Teil 2: Statistik

V. Tassion (based on a previous version from M. Schweizer)

D-INFK Frühling 2022 (Aktualisiert: 24. Mai 2022)

Statistische Grundideen

Ausgangssituation: Man hat beobachtete Daten und will daraus Rückschlüsse ziehen auf den zugrundeliegenden Mechanismus, der diese Daten generiert hat.

Ein häufiger erster Schritt ist eine *graphische Aufbereitung* der Daten; das ist oft nützlich, um einen ersten Eindruck und erste Ideen zu bekommen. Die entsprechenden Methoden gehören zur *deskriptiven* oder *beschreibenden Statistik*; diese Aspekte werden hier aber nicht behandelt.

Wir befassen uns im Folgenden mit der *induktiven Statistik*. Die Grundidee dabei ist relativ einfach. Man fasst die Daten $x_1, \ldots, x_n$ auf als Realisierungen/realisierte Werte $X_1(\omega), \ldots, X_n(\omega)$ von Zufallsvariablen $X_1, \ldots, X_n$, und sucht dann (unter geeigneten Zusatzannahmen) Aussagen über die Verteilung von $X_1, \ldots, X_n$.

Wichtig: Man muss immer sauber unterscheiden zwischen den *Daten* $x_1, \ldots, x_n$ (bezeichnet mit kleinen Buchstaben; $x_1, \ldots, x_n$ sind also in der Regel *Zahlen*) und dem generierenden *Mechanismus* $X_1, \ldots, X_n$ (bezeichnet mit grossen Buchstaben; $X_1, \ldots, X_n$ sind *Zufallsvariablen*, also *Funktionen* auf einem Ω). Das wird leider nicht immer eingehalten, obwohl schon der amerikanische Philosoph und Psychologe William James (1842–1910) im 19. Jahrhundert auf diesen Punkt hinwies: "We must be careful not to confuse data with the abstractions we use to analyze them."

Terminologie: Die Gesamtheit der Beobachtungen $x_1, \ldots, x_n$ oder Zufallsvariablen $X_1, \ldots, X_n$ nennt man oft eine *Stichprobe*; die Anzahl n heisst dann der *Stichprobenumfang*.

Ausgangspunkt unserer Betrachtungen ist in der Regel ein Datensatz $x_1, \ldots, x_n$ aus einer Stichprobe $X_1, \ldots, X_n$, für die wir ein Modell suchen. Dieses ist beschreibbar durch einen (möglicherweise hochdimensionalen) *Parameter* $\theta \in \Theta$, und um Begriffe und Notationen sauber definieren und benutzen zu können, muss man genauer spezifizieren, in welcher

Art wahrscheinlichkeitstheoretische Aussagen vom Parameter θ abhängen. Dazu betrachtet man simultan eine ganze *Familie von Wahrscheinlichkeitsräumen*; man hat typisch einen festen Grundraum $(\Omega, \mathcal{F})$ und für jeden Parameter θ aus dem *Parameterraum* Θ ein Wahrscheinlichkeitsmass $\mathbb{P}_\theta$ auf $(\Omega, \mathcal{F})$, also einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$ für jedes $\theta \in \Theta$. Anschaulich kann man sich vorstellen, dass jemand ("die Natur") einen Parameter $\theta \in \Theta$ und damit einen konkreten stochastischen Mechanismus $\mathbb{P}_\theta$ wählt. Als Statistiker weiss man aber nicht, welches θ aus Θ gewählt wurde; man betrachtet also die Daten $x_1, \dots, x_n$ als Ergebnisse von Zufallsvariablen $X_1, \dots, X_n$ unter dem unbekannten Mechanismus $\mathbb{P}_\theta$ und versucht, daraus Rückschlüsse über θ zu ziehen.

Beispiel:

Betrachten wir die obigen Ideen einmal für die Situation eines Münzwurfmodells. Wir haben eine Münze, von der wir nicht recht wissen, wie sie sich beim Werfen verhält. Wir nehmen an, dass die Münze bei jedem Wurf Kopf oder Zahl mit (immer der gleichen) Wahrscheinlichkeit p bzw. $1 - p$ produziert; wir kennen aber p nicht und betrachten es deshalb als einen unbekannten Parameter. Hier ist also $\Theta = [0, 1]$, der Parameter $\theta = p$ entspricht der Erfolgswahrscheinlichkeit für Kopf bei einem einzelnen Münzwurf, und das Wahrscheinlichkeitsmass $\mathbb{P}_\theta = \mathbb{P}_p$ beschreibt das Münzwurfmodell mit dem Erfolgsparameter p. Wir nehmen zusätzlich auch noch an, dass die einzelnen Würfe jeweils (d.h. in jedem Modell $\mathbb{P}_\theta$) unabhängig sind.

Nun werfen wir unsere Münze n Mal, schreiben jeweils 0 für Zahl, 1 für Kopf und erhalten so Daten $x_1, \dots, x_n$ aus einer Stichprobe $X_1, \dots, X_n$. Unser Modell (oder genauer unsere Modellfamilie) ist

$$\text{die } X_i \text{ sind unter } \mathbb{P}_p \text{ i.i.d.} \sim \text{Ber}(p);$$

der Parameter ist also wie erwähnt $\theta = p \in [0, 1] = \Theta$, und das Modell $\mathbb{P}_p$ beschreibt unabhängige Münzwürfe mit Erfolgsparameter p für "Kopf". Unser Ziel ist es, aus den Daten Rückschlüsse über p zu ziehen. (Zum Beispiel möchten wir vermutlich wissen, ob die Münze wohl fair ist oder nicht.)

In vielen Fällen ist der Parameterraum Θ eine Teilmenge von $\mathbb{R}^m$; wenn man zu gegebenen Daten ein passendes Modell finden und darüber gewisse statistische Aussagen machen möchte, so spricht man dann von einer *parametrischen statistischen Analyse*. Allgemein gehören dazu die folgenden Etappen:

1) Beschreibende Statistik der Daten: In diesem Schritt versucht man mit graphischen Methoden, aufgrund der Daten eine erste Idee für die Wahl einer geeigneten Modellierung zu finden. (Diesen Schritt werden wir hier nicht weiter erklären.)

2) Wahl eines (parametrischen) Modells: Hier spezifiziert man die Parametermenge Θ und die Familie $(\mathbb{P}_\theta)_{\theta\in\Theta}$ von Modellen, mit denen man arbeiten will.

3) Schätzung der Parameter: Aufgrund der Daten will man ein möglichst gut passendes Modell wählen. Dazu benutzt man einen *Schätzer*; die zugehörige *Schätzfunktion* ist eine Abbildung, die gegebenen Daten $x_1, \ldots, x_n$ einen Parameter $\theta \in \Theta$ zuordnet.

4) Kritische Modellüberprüfung (Anpassungstest): Hier fragt man, ob die Daten zu dem gewählten Parameter θ bzw. Modell $\mathbb{P}_\theta$ gut passen; das macht man mit einem geeigneten *statistischen Test*.

5) Aussagen über Zuverlässigkeit der Schätzungen: Statt eines einzigen Parameterwertes kann man auch versuchen, einen Bereich in Θ so zu spezifizieren, dass alle zugehörigen Modelle $\mathbb{P}_\theta$ in einem zu präzisierenden Sinn gut zu den Daten passen; man spricht dann von einem *Konfidenzbereich*.

Kapitel 1

Schätzer

Ziel: Überblick über grundlegende Ideen und Methoden zur Schätzung von Parametern.

Setup:

- Parameterraum $\Theta \subset \mathbb{R}$,
- Grundraum Ω,
- sigma-Algebra $\mathcal{F}$,
- $(\mathbb{P}_\theta)_{\theta \epsilon \Theta}$ Familie von Wahrscheinlichkeitsmasse auf $(\Omega, \mathcal{F})$,
- $X_1, \ldots, X_n$ Zufallsvariablen auf $(\Omega, \mathcal{F})$.

1.1 Grundbegriffe

Wir suchen dann für der Parameter θ ein *Schätzer* T aufgrund unserer Stichprobe $(X_1, \ldots, X_n)$.

Definition 1.1. *Ein* ***Schätzer*** *ist eine Zufallsvariable* $T : \Omega \to \mathbb{R}$ *der Form*

$$T = t(X_1, \ldots, X_n),$$

wobei $t : \mathbb{R}^n \to \mathbb{R}$.

Einsetzen von Daten $x_i = X_i(\omega)$, $i = 1, \ldots, n$, liefert dann *Schätzwerte* $T(\omega) = t(x_1, \ldots, x_n)$ für θ.

Es ist wichtig, die Konzepte "Schätzer" und "Schätzwert" sauber auseinanderzuhalten. Ein Schätzwert ist eine *Zahl* $T(\omega) = t(X_1(\omega), \ldots, X_n(\omega)) = t(x_1, \ldots, x_n)$; wie die Daten x_i ist das die Realisation $T(\omega)$ (im von uns betrachteten konkreten Experiment ω) der Zufallsvariable T.

Beispiel: tea tasting lady

Eine englische Lady behauptet, bei Tee mit Milch anhand des Geschmacks unterscheiden zu können, ob zuerst die Milch oder zuerst der Tee in die Tasse eingegossen worden ist. Wie kann man überprüfen, ob das stimmen kann?

