Grundidee und Beispiele statistischer Tests

8

Nachdem wir uns bisher vor allem mit der Schätzung unbekannter Parameter beschäftigt haben, widmen wir uns nun der Grundidee statistischer Hypothesentests. Es geht dabei wieder darum, die Unwahrscheinlichkeit von Beobachtungen zu beurteilen. Diese Denkweise kennen wir schon von den Konfidenzintervallen aus Kap. 6. Im Folgenden führen wir die Schritte des statistischen Testens anhand des Einführungsbeispiels aus Kap. 1 ausführlich ein.

8.1 Idee des statistischen Hypothesentests am Einführungsbeispiel

Wir betrachten zunächst nochmals das Einführungsbeispiel, in dem der Anteil aller Studierenden geschätzt werden sollte, die eine Party besuchen werden. Nehmen wir an, dass aus allen Studierenden rein zufällig $n = 60$ Personen befragt wurden, ob sie die Party besuchen werden. Ihre Antwort ist im Vektor $\mathbf{x} = (x_1, \ldots, x_n)^t$ festgehalten. Ein Eintrag ist mit 1 kodiert, wenn die Person angegeben hat, die Party besuchen zu wollen, und mit 0 sonst. Angenommen, es möchten $\sum x_i = 35$, also ein Anteil von etwa $\hat{p}(\mathbf{x}) \approx 0.58$, die Party besuchen. Da der bisherige Raum aber nur für einen Anteil von etwa 40 % der Studierenden ausgelegt ist, stellt sich die Frage, ob die Abweichung $|\hat{p}(\mathbf{x}) - p^{(0)}|$ vielleicht durch Zufall zu erklären ist, oder ob die Beobachtungen Anlass geben, am Wert $p^{(0)} = 0.4$ zu zweifeln. Dieser Art von Fragen gehen *statistischen Hypothesentests* nach. Sie verwenden folgende Schritte:

1. *Wahl des Modells:* Man wählt ein zu den Beobachtungen passendes statistisches Modell. Wir wählen das Bernoullimodell, in dem $\mathfrak{X} = (X_1, \ldots, X_n)^t$ unabhängige, identisch Bernoulli-verteilte Komponenten hat, d. h., insbesondere ist $X_1 \sim ber(p)$, mit $p \in \Theta = (0,1)$. Da der Anteil an Partybesuchern in der Stichprobe weder 0 noch 1 ist, können wir $p = 0$ und $p = 1$ sowieso ausschließen.

M. Messer und G. Schneider, *Statistik,* https://doi.org/10.1007/978-3-662-59339-4_8

2. *Formulierung von Hypothesen:* Im Rahmen des Modells formulieren wir nun Hypothesen. Die Behauptung, der wahre Parameter sei $p^{(0)} = 0.4$, nennen wir die zu testende *Nullhypothese* H_0. Sämtliche anderen Möglichkeiten für p bilden die sogenannte Alternativhypothese H_A, kurz:

$$\begin{aligned} H_0: &\quad p \in \{0.4\}, \\ H_A: &\quad p \in \Theta \backslash \{0.4\}. \end{aligned}$$

   Sprechweise: Wir testen die Nullhypothese, dass der wahre Anteil der Partybesucher in der Population bei $p^{(0)} = 0.4$ liegt.
3. *Wahl einer Teststatistik:* Wir wählen eine Statistik T, die die Abweichung der Beobachtungen $\mathbf{x}$ von der Behauptung quantifizieren soll. Wir wählen

$$T(\mathbf{x}) := \hat{p}(\mathbf{x}) - p^{(0)},$$

   für $\mathbf{x} \in \{0, 1\}^n$. Der Bildraum Γ von T ist gegeben durch $\Gamma = \{-p^{(0)}, -p^{(0)} + 1/n, \ldots, -p^{(0)} + 1\} = \{-0.4, \ldots, 0.6\}$, siehe Abb. 8.1. Die Denkweise ist, dass die Statistik einen ‚extremen' (hier: einen betragsmäßig großen) Wert annehmen soll, wenn die Beobachtungen nicht mit der Behauptung verträglich sind. Im Kontext der Hypothesentests nennen wir eine Statistik eine *Teststatistik*. Für unsere Beobachtungen gilt $T(\mathbf{x}) \approx 0.18$.
4. *Wahl eines Ablehnungsbereichs:* Wir teilen nun den Bildraum Γ von T auf in einen Ablehnungsbereich $\mathcal{R}$ und einen Rest $\mathcal{R}^c = \Gamma \backslash \mathcal{R}$. Die Idee ist, dass damit eine Vorschrift eingeführt wird, auf Basis derer wir entscheiden, ob eine Abweichung als extrem klassifiziert wird: Fällt die auf den Beobachtungen basierende Statistik $T(\mathbf{x})$ in $\mathcal{R}$, dann sagt man, die Nullhypothese wird abgelehnt.

$$\begin{aligned} T(\mathbf{x}) \in \mathcal{R} &\quad \Rightarrow \quad H_0 \text{ wird abgelehnt.} \\ T(\mathbf{x}) \notin \mathcal{R} &\quad \Rightarrow \quad H_0 \text{ wird nicht abgelehnt.} \end{aligned}$$

   Damit diese Denkweise Sinn ergibt, ist der Ablehnungsbereich an die Randbereiche des Bildraums Γ zu legen. Denn falls die Beobachtungen nicht mit der Nullhypothese verträglich sind, so wird die Teststatistik $T(x)$ ja extreme Werte annehmen. Wir wählen hier $\mathcal{R} := \{-p^{(0)}, \ldots, -q\} \cup \{q, \ldots, -p^{(0)} + 1\}$, siehe Abb. 8.1, roter Bereich. Für unsere Beobachtungen gilt damit, dass wir die Nullhypothese ablehnen, falls $T(\mathbf{x}) = 0.18 > q$ gilt.
5. *Wahl des Signifikanzniveaus:* Wir wählen $\alpha \in (0, 1)$ und stellen eine zusätzliche Bedingung an den Ablehnungsbereich: $\mathcal{R}$ soll möglichst groß, d. h., q möglichst klein sein, sodass gerade noch gilt, dass

