Kapitel 4

Statistische Tests

4.1 Grundbegriffe

Wir betrachten wieder ein parametrisches Modell $\{P_\theta : \theta \in \Theta\}$ und eine zugehörige Zufallsstichprobe $X_1, \ldots, X_n$. Wir wollen nun die Beobachtung der $X_1, \ldots, X_n$ verwenden, um bestimmte Aussagen über die zugrundeliegende Verteilung (genauer: den zugrundeliegenden Parameter θ) zu testen. Wir verwenden folgende Terminologie:

Eine *Hypothese* ist eine Aussage über den Parameter θ. Modelliert wird eine Hypothese durch eine Teilmenge Θ_0 von Θ. Man sagt, die Hypothese *trifft zu*, falls $\theta \in \Theta_0$. Eine Hypothese heißt *einfach*, falls Θ_0 einelementig ist, ansonsten sagt man, die Hypothese ist *zusammengesetzt*. Die Negation der Hypothese, also die Aussage $\theta \notin \Theta_0$ – äquivalent, die Aussage $\theta \in \Theta_1 := \Theta \setminus \Theta_0$ – heißt *Alternative*. Manchmal nennt man die Hypothese auch *Nullhypothese* (und schreibt $H_0 : \theta \in \Theta_0$) und die Alternative *Alternativhypothese* und schreibt $H_1 : \theta \in \Theta_1$).

Ein *Test (der Hypothese $\theta \in \Theta_0$ gegen die Alternative $\theta \in \Theta_1$)* ist eine Entscheidungsregel, die für jede Realisierung $x_1, \ldots, x_n$ der Stichprobe $X_1, \ldots, X_n$ festlegt, ob die Hypothese oder die Alternative gewählt wird. Man hat also eine disjunkte Zerlegung des Stichprobenraumes M^n in Teilmengen K_0 und K_1, sodass wir uns für die Hypothese entscheiden, wenn $(X_1, \ldots, X_n) \in K_0$ (wir sagen "*die Hypothese wird akzeptiert*") und wir uns für die Alternative entscheiden, wenn $(X_1, \ldots, X_n) \in K_1$ (wir sagen "*die Hypothese wird verworfen*").

K_0 heißt auch *Annahmebereich*, K_1 *kritischer Bereich*.

Beispiel 4.1.1. Wir betrachten erneut das Werfen einer Reißzwecke und möchten untersuchen, ob es gleich wahrscheinlich ist, auf der flachen Seite oder mit der Spitze schräg nach unten liegen zu bleiben. Als parametrisches Modell betrachten wir $X \sim Bin(1, p)$, wobei wir $X = 1$ als "flache Seite" und $X = 0$ als "mit der Spitze schräg nach unten" interpretieren. Wir wollen die Hypothese $p = \frac{1}{2}$ (dies ist eine einfache Hypothese) gegen die Alternative $p \neq \frac{1}{2}$ testen.

Wir beobachten Zufallsvariablen $X_1, \ldots, X_{1000} \sim Bin(1, p)$ und bilden den Mittelwert $\bar{X}$.

Ein möglicher Test akzeptiert die Hypothese, wenn $\bar{X} = \frac{1}{2}$ ist, und verwirft sie sonst (Test T_1).

Eine andere Möglichkeit wäre es, die Hypothese zu akzeptieren, wenn $\bar{X} \in [0.47, 0.53]$ liegt, und sie sonst zu verwerfen (Test T_2).

Schließlich wäre es auch möglich, die Hypothese immer zu akzeptieren (Test T_3).

Offensichtlich sind nicht alle Tests in obigem Beispiel gleich gut. Bei Test T_1 ist das Problem, dass selbst wenn der wahre Parameter $p = \frac{1}{2}$ ist, nicht notwendigerweise $\bar{X} = \frac{1}{2}$ sein muss. Genauer gesagt besitzt dieses Ereignis (sofern $p = \frac{1}{2}$) gerade Wahrscheinlichkeit $\binom{1000}{500}\frac{1}{2}^{500}\frac{1}{2}^{500} \approx 0,025$. Beachte, dass falls der Stichprobenumfang n ungerade ist, $\bar{X} = \frac{1}{2}$ Wahrscheinlichkeit 0 besitzt. Offensichtlich ist, wenn die Hypothese zutrifft, die Wahrscheinlichkeit, dass der Test die Hypothese akzeptiert, bei Test T_2 größer. Bei Test T_3 ist sie sogar noch größer. Allerdings irrt Test T_3 immer, wenn die Alternative richtig gewesen wäre.

Um diese Phänomene genauer zu untersuchen, führen wir folgende Begriffe ein.