Um zunächst einmal Daten zu bekommen, stellen wir der Lady an n Tagen die Aufgabe, zwei Tassen (je eine vom Typ 1 und Typ 2) zu klassifizieren; sie soll also angeben, in welche der Tassen zuerst Milch eingegossen worden ist. Wir notieren uns dabei die Ergebnisse $x_1, \ldots, x_n \in \{0, 1\}$ (falsch bzw. richtig klassifiziert) und fassen wie üblich diese Daten als Realisationen von Zufallsvariablen $X_1, \ldots, X_n$ auf. Dann ist $S_n = \sum_{i=1}^n X_i$ die (zufällige) Anzahl der korrekt klassifizierten Tassenpaare, und $s_n = \sum_{i=1}^n x_i$ ist die beobachtete Anzahl von Erfolgen.

Als Modelle nehmen wir nun an, dass die X_i unter $\mathbb{P}_\theta$ i.i.d. $\sim \mathrm{Ber}(\theta)$ mit $\theta \in \Theta = [0, 1]$ sind. Dann ist natürlich $S_n \sim \mathrm{Bin}(n, \theta)$ unter $\mathbb{P}_\theta$, d.h. im Modell $\mathbb{P}_\theta$, das zu θ gehört, ist die Anzahl S_n der Erfolge binomialverteilt mit Parametern n und θ.

Weil wir den Parameter θ nicht kennen, liegt es nahe, zuerst einmal dafür einen Schätzer zu suchen. Eine erste Möglichkeit wäre, einfach das letzte Ergebnis zu nehmen; unser erster Schätzer $\hat{T}$ für θ wäre also $\hat{T} = X_n$. Obwohl das absurd aussieht, werden wir sehen, dass dieser Schätzer durchaus die eine oder andere vernünftige Eigenschaft hat.

Ein zweiter naheliegender Schätzer wäre die durchschnittliche Anzahl der Erfolge der Lady bei ihren n Versuchen; unser zweiter Schätzer wäre also $T = \overline{X}_n = \frac{1}{n} S_n$.

Für gegebene Daten $x_1, \ldots, x_n$ gibt uns das dann zwei Schätzwerte $\hat{t}(x_1, \ldots, x_n) = x_n$ und $t(x_1, \ldots, x_n) = \overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$, die wir konkret berechnen könnten.

◇

1.2 Bias

Wie die X_i ist der Schätzer T ein Zufallsvariable, deren Verteilung (unter $\mathbb{P}_\theta$) von dem unbekannten Parameter θ abhängt. Nur selten kann man diese Verteilung explizit bestimmen; siehe später die Resultate für die Normalverteilung. Bevor wir zumindest eine systematische Methode zur *Bestimmung* von Schätzern betrachten, führen wir einige allgemeine wünschenswerte *Eigenschaften* auf.

Definition 1.2. *Ein Schätzer T heisst **erwartungstreu** für θ, falls für alle $\theta \in \Theta$ gilt*

$$\mathbb{E}_\theta[T] = \theta.$$

Interpretation Im Mittel (über alle denkbaren Realisationen ω) schätzt T also richtig, und zwar unabhängig davon, welches Modell $\mathbb{P}_\theta$ zu Grunde liegt.

Definition 1.3. *Sei $\theta \in \Theta$, und T ein Schätzer. Der **Bias** (oder erwartete Schätzfehler) von T im Modell $\mathbb{P}_\theta$ ist definiert als*

$$\mathbb{E}_\theta[T] - \theta.$$

Der mittlere quadratische Schätzfehler ("mean squared error", MSE) von T im Modell $\mathbb{P}_\theta$ ist definiert als

$$\mathrm{MSE}_\theta[T] := \mathbb{E}_\theta[(T - \theta)^2].$$

Erwartungstreu (auf Englisch "*unbiased*") bedeutet also, dass der Bias identisch Null ist.

Bemerkung: Man kann den MSE zerlegen als

$$\mathrm{MSE}_\theta[T] = \mathbb{E}_\theta[(T - \theta)^2] = \mathrm{Var}_\theta[T] + (\mathbb{E}_\theta[T] - \theta)^2,$$

also in die Summe aus der Varianz des Schätzers T und dem Quadrat des Bias. Für erwartungstreue Schätzer sind Varianz und MSE dasselbe.

Beispiel: tea tasting lady

Beide oben angegebenen Schätzer $\hat{T} = X_n$ und $T = \overline{X}_n$ sind erwartungstreu. Unter $\mathbb{P}_\theta$ ist ja $\hat{T} = X_n \sim Be(\theta)$, also

$$\mathbb{E}_\theta[\hat{T}] = \mathbb{E}_\theta[X_n] = \theta, \text{ und}$$
$$\mathrm{MSE}_\theta[\hat{T}] = \mathrm{Var}_\theta(X_n) = \theta(1-\theta).$$

Obwohl aber $\hat{T}$ erwartungstreu ist, wird er kaum je das richtige Ergebnis für den Parameter θ liefern; für jede konkrete Realisierung ω hat ja $\hat{T}(\omega) = X_n(\omega)$ den Wert 0 oder 1, und nur im theoretischen Mittel über alle ω erhalten wir θ.

Unser zweiter Schätzer für θ ist $T = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$; er ist auch erwartungstreu, denn

$$\mathbb{E}_\theta[T] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta[X_i] = \theta \quad \text{für alle } \theta,$$

und (wegen Unabhängigkeit unter $\mathbb{P}_\theta$)

$$\mathrm{Var}_\theta[T] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}_\theta[X_i] = \frac{1}{n}\theta(1-\theta),$$

weil alle $X_i \sim \mathrm{Ber}(\theta)$ unter $\mathbb{P}_\theta$ sind.

Wegen

$$\mathrm{Var}_\theta[\hat{T}] = \mathrm{Var}_\theta[X_n] = \theta(1-\theta)$$

hat T kleinere Varianz als $\hat{T}$. Das ist einleuchtend, weil T die Information in der Stichprobe $X_1, \ldots, X_n$ viel besser ausnutzt. ◇

1.3 Die Maximum-Likelihood-Methode (ML-Methode)

In diesem Abschnitt stellen wir eine Methode vor, um systematisch Schätzer zu bestimmen. Diese Methode liefert in sehr vielen Situationen Ergebnisse, die sowohl plausibel sind als auch gute Eigenschaften haben.

Ausgangspunkt im Folgenden ist immer eine von zwei Situationen, je nachdem, ob wir es mit diskreten oder mit stetigen Zufallsvariablen zu tun haben. Wir schreiben oft kurz $\vec{X} = (X_1, \ldots, X_n)$. In jedem Modell $\mathbb{P}_\theta$ sind $X_1, \ldots, X_n$ entweder diskret mit gemeinsamer Gewichtsfunktion $p_{\vec{X}}(x_1, \ldots, x_n; \theta)$ oder stetig mit gemeinsamer Dichtefunktion $f_{\vec{X}}(x_1, \ldots, x_n; \theta)$. Meistens sind sogar die X_i unter $\mathbb{P}_\theta$ i.i.d. mit individueller Gewichts-

funktion $p_X(x;\theta)$ bzw. Dichtefunktion $f_X(x;\theta)$; dann ist also die gemeinsame Gewichtsfunktion

$$p_{\vec{X}}(x_1,\ldots,x_n;\theta) = \prod_{i=1}^{n} p_X(x_i;\theta)$$

bzw. die gemeinsame Dichtefunktion

$$f_{\vec{X}}(x_1,\ldots,x_n;\theta) = \prod_{i=1}^{n} f_X(x_i;\theta).$$

Anschaulich ist

$$p_{\vec{X}}(x_1,\ldots,x_n;\theta) = P_\theta[X_1 = x_1,\ldots,X_n = x_n]$$

gerade die Wahrscheinlichkeit im Modell P_θ, dass unsere Stichprobe $X_1,\ldots,X_n$ die Werte $x_1,\ldots,x_n$ liefert, und $f_{\vec{X}}(x_1,\ldots,x_n;\theta)$ ist das übliche stetige Analogon.

Definition 1.4. *Die* ***Likelihood-Funktion*** *ist*

$$L(x_1,\ldots,x_n;\theta) := \begin{cases} p_{\vec{X}}(x_1,\ldots,x_n;\theta) & \textit{im diskreten Fall,} \\ f_{\vec{X}}(x_1,\ldots,x_n;\theta) & \textit{im stetigen Fall.} \end{cases}$$

Die Funktion $\log L(x_1,\ldots,x_n;\theta)$ heisst **log-Likelihood-Funktion**. Sie hat gegenüber der Likelihood-Funktion den Vorteil, dass sie im i.i.d.-Fall durch eine Summe (statt ein Produkt) gegeben und damit zum Rechnen oft wesentlich einfacher ist.

Nach diesen allgemeinen Vorbereitungen wenden wir uns nun der erwähnten Methode zur Bestimmung von Schätzern zu. Für eine Stichprobe $X_1,\ldots,X_n$ gibt uns die Likelihoodfunktion $L(x_1,\ldots,x_n;\theta)$ zumindest im diskreten Fall die Wahrscheinlichkeit im Modell P_θ, dass unsere Stichprobe gerade die Werte $x_1,\ldots,x_n$ liefert. Um eine möglichst gute Anpassung des Modells an die Daten zu erreichen, wollen wir diese Wahrscheinlichkeit möglichst gross machen, indem wir den Parameter geschickt wählen.