$$\mathbb{P}_{0.4}(T(\mathfrak{X}) \in \mathcal{R}) \leq \alpha. \tag{8.1}$$

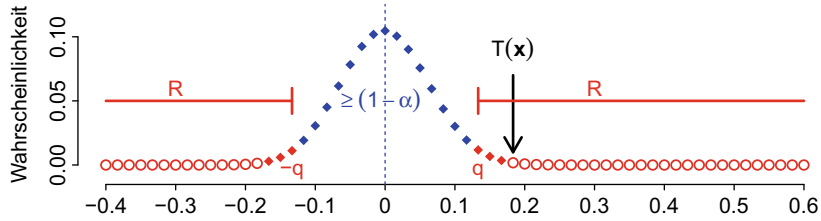


Abb. 8.1 Ablehnungsbereich im Einführungsbeispiel mit $\alpha = 0.05$

Interpretation: Wenn die Nullhypothese stimmt, dann machen wir in höchstens $(\alpha \cdot 100)$ % der Fälle einen Fehler und lehnen die Nullhypothese fälschlicherweise ab. Wir nennen α das Signifikanzniveau des Tests. Es ist vor Betrachtung der Beobachtungen zu wählen und quantifiziert, wie konservativ wir bei der Beurteilung der Unverträglichkeit der Beobachtungen mit der Nullhypothese sind. Prominente Kandidaten sind $\alpha = 0.05, 0.01$ oder 0.001. Die Bedingung (8.1) ist also über die Verteilung der Teststatistik $T(\mathfrak{X})$ unter der Nullhypothese gegeben: In Abb. 8.1 fangen also die blauen Gewichte gerade mindestens $1 - \alpha$ der Masse dieser Verteilung ein.

5. *Auswertung der Beobachtungen:* Bei unserer Betrachtung ergibt sich mit $\alpha = 0.05$ ein Wert von $q = 8/n \approx 0.13$, und damit fällt $T(\mathbf{x}) = 0.18$ in den Ablehnungsbereich $\mathcal{R}$ (siehe auch Abb. 8.1). Sprechweise: Die Nullhypothese wird auf dem 5 % Niveau abgelehnt. Bei einer Wahl von $\alpha = 0.01$ bzw. $\alpha = 0.001$ wären wir strenger, denn q ergäbe sich als $10/n \approx 0.17$ bzw. $13/n \approx 0.22$ d. h., der Ablehnungsbereich würde sukzessive kleiner. Mit $T(\mathbf{x}) = 0.18$ würde die Nullhypothese auch auf dem 1 %-Niveau, nicht aber auf dem 0.1 %-Niveau abgelehnt.

Wichtig ist, dass wir erkennen, dass das Prozedere eines Hypothesentests keine Wahrheiten (nicht einmal Wahrscheinlichkeiten) über die Hypothesen liefert, sondern lediglich eine Entscheidungsregel beschreibt, anhand derer wir die Verträglichkeit der Beobachtungen mit der Nullhypothese klassifizieren. Außerdem betonen wir, dass durch die Bedingung (8.1) insbesondere Kenntnis über die Verteilung der Teststatistik unter der Nullhypothese vorausgesetzt wird. Wir formulieren den Hypothesentest nun formal.

Definition 8.1 (Hypothesentest)
Es sei ein statistisches Modell gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ *mit Bildraum* $\mathcal{X}$ *und eine Verteilungsfamilie* $(\nu_\vartheta)_{\vartheta \in \Theta}$. *Es sei* $\Theta_0 \subseteq \Theta$ *und* $\alpha \in (0, 1)$. *Zudem sei* $T : \mathcal{X} \to \Gamma$ *eine Statistik und* $\mathcal{R} \subseteq \Gamma$. *Dann heißt* T *eine Teststatistik eines Tests der Nullhypothese*

$$H_0 : \vartheta \in \Theta_0$$

zum Signifikanzniveau α mit Ablehnungsbereich $\mathcal{R}$, falls für alle $\vartheta \in \Theta_0$ gilt, dass

$$\mathbb{P}_\vartheta(T(\mathfrak{X}) \in \mathcal{R}) \leq \alpha. \tag{8.2}$$

Wir denken meist konkret an reellwertige Teststatistiken, d. h., $\Gamma \subseteq \mathbb{R}$. Wir nennen zudem $H_A : \vartheta \in \Theta \backslash \Theta_0$ die Alternativhypothese. Die Formulierung von Hypothesen ist also eine Zerlegung von Θ in Θ_0 und $\Theta \backslash \Theta_0$. Somit denken wir uns die Verteilungsfamilie $(\nu_\vartheta)_{\vartheta \in \Theta}$ aufgeteilt in zwei Teilfamilien $(\nu_\vartheta)_{\vartheta \in \Theta_0}$ und $(\nu_\vartheta)_{\vartheta \in \Theta \backslash \Theta_0}$. Im vorherigen Beispiel hatten wir es mit der einelementigen Nullhypothese $H_0 : p^{(0)} \in \{0.4\}$ zu tun.

Die definierende Eigenschaft (8.2) eines Hypothesentests besagt, dass die Teststatistik höchstens mit Wahrscheinlichkeit α in den Ablehnungsbereich $\mathcal{R}$ fallen darf, und zwar unter allen Kandidatenverteilungen ν_ϑ, die der Nullhypothese $\vartheta \in \Theta_0$ zugeordnet sind. Da die Wahl des Ablehnungsbereichs $\mathcal{R}$ auch von α abhängt, schreiben wir manchmal auch $\mathcal{R}(\alpha)$.

Notation Gilt eine Aussage über eine Statistik unter allen Verteilungen, die der Nullhypothese $H_0 : (\nu_\vartheta)_{\vartheta \in \Theta_0}$ zugeordnet sind, so deuten wir das häufig durch die Formulierung des Index H_0 an, und die Sprechweise ist dann, dass die entsprechende Eigenschaft unter H_0 gilt. Beispielsweise schreiben wir (8.2) als $\mathbb{P}_{H_0}(T(\mathfrak{X}) \in \mathcal{R}) \leq \alpha$ und sagen, dass die Teststatistik unter der Nullhypothese mit Wahrscheinlichkeit höchstens α in den Ablehnungsbereich $\mathcal{R}$ fällt. Diese Schreib- und Sprechweise vererbt sich dann auf sämtliche Kenngrößen der Verteilung wie etwa den Erwartungswert $\mathbb{E}_{H_0}[\cdot]$ oder die Varianz $\mathbb{V}ar_{H_0}(\cdot)$. Analog schreiben wir $\overset{H_0}{\sim}$ für die Gleichheit in Verteilung, sowie $\xrightarrow{d_{H_0}}$ für die Verteilungskonvergenz.