Definition 4.1.2. Verwirft ein Test die Hypothese, obwohl sie richtig gewesen wäre, so sagt man, es liegt ein *Fehler erster Art* vor. Akzeptiert ein Test die Hypothese, obwohl die Alternative richtig gewesen wäre, so sagt man, es liegt ein *Fehler zweiter Art* vor.

Hat man nun einen Test mit Annahmebereich K_0 und kritischem Bereich K_1 gegeben, so heißt

$$G(\theta) := \mathbb{P}_\theta((X_1, \ldots, X_n) \in K_1)$$

die *Gütefunktion* des Tests. Die Zahl

$$\alpha := \sup\{G(\theta) : \theta \in \Theta_0\}$$

heißt *Umfang* des Tests. Ist zudem $G(\theta) \geq \alpha$ für $\theta \in \Theta_1$, so sagt man der Test sei *unverfälscht*.

Interpretation: Die Gütefunktion gibt an, mit welcher Wahrscheinlichkeit die Hypothese verworfen wird, wenn der wahre Parameter θ ist. Für $\theta \in \Theta_0$ ist also $G(\theta)$ gerade die Wahrscheinlichkeit, einen Fehler erster Art zu begehen. Bei einem Test zum Umfang α ist also die Wahrscheinlichkeit einen Fehler erster Art zu begehen höchstens α.

Beispiel 4.1.3. Wir betrachten wiederum die Tests T_3 und T_2 aus Beispiel 4.1.1. Es sei G_j die Gütefunktion des Tests T_j.

Die Gütefunktion von T_3 gegeben durch $G_3(p) \equiv 0$.

Für den Test T_2 ist die Gütefunktion gegeben durch

$$G_2(p) = 1 - \sum_{k=470}^{530} \binom{1000}{k} p^k (1-p)^{1000-k}.$$

Natürlich ist es wiederum schwer, diese Funktion explizit auszurechnen. Sie kann mit Hilfe des Zentralen Grenzwertsatzes approximiert werden.

Wir bemerken, dass Fehler erster Art und Fehler zweiter Art unterschiedlich behandelt werden. Bei einem Test zum Signifikanzniveau α ist die Wahrscheinlichkeit einen Fehler erster Art zu begehen begrenzt. Allerdings gibt es keine Einschänkungen hinsichtlich des Fehlers zweiter Art. In Anwendungen ist es jedoch häufig der Fall, dass ein Fehler schwerwiegender ist als der andere.

Man denke etwa daran, ein neues Medikament auf Nebenwirkungen zu testen: Es ist wesentlich schwerwiegender vorhandene Nebenwirkungen nicht zu entdecken als falschen Alarm zu schlagen (und nichtvorhandene Nebenwirkungen zu erkennen).

Diese Asymmetrie sollte man bei der Wahl von Hypothese und Alternative beachten: die Hypothese sollte so gewählt werden, dass der schwerwiegendere Fehler der Fehler erster Art ist. Bei den Medikamenten sollte man also die Hypothese "Das Medikament hat Nebenwirkungen" gegen die Alternative "Das Medikament hat keine Nebenwirkungen" testen.

Problem: In der Regel existieren nur für eine der beiden Wahlen Tests.

4.2 Tests für den Erwartungswert einer Normalverteilung

Wir betrachten nun einige "Standardtests", die den Erwartungswert μ einer Normalverteilung betreffen. Man unterscheidet hierbei *einseitige Tests*, bei denen die Hypothese $H_0 : \mu \leq \mu_0$ gegen die Alternative $H_1 : \mu > \mu_0$ (resp. die Hypothese $H_0 : \mu \geq \mu_0$ gegen die Alternative $H_1 : \mu < \mu_0$) getestet wird und *zweiseitige Tests*, bei denen die Hypothese $H_0 : \mu = \mu_0$ gegen die Alternative $H_1 : \mu \neq \mu_0$ getestet wird.

Eine weitere Unterscheidung ist, ob die Varianz bekannt oder unbekannt ist.

In allen hier diskutierten Tests verwenden wir eine sogenannte *Teststatistik* T und entscheiden uns für oder gegen die Hypothese, abhängig davon, wo T liegt.

Als Teststatistik treten $T := \sqrt{n}\frac{\bar{X}-\mu_0}{\sigma_0}$ bei bekannter Varianz σ_0^2 und $T := \sqrt{n}\frac{\bar{X}-\mu_0}{s}$ bei unbekannter Varianz auf. Ist der wahre Parameter μ_0 (also unter Nullhypothese beim zweiseitigen Test), so ist $T \sim \mathcal{N}(0,1)$ resp. $T \sim t_{n-1}$.