Definition 1.5. *Für jedes $x_1,\ldots,x_n$, sei $t_{ML}(x_1,\ldots,x_n) \in \mathbb{R}$ der Wert, der $\theta \mapsto L(x_1,\ldots,x_n;\theta)$ als Funktion von θ maximiert. D.h.,*

$$L(x_1,\ldots,x_n;t_{ML}(x_1,\ldots,x_n)) = \max_{\theta\in\Theta} L(x_1,\ldots,x_n;\theta).$$

Ein ***Maximum-Likelihood-Schätzer (ML-Schätzer)*** T_{ML} *für θ wird definiert*

durch

$$T_{\mathrm{ML}} = t_{\mathrm{ML}}(X_1, \ldots, X_n).$$

Meistens sind $X_1, \ldots, X_n$ i.i.d. unter $\mathbb{P}_\theta$; die Likelihood-Funktion L ist dann ein Produkt, und es ist bequemer, statt L die log-Likelihood-Funktion $\log L$ zu maximieren, weil diese eine Summe ist. Statt zu maximieren sucht man ferner meistens nur Nullstellen der Ableitung (nach θ).

Bemerkung:

In den Rechnungen arbeitet man oft mit $L(x_1, \ldots, x_n; \theta)$, insbesondere beim Maximieren über θ. Das optimale θ^* ist dann eine Funktion $t_{\mathrm{ML}}(x_1, \ldots, x_n)$ von $x_1, \ldots, x_n$. Damit der resultierende Schätzer T_{ML} von der Stichprobe $X_1, \ldots, X_n$ abhängt, muss dann aber $x_1, \ldots, x_n$ durch $X_1, \ldots, X_n$ ersetzt werden, d.h. der Maximum-Likelihood-Schätzer ist $T_{\mathrm{ML}} = t_{\mathrm{ML}}(X_1, \ldots, X_n)$.

Beispiel 1: Bernoulli-Verteilung.

Im Modell $\mathbb{P}_p$ seien

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathrm{Ber}(p).$$

Hier ist $\theta = p$, und wir wollen also für eine unbekannte Münze den Erfolgsparameter schätzen. Diese Fragestellung haben wir schon im letzten Abschnitt im Beispiel mit der tea tasting lady angetroffen.

Die Gewichtsfunktion einer $\mathrm{Ber}(\theta)$-Verteilung ist

$$p_X(x; \theta) = P_\theta[X = x] = \theta^x (1-\theta)^{1-x} \text{ für } x \in \{0, 1\}.$$

Weil die X_i i.i.d. sind, ist also

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} p_X(x_i; \theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}$$

und

$$\log L(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} x_i \log \theta + \left(n - \sum_{i=1}^{n} x_i \right) \log(1-\theta).$$

Wir wollen das über θ maximieren und setzen dazu die entsprechende Ableitung Null. Die Ableitung nach θ ist

$$\frac{\partial}{\partial \theta} \log L(x_1, \ldots, x_n; \theta) = \frac{1}{\theta} \sum_{i=1}^{n} x_i - \frac{1}{1-\theta} \left(n - \sum_{i=1}^{n} x_i \right),$$

und das ist 0 für

$$(1-\theta)\sum_{i=1}^{n} x_i = \theta\left(n - \sum_{i=1}^{n} x_i\right),$$

d.h. für $\theta = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Der ML-Schätzer für θ bzw. p ist hier also

$$T = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n.$$

$\diamond$

1.4 Modelle mit mehreren Parametern

Unsere aktuelle Theorie begrenzt sich auf stochastische Modelle mit einem paramter θ in $\mathbb{R}$. Verschiedene Situationen verlangen aber Modelle mit mehreren Parametern $\theta_1, \theta_2, \ldots, \theta_m$, $m \geq 2$. Um solche Modelle besser zu verstehen, entwickeln wir nun eine allgemeinere Theorie.

Als ersten Schritt betrachten wir folgende Parametermenge

$$\Theta \subset \mathbb{R}^m,$$

wobei m gerade der Anzahl an Parametern entspricht und das (stochastische) Modell gegeben ist durch eine Familie von Massen $(\mathbb{P}_\theta)_{\theta\in\Theta}$. Dabei interessieren wir uns für eine Schätzung der Parameter $\theta = (\theta_1, \ldots, \theta_n)$. Es ist anzumerken, dass sich alle vorherigen Definitionen in dieses Setup problemlos übertragen lassen.

Beispiel 2: Normalverteilung.

Im Modell $\mathbb{P}_\theta$ seien

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma^2).$$

Hier ist die Dimension des unbekannten Parameters $m = 2$, wir haben $\theta = (\mu, \sigma^2) = (\mu, v)$, und wir wollen μ und $\sigma^2 = v$ schätzen.

Die Dichtefunktion von X_i unter $\mathbb{P}_\theta$ ist

$$f_X(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-\mu)^2}{2v}}.$$

Weil die X_i i.i.d. sind, ist also

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_X(x_i;\theta)$$

und

$$\log L(x_1,\dots,x_n;\theta) = \sum_{i=1}^{n} \log f_X(x_i;\theta) = -n\frac{1}{2}(\log 2\pi + \log v) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2v}.$$

Die Ableitungen nach μ und v sind

$$\frac{\partial}{\partial \mu} \log L(x_1,\dots,x_n;\theta) = 2\sum_{i=1}^{n} \frac{x_i-\mu}{2v},$$
$$\frac{\partial}{\partial v} \log L(x_1,\dots,x_n;\theta) = -\frac{n}{2}\frac{1}{v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i-\mu)^2,$$

und sie werden beide gleichzeitig 0 für

$$0 = \sum_{i=1}^{n}(x_i-\mu) = \sum_{i=1}^{n} x_i - n\mu, \text{ d.h. für } \mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}_n,$$
$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2.$$

Der ML-Schätzer für $\theta = (\mu, \sigma^2)$ ist also

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n,$$
$$T_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X}_n)^2,$$

wobei die zweite Gleichheit durch Ausquadrieren folgt. ◇

Der Schätzer $T = (T_1, T_2)$ im obigen Beispiel ist ganz allgemein auch der sogenannte Momentenschätzer für

$$(\mathbb{E}_\theta[X], \mathrm{Var}_\theta[X])$$

in jedem Modell P_θ, wo $X_1,\dots,X_n$ i.i.d. sind. Dieser Schätzer hat aber den allgemeinen Nachteil, dass er nicht erwartungstreu für $(\mathbb{E}_\theta[X], \mathrm{Var}_\theta[X])$ ist. Zwar ist für jedes θ

$$\mathbb{E}_\theta[T_1] = \mathbb{E}_\theta[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta[X_i] = \mathbb{E}_\theta[X];$$

aber

$$\mathbb{E}_\theta[(\overline{X}_n)^2] = \frac{1}{n^2}\sum_{i,k=1}^{n}\mathbb{E}_\theta[X_iX_k] = \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}_\theta[X_i^2] + \sum_{i\neq k}\mathbb{E}_\theta[X_iX_k]\right),$$

und wegen Unabhängigkeit ist für $i \neq k$

$$\mathbb{E}_\theta[X_iX_k] = \mathbb{E}_\theta[X_i]\mathbb{E}_\theta[X_k] = (\mathbb{E}_\theta[X])^2.$$

Also ist

$$\begin{aligned}\mathbb{E}_\theta[T_2] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta[X_i^2] - \left(\frac{1}{n}\mathbb{E}_\theta[X^2] + \frac{n^2-n}{n^2}(\mathbb{E}_\theta[X])^2\right)\\ &= \left(1-\frac{1}{n}\right)\left(\mathbb{E}_\theta[X^2] - (\mathbb{E}_\theta[X])^2\right)\\ &= \frac{n-1}{n}\mathrm{Var}_\theta[X].\end{aligned}$$

Um einen erwartungstreuen Schätzer T' für $(\mathbb{E}_\theta[X], \mathrm{Var}_\theta[X])$ zu haben, benutzt man deshalb meistens

$$\begin{aligned}T_1' &= T_1 = \overline{X}_n\\ T_2' &= \frac{n}{n-1}T_2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{n}{n-1}(\overline{X}_n)^2.\end{aligned}$$

Für T_2' benutzt man oft auch die Notation

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

und nennt S^2 die **empirische Stichprobenvarianz**.

Kapitel 2

Konfidenzintervalle

Grundidee: Wie in Abschnitt 1 suchen wir aus einer Familie $(\mathbb{P}_\theta)_{\theta\in\Theta}$ von Modellen eines, das zu unseren Daten $x_1, \ldots, x_n$ passt. Ein Schätzer für θ gibt uns dabei einen einzelnen zufälligen möglichen Parameterwert. Weil es schwierig ist, mit diesem einen Wert den richtigen (aber unbekannten) Parameter zu treffen, suchen wir nun stattdessen eine **(zufällige) Teilmenge des Parameterbereichs**, die hoffentlich den wahren Parameter enthält.

Setup:

- Parameterraum $\Theta \subset \mathbb{R}$,
- Grundraum Ω,
- sigma-Algebra $\mathcal{F}$,
- $(\mathbb{P}_\theta)_{\theta\in\Theta}$ Familie von Wahrscheinlichkeitsmasse auf $(\Omega, \mathcal{F})$,
- $X_1, \ldots, X_n$ Zufallsvariablen auf $(\Omega, \mathcal{F})$.