Wir bemerken: Ist die Teststatistik unter H_0 stetig und unter allen mit H_0 assoziierten Verteilungen identisch verteilt, so können wir $\mathcal{R}$ so wählen, dass in (8.2) Gleichheit gilt. Ist beispielsweise der Bildraum der Statistik reellwertig, so könnte man $\mathcal{R} = (-\infty, q_{\alpha/2}] \cup [q_{1-\alpha/2}, \infty)$ wählen, wobei q_α das α-Quantil der Verteilung der Teststatistik unter H_0 bezeichnet. Denn bei dieser Wahl fällt die Teststatistik offenbar genau mit Wahrscheinlichkeit α in den Ablehnungsbereich $\mathcal{R}$, wenn H_0 stimmt.

Der P-Wert und ein- und zweiseitiges Testen Im Kontext von Hypothesentests ist die Wahl des Signifikanzniveaus α willkürlich. Daher wird in der Praxis häufig der sogenannte *P-Wert* bestimmt, der ohne die explizite Wahl von α auskommt. Der P-Wert ist eine Statistik. Intuitiv beschreibt er die Wahrscheinlichkeit, einen ‚mindestens so extremen' Wert der zugrunde liegenden Teststatistik $T(\mathfrak{X})$ zu beobachten wie $T(\mathbf{x})$, wenn die Nullhypothese stimmt. Ist diese Wahrscheinlichkeit klein, so interpretieren wir die Beobachtungen $\mathbf{x}$ als nur schwer mit der Nullhypothese verträglich. In Abb. 8.1 ist der P-Wert gegeben durch

die Summe sämtlicher Gewichte, die durch einen Kreis gekennzeichnet sind. Wir erkennen direkt, dass er kleiner ist als der α-Fehler von 5 %, was der Summe aller roten Gewichte entspricht.

Wir definieren den P-Wert für zwei Standardfälle. Diese Formulierungen hängen insbesondere von der konzeptionellen Wahl des Ablehnungsbereichs $\mathcal{R}$ ab, genauer davon, wo $\mathcal{R}$ innerhalb des Bildraums Γ der Teststatistik T positioniert ist. Das ergibt Sinn, denn $\mathcal{R}$ gibt ja an, ob die Auswertung $T(\mathbf{x})$ als extrem klassifiziert wird, und auch der P-Wert ist über die Denkweise der extremen Beobachtung motiviert.

Wir nehmen an, dass der Bildraum Γ der Teststatistik $T(\mathfrak{X})$ ein Intervall bildet. Der P-Wert wird über die Verteilung von $T(\mathfrak{X})$ formuliert, und dafür gehen wir davon aus, dass diese Verteilung unter allen mit der Nullhypothese H_0 assoziierten Kandidatenverteilungen gleich ist. Der P-Wert wird typischerweise in einer der beiden folgenden Weisen definiert:

1. Der Ablehnungsbereich $\mathcal{R}$ liegt an beiden Rändern von Γ, d. h., sowohl extrem kleine als auch extrem große Werte von $T(\mathbf{x})$ sprechen gegen die Nullhypothese. In diesem Fall sprechen wir von einem *zweiseitigen* Test, und der P-Wert ist definiert durch

$$P(\mathbf{x}) := \mathbb{P}_{H_0}(|T(\mathfrak{X}) - m_{H_0}| \geq |T(\mathbf{x}) - m_{H_0}|), \tag{8.3}$$

   wobei m_{H_0} den Median der Verteilung von $T(\mathfrak{X})$ unter H_0 bezeichne.
2. Der Ablehnungsbereich $\mathcal{R}$ liegt an genau einem der Ränder von Γ. In diesem Fall sprechen wir von einem *einseitigen* Test. Liegt $\mathcal{R}$ am linken Rand von Γ, so setzen wir

$$P(\mathbf{x}) := \mathbb{P}_{H_0}(T(\mathfrak{X}) \leq T(\mathbf{x})).$$

   Liegt $\mathcal{R}$ am rechten Rand von Γ, so definieren wir analog $P(\mathbf{x}) := \mathbb{P}_{H_0}(T(\mathfrak{X}) \geq T(\mathbf{x}))$.

In unserem obigen Beispiel haben wir es mit einem zweiseitigen Test zur einelementigen Nullhypothese $H_0 : p \in \{0.4\}$ zu tun, und da der Median $m_{H_0} = 0$ von $T(\mathfrak{X})$ unter H_0 verschwindet, schreibt sich der P-Wert als $p(\mathbf{x}) = \mathbb{P}_{0.4}(|T(\mathfrak{X})| \geq |T(\mathbf{x})|) \approx 0.002$.

Es ist $P(\mathbf{x}) \leq \alpha$ äquivalent dazu, dass $T(\mathbf{x}) \in \mathcal{R}$, und folglich lehnen wir H_0 genau dann ab, wenn der P-Wert kleiner als α ist. Wenn $P(\mathbf{x}) \leq \alpha$, so verwendet man häufig die Formulierung: Die beobachtete Diskrepanz zur Nullhypothese war zum Niveau α signifikant. Ist sogar $P(\mathbf{x}) \leq 0.01$, so spricht man auch von einer ‚hoch signifikanten' Diskrepanz. Ist $P(\mathbf{x}) > \alpha$, so sagt man analog, dass die Diskrepanz (auf dem $(\alpha \cdot 100)$ % Niveau) nicht signifikant war.

α- und β-Fehler und die Testmacht Im Kontext eines Hypothesentests, siehe Definition 8.1, nennen wir

$$\sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta(T(\mathfrak{X}) \in \mathcal{R})$$

den α-*Fehler* (oder auch Fehler erster Art). Er bezeichnet die Wahrscheinlichkeit, die Nullhypothese fälschlicherweise zu verwerfen. Nach Konstruktion des Tests ist der α-Fehler durch das Signifikanzniveau α beschränkt.

