzbereiche für p . Entwickelt wurde diese Methode von Edwin B. Wilson².

Beispiel

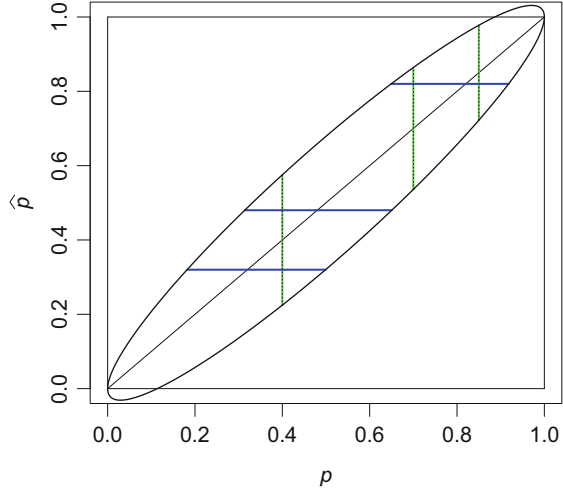
Abbildung 2.4 zeigt für $n = 30$ und $\alpha = 0,05$ die Kurven $p \mapsto p \pm c \sqrt{p(1-p)}$ mit $c = c_{\alpha/2,n}$, welche gemeinsam eine Ellipse ergeben. Für drei verschiedene Zahlen $p \in (0, 1)$ werden die Intervalle $[p \pm c \sqrt{p(1-p)}]$ als vertikale Linien gezeichnet. Außerdem sieht man für drei verschiedene Schätzwerte $\hat{p} \in (0, 1)$ die entsprechenden Konfidenzintervalle (2.2) als horizontale Linien.

Für Praktiker stellt sich die Frage, in welchen Situationen man nun Wilsons Methode anwenden darf. Eine einfache Antwort wäre „nie“, denn heutzutage stellt die Berechnung der exakten Schranken kein Problem dar. Erfahrungsgemäß liefern die exakte und Wilsons Methode ähnliche Resultate, wenn $n\hat{p}(1-\hat{p}) \geq 5$.

Walds Methode Wir beschreiben eine noch weitverbreitete und recht einfache Methode, einen Spezialfall eines viel allgemeineren Rezeptes von Abraham Wald³. Neben dem

² Edwin B. Wilson (1879–1964): US-amerikanischer Mathematiker mit vielfältigen Arbeitsgebieten.

³ Abraham Wald (1902–1950): rumänisch-US-amerikanischer Mathematiker, der u. a. sequenzielle Verfahren, d.h. Verfahren mit datenabhängigem Stichprobenumfang, entwickelte.

Abb. 2.4 Wilsons Methode

Zentralen Grenzwertsatz, der uns Aussage (2.1) liefert, gilt auch folgende Ungleichung für \hat{p} :

$$\mathbb{E} \left| \frac{\hat{p}(1 - \hat{p})}{p(1 - p)} - 1 \right| \leq \frac{\mathbb{E} |\hat{p} - p|}{p(1 - p)} \leq \frac{1}{\sqrt{np(1 - p)}}.$$

Beide Tatsachen zusammen implizieren, dass man in (2.1) den Term $\sqrt{p(1 - p)/n}$ durch $\sqrt{\hat{p}(1 - \hat{p})/n}$ ersetzen darf; siehe auch Aufgabe 26(b) in Abschn. 4.5. Man kann also mit einer Sicherheit von ca. $1 - \alpha$ davon ausgehen, dass eine der folgenden Ungleichungen erfüllt ist:

$$\begin{aligned} p &\geq \hat{p} - c_{\alpha,n} \sqrt{\hat{p}(1 - \hat{p})} \quad \text{bzw.} \\ p &\leq \hat{p} + c_{\alpha,n} \sqrt{\hat{p}(1 - \hat{p})} \quad \text{bzw.} \\ p &\in [\hat{p} \pm c_{\alpha/2,n} \sqrt{\hat{p}(1 - \hat{p})}]. \end{aligned}$$

Die Konfidenzschranken auf der rechten Seite ergeben sich auch aus Wilsons Schranken, wenn man dort alle Terme c^2 durch null ersetzt.

Zwar sind Walds Schranken wesentlich einfacher als die von Wilson, allerdings kann das tatsächliche Vertrauensniveau mit Walds Methode auch drastisch kleiner sein als das angestrebte $1 - \alpha$, wenn p nahe bei null oder eins ist. Wir betrachten die tatsächlichen Überdeckungswahrscheinlichkeiten $\mathbb{P}_p(p \in C(H))$ als Funktion von $p \in (0, 1)$. Dabei steht $C(H)$ für das Konfidenzintervall $C_{\text{Wilson}}(H)$ nach Wilsons Methode oder $C_{\text{Wald}}(H)$ nach Walds Methode. In beiden Fällen ist die Funktion

$$(0, 1) \ni p \mapsto \mathbb{P}_p(p \in C(H))$$

symmetrisch um 0,5. Daher zeigen wir in Abb. 2.5 für $n = 100$ und $\alpha = 0,05$ die Funktion $p \mapsto \mathbb{P}_p(p \in C_{\text{Wilson}}(H))$ auf $(0, 0,5]$ und die Funktion $p \mapsto \mathbb{P}_p(p \in C_{\text{Wald}}(H))$

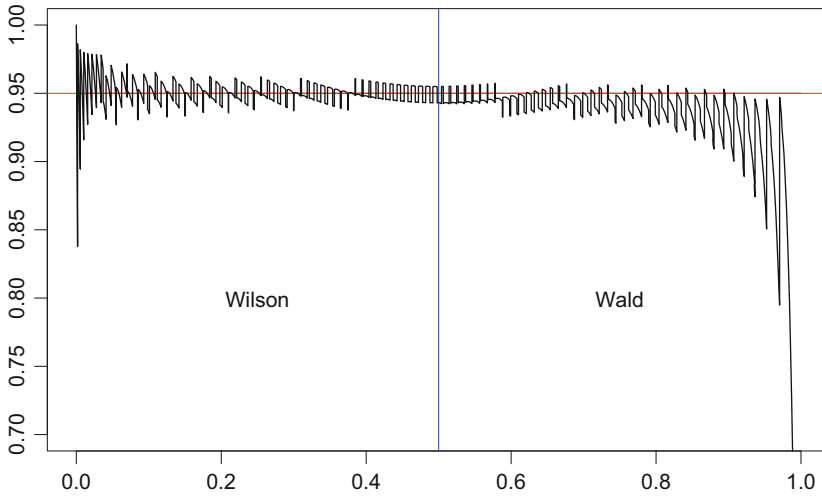


Abb. 2.5 Überdeckungswahrscheinlichkeiten des Wilson- bzw. Wald-Konfidenzintervalls, wenn $n = 100$ und $\alpha = 0,05$

auf $[0,5, 1)$. Auf der vertikalen Achse wird nur der Bereich $[0,7, 1]$ gezeigt. Tatsächlich konvergiert $\mathbb{P}_p(p \in C_{\text{Wald}}(H))$ gegen 0 für $p \rightarrow 1$.

Obere Konfidenzschranken für $|p - p_0|$

Mit unserem $(1 - \alpha)$ -Konfidenzintervall $[a_{\alpha/2}(H), b_{\alpha/2}(H)]$ für p kann man gegebenenfalls mit einer Sicherheit von $1 - \alpha$ nachweisen, dass p von einem vorgegebenen Wert p_0 abweicht. Wenn nämlich das Konfidenzintervall den Wert p_0 nicht enthält, können wir sogar mit einer Sicherheit von $1 - \alpha$ das Vorzeichen von $p - p_0$ und eine untere Schranke für die Abweichung $|p - p_0|$ angeben.