2.1 Definition

Im vorangegangenen Kapitel haben wir eine Methode zur Schätzung unbekannter Parameter mittels Formeln kennengelernt. Eine naheliegende Frage ist: Wie reichhalting sind diese Schätzer? Werfen wir zum Beispiel eine Münze n mal, ohne die Wahrscheinlichkeit p von Kopf zu kennen. Falls wir 70 Mal Kopf erhalten, ist der Maximum-Likelihood-Schätzer für p $T_{ML} = 0.7$. Wie weit liegt T_{ML} von dem wahren Wert p entfernt? Um diese Art von Frage zu beantworten Frage zu beantworten, führen wir den Begriff der Konfidenzintervalle ein.

Definition 2.1. *Sei $\alpha \in [0,1]$. Ein **Konfidenzintervall für θ mit Niveau** $1-\alpha$ ist ein Zufallsintervall $I = [A,B]$, sodass gilt*

$$\forall \theta \in \Theta \qquad \mathbb{P}_\theta[A \le \theta \le B] \ge 1-\alpha, \tag{2.1}$$

wobei A, B Zufallsvariablen der Form $A = a(X_1,\dots,X_n)$, $B = b(X_1,\dots,X_n)$ mittels $a, b : \mathbb{R}^n \to \mathbb{R}$ sind.

Bemerkung 2.2. *In Eq. (2.1), ist der Paramter θ deterministisch und nicht zufällig. Die stochastischen Elemente sind gerade die Schranken $A = a(X_1,\dots,X_n)$ und $B = b(X_1,\dots,X_n)$.*

Example 1: *Konfidenzintervall eines normalen Modells mit Varianz 1 und unbekannten Mittelwerts*

Seien $X_1,\dots,X_n$ n u.i.v. normalverteilte Zufallsvariablen mit Parametern m und $\sigma^2 = 1$. Wir betrachten somit ein stochastisches Modell mit bekannter Varianz ($\sigma^2 = 1$) aber unbekanntem Mittelwert m. Man kann zeigen, dass der Maximum-Likelihood Schätzer gegeben ist durch

$$T = T_{ML} = \frac{X_1 + \cdots + X_n}{n}.$$

Wir suchen nun für m Konfidenzintervalle der folgenden Form

$$I = [T - \frac{c}{\sqrt{n}}, T + \frac{c}{\sqrt{n}}],$$

wobei $c > 0$ eine von n unabhängige Konstante ist. Zuerst betrachten wir

$$\mathbb{P}_\theta[T - \frac{c}{\sqrt{n}} \le m \le T + \frac{c}{\sqrt{n}}] = \mathbb{P}_\theta[-c \le Z \le c],$$

wobei $Z = \sqrt{n}(T - m) = \frac{X_1 + \cdots + X_n - nm}{\sqrt{n}}$ eine standardnormalverteilte Zufallsvariable ist. Somit können wir die obige Wahrscheinlichkeit explizit bestimmen

$$\begin{aligned}\mathbb{P}_\theta[-c \le Z \le c] &= \mathbb{P}_\theta[Z \le c] - \underbrace{\mathbb{P}_\theta[X < -c]}_{=1-\mathbb{P}_\theta[Z \le c]} \\ &= 2\Phi(c) - 1,\end{aligned}$$

wobei $\Phi(c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} e^{-x^2/2} dx$. Nachlesen in einer Tabelle der Standardnormalverteilung ergibt $2\Phi(1.96) - 1 \ge 0.95$. Somit erhalten wir mittels Wahl von $c = 1.96$ nach obiger Rechnung

$$\mathbb{P}_\theta[T - \frac{1.96}{\sqrt{n}} \le m \le T + \frac{1.96}{\sqrt{n}}] \ge \frac{95}{100}.$$

Somit gilt schließlich, dass

$$I = [T - \frac{1.96}{\sqrt{n}}, T + \frac{1.96}{\sqrt{n}}]$$

nach Definition ein Konfidenzintervall für m mit Niveau 95% ist.

Was bedeutet dies genau?

Stellen wir uns n Messungen zu einer physikalischen Größe vor. Man möchte zum Beispiel bei Raumtemperatur die Temperatur ermitteln, bei der Wasser anfängt zu kochen. Die Eigenschaften des Thermometers legen nahe, dass jede Messung durch eine normalverteilte Zufallsvariable $\mathcal{N}(m, 1)$ modelliert werden kann, wobei m die Temperatur ist, bei der das Wasser anfängt zu kochen. Man führt einige Messung $x_1 = 99.2$, $x_2 = 98.7, \cdots$ durch. Nach $n = 100$ aufeinanderfolgenden Versuchen berechnest man den empirischen Durchschnitt $\widehat{m}(x) = \frac{x_1 + \cdots + x_n}{n} = 99.106$. Das obige festgelegte Konfidenzintervall besagt, dass unter der Voraussetzung eines korrekten stochastischen Modells der reale Wert m mit 95%iger Wahrscheinlichkeit im obigen Intervall liegt

$$[99.106 - 0.196, 99.106 + 0.196] = [98.910, 99.302].$$

Was sind die Knackpunkte?

Im obigen Beispiel ist die wichtigste Beobachtung, dass die Zufallsvariable $Z = \sqrt{n}(T - m)$ für jeden Wert des unbekannten Parameters immer normalverteilt ist. Allgemein kann man Konfidenzintervalle für einen Parameter θ versuchen zu erhalten, in dem man zuerst einen Schätzer T für θ ermittelt. Anschließend versucht man eine Zufallsvariable der Form $Z = f(T, \theta)$ zu finden, deren Verteilung bestimmt werden kann und nicht

von θ abhängt. Dies ist im Allgemeinen einfacher, wenn die Zufallsvariablen $X_1, \ldots, X_n$ normalverteilt sind, da Operationen auf normalverteilten Zufallsvariablen gut nachvollziehbar sind. Zum Beispiel haben wir oben verwendet, dass die Summe von unabhängigen normalverteilter Zufallsvariablen ebenfalls normalverteilt ist. Im nächsten Abschnitt werden wir neue Verteilungen einführen, welche sich aus Operationen mittels normalverteilten Zufallsvariablen ergeben.

2.2 Verteilungsaussagen

In vielen Situationen ist es nützlich oder nötig, die Verteilung (unter $\mathbb{P}_\theta$, für jedes $\theta \in \Theta$ oder für gewisse θ) eines Schätzers zu kennen. Exakte allgemeine Aussagen gibt es dazu nicht viele; für die Normalverteilung folgt das weiter unten in Satz 2.7.

Definition 2.3. *Eine stetige Zufallsvariable X heisst χ^2-**Verteilt mit** m **Freiheitsgraden** falls ihre Dichte gegeben ist durch*

$$f_X(y) = \frac{1}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})} y^{\frac{m}{2}-1} e^{-\frac{1}{2}y} \text{ für } y \geq 0.$$

Wir schreiben dann $X \sim \chi^2_m$.

Dabei ist die sogenannte Gamma-Funktion für $v \geq 0$ definiert durch

$$\Gamma(v) := \int_0^\infty t^{v-1} e^{-t} dt.$$

Es gilt $\Gamma(n) = (n-1)!$ für $v = n \in \mathbb{N}$.

Bemerkung: Die χ^2 Verteilung mit m Freiheitsgraden ist der Spezialfall einer $Ga(\alpha, \lambda)$-Verteilung mit $\alpha = \frac{m}{2}$ und $\lambda = \frac{1}{2}$. Für $m = 2$ ergibt das eine Exponentialverteilung mit Parameter $\frac{1}{2}$.

Satz 2.4. *Sind die Zufallsvariablen $X_1, \ldots, X_m$ u.i.v. $\sim \mathcal{N}(0,1)$, so ist die Summe $Y := \sum_{i=1}^m X_i^2 \sim \chi^2_m$.*

Definition 2.5. *Eine stetige Zufallsvariable X heisst t-**verteilt mit** m **Freiheitsgraden** falls ihre Dichte gegeben ist durch*

$$f_X(x) = \frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\Gamma(\frac{m}{2})} \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}} \text{ für } x \in \mathbb{R}.$$

Wir schreiben dann $X \sim t_m$.