Tests bei bekannter Varianz

Gegeben eine Stichprobe $X_1, \ldots, X_n$ zur Normalverteilung $\mathcal{N}(\mu, \sigma_0^2)$ bei *bekannter Varianz* σ_0^2, wollen wir Tests zu vorgegebenem Umfang $\alpha \in (0,1)$ konstruieren. Wir betrachten zunächst den zweiseitigen Test für $H_0 : \mu = \mu_0$ gegen die Alternative $H_1 : \mu \neq \mu_0$.

Wir akzeptieren H_0, wenn $|T| \leq c$ für ein geeignetes $c > 0$ ist, und lehnen H_0 sonst ab.

Damit der Test Umfang α hat, muss $\mathbb{P}_{\mu_0}(|T| > c) = \alpha$ gelten. Unter der Nullhypothese $\mu = \mu_0$ ist $T \sim \mathcal{N}(0,1)$. Daher wählen wir $c := \Phi^{-1}(1-\alpha/2)$.

Beispiel 4.2.1. Eine Maschine füllt 200 g Packungen mit Müsli ab. Aus Erfahrungen ist bekannt, dass das tatsächliche Gewicht, das von der Maschine abgefüllt wird, normalverteilt mit Varianz 4 ist. Um zu überprüfen, ob die Maschine korrekt eingestellt ist, werden 10 Packungen nachgewogen, was ein Durchschnittsgewicht von 197 g ergibt. Es soll nun zum Signifikanzniveau von 5% getestet werden, ob $\mu = 200$ plausibel ist.

Lösung: Es ist $\Phi^{-1}(0,975) = 1,96$. Für die Teststatistik T ergibt sich $T = \sqrt{10}\frac{197-200}{2} = -4,743$. Die Hypothese wird also verworfen; die Maschine ist nicht richtig eingestellt.

Nun betrachten wir einen einseitigen Test, bei dem wir $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$ testen (Um $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$ zu testen, geht man analog vor).

Wir betrachten wieder $T = \sqrt{n}\frac{\bar{X}-\mu_0}{\sigma_0}$ und akzeptieren H_0 falls $T \leq c$ für ein (neues) geeignetes c. Dieses ist so zu wählen, dass

$$\sup\{G(\mu) \mid \mu \leq \mu_0\} = \alpha.$$

Wir berechnen die Gütefunktion in Abhängigkeit von c:

$$\begin{aligned} G(\mu) = \mathbb{P}_\mu((X_1,\ldots,X_n) \in K_1) = \mathbb{P}_\mu\big(\sqrt{n}\tfrac{\bar{X}-\mu_0}{\sigma_0} > c\big) &= \mathbb{P}_\mu\big(\sqrt{n}\tfrac{\bar{X}-\mu}{\sigma_0} > c + \sqrt{n}\tfrac{\mu_0-\mu}{\sigma_0}\big) \\ &= 1 - \Phi\big(c + \sqrt{n}\tfrac{\mu_0-\mu}{\sigma_0}\big) \end{aligned}$$

Wir stellen fest, dass die Gütefunktion monoton wachsend ist. Somit

$$\alpha = \sup\{G(\mu) \mid \mu \leq \mu_0\} = G(\mu_0) = 1 - \Phi(c)$$

und wir müssen $c = \Phi^{-1}(1-\alpha)$ wählen. Weiter ist der Test unverfälscht.

Beispiel 4.2.2. Es werden Briefumschläge für Luftpost produziert, die im Schnitt nicht mehr als 2 g wiegen dürfen. Eine Stichprobe von 20 Briefumschlägen ergibt ein Durchschnittsgewicht von 2,01 g. Ferner ist bekannt, dass das Gewicht normalverteilt mit Erwartungswert μ und Varianz $0,03^2$ ist. Es soll zu einem Signifikanzniveau von 1% getestet werden, ob $\mu \leq 2$.

Für die Teststatistik ergibt sich $T = \sqrt{20}\frac{2{,}01-2}{0{,}03} \approx 1{,}49$. Da $\Phi(0{,}99) = 2{,}33$ wird die Hypothese akzeptiert. Es ist also denkbar, dass das erwartete Gewicht unter $2g$ liegt.