In manchen Anwendungen möchte man aber zeigen, dass der unbekannte Parameter p *nahe* an dem speziellen Wert p_0 liegt, auch wenn nicht auszuschließen ist, dass $p \neq p_0$. Aus obigem Konfidenzintervall ergibt sich folgende Aussage: Mit einer Sicherheit von $1 - \alpha$ ist $|p - p_0|$ nicht größer als

$$\max\{|p' - p_0| : a_{\alpha/2}(H) \leq p' \leq b_{\alpha/2}(H)\} = \max\{b_{\alpha/2}(H) - p_0, p_0 - a_{\alpha/2}(H)\}.$$

Doch diese Schranke ist zu konservativ. Eine bessere Schranke ergibt sich, wenn man das $(1 - \alpha)$ -Konfidenzintervall

$$[\min(a_{\alpha}(H), p_0), \max(b_{\alpha}(H), p_0)]$$

für p berechnet. Man kombiniert also die untere und obere $(1 - \alpha)$ -Vertrauensschranke für p ohne vorherige Halbierung von α , erzwingt aber, dass p_0 im Konfidenzintervall liegt.

Hinter dieser Konstruktion steckt ein allgemeines Prinzip, das in Aufgabe 12 behandelt wird. Für den Abstand $|p - p_0|$ ergibt sich die obere $(1 - \alpha)$ -Vertrauensschranke

$$\max\{b_\alpha(H) - p_0, p_0 - a_\alpha(H)\}.$$

2.3 Chiquadrat-Anpassungstest und Alternativen

In manchen Anwendungen interessiert man sich für die Frage, ob der Vektor $\mathbf{p} = (p_k)_{k=1}^K$ mit einem bestimmten Vektor $\mathbf{p}^0 = (p_k^0)_{k=1}^K$ (Nullhypothese) übereinstimmt.

Beispiele

- Ein Spielzeughersteller produziert Würfel. Nun soll überprüft werden, ob mit einem neu produzierten Würfel alle sechs Zahlen die gleiche Wahrscheinlichkeit haben. Hier ist $K = 6$, $x_k = k$ und $p_k^0 = 1/6$ für alle k . Aus Sicht des Herstellers ist es wünschenswert, dass die tatsächlichen Wahrscheinlichkeiten p_k möglichst nahe an den Werten p_k^0 sind.
- Der Roulettetisch einer Spielbank soll überprüft werden. Die Frage ist, ob alle 37 möglichen Ausgänge $0, 1, \dots, 36$ die gleiche Wahrscheinlichkeit $p_k^0 = 1/37$ haben. Ein Kontrolleur der Spielbank möchte allfällige Abweichungen der p_k von den p_k^0 möglichst zuverlässig erkennen.
- Bei der Befragung der Vorlesungsteilnehmenden wurden diese u. a. dazu aufgefordert, eine „Zufallsziffer“ aus $\{0, 1, \dots, 9\}$ zu wählen. Die Frage ist, ob und welche p_k deutlich von $p_k^0 = 1/10$ abweichen.
- Bei einer anderen Befragung von Vorlesungsteilnehmenden wurden diese aufgefordert, jeweils eine „rein zufällige“ 0-1-Sequenz der Länge 10 aufzuschreiben. Als Merkmal betrachten wir für jede der n Sequenzen die Anzahl X von Wechslen, also $X \in \{0, 1, \dots, 9\}$; siehe auch Beispiel 1.7. Unter der Nullhypothese, dass die Sequenzen wirklich rein zufällig erzeugt werden, ist p_k gleich

$$p_k^0 := \binom{9}{k-1} 2^{-9}.$$

Chiquadrat-Test

Wir möchten nun einen Test der Nullhypothese, dass $\mathbf{p} = \mathbf{p}^0$, konstruieren. Das heißt, wir möchten gegebenenfalls die Arbeitshypothese, dass $\mathbf{p} \neq \mathbf{p}^0$, mit einer gewissen Sicherheit nachweisen.

Teststatistik Um die obige Nullhypothese zu testen, benötigen wir eine Teststatistik $T = T(\mathbf{H})$, welche die augenscheinliche Abweichung von der Nullhypothese quantifiziert: Jeder Wert \hat{p}_k wird mit seinem hypothetischen Wert p_k^0 verglichen, und wir bilden die Summe

$$T := n \sum_{k=1}^K \frac{(\hat{p}_k - p_k^0)^2}{p_k^0} = \sum_{k=1}^K \frac{(H_k - np_k^0)^2}{np_k^0}.$$

Dies ist Karl Pearsons *Chiquadrat-Teststatistik*. Warum die speziellen Gewichtungsfaktoren $1/p_k^0$ auftreten, werden wir später noch sehen. Zunächst kann man schnell aus Lemma 2.1 ableiten, dass

$$\mathbb{E}(T) = K - 1 \quad \text{falls } \mathbf{p} = \mathbf{p}^0.$$

Exakter Test Unter der Nullhypothese hat die Teststatistik T eine bestimmte Verteilungsfunktion G_0 , nämlich

$$G_0(x) = \sum_{\mathbf{h} \in \mathbb{N}_0^K} 1_{[T(\mathbf{h}) \leq x]} f_{n, \mathbf{p}^0}(\mathbf{h})$$

für $x \in \mathbb{R}$; siehe Lemma 2.1. Bei Verletzung der Nullhypothese tendiert T zu großen Werten. Daher möchten wir die Nullhypothese verwerfen, wenn T „verdächtig groß“ ist. Falls also der (*rechtsseitige*) P -Wert

$$1 - G_0(T-)$$

kleiner oder gleich α ist, verwerfen wir die Nullhypothese auf dem Niveau α . Mit anderen Worten, wir behaupten dann mit einer Sicherheit von $1 - \alpha$, dass $\mathbf{p} \neq \mathbf{p}^0$. Im Falle eines P -Wertes größer als α machen wir keine definitive Aussage. Gerechtfertigt wird dieses Vorgehen durch Lemma 1.3 in Kap. 1.

Monte-Carlo-Tests Die explizite Berechnung des obigen P -Wertes $1 - G_0(T-)$ ist in der Regel sehr oder sogar zu aufwendig. Eine Alternative zum exakten P -Wert $1 - G_0(T-)$ kann man wie folgt generieren: Man simuliert mit dem Computer m stochastisch unabhängige, nach $\text{Mult}(n, \mathbf{p}^0)$ verteilte Zufallsvektoren $\mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \dots, \mathbf{H}^{(m)}$ und berechnet die entsprechenden Teststatistiken $T_s = T(\mathbf{H}^{(s)})$. Dann bestimmt man den Monte-Carlo- P -Wert

$$\frac{\#\{s \in \{1, \dots, m\} : T_s \geq T\} + 1}{m + 1}.$$

Ist dieser P -Wert kleiner oder gleich α , behaupten wir mit einer Sicherheit von $1 - \alpha$, dass die Nullhypothese nicht zutrifft. Eine theoretische Rechtfertigung dieses Verfahrens liefert die nachfolgende „Monte-Carlo-Version“ von Lemma 1.3.