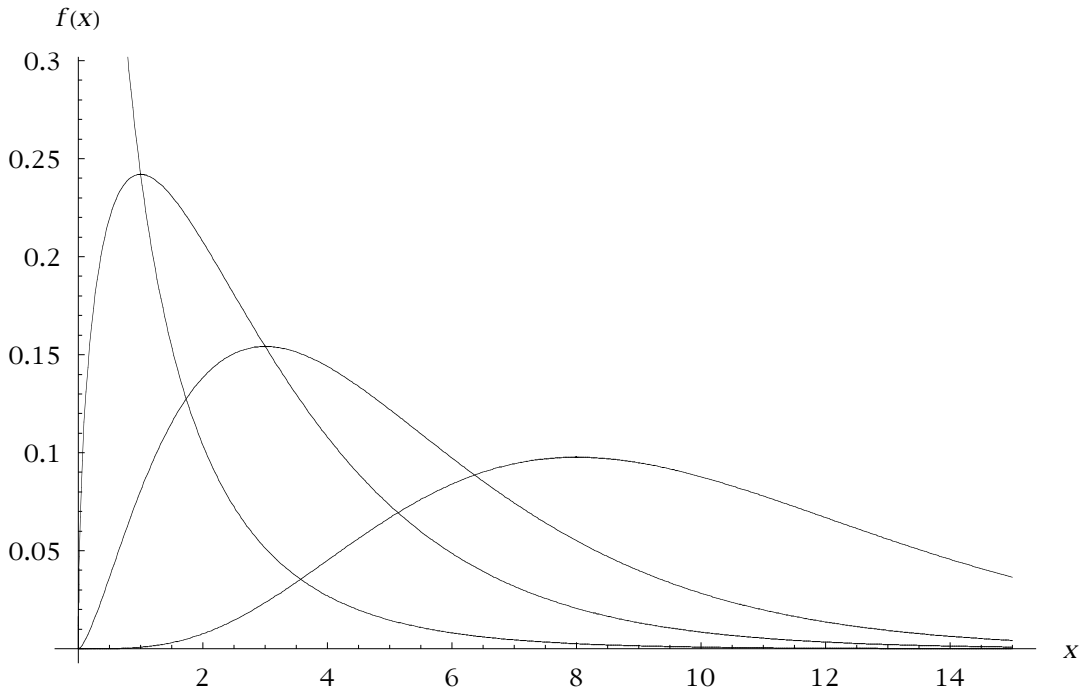


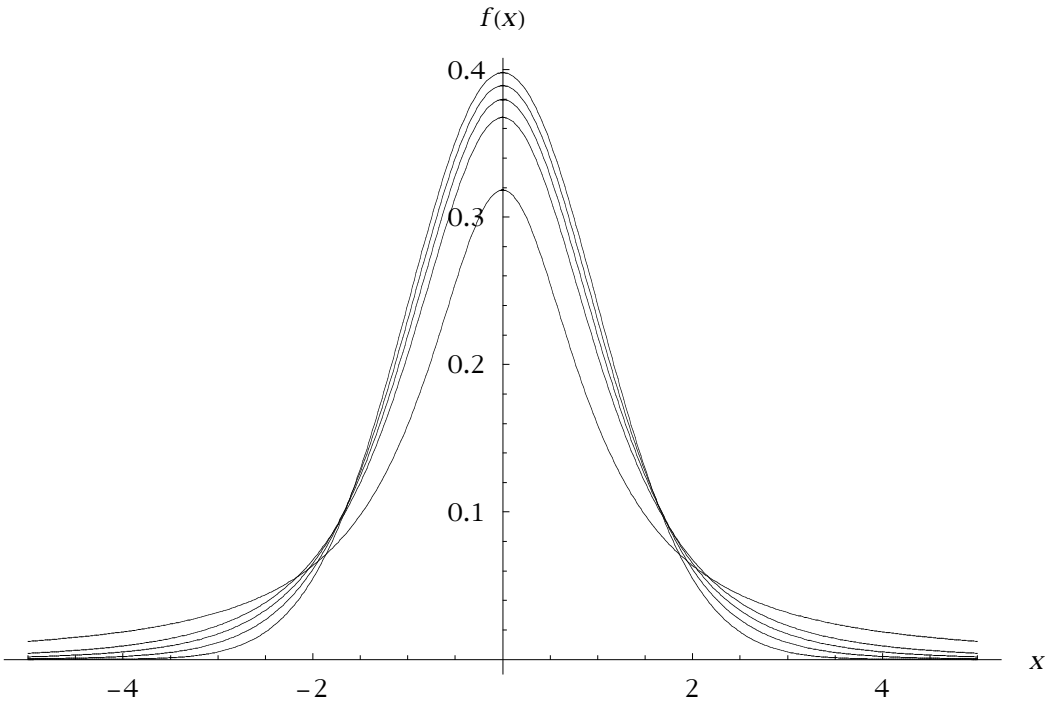
Abbildung 2.1: Graphische Darstellung der Dichten von χ^2-Verteilungen mit Anzahl der Freiheitsgrade (von oben nach unten) $m = 1, 3, 5, 10$.

Bemerkung: Für $m = 1$ ist das eine Cauchy-Verteilung, und für $m \to \infty$ erhält man asymptotisch eine $\mathcal{N}(0,1)$-Verteilung. Wie die $\mathcal{N}(0,1)$-Verteilung ist die t-Verteilung symmetrisch um 0; sie ist aber langschwänziger (d.h. ihre Dichte geht langsamer gegen 0, wenn das Argument gegen $\pm\infty$ geht), und zwar umso mehr, je kleiner m ist.

Satz 2.6. *Sind X und Y unabhängig mit $X \sim \mathcal{N}(0,1)$ und $Y \sim \chi^2_m$, so ist der Quotient*

$$Z := \frac{X}{\sqrt{\frac{1}{m}Y}}$$

t-verteilt mit m Freiheitsgraden.



Graphische Darstellung der Dichten von t-Verteilungen
mit Anzahl der Freiheitsgrade (von unten nach oben) $m = 1, 3, 5, 10, 100$.

2.3 Normal mit σ und m unbekannt

Für normalverteilte Stichproben hat man exakte Aussagen; wir werden das auch bei der Diskussion von Tests später noch ausgiebig benutzen. Wir erinnern an die Notationen

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

für das Stichprobenmittel und die Stichprobenvarianz.

Satz 2.7. *Seien $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$. Dann $\overline{X}_n$ und S^2 sind unabhängig.*

Beweis. Siehe [Rice, Abschnitt 6.3]. □

Beispiel (Strausseneier):

Die Australier Mr. Smith und Dr. Thurston streiten sich noch immer über das Gewicht von Strausseneiern. Sie haben von ihrer Afrikareise $n = 8$ Eier mitgebracht, deren Gewichte (in g) 1090, 1150, 1170, 1080, 1210, 1230, 1180, 1140 betragen. Diese Daten fassen sie auf als Realisationen von Zufallsvariablen $X_1, \dots, X_n$, die alle unter $\mathbb{P}_\theta$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ sind. Gesucht sind nun Konfidenzbereiche für die unbekannten Parameter μ und σ^2. Im Gegensatz zum (ersten) Beispiel im letzten Abschnitt wird hier σ^2 auch als unbekannt angenommen.

Die offensichtlichen **Schätzer** für μ und σ^2 sind das Stichprobenmittel $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ und die Stichprobenvarianz $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Es liegt nahe, als Konfidenzbereich jeweils ein Intervall um diese Schätzer herum anzusetzen, und wir machen das zuerst für μ.

Machen wir den Ansatz

$$C(X_1, \dots, X_n) = \big[\overline{X}_n - \cdots, \overline{X}_n + \cdots\big],$$

so wollen wir erreichen, dass gilt

$$1 - \alpha \le \mathbb{P}_\theta[C(X_1, \dots, X_n) \ni \theta] = \mathbb{P}_\theta\big[[\overline{X}_n - \cdots, \overline{X}_n + \cdots] \ni \mu\big] = \mathbb{P}_\theta\big[|\overline{X}_n - \mu| \le \cdots\big].$$

Nach Satz 2.6 und Satz 2.7 ist für jedes $\theta \in \Theta$

$$\frac{\overline{X}_n - \mu}{S/\sqrt{n}} \sim t_{n-1} \text{unter } \mathbb{P}_\theta;$$

also wollen wir

$$1 - \alpha \le \mathbb{P}_\theta\left[\left|\frac{\overline{X}_n - \mu}{S/\sqrt{n}}\right| \le \frac{\cdots}{S/\sqrt{n}}\right].$$

Um ein möglichst kurzes Intervall zu erhalten, erfüllen wir diese Bedingung mit Gleichheit, und dann brauchen wir gerade

$$\frac{\cdots}{S/\sqrt{n}} = t_{n-1,1-\frac{\alpha}{2}}.$$

Also erhalten wir als Konfidenzintervall für μ zum Niveau $1-\alpha$

$$C(X_1,\dots,X_n) = \left[\overline{X}_n - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X}_n + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right].$$

Die konkreten Realisierungen in unserem Beispiel für $1-\alpha = 99\%$ sind

$$\overline{x}_n = 1156.25, \qquad s = 52.90, \qquad t_{7,0.995} = 3.499$$

und damit

$$C(x_1,\dots,x_n) = [1090.81, 1221.69].$$

Wir sehen, dass sowohl 1100 als auch 1200 in diesem realisierten Intervall liegen. Aufgrund der vorliegenden Daten sind also beide Behauptungen (diejenige von Dr. Thurston und die von Mr. Smith) plausibel.