Tests bei unbekannter Varianz

In Anwendungen ist in der Regel die Varianz σ^2 nicht bekannt. Daher verwendet man die Teststatistik $T := \sqrt{n}\frac{\bar{X}-\mu_0}{s}$. Man konstruiert nun Tests wie im Fall von bekannter Varianz, ersetzt aber die Quantile der Standardnormalverteilung durch entsprechende Quantile der t_{n-1}-Verteilung. Genauer konstruiert man Tests zum Signifikanzniveau α wie folgt:

- Beim zweiseitigen Test von $H_0 : \mu = \mu_0$ gegen die Alternative $H_1 : \mu \neq \mu_0$ lehnt man die Hypothese ab, sofern $|T| \geq c$ ist, wobei c das $1 - \alpha/2$-Quantil der t_{n-1}-Verteilung ist.

- Beim einseitigen Test von $H_0 : \mu \leq \mu_0$ gegen die Alternative $H_1 : \mu > \mu_0$ lehnt man die Hypothese ab, sofern $T > c$, wobei c das $1 - \alpha$-Quantil der t_{n-1}-Verteilung ist.

Beispiel 4.2.3. Ein Diätkonzept, bei dem die Teilnehmer durchschnittlich $7{,}0kg$ abnehmen, wird überarbeitet. Um den Erfolg der Überarbeitung zu kontrollieren, werden die Gewichtsabnahmen von 28 Teilnehmern des neuen Konzepts erfasst. Die so erhobenen Daten werden als Zufallsstichprobe einer $\mathcal{N}(\mu, \sigma^2)$-verteilten Zufallsvariable angesehen. Sodann wird die Hypothese $\mu \leq 7{,}0$ gegen die Alternative $\mu > 7{,}0$ zu einem Signifikanzniveau von 2,5% getestet.

Die durchschnittliche Gewichtsabnahme betrug $7{,}3kg$ bei einer Stichprobenstandardabweichung von $0{,}4kg$. Somit ergibt sich für die Teststatistik $T = \sqrt{28}\frac{7{,}3-7{,}0}{0{,}4} = 3{,}97$. Das 97,5% -Quantil der t_{27}-Verteilung ist 2,052, sodass die Hypothese verworfen wird. Somit war die Überarbeitung in der Tat erfolgreich.

Test für verbundene Stichprobem

Bei dieser Art Testprobleme haben wir eine Folge von *Paaren* von Beobachtungen, also zwei Stichproben $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_n$, so dass die Paare $(X_1, Y_1), \ldots, (X_n, Y_n)$ zwar voneinander unabhängig sind, die beiden Komponenten eines Paares jedoch i.d.R. abhängig sind. Beispielsweise kann man sich vorstellen, dass eine bestimmte Größe (etwa die Länge eines Werkstückes) mit zwei verschiedenen Methoden gemessen wird (die eine Methode liefert $X_1, \ldots, X_n$, die andere Methode $Y_1, \ldots, Y_n$), die es zu vergleichen gilt. Man spricht in diesem Falle von *Verbundenen Stichproben.*

Manchmal kann man hierbei einen der obigen Tests auf die Differenzen $X_1 - Y_1, \ldots, X_n - Y_n$ anwenden. Wir geben ein Beispiel.

Beispiel 4.2.4. Um zu untersuchen, ob eine geplante Vereinfachung des Steuerrechts zu Mindereinnahmen des Staates führt, gehen wir wie folgt vor:

Wir wählen $n = 100$ Steuererklärungen des vergangenen Jahres zufällig aus und berechnen die Steuer nach dem geltenden Steuerrecht und nach dem vorgeschlagenen neuen Steuerrecht. Nennen wir für Steuererklärung k die Steuerschuld nach dem vorgeschlagenen Recht X_k und die Steuerschuld nach geltendem Recht Y_k. Wir bilden nun die Differenz $Z_k := X_k - Y_k$. Bei den untersuchten Steuererklärungen ergibt sich $\bar{Z} = 120$ bei einer Stichprobenstandardabweichung von 725.

Wir nehmen an, dass $Z_1, \ldots, Z_{100}$ normalverteilt sind und testen $H_0 : \mu \leq 0$ ("Mindereinnahmen") gegen die Alternative $H_1 : \mu > 0$. Beachte, dass wir hier wiederum die Hypothese und die Alternative so gewählt haben, dass der schwerwiegendere Irrtum (Es gibt

Mindereinnahmen für den Staat, aber unser Test sagt, es gibt keine) der Fehler erster Art ist.

Als Signifikanzniveau wählen wir $\alpha = 0,05$. Es ist dann $T = \sqrt{100}\frac{120-0}{725} \approx 1,655$. Weiterhin ist das 95%-Quantil der t_{99}-Verteilung gegeben durch $c = 1,660$. Weil $T \leq c$ ist, akzeptiert der Test die Nullhypothese. Mindereinnahmen sind also nicht auszuschließen.