Lemma 2.6 Seien T_0, T_1, \dots, T_m reellwertige Zufallsvariablen mit folgender Eigenschaft: Für jede Permutation σ von $\{0, 1, \dots, m\}$ sind $(T_{\sigma(0)}, T_{\sigma(1)}, \dots, T_{\sigma(m)})$ und (T_0, T_1, \dots, T_m) identisch verteilt. Für die Zufallsgröße

$$\hat{\pi} := \frac{\#\{s \in \{0, 1, \dots, m\} : T_s \geq T_0\}}{m + 1},$$

und beliebige $\alpha \in (0, 1)$ ist dann

$$\mathbb{P}(\hat{\pi} \leq \alpha) \leq \frac{\lfloor (m+1)\alpha \rfloor}{m+1} \leq \alpha.$$

Die vorletzte Ungleichung ist eine Gleichung, wenn die Werte T_0, T_1, \dots, T_m fast sicher paarweise verschieden sind.

Die Eigenschaft eines Zufallstupels (T_0, T_1, \dots, T_m) , dass seine Verteilung unter beliebigen Permutationen seiner Komponenten unverändert bleibt, wird uns noch mehrmals begegnen, insbesondere in Abschnitt 8.2 über Permutationstests. Sie ist beispielsweise erfüllt, wenn die Zufallsvariablen T_0, T_1, \dots, T_m stochastisch unabhängig und identisch verteilt sind.

Beweis von Lemma 2.6 Aus der Voraussetzung an die Zufallsgrößen T_0, T_1, \dots, T_m folgt, dass die $m+1$ Zufallsvariablen $\hat{\pi}_0, \hat{\pi}_1, \dots, \hat{\pi}_m$ mit

$$\hat{\pi}_j := \frac{\#\{s \in \{0, \dots, m\} : T_s \geq T_j\}}{m+1}$$

identisch verteilt sind. Daher ist $\mathbb{P}(\hat{\pi} \leq \alpha) = \mathbb{P}(\hat{\pi}_0 \leq \alpha)$ gleich

$$\frac{1}{m+1} \sum_{j=0}^m \mathbb{P}(\hat{\pi}_j \leq \alpha) = \frac{1}{m+1} \sum_{j=0}^m \mathbb{E}(1_{[\hat{\pi}_j \leq \alpha]}) = \frac{1}{m+1} \mathbb{E}\left(\sum_{j=0}^m 1_{[\hat{\pi}_j \leq \alpha]}\right).$$

Nun genügt es zu zeigen, dass stets

$$\sum_{j=0}^m 1_{[\hat{\pi}_j \leq \alpha]} \leq \lfloor (m+1)\alpha \rfloor$$

mit Gleichheit, falls die $m+1$ Zahlen T_0, T_1, \dots, T_m paarweise verschieden sind. Zu diesem Zweck seien $t_0 \leq t_1 \leq \dots \leq t_m$ die der Größe nach sortierten Werte T_0, T_1, \dots, T_m . Dann ist $\sum_{j=0}^m 1_{[\hat{\pi}_j \leq \alpha]}$ gleich

$$\begin{aligned} & \#\left\{j \in \{0, \dots, m\} : \underbrace{\#\{s \in \{0, \dots, m\} : t_s \geq t_j\}}_{\geq m+1-j} \leq (m+1)\alpha\right\} \\ & \leq \#\{j \in \{0, \dots, m\} : m+1-j \leq (m+1)\alpha\} \\ & = \#\{k \in \{1, \dots, m+1\} : k \leq (m+1)\alpha\} \\ & = \lfloor (m+1)\alpha \rfloor. \end{aligned}$$

Die vorangehenden Ungleichungen sind Gleichungen, wenn $t_0 < t_1 < \dots < t_m$. □

Monte-Carlo-Tests sind sehr einfach zu implementieren, treffen aber nicht bei allen Anwendern auf Gegenliebe, da der resultierende P-Wert nicht nur von den Daten, sondern auch von den Simulationen der $\mathbf{H}^{(s)}$ abhängt. Andererseits kann man leicht zeigen, dass sich der exakte P-Wert und der Monte-Carlo-P-Wert $\hat{\pi}$ bei großem m nur wenig unterscheiden, siehe Aufgabe 15.

Chiquadrat-Verteilungen und approximativer Test Historisch gesehen, wurde der nachfolgend beschriebene Test zuerst vorgeschlagen, da in den Anfangszeiten der Statistik rechenintensive Verfahren wie der exakte Test oder seine Monte-Carlo-Variante nicht praktikabel waren. Zunächst definieren wir eine Familie von Verteilungen, die vielerorts in der Statistik auftauchen:

Definition (Chiquadrat-Verteilungen)

Die *Chiquadrat-Verteilung* mit $l \in \mathbb{N}$ Freiheitsgraden ist definiert als die Verteilung von $\sum_{j=1}^l Z_j^2$. Dabei sind Z_1, Z_2, \dots, Z_l stochastisch unabhängig und standardnormalverteilt. Als Symbol für diese Verteilung verwendet man χ_l^2 .

In unserem speziellen Testproblem taucht die Chiquadrat-Verteilung als Approximation für die tatsächliche Verteilungsfunktion G_0 von T unter der Nullhypothese auf:

Satz 2.7 (Chiquadrat-Approximation) Sei F_{K-1} die (stetige) Verteilungsfunktion von χ_{K-1}^2 . Dann gilt:

$$\sup_{c \geq 0} |G_0(c) - F_{K-1}(c)| \rightarrow 0 \quad \text{für} \quad \min_{k=1, \dots, K} np_k^0 \rightarrow \infty.$$

Man beachte, dass die Zahl $K - 1$ der Freiheitsgrade gleich der *Anzahl von Ausprägungen minus eins* ist. Für unser Testproblem liefert Satz 2.7 den *approximativen P-Wert*

$$1 - F_{K-1}(T).$$

Eine grobe Faustregel, die in manchen Lehr- und Handbüchern propagiert wird, besagt: Wenn $\min_{k=1, \dots, K} np_k^0 \geq 5$, ist diese Approximation zuverlässig.

Illustration der Approximation In Abb. 2.6 illustrieren wir die Approximation von G_0 durch F_{K-1} in zwei Spezialfällen mit $K = 10$. Die beiden oberen Bilder zeigen die Verteilungsfunktionen G_0 (Treppenfunktion) und F_9 (glatte Funktion) im Falle von $p_k^0 = 1/10$ für $k = 1, 2, \dots, 10$ und $n = 20$ (links) bzw. $n = 50$ (rechts). Die Kenngröße $\min_k np_k^0$ ist hier gleich $n/10$, und in der Tat ist die Approximation für $n = 50$ sehr gut. Für die beiden unteren Bilder verwendeten wir $p_k^0 = 2^{-9} \binom{9}{k-1}$ und $n = 20$ (links) bzw. $n = 100$ (rechts). Hier ist $\min_k np_k^0 = n/512$, und in der Tat sind die Unterschiede zwischen G_0 und F_9 auch für $n = 100$ noch deutlich sichtbar.