Um ein Konfidenzintervall für σ^2 zu konstruieren, benutzen wir die ebenfalls aus Satz [to add] bekannte Tatsache, dass

$$\frac{1}{\sigma^2}(n-1)S^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \sim \chi^2_{n-1} \text{unter } \mathbb{P}_\theta.$$

Also ist, mit der Notation $\chi^2_{m,\gamma}$ für das γ-Quantil einer χ^2_m-Verteilung,

$$1-\alpha = \mathbb{P}_\theta\left[\chi^2_{n-1,\frac{\alpha}{2}} \le \frac{1}{\sigma^2}(n-1)S^2 \le \chi^2_{n-1,1-\frac{\alpha}{2}}\right] = \mathbb{P}_\theta\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right],$$

und das Konfidenzintervall für σ^2 zum Niveau $1-\alpha$ wird

$$C(X_1,\dots,X_n) = \left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right].$$

Die konkreten Realisierungen in unserem Beispiel für $1-\alpha = 95\%$ sind

$$s^2 = 2798.21, \qquad \chi^2_{7,0.025} = 1.69, \qquad \chi^2_{7,0.975} = 16.01$$

und damit

$$C(x_1,\dots,x_n) = [1223.45, 11'590.23].$$

Übersetzt für $\sigma = \sqrt{\sigma^2}$ erhalten wir als realisiertes Konfidenzintervall $[34.98, 107.66]$. ⋄

Bemerkung: Im obigen Beispiel haben wir exakte Konfidenzintervalle erhalten, weil wir genügend genaue Verteilungsaussagen zur Verfügung haben. In allgemeinen Situationen kann man oft nur approximative Konfidenzintervalle mit Hilfe des zentralen Grenzwertsatzes bekommen; siehe Abschnitt 2.2.

2.4 Approximative Konfidenzintervalle

Einen allgemeinen **approximativen Zugang** liefert der zentrale Grenzwertsatz. Oft ist ein Schätzer T eine Funktion einer Summe $\sum_{i=1}^n Y_i$, wobei die Y_i im Modell P_θ i.i.d. sind; das einfachste Beispiel ist $T = \overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Nach dem zentralen Grenzwertsatz ist dann für grosse n

$$\sum_{i=1}^n Y_i \quad \text{approximativ normalverteilt unter } \mathbb{P}_\theta$$

mit Parametern $\mu = n\mathbb{E}_\theta[Y_i]$ und $\sigma^2 = n\mathrm{Var}_\theta[Y_i]$. Das kann man benutzen, um für die Verteilung von T approximative Aussagen zu bekommen und damit gewisse Fragen zumindest approximativ zu beantworten.

Beispiel (tea tasting lady): Nehmen wir nochmals an, dass die Lady in $n = 10$ Versuchen insgesamt 6 Tassenpaare richtig klassifiziert hat. Wie können wir dann einen Konfidenzbereich für ihre Erfolgswahrscheinlichkeit bekommen?

Allgemein ist in jedem Modell $\mathbb{P}_\theta$ die Anzahl S_n der Erfolge $Bin(n,\theta)$-verteilt, und wir suchen einen Konfidenzbereich für den unbekannten Parameter θ. Wir wollen den zentralen Grenzwertsatz benutzen, um einen approximativen Konfidenzbereich zu bekommen. Nach dem ZGS ist

$$S_n^* := \frac{S_n - n\theta}{\sqrt{n\theta(1-\theta)}} \overset{\text{Approx}}{\approx} \mathcal{N}(0,1) \text{unter } \mathbb{P}_\theta.$$

Also gilt

$$\mathbb{P}_\theta\left[\left|\frac{S_n - n\theta}{\sqrt{n\theta(1-\theta)}}\right| \le z_{1-\frac{\alpha}{2}}\right] = \mathbb{P}_\theta[|S_n^*| \le z_{1-\frac{\alpha}{2}}] \approx 1-\alpha,$$

und wir können versuchen, diese Ungleichung nach θ aufzulösen. Wir wollen also

$$|S_n - n\theta| \le z_{1-\frac{\alpha}{2}}\sqrt{n\theta(1-\theta)} \qquad \text{oder} \qquad (S_n - n\theta)^2 \le z_{1-\frac{\alpha}{2}}^2 n\theta(1-\theta),$$

aber das wird eher kompliziert. (Man kann die Ungleichung durch eine Gleichung ersetzen und die resultierende quadratische Gleichung für θ lösen; das ergibt mit der Abkürzung $z := z_{1-\frac{\alpha}{2}}$ die zwei Lösungen

$$\hat\theta_\pm = \frac{2S_n + z^2 \pm \sqrt{(2S_n+z^2)^2 - 4S_n^2(1+\frac{z^2}{n})}}{2n(1+\frac{z^2}{n})},$$

und das resultierende approximative Konfidenzintervall ist dann $[\hat\theta_-, \hat\theta_+]$.)

Alternativ kann man wie folgt vorgehen:

Methode 1: Wir gehen davon aus, dass $\theta(1-\theta) \approx \frac{1}{4}$ ist und setzen das ein. Dann wollen wir also

$$|S_n - n\theta| \leq z_{1-\frac{\alpha}{2}} \sqrt{\frac{n}{4}},$$

und das approximative Konfidenzintervall für θ ergibt sich als

$$\left[\overline{S}_n - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}, \overline{S}_n + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right].$$

Methode 2: Wir benutzen den ZGS, um zuerst

$$\overline{S}_n \overset{\text{Approx}}{\approx} \mathcal{N}\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$

zu erhalten. Im approximativen Konfidenzintervall für θ mit Grenzen

$$E_\theta[\overline{S}_n] \pm z_{1-\frac{\alpha}{2}} \sqrt{\mathrm{Var}_\theta[\overline{S}_n]} = \theta \pm \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \sqrt{\theta(1-\theta)}$$

ersetzen wir dann θ durch seinen Schätzer $\overline{S}_n$ und erhalten so das "doppelt approximative" Konfidenzintervall

$$\left[\overline{S}_n - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \sqrt{\overline{S}_n(1-\overline{S}_n)}, \overline{S}_n + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \sqrt{\overline{S}_n(1-\overline{S}_n)}\right].$$

Für $1-\alpha = 95\%$ ist $z_{1-\frac{\alpha}{2}} = 1.96$. Für $n = 10$ und $s_n = 6$ ergeben sich dann als Schätzwert für θ der Wert 0.6 und die realisierten Intervalle $[0.290, 0.910]$ mit Methode 1 und $[0.296, 0.904]$ mit Methode 2. (Mit dem Lösen der quadratischen Gleichung erhält man das realisierte Intervall $[0.3127, 0.8318]$.) Das zeigt zum einen, dass man mit so wenigen Daten (nicht überraschend) keine präzisen Aussagen erwarten kann, und zum anderen auch, dass die zusätzlichen Approximationen hier doch deutliche Abweichungen liefern.

Hätte man stattdessen $n = 100$ Versuche mit $s_n = 60$ Erfolgen, so wäre der Schätzwert für θ unverändert 0.6. Die realisierten approximativen Konfidenzintervalle sind aber wesentlich enger — wir erhalten $[0.502, 0.698]$ mit Methode 1 und $[0.504, 0.696]$ mit Methode 2. (Das Lösen der quadratischen Gleichung liefert das realisierte Intervall $[0.502, 0.691]$.) Hier ergeben also alle drei Ansätze sehr ähnliche Resultate.

$\diamond$

Kapitel 3

Tests

Ziel: Überblick über grundlegende Ideen und Methoden sowie einige Beispiele zum Testen von Hypothesen.

3.1 Null- und Alternativhypothese

Ausgangspunkt ist wie im letzten Abschnitt eine Stichprobe $X_1, \ldots, X_n$. Wir betrachten wieder eine Familie von Wahrscheinlichkeiten P_θ mit $\theta \in \Theta$, die unsere möglichen Modelle beschreiben. Wie bisher kann θ ein- oder mehrdimensional sein. Wir haben schon eine Vermutung, wo in Θ der richtige (aber unbekannte) Parameter θ liegen könnte, und wollen diese mit Hilfe der Daten überprüfen ("testen"). Das Grundproblem ist also, eine Entscheidung zwischen zwei konkurrierenden Modellklassen zu treffen — der **Nullhypothese** $\Theta_0 \subseteq \Theta$ und der **Alternativhypothese** $\Theta_A \subseteq \Theta$, wobei $\Theta_0 \cap \Theta_A = \varnothing$ ist. Meist schreibt man das als

$$\text{Nullhypothese } H_0 : \theta \in \Theta_0,$$
$$\text{Alternativhypothese } H_A : \theta \in \Theta_A.$$

Ist keine explizite Alternative spezifiziert, so hat man $\Theta_A = \Theta_0^c = \Theta \backslash \Theta_0$. Null- und/oder Alternativhypothese heissen **einfach**, falls Θ_0 bzw. Θ_A aus einem einzelnen Wert, θ_0 bzw. θ_A, bestehen, also z.B. $\Theta_0 = \{\theta_0\}$ ist; sonst heissen sie **zusammengesetzt**. Expliziter formuliert ist also die Nullhypothese

$$H_0 : \text{"der wahre (aber unbekannte) Parameter } \theta \text{ liegt in der Menge } \Theta_0\text{"}$$

und die Alternativhypothese

$$H_A : \text{“der wahre Parameter liegt in } \Theta_A\text{”}.$$

Beispiel: Tea testing lady. Eine englische Lady behauptet, bei Tee mit Milch anhand des Geschmacks unterscheiden zu können, ob zuerst die Milch oder zuerst der Tee in die Tasse eingegossen worden ist. Wie kann man überprüfen, ob das stimmen kann?

Wie im letzten Abschnitt stellen wir der Lady an n Tagen die Aufgabe, zwei Tassen (je eine vom Typ 1 und Typ 2) zu klassifizieren; sie soll also angeben, in welche der Tassen zuerst Milch eingegossen worden ist. Wir notieren die Ergebnisse $x_1, \ldots, x_n \in \{0, 1\}$ (falsch bzw. richtig klassifiziert) und fassen wie üblich diese Daten als Realisationen von Zufallsvariable n $X_1, \ldots, X_n$ auf. Dann ist $S_n = \sum_{i=1}^{n} X_i$ die Anzahl der korrekt klassifizierten Tassenpaare.