Für $\vartheta \in \Theta \backslash \Theta_0$ setzen wir

$$\beta(\vartheta) := \mathbb{P}_\vartheta(T(\mathfrak{X}) \notin \mathcal{R}).$$

Diese Größe, oft auch β-Fehler oder Fehler zweiter Art genannt, bezeichnet also die Wahrscheinlichkeit, dass die Nullhypothese fälschlicherweise *nicht* abgelehnt wird, obwohl die wahre Verteilung der Alternativhypothese zugeordnet ist. Die Größe $1 - \beta(\vartheta)$ heißt auch die *Testmacht* unter ϑ.

Asymptotische Tests In der Praxis kommt es häufig vor, dass man die Verteilung der Teststatistik unter der Nullhypothese nicht ausrechnen kann. Häufig lässt sie sich aber durch bekannte Verteilungen approximieren. Das führt dann zur Konstruktion *asymptotischer Tests*. Dazu betrachten wir ein Modell bestehend aus $\mathfrak{X}_\infty = (X_1, X_2, \ldots)^t$ und Verteilungsfamilie $(\nu_\vartheta)_{\vartheta \in \Theta}$. Für $n = 1, 2, \ldots$ betrachte man dann im zugehörigen Modell bestehend aus den ersten n Komponenten $(\mathfrak{X}_n = (X_1, \ldots, X_n)^t)$ eine Statistik T_n mit Bildraum Γ. Für einen asymptotischen Test ersetze man die Eigenschaft (8.2) durch

$$\lim_{n \to \infty} \mathbb{P}_{H_0}(T_n(\mathfrak{X}_n) \in \mathcal{R}) \leq \alpha. \tag{8.4}$$

Wir bemerken, dass weder der Ablehnungsbereich $\mathcal{R}$ noch das Signifikanzniveau α von n abhängen.

Hinsichtlich (8.4) denken wir häufig daran, dass die Folge von Statistiken unter H_0 in Verteilung konvergiert, $T_n(\mathfrak{X}_n) \xrightarrow{d_{H_0}} T_\infty$ für $n \to \infty$. In der Praxis wird man dann die auf den Beobachtungen basierende Teststatistik $T_n(\mathbf{x}_n)$ anhand der Verteilung von T_∞ beurteilen, anstatt der nicht handhabbaren Verteilung der Teststatistik $T_n(\mathfrak{X}_n)$.

Der Vergleich von $T_n(\mathbf{x}_n)$ mit der Grenzverteilung motiviert auch die Formulierung eines Analogons zum P-Wert: Für den zweiseitigen Test setzen wir

$$P_n(\mathbf{x}_n) := \mathbb{P}_{H_0}(|T_\infty - m_\infty| \geq |T_n(\mathbf{x}_n) - m_\infty|), \tag{8.5}$$

wobei m_∞ den Median der Verteilung von T_∞ bezeichne. Für die einseitigen Tests setzen wir entsprechend $P_n(\mathbf{x}_n) := \mathbb{P}_{H_0}(T_\infty \leq T_n(\mathbf{x}_n))$, bzw. $P_n(\mathbf{x}_n) := \mathbb{P}_{H_0}(T_\infty \geq T_n(\mathbf{x}_n))$. Wir verstehen dann $P_n(\mathbf{x}_n)$ als Approximation des P-Wertes aus (8.3) und nennen ihn der Einfachheit halber ebenfalls kurz den P-Wert.

In Abschn. 8.2 lernen wir einen solchen asymptotischen Tests kennen.

8.2 Einstichprobentest eines behaupteten Erwartungswerts

Wir diskutieren die Konzepte des statistischen Tests an einem zweiten Beispiel. Hier wird die Nullhypothese getestet, dass Beobachtungen aus einer Verteilung stammen, die einen behaupteten Erwartungswert besitzt.

Sie schauen zufällig ein Video, in dem es um Blauwale geht. Der Sprecher lässt verlauten, dass die mittlere Länge der Wale bei 20 m läge. Ihnen kommt das zu kurz vor, und Sie denken für einen Moment: ‚Ach, wie schön wäre es, die Weltmeere zu besegeln und echte Wale zu messen'. Sie verwerfen das ganz schnell wieder, weil Sie in den Mühlen der Klausuren stecken, aber führen immerhin ein Telefonat mit einer Freundin an einem Institut für Marine Biodiversitätsforschung, die Ihnen die Körperlängen $\mathbf{x}_n = (x_1, \ldots, x_n)^t$ von $n = 16$ Blauwalen durchgibt.

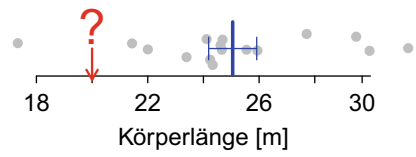
Sie stellen diese Beobachtungen grafisch dar, siehe Abb. 8.2, und stellen fest, dass die Werte typischerweise oberhalb des behaupteten Wertes 20 m liegen. Der Mittelwert $\bar{x}_n$ ist ebenfalls größer als 20 m. (Die Einheit m wird im Folgenden unterdrückt.) Sie fragen sich: Ist die Abweichung $|\bar{x}_n - 20|$ ‚leicht' durch Zufall zu erklären, oder geben mir die Beobachtungen Anlass, an dem behaupteten Wert 20 zu zweifeln? Sie führen einen statistischen Test durch.