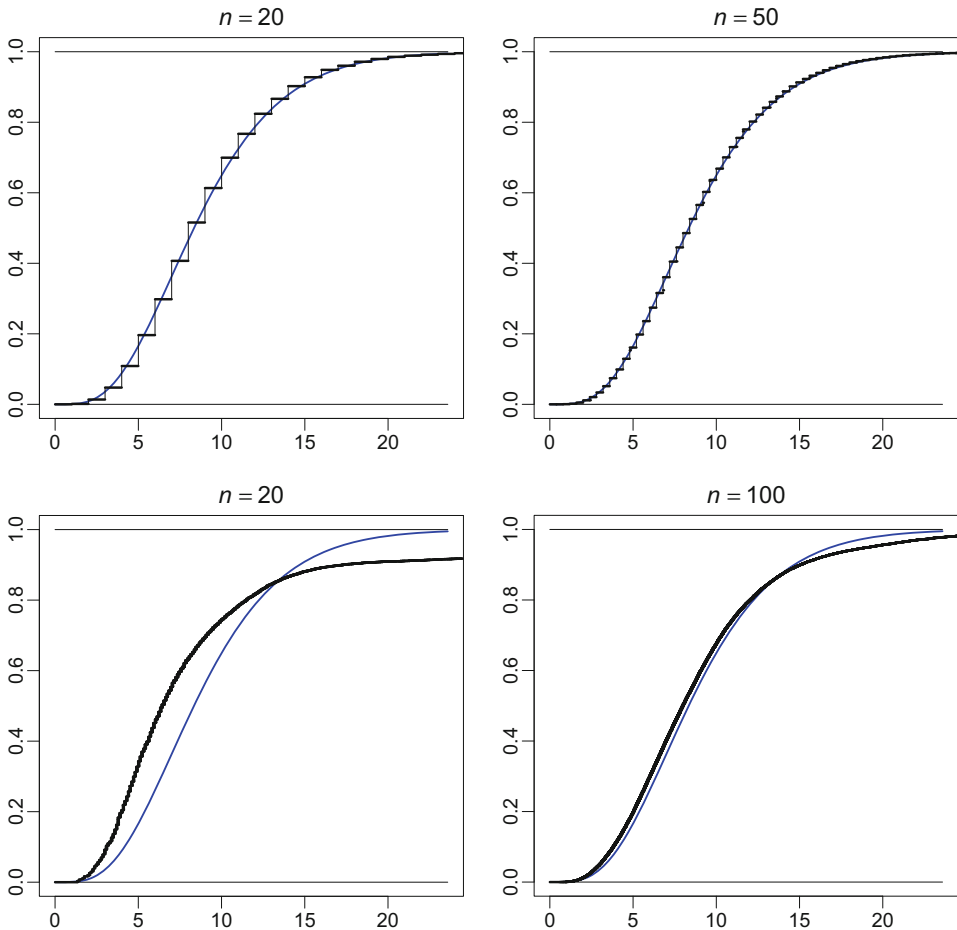


Abb. 2.6 Zur Approximation des Chiquadrat-Tests

Beispiel („Zufallsziffern“)

Für die Daten in Beispiel 2.2 möchten wir nun die Nullhypothese, dass alle p_k gleich 0,1 sind, auf dem Niveau $\alpha = 0,01$ testen. Die χ^2 -Teststatistik ist gleich

$$T = 262 \sum_{k=1}^{10} \frac{(\hat{p}_j - 0,1)^2}{0,1} \approx 122,580.$$

Wegen $\min_k np_k^0 = 26,2$ vertrauen wir der Approximation von G_0 durch F_9 ; siehe auch Abb. 2.6. Der approximative P-Wert ist hier gleich $1 - F_9(122,580) < 10^{-4}$, und auch die Monte-Carlo-Methode liefert extrem kleine P-Werte. Wir können also mit einer Sicherheit von 99 % behaupten, dass p keine Gleichverteilung auf den zehn Ziffern darstellt.

Begründung von Satz 2.7 Die χ^2 -Teststatistik T ist gleich $\|Y\|^2$ mit dem Zufallsvektor

$$Y := \sqrt{n} \left(\frac{\hat{p}_k - p_k^0}{\sqrt{p_k^0}} \right)_{k=1}^K.$$

Dieser Zufallsvektor liegt in der $(K - 1)$ -dimensionalen Ebene

$$\mathbb{H} := \left\{ y \in \mathbb{R}^K : \sum_{k=1}^K y_k \sqrt{p_k^0} = 0 \right\}.$$

Aus dem multivariaten Zentralen Grenzwertsatz folgt, dass der Zufallsvektor Y approximativ standardnormalverteilt auf \mathbb{H} ist, wenn $p = p^0$ und $\min_k n p_k^0 \rightarrow \infty$. Das heißt, Y ist approximativ verteilt wie $\sum_{j=1}^{K-1} Z_j b_j$ mit stochastisch unabhängigen, nach $\mathcal{N}(0, 1)$ verteilten Zufallsvariablen Z_1, Z_2, \dots, Z_{K-1} und einer Orthonormalbasis b_1, b_2, \dots, b_{K-1} von \mathbb{H} . Dies bedeutet aber, dass $T = \|Y\|^2$ approximativ verteilt ist wie

$$\left\| \sum_{j=1}^{K-1} Z_j b_j \right\|^2 = \sum_{j=1}^{K-1} Z_j^2 \sim \chi_{K-1}^2. \quad \square$$

Alternatives Verfahren

Der zuvor beschriebene Chiquadrat-Test hat zwei Schwächen: Wenn der Test die Nullhypothese, dass $p = p^0$, ablehnt, hat man noch keinerlei Information darüber, welche Komponenten p_k in welche Richtung von p_k^0 abweichen. In anderen Situationen möchte man vielleicht nachweisen bzw. quantifizieren, dass p „ziemlich nahe“ an p^0 ist.

Eine mögliche Alternative zu statistischen Tests ist die Berechnung eines Konfidenzintervalls $[\tilde{a}_k, \tilde{b}_k]$ für p_k , *simultan für alle* $k = 1, \dots, K$. Genauer gesagt, möchte man mit den gegebenen Daten Konfidenzschranken $\tilde{a}_k = \tilde{a}_k(H)$ und $\tilde{b}_k = \tilde{b}_k(H)$ berechnen, sodass für ein vorgegebenes α gilt:

$$\mathbb{P}(p_k \in [\tilde{a}_k, \tilde{b}_k] \text{ für } k = 1, \dots, K) \geq 1 - \alpha.$$

Mit anderen Worten: Man berechnet für den Parametervektor p ein *Konfidenzrechteck*

$$C(H) = [\tilde{a}_1, \tilde{b}_1] \times [\tilde{a}_2, \tilde{b}_2] \times \dots \times [\tilde{a}_K, \tilde{b}_K]$$

derart, dass

$$\mathbb{P}_p(p \in C(H)) \geq 1 - \alpha \quad \text{für beliebige } p.$$

Dann kann man mit einer Sicherheit von $1 - \alpha$ davon ausgehen, dass *jeder* Parameter p_k in dem entsprechenden Intervall $[\tilde{a}_k, \tilde{b}_k]$ liegt. Insbesondere lässt sich dann prüfen, ob jeder hypothetische Parameter p_k^0 in dem entsprechenden Intervall $[\tilde{a}_k, \tilde{b}_k]$ liegt.

Diese Sicherheit erreicht man durch eine sogenannte Bonferroni-Korrektur⁴: Für jeden einzelnen Parameter p_k berechnet man ein $(1 - \alpha/K)$ -Vertrauensintervall $[\tilde{a}_k, \tilde{b}_k]$, ersetzt also α durch α/K . Dann ist

$$\begin{aligned} & \mathbb{P}(p_k \in [\tilde{a}_k, \tilde{b}_k] \text{ für } k = 1, \dots, K) \\ &= 1 - \mathbb{P}(p_k \notin [\tilde{a}_k, \tilde{b}_k] \text{ für mind. ein } k \in \{1, \dots, K\}) \\ &\geq 1 - \sum_{k=1}^K \mathbb{P}(p_k \notin [\tilde{a}_k, \tilde{b}_k]) \\ &\geq 1 - \sum_{k=1}^K \alpha/K \\ &= 1 - \alpha. \end{aligned}$$

Der Vorteil dieser Methode ist, dass man möglicherweise Aussagen über die Abweichung bestimmter Parameter p_k von p_k^0 machen kann, insbesondere über die Richtung der Abweichung. Allerdings gibt es auch Datenbeispiele, bei denen der χ^2 -Anpassungstest die Nullhypothese verwirft, obwohl $p_k^0 \in [\tilde{a}_k, \tilde{b}_k]$ für alle $k = 1, \dots, K$.

