Kapitel 4

Statistische Tests

Wir betrachten wieder ein parametrisches Modell $\{F_\theta : \theta \in \Theta\}$ und eine zugehörige unabhänging identisch verteilte 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 $H_1 : \theta \in \Theta_1$).
Es gilt also: $\Theta_0 \cup \Theta_1 = \Theta$ und $\Theta_0 \cap \Theta_1 = \emptyset$.

Ein *Test (der Hypothese $\theta \in \Theta_0$ gegen die Alternative $\theta \in \Theta_1$)* ist eine Entscheidungsregel $\varphi : \mathbb{R}^n \to \{0, 1\}$, die für jede Realisierung $x_1, \ldots, x_n$ der Stichprobe $X_1, \ldots, X_n$ festlegt, ob die Hypothese ($\varphi(x_1, \ldots, x_n) = 0$)oder die Alternative ($\varphi(x_1, \ldots, x_n) = 1$) gewählt wird. Man hat also eine disjunkte Zerlegung des Stichprobenraumes $\mathbb{R}^n$ in eine Teilmenge K und K^C, sodass wir uns für die Alternative H_1 entscheiden, wenn $(x_1, \ldots, x_n) \in K$ (wir sagen "*die Hypothese wird verworfen*").

K heißt *kritischer Bereich/ Ablehnbereich*.

Beispiel 4.0.1. Wir betrachten 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 4.0.2. 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 kritischem Bereich K gegeben, so heißt $\alpha_n : \Theta \to [0,1]$ mit

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

die *Gütefunktion* des Tests. Die Zahl

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

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

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

Beispiel 4.0.3. Wir betrachten wiederum die Tests T_3 und T_2 aus Beispiel 4.0.1. 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.

4.1 Tests für den Erwartungswert einer Normalverteilung

Wir betrachten nun einige "Standardtests", die den Erwartungswert μ einer Normalverteilung betreffen.

Eine 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 $T := \sqrt{n}\frac{\bar{X}_n-\mu_0}{S_X^2}$ bei unbekannter Varianz auf. Ist der wahre Parameter μ_0, so ist $T \sim \mathcal{N}(0,1)$ resp. $T \sim t_{n-1}$. Zur Erinnerung:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Tests bei bekannter Varianz: Gaußtest

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 Signifikanzniveau $\alpha \in (0,1)$ konstruieren. Wir betrachten 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 Signifikanzniveau 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 $c := \Phi^{-1}(1-\alpha/2) = z_{1-\alpha/2}$.

Beispiel 4.1.1. 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.

Zusammengefasst heißt das:

Gaußtest

Es seien $X_1, \ldots, X_n$ unabhänging identisch verteilte Zufallsvariablen mit $X_i \sim N(\mu, \sigma^2)$, wobei $\theta = \mu$ unbekannt ist und σ^2 bekannt ist. Sei das Signifikanzniveau $\alpha \in (0,1)$ fest.

Für das Testproblem

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

heißt er Test $\varphi : \mathbb{R}^n \to \{0,1\}$ mit

$$\varphi(x_1, \ldots, x_n) = \begin{cases} 1 & (x_1, \ldots, x_n) \in K \\ 0 & \text{sonst} \end{cases}$$

mit kritischem Bereich

$$K = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n | \, |T| = \left| \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{\sigma^2} \right| > z_{1-\alpha/2} \right\}$$

Gaußtest. $z_{1-\alpha/2}$ ist das $1-\alpha/2$-Quantil der Standardnormalverteilung.
φ ist ein unverfälschter Test zum Signifikanzniveau α.

Tests bei unbekannter Varianz: t-Test

Es seien $X_1, \dots, X_n$ unabhänging identisch verteilte Zufallsvariablen mit $X_i \sim N(\mu, \sigma^2)$, wobei $\theta = (\mu, \sigma^2)$ unbekannt sind. Sei das Signifikanzniveau $\alpha \in (0, 1)$ fest.
Für das Testproblem

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

heißt er Test $\varphi : \mathbb{R}^n \to \{0, 1\}$ mit

$$\varphi(x_1, \dots, x_n) = \begin{cases} 1 & (x_1, \dots, x_n) \in K \\ 0 & \text{sonst} \end{cases}$$

mit kritischem Bereich

$$K = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n |\, |T| = \left| \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{s_x^2} \right| > t_{n-1,1-\alpha/2} \right\}$$

t-Test. $t_{n-1,1-\alpha/2}$ ist das $1 - \alpha/2$-Quantil der t-Verteilung mit $n - 1$ Freiheitsgraden.
φ ist ein unverfälschter Test zum Signifikanzniveau α.

4.2 Tests für die Varianz einer Normalverteilung

Es seien $X_1, \dots, X_n$ unabhänging identisch verteilte Zufallsvariablen mit $X_i \sim N(\mu, \sigma^2)$, wobei $\theta = (\mu, \sigma^2)$ unbekannt sind. Sei das Signifikanzniveau $\alpha \in (0, 1)$ fest.
Für das Testproblem

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs.} \quad H_1 : \sigma^2 \neq \sigma_0^2$$

heißt er Test $\varphi : \mathbb{R}^n \to \{0, 1\}$ mit

$$\varphi(x_1, \dots, x_n) = \begin{cases} 1 & (x_1, \dots, x_n) \in K \\ 0 & \text{sonst} \end{cases}$$

mit kritischem Bereich

$$K = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n | T = \frac{(n-1)s_x^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha/2} \text{ oder } T = \frac{(n-1)s_x^2}{\sigma_0^2} < \chi^2_{n-1,\alpha/2} \right\}$$

Test für die Varianz der Normalverteilung. $\chi^2_{n-1,1-\alpha/2}$ ist das $1 - \alpha/2$-Quantil der χ^2-Verteilung mit $n - 1$ Freiheitsgraden.
φ ist ein unverfälschter Test zum Signifikanzniveau α.