1. *Wahl eines statistisches Modells:* Es seien $X_1, X_2, \ldots$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim \nu_\vartheta$, und ν_ϑ ist Mitglied der Familie $(\nu_\vartheta)_{\vartheta \in \Theta}$ der quadratintegrierbaren Verteilungen mit positiver Varianz. Für $n = 1, 2, \ldots$ sei ein statistisches Modell beschrieben durch den Vektor der ersten n Zufallsvariablen $\mathfrak{X}_n = (X_1, \ldots, X_n)^t$.
2. *Formulierung der Nullhypothese:* Sei $\Theta_0 = \{\vartheta \in \Theta \mid \mathbb{E}_\vartheta[X_1] = 20\}$, also

$$H_0 : \quad \vartheta \in \Theta_0.$$

   Die Nullhypothese ist also mit der Teilfamilie $(\nu_\vartheta)_{\vartheta \in \Theta_0}$ assoziiert, deren Mitglieder einen Erwartungswert von X_1 von 20 haben.
3. *Wahl einer Teststatistik:* Wieder betrachten Sie die Abweichung vom Mittelwert $|\bar{x}_n - 20|$. Das ist sinnvoll, da diese Statistik sensitiv gegenüber einer Abweichung von der Nullhypothese ist: Stammen die Beobachtungen aus einer Verteilung mit Erwartungswert 20, so ist die Abweichung typischerweise betragsmäßig klein und hat Erwartungswert null. Ist andererseits der wahre Erwartungswert zum Beispiel 30, so ist die Statistik in Erwartung von null verschieden, denn wir zentrieren ja mit 20.

Abb. 8.2 Einstichprobentest



Allerdings müssen Sie für die Konstruktion eines Tests auch die Verteilung der Statistik $|\bar{X}_n - 20|$ unter der Nullhypothese kontrollieren, um den Ablehnungsbereich zu wählen, siehe (8.2). Das Problem ist, dass die Verteilung dieser Statistik von der Varianz der Verteilung von X_1 abhängt und daher nicht für sämtliche Verteilungen der Nullhypothese gleich ist. Sie wählen daher lieber folgende standardisierte Statistik

$$T_n(\mathbf{x}_n) := \frac{\bar{x}_n - 20}{s_n(\mathbf{x}_n)/\sqrt{n}},$$

wobei $s_n(\mathbf{x}_n)$ die empirische korrigierte Stichprobenstandardabweichung aus (3.3) bezeichnet. Sie wissen: Unter H_0 gilt für $n \to \infty$

$$T_n(\mathfrak{X}_n) := \frac{\bar{X}_n - 20}{s_n(\mathfrak{X}_n)/\sqrt{n}} \xrightarrow{d_{H_0}} T_\infty \sim N(0,1), \tag{8.6}$$

und Sie konstruieren einen asymptotischen Test. Es sei bemerkt, dass die Konvergenz unter allen Verteilungen aus $(\nu_\vartheta)_{\vartheta\in\Theta_0}$ gilt. Das folgt aus dem Zentralen Grenzwertsatz 2.11 und dem Satz von Slutsky 2.12.

4. *Wahl eines Signifikanzniveaus und eines Ablehnungsbereichs:* Sie wählen ein Signifikanzniveau $\alpha \in (0,1)$. Aus der obigen Überlegung, dass betragsmäßig große Werte von T gegen H_0 sprechen, konstruieren Sie den Ablehnungsbereich

$$\mathcal{R} = (-\infty, -q_{1-\alpha/2}] \cup [q_{1-\alpha/2}, \infty),$$

wobei $q_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der Standardnormalverteilung bezeichnet. Nach (8.6) ist damit die Forderung des asymptotischen Tests (8.4) erfüllt, denn es gilt $\lim_{n\to\infty} \mathbb{P}(T_n(\mathfrak{X}_n) \in \mathcal{R}) = \mathbb{P}(T_\infty \in \mathcal{R}) = \alpha$.

5. *Auswertung der Beobachtungen*: Die Längen Ihrer $n = 16$ Blauwale haben einen Mittelwert von $\bar{x}_n \approx 25\,\text{m}$ und eine Standardabweichung von $s_n(\mathbf{x}_n) \approx 3.5\,\text{m}$. Damit ist der Standardfehler des Mittelwertes, d. h. der Nenner von $T_n(\mathbf{x}_n)$, etwa $sem(\mathbf{x}_n) = s(\mathbf{x}_n)/\sqrt{n} \approx 0.9\,\text{m}$, und mit dem behaupteten Populationsmittelwert von $\mu^{(0)} = 20\,\text{m}$ ist $T_n(\mathbf{x}_n) \approx 5.6$ (Abb. 8.3).
Die Statistik fällt also in den Ablehnungsbereich $\mathcal{R}$ zum Niveau $\alpha = 0.05$, und Sie lehnen die Nullhypothese ab. Sie können sagen: Die beobachtete Abweichung des Mittelwertes $\bar{x}_n$ von 20 war auf dem 5 %-Niveau signifikant. Interpretation: Wenn die Nullhypothese stimmt, ist etwas Unwahrscheinliches eingetreten, das nur in etwa 5 % der Fälle durch Zufall passieren würde.
Zur Berechnung des P-Wertes bemerken wir, dass der Median von T_∞ verschwindet, sodass

$$P_n(\mathbf{x}_n) := \mathbb{P}_{H_0}(|T_\infty| \geq |T_n(\mathbf{x}_n)|) \approx \mathbb{P}_{H_0}(|T_\infty| \geq 5.6) < 10^{-7}.$$

Dieser P-Wert ist winzig und sagt Ihnen: Unter der Nullhypothese tritt ein solch extremes Ereignis durch Zufall in weniger als einem von zehn Millionen Fällen auf! Für jedes noch

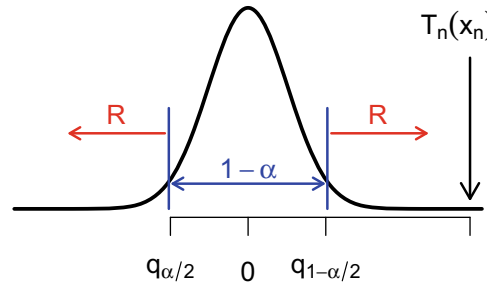


Abb. 8.3 Ablehnungsbereiche bei normalverteilter Teststatistik

so kleine α, das größer ist als 10^{-7}, würde die Nullhypothese abgelehnt! In diesem Sinne passen die Beobachtungen also sehr schlecht zur Nullhypothese und geben im gegebenen Modell durchaus Anlass, an der Nullhypothese zu zweifeln.