Beispiel („Zufallsziffern“)

Für die Daten in Beispiel 2.2 berechnen wir nun Vertrauensintervalle für die zehn Parameter p_k mit Konfidenzniveau $(1 - \alpha/10) = 0,995$, $\alpha = 5\%$. Genauer gesagt, berechnen wir für jedes p_k die exakten einseitigen $(1 - \alpha/20)$ -Konfidenzschranken $\tilde{a}_k = a_{\alpha/20}(H_k)$ und $\tilde{b}_k = b_{\alpha/20}(H_k)$:

x_k	0	1	2	3	4	5	6	7	8	9
\tilde{a}_k	0,009	0,005	0,017	0,072	0,052	0,046	0,060	0,194	0,099	0,030
\tilde{b}_k	0,074	0,063	0,095	0,189	0,157	0,148	0,171	0,350	0,229	0,119

Insbesondere kann man mit einer Sicherheit von 95 % behaupten, dass die Wahrscheinlichkeiten der Ziffern 0, 1, 2 strikt kleiner und diejenige der Ziffer 7 strikt größer sind als 0,1.

Möchte man ausschließlich untermauern, dass \mathbf{p} nahe an \mathbf{p}^0 ist, kann man die Konfidenzintervalle $[\tilde{a}_k, \tilde{b}_k]$ auch wie folgt konstruieren: Sind $\tilde{a}_k^* = \tilde{a}_k^*(H)$ und $\tilde{b}_k^* = \tilde{b}_k^*(H)$ eine untere bzw. obere $(1 - \alpha/K)$ -Konfidenzschranke für p_k , dann ist

$$[\tilde{a}_k, \tilde{b}_k] := [\min(\tilde{a}_k^*, p_k^0), \max(\tilde{b}_k^*, p_k^0)]$$

ein $(1 - \alpha/K)$ -Vertrauensintervall für p_k , welches per Konstruktion stets den Wert p_k^0 enthält.

⁴ Carlo E. Bonferroni (1892–1960): italienischer Mathematiker, der Wahrscheinlichkeitsungleichungen in der Versicherungsmathematik und Statistik einsetzte.

Beispiel (Mendels Gesetz)

In einem Kreuzungsexperiment soll Mendels Vererbungsgesetz verifiziert werden. Von zwei Pflanzen werden durch Kreuzung $n = 400$ Tochterpflanzen erzeugt, die in Bezug auf ein bestimmtes Merkmal (Gen) vom Typ „AA“, „AB“ oder „BB“ sein können. Wenn beide Elternpflanzen vom Typ „AB“ sind, sagt Mendels Gesetz voraus, dass der Typ einer Tochterpflanze wie folgt verteilt ist:

$$(p_{AA}^0, p_{AB}^0, p_{BB}^0) = (1/4, 1/2, 1/4).$$

Angenommen, das Experiment liefert nun

$$(H_{AA}, H_{AB}, H_{BB}) = (106, 178, 116),$$

also

$$(\hat{p}_{AA}, \hat{p}_{AB}, \hat{p}_{BB}) = (0,265, 0,445, 0,290).$$

Nun berechnen wir nach der exakten Methode für die drei Parameter p_{AA} , p_{AB} und p_{BB} jeweils eine untere und eine obere $(1 - \alpha/3)$ -Vertrauensschranke, wobei $\alpha = 0,05$:

Typ	AA	AB	BB
Untere Schranke	0,2190	0,3915	0,2424
Obere Schranke	0,3151	0,4994	0,3412

Wir können also mit einer Sicherheit von $1 - \alpha = 95\%$ behaupten, dass

$$(p_{AA}, p_{AB}, p_{BB}) \in [0,2190, 0,3151] \times [0,3915, \mathbf{0,5}] \times [0,2424, 0,3412].$$

Insbesondere können wir mit einer Sicherheit von 95 % behaupten, dass die maximale Abweichung der tatsächlichen Wahrscheinlichkeiten von den Mendel'schen Werten höchstens gleich 0,1085 ist.

Abwandlung des Chiquadrat-Tests

Der χ^2 -Test in der üblichen Formulierung dient dem Nachweis, dass $\mathbf{p} \neq \mathbf{p}^0$. Man kann ihn aber auch dazu verwenden, „geschönte Daten“ aufzuspüren. Das heißt, man kann darauf achten, ob der Vektor $\hat{\mathbf{p}}$ verdächtig *nahe* an \mathbf{p}^0 ist. Zu diesem Zweck berechne man einfach den *linksseitigen* P-Wert

$$G_0(T)$$

bzw. die Monte-Carlo-Approximation

$$\frac{\#\{s \in \{1, \dots, m\} : T_s \leq T\} + 1}{m + 1}$$

bzw. die Approximation

$$F_{K-1}(T)$$

mit der Verteilungsfunktion F_{K-1} von χ_{K-1}^2 . Wenn dieser P-Wert kleiner oder gleich α ist, kann man mit einer Sicherheit von $1 - \alpha$ behaupten, dass die beobachteten absoluten Häufigkeiten *keine* Realisation eines Zufallsvektors mit Verteilung $\text{Mult}(n, \mathbf{p}^0)$ darstellen.

Beispiel

Wir greifen noch einmal das vorangehende Beispiel zu Mendels Vererbungsgesetz auf. Angenommen, ein Experimentator behauptet, sein Experiment habe $(H_{AA}, H_{AB}, H_{BB}) = (102, 199, 99)$ ergeben. Dies würde verdächtig gut zu Mendels Gesetz passen. In der Tat ist hier $T = 0,055$, und der approximative linksseitige P-Wert ist gleich $F_2(0,055) \approx 0,0271$. (Wir verwenden die χ^2 -Approximation, da $\min_k np_k = 100$.) Es sind also Zweifel am Bericht des Experimentators erlaubt. Denkbar wäre beispielsweise, dass er die Daten manipuliert oder aus mehreren Experimenten das schönste ausgewählt hat.

2.4 Übungsaufgaben

1. (Punktschätzung von p) Sei H eine Zufallsvariable mit Verteilung $\text{Bin}(n, p)$, wobei $n \in \mathbb{N}$ gegeben, aber $p \in [0, 1]$ unbekannt ist. Betrachten Sie für $c \geq 0$ den Schätzer

$$\hat{p}_c := \frac{H + c/2}{n + c}.$$

Für $c = 0$ ergibt dies den Standardschätzer $\hat{p} = H/n$, und für $c > 0$ wird letzterer zum Wert $1/2$ hin verschoben.

- (a) Bestimmen Sie Bias, Varianz und mittleren quadratischen Fehler von \hat{p}_c . Letztlich sollten Sie sehen, dass $\text{MSE}_p(\hat{p}_c)$ eine Funktion von n , c und $|p - 1/2|$ ist.
- (b) Skizzieren Sie die Funktion $p \mapsto \text{MSE}_p(\hat{p}_c)$ für $n = 25$ und $c = 0, 1, 2, \dots, 7$.
- (c) Für welchen Wert $c = c(n)$ ist der maximale mittlere quadratische Fehler,

$$\max_{0 \leq p \leq 1} \text{MSE}_p(\hat{p}_c),$$

möglichst klein?