Als Modelle nehmen wir wieder an, dass die X_i unter $\mathbb{P}_\theta$ i.i.d. $\sim \text{Ber}(\theta)$ mit $\theta \in \Theta = [0, 1]$ sind. Dann ist natürlich $S_n \sim \text{Binomial}(n, \theta)$ unter $\mathbb{P}_\theta$, d.h. im Modell P_θ, das zu θ gehört, ist die Anzahl S_n der Erfolge binomialverteilt mit Parametern n und θ.

Als Skeptiker zweifeln wir an den Fähigkeiten der Lady; wir wählen deshalb als (einfache) Nullhypothese $H_0 : \theta = \frac{1}{2}$, d.h. $\Theta_0 = \{\frac{1}{2}\}$ (“zufälliges Raten — das kann jeder”). Die (zusammengesetzte) Alternativhypothese, dass die Lady besondere Fähigkeiten hat, ist dann

$$H_A : \theta > \frac{1}{2}, \text{d.h. } \Theta_A = \left(\frac{1}{2}, 1\right].$$

Um weiterzukommen, müssen wir nun die Entscheidungsfindung anhand der Daten formalisieren. Auf das Beispiel selbst kommen wir später zurück. ◇

3.2 Test und Entscheidung

Definition 3.1. *Ein **Test** ist ein Paar (T, K), wobei*

- *T eine Zufallsvariable der Form $T = t(X_1, \ldots, X_n)$ ist, und*
- *$K \subseteq \mathbb{R}$ eine (deterministische) Teilmenge von $\mathbb{R}$ ist.*

*Die Zufallsvariable $T = t(X_1, \ldots, X_n)$ heisst dann **Teststatistik**, und K heisst **kritischen Bereich** oder **Verwerfungsbereich**.*

Gegeben seien n Beobachtungen $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$. Ein statistischer Test ermöglicht uns eine systematische Annahme oder Ablehnung der Nullhyphothese H_0. Dabei berechnen wir zuerst die Teststatistik $T(\omega) = t(X_1(\omega), \ldots, X_n(\omega))$ und gehen dann wie folgt vor

Entscheidungsregel:

- die Hypothese H_0 wird verworfen, falls $T(\omega) \in K$,
- die Hypothese H_0 wird nicht verworfen bzw. angenommen , falls $T(\omega) \notin K$.

Beachte: Die Entscheidung des Tests hängt via $T(\omega)$ von der Realisierung ω ab. Weil T eine Zufallsvariable ist, ist die Menge $\{T \in K\}$ ein Ereignis, und wir können ihre Wahrscheinlichkeit $\mathbb{P}_\theta[T \in K]$ in jedem Modell $\mathbb{P}_\theta$ betrachten.

Die Entscheidung bei einem Test kann auf zwei verschiedene Arten falsch herauskommen:

1) Bei einem **Fehler 1. Art** wird die Nullhypothese zu Unrecht verworfen, d.h. obwohl sie richtig ist. Das passiert für $\theta \in \Theta_0$ und $T \in K$; deshalb heisst $\mathbb{P}_\theta[T \in K]$ für $\theta \in \Theta_0$ die Wahrscheinlichkeit für einen Fehler 1. Art.

2) Bei einem **Fehler 2. Art** wird die Nullhypothese zu Unrecht *nicht* verworfen, d.h. man akzeptiert die Nullhypothese (verwirft sie nicht), obwohl sie falsch ist. Das passiert für $\theta \in \Theta_A$ und $T \notin K$, und deshalb heisst $\mathbb{P}_\theta[T \notin K] = 1 - \mathbb{P}_\theta[T \in K]$ für $\theta \in \Theta_A$ die Wahrscheinlichkeit für einen Fehler 2. Art.

Beispiel: tea testing lady.

Ein Fehler 1. Art besteht hier darin, die Nullhypothese des zufälligen Ratens abzulehnen, obwohl sie richtig ist. Anders gesagt glaubt man hier bei einem Fehler 1. Art an verborgene Fähigkeiten der Lady, obwohl ihre Ergebnisse durch zufälliges Raten entstehen (könnten). Bei einem Fehler 2. Art dagegen glaubt man nicht an die Fähigkeiten der Lady, obwohl diese durchaus vorhanden sind. ◇

3.3 Signifikanzniveau und Macht

Bei der Auswahl eines geeigneten Tests ist insbesondere die Minimierung von Fehlern 1. Art entscheidend. Ein Fehler 1. Art tritt ein, falls wir H_0 ablehnen (aufgrund von

$\mathbb{P}_\theta[T \in K]$), obwohl H_0 erfüllt ist.

Wir würden somit gerne Tests benutzen, welche eine geringe Chance auf einen Fehler 1. Art aufweisen. Hierfür definieren wir das Signifikanzniveau eines Test.

Definition 3.2. *Sei $\alpha \in (0,1)$. Ein Test (T, K) besitzt **Signifikanzniveau** α, falls*

$$\forall \theta \in \Theta_0 \quad \mathbb{P}_\theta[T \in K] \leq \alpha.$$

Als zweites Ziel wollen wir zudem Fehler 2. Art vermeiden. Dies führt direkt zur Definition der Macht.

Definition 3.3. *Die **Macht** eines Tests (T, K) wird definiert als folgende Funktion*

$$\beta : \Theta_A \to [0,1], \qquad \theta \mapsto \beta(\theta) := \mathbb{P}_\theta[T \in K],$$

Unser erstes Ziel ist wie gesagt die Minimierung Fehler 1. Art. Hierfür fixieren wir einen Parameter α, und designen einen Test zum Signifikanzniveau α.

Als zweites Ziel wollen wir einen Fehler 2. Art vermeiden. Nachdem wir einen Test mit Signifikanzniveau α gefunden haben, suchen wir nach dem Test mit der grössten Macht. Äquivalent formuliert: Minimiere die Grösse $1 - \beta(\theta) = \mathbb{P}_\theta[T \notin K]$ für $\theta \in \Theta_A$, also die Wahrscheinlchkeit für einen Fehler 2. Art.

Das obige asymmetrische Vorgehen macht es schwieriger, die Nullhypothese zu verwerfen als sie beizubehalten. Ein seriöser Test wird deshalb als Nullhypothese immer die Negation der eigentlich gewünschten Aussage benutzen. Gelingt es dann nämlich, das trotz der erschwerten Bedingungen zu verwerfen, so kann man viel eher zuversichtlich sein, tatsächlich einen Effekt gefunden zu haben.

Aus der asymmetrischen Behandlung von H_0 und H_A folgt auch, dass die Entscheidung bei einem Test davon abhängt, was man als Nullhypothese und was als Alternativhypothese wählt. Es kann also passieren, dass die gleiche inhaltliche Frage zu unterschiedlichen Entscheidungen führt, wenn man bei ihrem Test Nullhypothese und Alternativhypothese vertauscht. Wir werden das später mit einem Beispiel explizit illustrieren.

Wichtig: Die Entscheidung bei einem Test ist nie ein Beweis, sondern immer nur eine **Interpretation** der Übereinstimmung zwischen Daten und vermutetem Modell. Ist $T(\omega) \in K$, so wird man die Nullhypothese ablehnen und wegen der Daten nicht mehr glauben, dass $\theta \in \Theta_0$ ist. Das kann (muss aber nicht) zur Konsequenz haben, dass man

eher glaubt, dass $\theta \in \Theta_A$ ist, so dass man die Alternativhypothese für plausibler hält. Ist $T(\omega) \notin K$, so wird man die Nullhypothese nicht verwerfen und sich im Glauben bestärkt fühlen, dass $\theta \in \Theta_0$ ist. **Wo aber θ tatsächlich liegt, weiss man genauso wenig wie vorher — ein Test liefert keinen Beweis!**

Beispiel: tea testing lady.

Unter $\mathbb{P}_\theta$ sind die Zufallsvariablen $X_1, \dots, X_n$ wieder i.i.d. $\sim \mathrm{Ber}(\theta)$ sowie $S_n = \sum_{i=1}^n X_i \sim \mathrm{Bin}(n, \theta)$. Hypothese und Alternative sind hier $H_0 : \theta = \frac{1}{2}$ und $H_A : \theta > \frac{1}{2}$.

Weil man für $\theta > \frac{1}{2}$ eher Einsen bei den X_i erwartet als für $\theta = \frac{1}{2}$, deuten viele Einsen bzw. ein grosser Wert von S_n eher auf H_A als H_0 hin. Ein plausibler Test könnte also die Teststatistik $T = S_n$ und einen kritischen Bereich der Form $K = (c, \infty)$ nehmen, d.h. man verwirft die Nullhypothese (des zufälligen Ratens), wenn die Lady viele Erfolge erzielt.

Als Nullhypothese haben wir hier $\theta = \frac{1}{2}$, d.h. "keine besonderen Fähigkeiten"; wir möchten zwar an diese Fähigkeiten eigentlich gerne glauben, aber natürlich nur, wenn sie wirklich von den Daten überzeugend gestützt werden. Wie oben erklärt machen wir es der Lady also bewusst schwer, um bei einem positiven Ergebnis zuversichtlich an ihre Fähigkeiten glauben zu können.