Im Rahmen dieses Beispiels fällt auch die Äquivalenz zum Konfidenzintervall aus Kap. 6 auf: Bei obigem Test wird die Nullhypothese, dass der wahre Erwartungswert gleich 20 ist, genau dann auf dem α-Niveau abgelehnt, wenn das asymptotische Konfidenzintervall für den Erwartungswert $\mathbb{E}_\vartheta[X_1]$

$$I_n(\mathbf{x}_n) := \left[\bar{x}_n - q_{\frac{\alpha}{2}} \cdot \frac{s_n(\mathbf{x}_n)}{\sqrt{n}}, \bar{x}_n + q_{1-\frac{\alpha}{2}} \cdot \frac{s_n(\mathbf{x}_n)}{\sqrt{n}}\right]$$

die 20 nicht überdeckt. Insbesondere ist auch die Interpretation die gleiche!

Wir fassen obigen asymptotischen, *nichtparametrischen Einstichprobentest* eines behaupteten Erwartungswerts in folgendem Lemma zusammen.

Lemma 8.2 (Nichtparametrischer Einstichprobentest)
Es seien $X_1, X_2, \ldots$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim \nu_\vartheta$, und ν_ϑ sei Mitglied der Familie $(\nu_\vartheta)_{\vartheta\in\Theta}$ der quadratintegrierbaren Verteilungen mit positiver Varianz. Für $n = 1, 2, \ldots$ ist dann ein statistisches Modell assoziiert durch den Vektor $\mathfrak{X}_n = (X_1, \ldots, X_n)^t$ der ersten n Zufallsvariablen. Es sei s_n wie in (3.3), und weiter sei $\alpha \in (0, 1)$ und q_α bezeichne das α-Quantil der $N(0, 1)$-Verteilung. Dann gilt:

1. *Asymptotisches Konfidenzintervall: Eine Folge $(I_n)_{n=1,2,\ldots}$ gegeben durch*

$$I_n(\mathbf{x}_n) := \left[\bar{x}_n - q_{\frac{\alpha}{2}} \cdot \frac{s_n(\mathbf{x}_n)}{\sqrt{n}}, \bar{x}_n + q_{1-\frac{\alpha}{2}} \cdot \frac{s_n(\mathbf{x}_n)}{\sqrt{n}}\right]$$

liefert ein asymptotisches $(1 - \alpha)$-Konfidenzintervall für $\mathbb{E}_\vartheta[X_1]$.

2. *Asymptotischer Einstichprobentest: Für* $\mu^{(0)} \in \mathbb{R}$ *sei eine Nullhypothese formuliert als*

$$H_0 : \vartheta \in \{\vartheta \in \Theta \,|\, \mathbb{E}_\vartheta[X_1] = \mu^{(0)}\}.$$

Dann bildet die Folge $(T_n)_{n=1,2,\ldots}$ *gegeben durch*

$$T_n(\mathbf{x}_n) := \frac{\bar{x}_n - \mu^{(0)}}{s_n(\mathbf{x}_n)/\sqrt{n}}$$

eine Folge von Teststatistiken für einen asymptotischen Test der Nullhypothese H_0 *zum Niveau* α *mit Ablehnungsbereich* $\mathcal{R} = (-\infty, q_{\alpha/2}] \cup [q_{1-\alpha/2}, \infty)$.

Beliebte Fehlinterpretationen Wie beim Konfidenzintervall gibt es wieder eine Reihe typischer Fehlinterpretationen und ungünstiger Sprechweisen.

1. *Nicht* folgern können wir ungeachtet des Ausgangs des Tests: Die Nullhypothese ist wahr/falsch/wahrscheinlich wahr/wahrscheinlich falsch etc., denn wir machen nur Aussagen über die Wahrscheinlichkeit der Beobachtung x unter der theoretischen Annahme, dass die Nullhypothese wahr ist!
2. Eine *falsche Sprechweise* ist: ‚Die Nullhypothese war signifikant'. Nullhypothesen sind theoretische Annahmen. Signifikant oder nicht signifikant sind die beobachteten *Abweichungen* von der Nullhypothese.
3. Eine *ungünstige Sprechweise* bei $T(\mathbf{x}_n) \notin \mathcal{R}$: ‚Die Nullhypothese wird angenommen'. Besser ist: Die Nullhypothese wird *nicht abgelehnt.* Grund: Die Nullhypothese sagt, die genaue mittlere Länge der Wale in der Population sei 20m, anders formuliert $20.00000000000000000000\ldots$m. Das ist von vorne herein eine absurde Behauptung. Die Nullhypothese ist im Falle von einelementigen Θ_0 praktisch immer falsch, und die vernünftige Frage ist eigentlich, ob die Beobachtungen einerseits mit der Nullhypothese einigermaßen verträglich sind oder andererseits genug Anlass geben, an der Nullhypothese zu zweifeln.
4. Wichtig ist auch, dass wir das α-Niveau *vor* der Durchführung des Tests und vor der Betrachtung der Beobachtungen festlegen sollten. Nach der Berechnung eines P-Wertes von zum Beispiel $P_n(\mathbf{x}_n) = 0.058$ das Signifikanzniveau auf $\alpha = 0.059$ zu setzen, um ablehnen zu dürfen, ist Selbstbetrug! Umgekehrt muss man aber auch bedenken, dass das festgesetzte α willkürlich gewählt wurde. Betrachtet man etwa einen statistischen Test, der zu einem P-Wert führt, der nur leicht kleiner als das vorgegebene α ist, so führt dies zur Ablehnung der Nullhypothese. Man interpretiert dann, dass die Diskrepanz der Beobachtungen und der Nullhypothese ‚groß' war. Wäre andererseits diese Diskrepanz nur minimal kleiner ausgefallen, in dem Sinne, dass sie in einem P-Wert leicht größer als

α resultiert hätte, so hätte man die Nullhypothese nicht ablehnen können. Bei Betrachtung dieser Grenzfälle erscheint die Konzeption eines statistischen Tests besonders brenzlig: Die P-Werte sind fast gleich, aber die Entscheidung ist eine andere. Daher ist es immer ein Zugewinn, zusätzlich den P-Wert im Auge zu behalten. So führen beispielsweise die P-Werte von 0.49 und 0.001 beide zur Ablehnung der Nullhypothese auf dem 5 % Niveau, aber die in der Stichprobe beobachtete Diskrepanz von der Nullhypothese ist im zweiten Fall sehr viel schwerer durch Zufall zu erklären, falls die Nullhypothese stimmt.