2. (Erwartungstreue Schätzung von $g(p)$) Seien H , n und p wie in Aufgabe 1, und sei $g : [0, 1] \rightarrow \mathbb{R}$ eine beliebige Funktion. Für $g(p)$ betrachten wir nun alle Schätzer der Form $\hat{g} = s(H)$ mit einer beliebigen Abbildung $s : \{0, 1, \dots, n\} \rightarrow \mathbb{R}$.
 - (a) Angenommen, der Schätzer $\hat{g} = s(H)$ ist erwartungstreu für $g(p)$. Zeigen Sie, dass $p \mapsto g(p)$ ein Polynom vom Grad höchstens n ist.
 - (b) Angenommen, $p \mapsto g(p)$ ist ein Polynom vom Grad höchstens n . Zeigen Sie, dass es einen erwartungstreuen Schätzer $\hat{g} = s(H)$ für $g(p)$ gibt.
Hinweis: Betrachten Sie für $k = 0, 1, \dots, n$ den speziellen Schätzer $\hat{g} := [H]_k$. Welche Größe $g(p)$ wird durch \hat{g} erwartungstreu geschätzt?
 - (c) Die vorangehenden Überlegungen illustrieren, dass Erwartungstreue eine auf den ersten Blick schöne, aber auch sehr restriktive Eigenschaft ist. Vergleichen Sie unter diesem Aspekt den erwartungstreuen Schätzer für $g(p) := (1 - p)^n$ mit dem naiven Schätzer $(1 - H/n)^n$.
3. (Implementierung der exakten Konfidenzschranken für p) Um exakte Konfidenzschranken für einen Binomialparameter p zu berechnen, muss man Gleichungen der Form

$$F_{n,p}(x) = \gamma$$

für vorgegebenes $n \in \mathbb{N}$, $x \in \{0, 1, \dots, n-1\}$ und $\gamma \in (0, 1)$ lösen. Der in Tab. 2.1 beschriebene Algorithmus löst obige Gleichung mit einer vorgegebenen Genauigkeit von $\delta > 0$. Das

Tab. 2.1 Zur Berechnung exakter Vertrauensschranken für p

```

Algorithmus( $p_1, p_2$ )  $\leftarrow$  BinoCB( $x, n, \gamma, \delta$ )
 $p_1 \leftarrow 0, F_1 \leftarrow 1$ 
 $p_2 \leftarrow 1, F_2 \leftarrow 0$ 
while  $p_2 - p_1 > \delta$  or  $F_1 - F_2 > \delta$  do
   $p_m \leftarrow (p_1 + p_2)/2, F_m \leftarrow F_{n, p_m}(x)$ 
  if  $F_m \geq \gamma$  then
     $p_1 \leftarrow p_m, F_1 \leftarrow F_m$ 
  else
     $p_2 \leftarrow p_m, F_2 \leftarrow F_m$ 
  end if
end while

```

Ergebnis sind zwei Zahlen $p_1, p_2 \in [0, 1]$ derart, dass $0 < p_2 - p_1 \leq \delta$, $F_{n, p_1}(x) \geq \gamma \geq F_{n, p_2}(x)$ und $F_{n, p_1}(x) - F_{n, p_2}(x) \leq \delta$.

Implementieren Sie diesen Algorithmus. Überprüfen Sie Ihr Programm anhand von Beispiel 2.4.

4. Beweisen Sie Lemma 2.5. Beschreiben Sie dann, wie man exakte Konfidenzschranken für einen unbekannten Parameter $\theta > 0$ berechnen kann, wenn man nur eine Zufallsvariable X mit Verteilungsfunktion F_θ beobachtet. Wie könnte man den Algorithmus in Tab. 2.1 an die hiesige Situation anpassen?
5. (Momente der Standardnormalverteilung) Sei Z eine standardnormalverteilte Zufallsvariable. Zeigen Sie mit einer Symmetrieüberlegung bzw. mit partieller Integration, dass $\mathbb{E}(Z^{2m-1}) = 0$ und

$$\mathbb{E}(Z^{2m}) = \prod_{i=1}^m (2i-1) \quad \text{für } m \in \mathbb{N}.$$

Eine alternative Herleitung wird in Aufgabe 13 behandelt.

6. (Ungleichungen für Wilsons und Walds Methode) Zeigen Sie, dass für $p, \hat{p} \in [0, 1]$ und $c > 0$ gilt:

$$\hat{p} \leq_{(\geq)} p +_{(-)} c \sqrt{p(1-p)}$$

genau dann, wenn

$$p \geq_{(\leq)} \frac{\hat{p} + c^2/2 -_{(+)} c \sqrt{\hat{p}(1-\hat{p}) + c^2/4}}{1 + c^2}.$$

Für welche Werte $\hat{p} \in [0, 1]$ ist Walds Intervall

$$[\hat{p} \pm c \sqrt{\hat{p}(1-\hat{p})}]$$

kürzer bzw. länger als Wilsons Intervall

$$\left[\frac{\hat{p} + c^2/2 \pm c \sqrt{\hat{p}(1-\hat{p}) + c^2/4}}{1 + c^2} \right] ?$$

7. (Beispiele zu Konfidenzbereichen für einen Binomialparameter p) Definieren Sie für die folgenden Anwendungssituationen jeweils einen geeigneten Wahrscheinlichkeitsparameter p und überlegen Sie, ob hierfür eine untere Konfidenzschranke, eine obere Konfidenzschranke oder ein Konfidenzintervall besonders geeignet wäre. Berechnen Sie dann diese Konfidenzbereiche mit $\alpha = 5\%$. Dabei können Sie entweder (i) exakte Schranken oder (ii) Wilsons Methode verwenden.
- (a) Wie verbreitet ist Flugangst? Anlässlich eines spektakulären „Fluchtversuches“ eines Flugpassagiers kurz vor dem Start äußerten sich 335 Schweizerinnen und Schweizer zu der Frage, ob sie unter Flugangst leiden. Ergebnis: 70 Personen antworteten mit „ja“.
 - (b) Möchte die Mehrheit der Wahlberechtigten gerne per Internet abstimmen? Man fragte 29 Personen, ob sie den Gang zur Urne, eine Briefwahl oder eine Onlinewahl bevorzugen würden. Ergebnis: 22 Personen bevorzugten die Onlinewahl.
 - (c) Ein Anbieter eines WLAN-Routers möchte untermauern, dass die meisten Kunden mit der neuen Installationssoftware und -broschüre gut zurechtkommen. Zu diesem Zweck recherchiert er über sein Callcenter, wie viele von 2500 Neukunden die Service-Hotline wegen Installationsproblemen in Anspruch nahmen. Ergebnis: 42 Kunden ließen sich wegen Problemen bei der Installation beraten.
 - (d) Eine Stadtregierung soll davon überzeugt werden, dass ein bestimmter Bereich der Innenstadt problematisch ist. Hierzu werden 250 Personen gefragt, ob sie sich nachts alleine in diese Gegend trauen würden. Ergebnis: 139 Personen verneinten diese Frage (keine Enthaltungen).
8. (Vergleich zweier Poissonparameter) In manchen Anwendungen betrachtet man zwei unabhängige, poissonverteilte Zufallsvariablen $Y_1 \sim \text{Poiss}(\lambda_1)$ und $Y_2 \sim \text{Poiss}(\lambda_2)$ mit unbekannten Parametern $\lambda_1, \lambda_2 > 0$. Die Frage ist, ob und inwiefern sich λ_1 und λ_2 unterscheiden. Anwendungsbeispiele sind der Vergleich zweier Zellkonzentrationen in biologisch-medizinischen Experimenten, der Vergleich der Radioaktivität zweier Substanzen in chemisch-physikalischen Experimenten oder der Vergleich zweier Schadensraten in der Versicherungsmathematik.
- (a) Zeigen Sie, dass die bedingte Verteilung von Y_1 , gegeben, dass $Y_1 + Y_2 = s$, eine Binomialverteilung mit Parametern s und $p := \lambda_1 / (\lambda_1 + \lambda_2)$ ist. Das heißt,

$$\mathbb{P}(Y_1 = k \mid Y_1 + Y_2 = s) = \binom{s}{k} p^k (1-p)^{s-k} \quad \text{für } k = 0, \dots, s.$$

- (b) Beschreiben Sie mithilfe von Teil (a), wie man Konfidenzschranken für λ_1/λ_2 berechnen könnte. Zu welchem Ergebnis kommen Sie, wenn beispielsweise $Y_1 = 14$ und $Y_2 = 21$?
9. (Wilsons Methode für Poissonparameter) Sei Y eine Zufallsvariable mit Verteilung $\text{Poiss}(\lambda)$, wobei $\lambda \geq 0$ ein unbekannter Parameter ist. Für λ kann man exakte Konfidenzschranken berechnen, doch wir wollen nun Wilsons Methode (für Binomialparameter) imitieren. Aus dem Zentralen Grenzwertsatz lässt sich ableiten, dass für beliebige Zahlen $-\infty \leq r < s \leq \infty$ gilt:

$$P\left(\frac{Y - \lambda}{\sqrt{\lambda}} \in [r, s]\right) \rightarrow \Phi(s) - \Phi(r) \quad \text{wenn } \lambda \rightarrow \infty.$$

Leiten Sie hieraus approximative $(1 - \alpha)$ -Konfidenzschranken bzw. -intervalle für λ ab.

10. (Stichprobenumfänge bei Schätzung eines Binomialparameters) Bisher betrachteten wir den Stichprobenumfang n als fest vorgegeben. Mitunter kann man vor der Datenerhebung überlegen, wie groß die Stichprobe eigentlich sein sollte. Als Beispiel betrachten wir $H \sim \text{Bin}(n, p)$ und das $(1 - \alpha)$ -Vertrauensintervall für p nach der Wilson-Methode.

- (a) Wie groß muss der Stichprobenumfang sein, damit die Länge des Vertrauensintervalls garantiert kleiner oder gleich $\delta > 0$ ist? Zu welchem Ergebnis gelangen Sie für $\alpha = 0,05$ und $\delta = 0,1$?
- (b) Von zwei vorgegebenen Werten $0 < p_1 < p_2 < 1$ soll das Vertrauensintervall höchstens einen enthalten. Wie groß muss n sein, damit dies gewährleistet ist? Tipp: Aufgabe 6. Zahlenbeispiel: Für die deutsche FDP ist ein Wähleranteil von $p_1 = 5\%$ oder darunter verheerend (wegen der „5 %-Hürde“), ein Wähleranteil von $p_2 = 15\%$ oder darüber ist schon ein Anlass zum Feiern. Wie groß muss der Stichprobenumfang sein, damit man mindestens einen dieser Fälle mit einer Sicherheit von ca. 99 % ausschließen kann?
11. (McNemar-Test) Sei $\mathbf{H} \sim \text{Mult}(n, \mathbf{p})$ mit unbekanntem Wahrscheinlichkeitsvektor $\mathbf{p} = (p_j)_{j=1}^K$. Die Frage ist nun, ob $p_1 \leq p_2$ (Nullhypothese) oder $p_1 > p_2$ (Alternativhypothese). Anstelle eines statistischen Tests konstruieren wir nun eine geeignete Konfidenzschranke für p_1/p_2 :
- (a) Zeigen Sie, dass H_1 bei gegebener Summe $H_1 + H_2$ binomialverteilt ist mit Parametern $H_1 + H_2$ und $\rho := p_1/(p_1 + p_2)$. Das heißt, für beliebige Zahlen $m \in \{0, 1, \dots, n\}$ und $x \in \{0, 1, \dots, m\}$ ist

$$\mathbb{P}(H_1 = x \mid H_1 + H_2 = m) = \binom{m}{x} \rho^x (1 - \rho)^{m-x}.$$

- (b) Beschreiben Sie nun, wie man mithilfe von Konfidenzschranken für einen Binomialparameter Konfidenzschranken für den Quotienten p_1/p_2 angeben kann.
- (c) Werten Sie nun das folgende fiktive Datenbeispiel aus: Für den Nachweis einer bestimmten Krankheit gibt es zwei konkurrierende medizinische Tests A und B. Die Arbeitshypothese lautet, dass Test A sensitiver ist als Test B. Das heißt, bei einer erkrankten Person ist $\mathbb{P}(\text{Test A positiv})$ größer als $\mathbb{P}(\text{Test B positiv})$. Nun werden bei insgesamt $n = 60$ erkrankten Personen beide Tests angewandt. Bei 57 Personen war Test A positiv, bei 50 Personen war Test B positiv, bei 48 Personen waren beide Tests positiv. Belegen diese Daten, dass Test A sensitiver ist als Test B?
- Hinweis: Bei jeder Person sind vier verschiedene Ausgänge denkbar. Benennen Sie diese vier Ausgänge und formulieren Sie die Arbeitshypothese mithilfe der entsprechenden Wahrscheinlichkeiten. Wenden Sie dann eine der einseitigen Konfidenzschranken aus Teil (b) an.
12. (Konfidenzschranken zum Nachweis geringer Abweichungen) Bisher konstruierten wir $(1 - \alpha)$ -Vertrauensintervalle für eine reelle Größe $g(\theta)$, indem wir eine untere $(1 - \alpha/2)$ -Vertrauensschranke und eine obere $(1 - \alpha/2)$ -Vertrauensschranke für $g(\theta)$ kombinierten. Wenn man primär zeigen möchte, dass $g(\theta)$ nahe an einem gegebenen Wert g_0 ist, kann man auch anders vorgehen:
- Seien $a_\alpha = a_\alpha(\text{Daten})$ und $b_\alpha = b_\alpha(\text{Daten})$ eine untere bzw. obere $(1 - \alpha)$ -Vertrauensschranke für $g(\theta)$, das heißt, für beliebige Parameter θ ist

$$\left. \begin{aligned} \mathbb{P}_\theta(g(\theta) \geq a_\alpha) \\ \mathbb{P}_\theta(g(\theta) \leq b_\alpha) \end{aligned} \right\} \geq 1 - \alpha.$$

- Zeigen Sie, dass $[\min(a_\alpha, g_0), \max(b_\alpha, g_0)]$ ein $(1 - \alpha)$ -Vertrauensintervall für $g(\theta)$ ist.
13. Um zu klären, ob bei Neugeborenen die relativen Anteile von Mädchen und Knaben unterschiedlich sind, wurden die Daten von $n = 429.440$ Neugeborenen ausgewertet. Darunter waren $H = 221.023$ Knaben.
- (a) Berechnen Sie nun mit Wilsons Methode ein 99 %-Vertrauensintervall für die Wahrscheinlichkeit p , dass ein Neugeborenes ein Knabe ist. Wie beantworten Sie die Ausgangsfrage?
- (b) Berechnen Sie eine obere 99 %-Vertrauensschranke für $|p - 0,5|$.