8.3 Dialog: Interpretation von Testergebnissen

Eine engagierte Studentin der Statistik (**S**) möchte ihrer Freundin aus der Medizin (**M**) bei der Analyse der Daten ihrer Doktorarbeit helfen. Die Medizinerin soll darin ein neues Behandlungsverfahren beurteilen, das an ihrer Klinik entwickelt wurde. Dazu misst sie u. a. den Anteil an behandelten Personen, bei denen innerhalb von zwei Tagen nach der Behandlung Kopfschmerzen auftraten. Bei dem älteren Standardverfahren, das schon Tausende Male angewandt wurde, lag dieser Anteil bei 50 %. Bei den mittlerweile 100 Patienten, die mit dem neuen Verfahren behandelt wurden, waren es dagegen nur 41 %. Die Medizinerin möchte nun gerne wissen, wie aussagekräftig ihre Ergebnisse sind.

M: Schau mal, mit dem neuen Verfahren bekommen viel weniger Patienten Kopfschmerzen! Aber mein Betreuer sagt, das können wir erst publizieren, wenn das auch ‚statistisch signifikant' ist. Kannst du mir dabei helfen, das herauszufinden?
S: Na klar! Deine Abweichung von den 50 % könnte ja im Prinzip bei diesen 100 Patienten auch durch Zufall zustande gekommen sein, selbst wenn in Wahrheit bei eurer Behandlung ebenfalls jeder Zweite Kopfschmerzen bekommt. Daher müssen wir die Nullhypothese testen, dass die Wahrscheinlichkeit für Kopfschmerzen beim neuen Verfahren auch bei 50 % liegt.
Die Statistikstudentin führt den Test analog zu Abschn. 8.1 zum Niveau $\alpha = 0.05$ durch mit auf den Beobachtungen $\mathbf{x}$ berechneter Teststatistik $T(\mathbf{x}) = 0.41 - 0.5 = -0.09$ und findet einen Ablehnungsbereich von $\mathcal{R} = \{-0.5, \ldots, -0.11\} \cup \{0.11, \ldots, 0.5\}$. Die Nullhypothese kann daher auf dem 5 %-Niveau nicht abgelehnt werden. Der P-Wert liegt bei etwa 0.09.
Die Medizinstudentin ist von diesem Ergebnis natürlich enttäuscht. Für ihre Doktorarbeit entwirft sie folgenden Satz.
M: Schau mal, kann ich das dann so schreiben? ‚Der Anteil an Patienten mit behandlungsbedingten Kopfschmerzen wird durch das neue Verfahren mit einer Wahrscheinlichkeit von 95 % nicht gesenkt.'
Die Statistikstudentin ist sich da nicht ganz sicher.

S: Ich weiß nicht. Das klingt irgendwie komisch, finde ich. Unser Dozent hat immer wieder vor falschen Formulierungen gewarnt. Das habe ich aber nicht ganz verstanden, da fragen wir lieber nochmal.
Beide besuchen also zusammen besagten Dozenten (**D**) und bitten ihn um Rat. Die Statistikstudentin erklärt:
S: Ich habe das Beispiel aus der Kopfschmerzstudie durchgerechnet, und das Ergebnis ist nicht signifikant. Bei der Rechnung bin ich mir ziemlich sicher, aber das mit den Formulierungen habe ich irgendwie noch nicht richtig verstanden, ich dachte, da frage ich lieber nochmal...
D: Sehr gut, dass Sie da nachfragen! Es ist nämlich sehr wichtig, genau auf die Formulierung zu achten. Sogar in Veröffentlichungen liest man leider viel zu oft falsche Formulierungen.
Der Dozent wendet sich an die Medizinerin:
D: Es ist natürlich verständlich: Sie möchten *wissen,* ob das neue Verfahren häufiger, seltener oder gleich oft zu Kopfschmerzen führt als das alte. Aber es ist ganz wichtig zu verstehen: Dummerweise kann ein statistischer Test diese Frage gar nicht beantworten!!
M: Wie bitte? Aber wozu benutzt man ihn denn dann überhaupt?
D: Wir können damit nur beurteilen, wie gut Daten zu einer Hypothese passen. Mehr können wir nicht.
An die Statistikstudentin gewandt fährt er fort:
S: Schauen Sie nochmal in das Kapitel zu den Konfidenzintervallen, das geht ganz genauso. Das Ergebnis Ihres Tests besagt, dass die Diskrepanz der Daten von der Nullhypothese statistisch nicht signifikant war. Und das bedeutet, dass die Daten nicht sehr deutlich gegen die Hypothese sprechen – die Abweichung von 41 % zu 50 % ist nicht überraschend groß, falls die Nullhypothese stimmt.
Die Medizinstudentin versucht, das Gehörte auf ihre Arbeit anzuwenden.
M: Okay, das verstehe ich so weit. Kann ich denn dann in meiner Arbeit schreiben: ‚Der Anteil an Patienten mit behandlungsbedingten Kopfschmerzen wird durch das neue Verfahren mit einer Wahrscheinlichkeit von 95 % nicht gesenkt.‘?
D: Nein, das ist leider falsch. Wir können nämlich leider gar keine Aussagen über die Hypothesen machen.
S: Aber zumindest Wahrscheinlichkeitsaussagen können wir doch über die Hypothesen treffen, oder?
D: Leider auch nicht. Denn es sind ja die Beobachtungen, die wir als zufällig verstehen. Wir können also nur Wahrscheinlichkeitsaussagen über die Beobachtungen machen. Wir fragen uns immer, wie wahrscheinlich eine mindestens so große Diskrepanz zwischen den Beobachtungen und der Nullhypothese ist, wenn die Nullhypothese stimmt. Die Hypothesen sind dabei theoretische Annahmen und haben keine Wahrscheinlichkeiten.
M: Aha, aber was schreibe ich denn dann in meinem Fall?
D: Zum Beispiel könnten Sie schreiben: ‚Bei 41 % der 100 Probanden traten innerhalb von zwei Tagen Kopfschmerzen auf‘ – diese Formulierung beschreibt zunächst die

Beobachtungen prägnant und bezieht sowohl das Ergebnis (die guten 41 %) als auch die gesamte Stichprobengröße von 100 mit ein. Zudem vermeiden Sie das Wort ‚behandlungsbedingt', da man ja die Ursache der Kopfschmerzen nicht genau kennt. Dann fehlt noch die Interpretation des Tests, die etwa so lauten könnte: ‚Dieser Anteil war statistisch nicht signifikant von 50 % verschieden ($P(\mathbf{x}) = 0.09$).' Neben der Aussage über die Signifikanz sieht man zusätzlich, dass der P-Wert nicht weit von 0.05 entfernt lag. Die beobachtete Diskrepanz war also zwar nicht unwahrscheinlich *genug,* um auf dem 5 %-Niveau signifikant zu sein, aber sie träte trotzdem durch Zufall seltener als in einem von zehn Versuchen auf, falls tatsächlich auch beim neuen Verfahren jeder Zweite Kopfschmerzen bekommt.