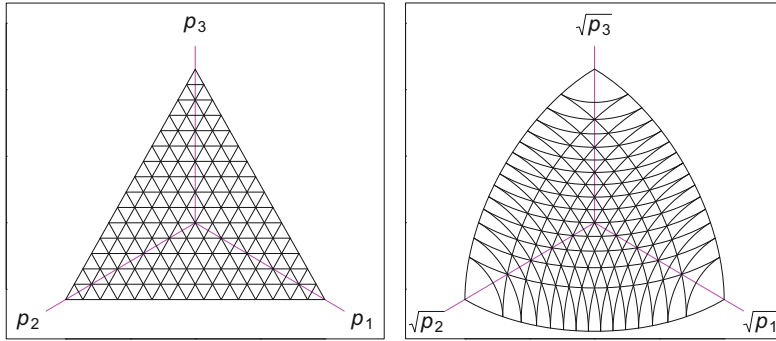


Abb. 2.7 Geometrische Betrachtung zur Chi-Quadrat-Statistik

14. (Geometrische Interpretation der Chi-Quadrat-Teststatistik) Für einen Wahrscheinlichkeitsvektor \mathbf{p} betrachten wir $\sqrt{\mathbf{p}} := (\sqrt{p_k})_{k=1}^K$. Dies definiert eine Abbildung $\mathbf{p} \mapsto \sqrt{\mathbf{p}}$ vom Einheits-simplex auf einen Ausschnitt der Einheitssphäre im \mathbb{R}^K ; siehe Abb. 2.7 für den Fall $K = 3$. Nun definieren wir für zwei Wahrscheinlichkeitsvektoren \mathbf{p}, \mathbf{q} folgende Größen:

$$T(\mathbf{p}, \mathbf{q}) := \sum_{k=1}^K \frac{(q_k - p_k)^2}{p_k}, \quad \tilde{T}(\mathbf{p}, \mathbf{q}) := 4 \|\sqrt{\mathbf{q}} - \sqrt{\mathbf{p}}\|^2$$

und

$$\delta(\mathbf{p}, \mathbf{q}) := \max_{k=1, \dots, K} \left| \frac{q_k}{p_k} - 1 \right|.$$

- (a) Zeigen Sie, dass im Falle von $\delta(\mathbf{p}, \mathbf{q}) > 0$ gilt:

$$1 - \frac{3\delta(\mathbf{p}, \mathbf{q})}{4} \leq \frac{T(\mathbf{p}, \mathbf{q})}{\tilde{T}(\mathbf{p}, \mathbf{q})} \leq 1 + \frac{\delta(\mathbf{p}, \mathbf{q})}{2}.$$

- (b) Angenommen, $\hat{\mathbf{p}} = n^{-1} \mathbf{H}$ mit $\mathbf{H} \sim \text{Mult}(n, \mathbf{p}^0)$. Zeigen Sie, dass

$$\mathbb{E}(\delta(\mathbf{p}^0, \hat{\mathbf{p}})^2) \leq \frac{K-1}{\min_{k=1, \dots, K} n p_k^0}.$$

15. Für eine Teststatistik $T = T(\text{Daten})$ betrachten wir den P-Wert

$$\pi := 1 - G_0(T-)$$

für eine gegebene Verteilungsfunktion G_0 sowie den Monte-Carlo-P-Wert

$$\hat{\pi} := \frac{\#\{s \in \{1, \dots, m\} : T_s \geq T\} + 1}{m + 1}.$$

Dabei sind T_1, T_2, \dots, T_m untereinander und von den Daten unabhängige, nach G_0 verteilte Zufallsvariablen. Nun vergleichen wir π und $\hat{\pi}$ bei gegebenen Daten, berücksichtigen also nur

den Zufall in den (simulierten) Variablen T_1, \dots, T_m und betrachten T als feste Zahl. Zeigen Sie, dass

$$\mathbb{E}((\hat{\pi} - \pi)^2) \leq \frac{1}{4m+1} \quad \text{falls } m \geq 2.$$

16. (*Leading digits*) Welcher Verteilung gehorcht die erste Ziffer einer Zahl? Wir betrachten folgende Stichprobe: Aus einem Ortsverzeichnis wurde zufällig eine Seite aufgeschlagen. Diese Seite enthält die Namen von 305 Ortschaften. In der Tabelle unten ist nun aufgeführt, in wie vielen Ortschaften die Einwohnerzahl mit der Ziffer 1, 2, ..., 9 beginnt.

Erste Ziffer	1	2	3	4	5	6	7	8	9
Häufigkeit	107	55	39	22	13	18	13	23	15

- (a) Testen Sie die Hypothese, dass diese Ziffern uniform verteilt sind auf der Menge $\{1, 2, \dots, 9\}$.
 (b) Testen Sie die Hypothese, dass diese Ziffern der Benford-Verteilung gehorchen, das heißt

$$\mathbb{P}(\text{Erste Ziffer} = k) = \log_{10}(1 + 1/k) \quad \text{für } k = 1, 2, \dots, 9.$$

17. (Benfords Gesetz) Hinter der Benford-Verteilung in der vorangehenden Aufgabe steht ein allgemeines Phänomen: Ist X eine Zufallsvariable mit stetiger Verteilungsfunktion F auf \mathbb{R} , und ist diese Verteilung „recht diffus“, dann ist die Zufallsvariable $Y := X - \lfloor X \rfloor$ „näherungsweise“ uniform verteilt auf $[0, 1)$. (Diese vage Aussage lässt sich mathematisch präzisieren.) Nun sei $Z > 0$ eine Zufallsvariable mit stetiger Verteilung auf $(0, \infty)$. Diese schreiben wir als Dezimalzahl, das heißt,

$$Z = Z_0.Z_1Z_2Z_3 \dots \cdot 10^W = (Z_0 + 10^{-1}Z_1 + 10^{-2}Z_2 + 10^{-3}Z_3 + \dots) \cdot 10^W$$

mit Ziffern $Z_0 \in \{1, \dots, 9\}$, $Z_1, Z_2, Z_3, \dots \in \{0, 1, \dots, 9\}$ und einem ganzzahligen Exponenten W . Wir gehen davon aus, dass $X = \log_{10}(Z)$ „recht diffus“ verteilt ist. Wie kann man nun aus dem oben beschriebenen Phänomen ableiten, dass

$$\mathbb{P}(Z_0 = k) \approx \log_{10}(1 + 1/k) \quad \text{für } k = 1, 2, \dots, 9?$$

Anmerkung: Benfords Gesetz wird beispielsweise bei Steuerprüfungen verwendet, um Manipulationen von Datenmaterial aufzuspüren.

18. Die folgende Tabelle enthält die Anzahl von Todesfällen in den USA in den 12 Monaten des Jahres 1966:

Januar	166.761	Juli	159.924
Februar	151.296	August	145.184
März	164.804	September	141.164
April	158.973	Oktober	154.777
Mai	156.455	November	150.678
Juni	149.251	Dezember	163.882

Die Frage ist nun, ob die Todesfallrate eines Monats proportional zu seiner zeitlichen Länge ist. Man kann mathematisch begründen, dass sich die Sterbemonate X_1, X_2, \dots, X_N der im

Jahre 1966 verstorbenen US-Amerikaner nach Bedingen auf N wie unabhängige und identisch verteilte Zufallsvariablen verhalten, und wir interessieren uns für die unbekannten Wahrscheinlichkeiten $p_k = \mathbb{P}(X_i = \text{Monat Nr. } k)$.

Formulieren und überprüfen Sie eine Nullhypothese mit den beiden zuvor beschriebenen Methoden, also mit dem χ^2 -Anpassungstest auf dem Niveau $\alpha = 0,01$ bzw. mit den simultanen 99 %-Konfidenzintervallen für die p_k . Wie interpretieren Sie die Ergebnisse?

Einführung in die Statistik

Dumbgen, L.

2016, X, 242 S. 48 Abb., 30 Abb. in Farbe., Softcover

ISBN: 978-3-0348-0003-7

A product of Birkhäuser Basel