Die Statistikstudentin ist mit dieser Erklärung zufrieden. Sie hat aber noch eine andere Idee.

S: Okay, das gefällt mir. Da fällt mir jetzt aber noch etwas ein: Gab es da nicht eine Möglichkeit, den P-Wert zu verkleinern, indem man nur ‚einseitig' testet? Wird der Unterschied dann vielleicht signifikant?

D: Vermutlich ist das hier nicht erlaubt, aber es ist trotzdem eine gute Frage. Schauen wir uns das doch mal genauer an: Ihr bisheriger Test untersuchte die Nullhypothese, dass der wahre Anteil genau 50 % ist. Extreme Abweichungen wurden durch starke Abweichungen des beobachteten Anteils von 50 % quantifiziert. Sowohl höhere als auch niedrigere Kopfschmerzquoten gelten als ‚extrem' und können zur Ablehnung der Nullhypothese führen – der Ablehnungsbereich teilt sich auf beide Seiten der Verteilung auf. Dies war ein sogenannter zweiseitiger Test – das übliche Verfahren.

S: Okay, d. h., extreme Werte, egal ob positiv oder negativ, führen zur Ablehnung der Nullhypothese.

D: Ganz genau. Nun könnte ein etwas praxisferner Mensch auf die Idee kommen, den Ablehnungsbereich nicht in die Ränder der Verteilung der Teststatistik zu legen, sondern zum Beispiel in die Mitte. Wählt er dort einen Bereich, der unter der Nullhypothese mit Wahrscheinlichkeit 5 % getroffen wird, so hat er tatsächlich einen Test zum Niveau 5 % konstruiert.

S: Das ergibt aber doch keinen Sinn, oder?

D: Stimmt, in der Regel nicht, denn Ihre Teststatistik ist ja gerade so konstruiert, dass extrem große oder extrem kleine Werte gegen die Nullhypothese sprechen. Damit hätten Sie dann, wenn die Nullhypothese nicht zutrifft, nur geringe Testmacht.

S: Also legt man den Ablehnungsbereich immer in beide Ränder des Wertebereichs der Teststatistik?

D: Meistens, aber nicht immer. Es kommt darauf an, welche Werte der Teststatistik als mit der Nullhypothese unverträglich gelten sollen. Wir könnten beispielsweise die Diskrepanz nicht mit T, sondern mit T^2 beschreiben. Wenn für T extrem negative und extrem positive Werte gegen die Nullhypothese sprechen, dann sind es bei T^2 nur noch extrem positive Werte. Der Ablehnungsbereich sollte dann also nur extrem positive Werte enthalten.

Die Statistikstudentin findet das nicht ganz überzeugend.
S: Klar, das ist aber ein etwas künstliches Beispiel, oder?
D: Stimmt, aber auch für T selbst kann es in seltenen Fällen sinnvoll sein, den Ablehnungsbereich nur auf einer Seite zu wählen, zum Beispiel nur in der linken Flanke der Verteilung bei den kleineren Anteilen. Wir sprechen dann auch manchmal von einem einseitigen Test. Um dann wieder mit Wahrscheinlichkeit höchstens 5 % abzulehnen, kann man den linken Ablehnungsbereich größer wählen als beim zweiseitigen Test. In Ihrem Fall hätten Sie beim einseitigen Testen tatsächlich einen P-Wert von 0.044 und könnten die Nullhypothese ablehnen.
M: Wunderbar, dann machen wir es doch so!
Doch der Dozent muss ihre Freude leider bremsen.
D: Langsam... bei der Wahl zwischen ein- und zweiseitigem Test müssen wir uns zuerst ein paar unangenehmen Fragen stellen. Was würden Sie zum Beispiel tun, wenn Sie im Gegensatz zu Ihrer verringerten Quote an Kopfschmerzpatienten eine erhöhte Quote beobachtet hätten, sagen wir 65 %? Das wäre beim zweiseitigen Test signifikant.
M: Dann wäre es auch gut – natürlich nicht gut für unser neues Verfahren, aber das hat ja vor allem andere Vorteile. Die Kopfschmerzen sind ja nur ein Nebeneffekt und gehen schnell wieder weg. Aber dieses Resultat müssten wir natürlich auch als signifikant publizieren.
D: Dann ist es so wie ich befürchtet hatte: Sie beurteilen sowohl extrem kleine als auch extrem große Werte der Teststatistik als nicht mit der Nullhypothese verträglich. Dann dürfen Sie leider nicht einseitig testen.
M: Stimmt, irgendwie wäre das geschummelt. Und wann hätte ich einseitig testen dürfen?
D: Wenn Sie zum Beispiel das neue Verfahren nur dann etablieren könnten, wenn es Anzeichen dafür gäbe, dass die Kopfschmerzquote gesunken ist. In diesem Fall gibt es für Sie zwei mögliche Ausgänge: Entweder Ihre Kopfschmerzquote ist überraschend gering, dann würden Sie die Nullhypothese ablehnen und Ihr Verfahren etablieren. Oder sie ist etwa gleich oder sogar höher. In beiden letzteren Fällen würden Sie nicht ablehnen und Ihr neues Verfahren verwerfen.
M: Also immer wenn ich auf der anderen Seite der Verteilung nicht ablehnen würde...
D: ... egal zu welch winzigem Niveau α es auch immer möglich wäre...
M: ... dann darf ich einseitig testen!
D: Ganz genau!