Um den kritischen Wert c zu einem Signifikanzniveau α zu bestimmen, brauchen wir die Wahrscheinlichkeiten $\mathbb{P}_\theta[T \in K] = \mathbb{P}_\theta[S_n > c]$ für $\theta = \frac{1}{2}$; für die Machtfunktion brauchen wir auch $\beta(\theta) = \mathbb{P}_\theta[T \in K] = \mathbb{P}_\theta[S_n > c]$ für $\theta > \frac{1}{2}$.

Allgemein formuliert bedeutet das, dass wir die Verteilung der Teststatistik T unter jedem $\mathbb{P}_\theta$ (d.h. in jedem Modell) brauchen, um solche Wahrscheinlichkeiten ausrechnen zu können. Das ist in der Regel nicht möglich; um aber zumindest das Signifikanzniveau einhalten zu können, brauchen wir wenigstens die Verteilung von T unter der Nullhypothese H_0, d.h. in jedem Modell $\mathbb{P}_\theta$ mit $\theta \in \Theta_0$ — und wenn wir diese Verteilung nicht exakt kennen, so brauchen wir sie mindestens approximativ.

Sei nun $n = 10$, d.h. wir lassen die Lady 10 Tage lang probieren. Die folgende Tabelle gibt dann die Binomial-Wahrscheinlichkeiten $\mathbb{P}_\theta[S_{10} > k]$ für verschiedene θ und k.

θ	$k = 7$	$k = 8$	$k = 9$	$k = 10$
0.7	0.3828	0.1493	0.0282	0
0.6	0.1673	0.0464	0.0060	0
0.5	0.0547	0.0107	0.0010	0

Damit wir ein Signifikanzniveau von α erhalten, muss $\mathbb{P}_{\frac{1}{2}}[S_{10} > c] \leq \alpha$ sein. Wählen wir $c = 7$, so ist das Niveau $\alpha = 0.0547$, also rund 5%. Auf diesem Niveau sind wir also bereit, bei 8 oder mehr Erfolgen der Lady unsere skeptische Nullhypothese zu verwerfen und an ihre Fähigkeiten zu glauben.

Die Macht des Tests erhalten wir auch aus der Tabelle; z.B. ist für das gewählte $c = 7$

$$\beta(0.6) = \mathbb{P}_{0.6}[S_{10} > 7] = 0.1673$$

oder $\beta(0.7) = 0.3828$. Wir sehen also, dass

$$1 - \beta(\theta) = 1 - \mathbb{P}_\theta[S_{10} > 7] = \mathbb{P}_\theta[S_{10} \leq 7]$$

für $\theta \in \Theta_A$ recht gross wird; der Test hat eine beträchtliche Wahrscheinlichkeit für einen Fehler 2. Art, d.h. für einen Unglauben an tatsächlich vorhandene Fähigkeiten. Das ist die Kehrseite unseres allgemeinen skeptischen Ansatzes; bei nur schwachen Indikationen (p grösser als $\frac{1}{2}$, aber nicht sehr viel grösser) kann es durchaus vorkommen, dass wir diese Fähigkeiten zu Unrecht nicht bemerken. ◇

Bemerkung 3.4. *Weil die Teststatistik T im obigen Beispiel diskret ist, kann ein vorgegebenes Niveau α in der Regel nicht genau eingehalten werden, d.h. es ist unmöglich, einen kritischen Bereich K mit $\mathbb{P}_{\theta_0}[T \in K] = \alpha$ zu finden. (Falls Θ_0 aus mehr als nur einem einzelnen θ_0 besteht, so ist das sowieso schwierig; im diskreten Fall ist das aber sogar schon für eine einfache Nullhypothese $\Theta_0 = \{\theta_0\}$ ein Problem.) Einen Ausweg bietet ein sogenannter* ***randomisierter Test****. Man wählt dazu $\gamma \in [0,1]$ so, dass gilt $\gamma\mathbb{P}_{\theta_0}[T > c] + (1-\gamma)P_{\theta_0}[T > c+1] = \alpha$ und entscheidet dann wie folgt: Ist $T > c$, so verwirft man H_0 mit Wahrscheinlichkeit γ, d.h. H_0 wird abgelehnt, falls erstens $T > c$ gilt und zweitens eine unabhängige $\mathcal{U}(0,1)$-verteilte Zufallsvariable einen Wert $\leq \gamma$ realisiert.*

Im obigen Beispiel würde man so das exakte Niveau $\alpha = 5\%$ mit $c = 7$ und

$$\gamma = \frac{\alpha - \mathbb{P}_{\theta_0}[T > c+1]}{\mathbb{P}_{\theta_0}[T > c] - \mathbb{P}_{\theta_0}[T > c+1]} = 0.893$$

erreichen.

Im obigen Beispiel haben wir die Wahl der Teststatistik $T = S_n$ und des kritischen Bereichs $K = (c, \infty)$ mit Plausibilitätsargumenten motiviert. Wir wollen nun kurz einen systematischen Ansatz erklären, der in vielen Situationen zu einem optimalen Test führt. Die Idee dazu geht auf Neyman und Pearson zurück, un wird in nächste Sektion beschreibt.

3.4 Konstruktion von Tests

Seien $\theta_0 \neq \theta_A$ zwei fixierte Zahlen. In diesem Abschnitt nehmen wir stets an, dass sowohl die Nullhypothese als auch die Alternativhyphothese von der einfachen Form

$$H_0 : \theta = \theta_0,$$
$$H_A : \theta = \theta_A,$$

ist. Ferner nehmen wir an, dass die Zufallsvariablen $X_1, \ldots, X_n$ entweder diskret, oder gemeinsam stetig unter $\mathbb{P}_{\theta_0}$ und unter $\mathbb{P}_{\theta_A}$ sind. Somit ist nach Annahme auch die Likelihood-Funktion $L(x_1, \ldots, x_n; \theta)$ für $\theta = \theta_0$ und $\theta = \theta_A$ wohldefiniert (Siehe Definition 1.4)

Definition 3.5. *Für jedes $x_1, \ldots, x_n$, definieren wir den **Likelihood-Quotienten** durch*

$$R(x_1, \ldots, x_n) := \frac{L(x_1, \ldots, x_n; \theta_A)}{L(x_1, \ldots, x_n; \theta_0)}.$$

Als Konvention setzten wir $R(x_1, \ldots, x_n) = +\infty$, falls $L(x_1, \ldots, x_n; \theta_0) = 0$.

Intuitiv gibt es nun direkt mehrere Aussagen. Für einen grossen Quotienten, ist der Zähler wesentlich grösser als der Nenner. Anschaulich bedeutet dies, dass die Beobachtungen $x_1, \ldots, x_n$ als Resultate im Modell $\mathbb{P}_{\theta_A}$ deutlich wahrscheinlicher sind als im Modell $\mathbb{P}_{\theta_0}$. Die Daten widersprechen θ_0 im Vergleich zu θ_A. Es liegt deshalb nahe, als Teststatistik $T := R(X_1, \ldots, X_n)$ und als kritischen Bereich $K := (c, \infty)$ zu wählen, wenn man θ_0 gegen θ_A testen will. Schließlich wird gerade die Hypothese H_0 verworfen, falls der Quotient R gross ist.

Definition 3.6. *Sei $c \geq 0$. Der **Likelihood-Quotienten-Test mit Parameter** c ist*

ein Test (T, K), wobei Teststatistik und Verwerfungsbereich gegeben sind durch

$$T = R(X_1, \ldots, X_n) \quad \textit{und} \quad K = (c, \infty].$$

Der **Likelihood-Quotienten-Test** ist dann im folgenden Sinn optimal: Jeder andere Test mit kleiner Signifikanzniveau hat auch kleinere Macht bzw. eine grössere Wahrscheinlichkeit für einen Fehler 2. Art.

Theorem 3.7 (Neyman–Pearson-Lemma)**.** *Sei $c \geq 0$. Sei (T, K) ein Likelihood-Quotienten-Test mit Parameter c und Signifikanzniveau $\alpha^* := \mathbb{P}_{\theta_0}[T > c]$. Ist (T', K') ein anderer Test mit Signifikanzniveau $\alpha \leq \alpha^*$, so gilt*

$$\mathbb{P}_{\theta_A}[T' \in K'] \leq \mathbb{P}_{\theta_A}[T \in K].$$

Beweis. Siehe [Krengel, Satz 6.2].

Die obige Situation mit einfacher Hypothese und Alternative ist so speziell, dass sie in der Praxis kaum je auftritt. Die Grundidee für den Test lässt sich aber verallgemeinern und liefert in gewissen (weniger restriktiven) Situationen immer noch gute oder optimale Tests, so dass man das Vorgehen mit gutem Gewissen als einen systematischen Ansatz empfehlen kann. Wie wir in Beispielen sehen werden, sind die resultierenden Tests oft auch intuitiv sehr einleuchtend.

Etwas genauer betrachtet man bei zusammengesetzten Hypothesen und Alternativen den sogenannten **verallgemeinerten Likelihood-Quotienten**

$$R(x_1, \ldots, x_n) := \frac{\sup_{\theta \in \Theta_A} L(x_1, \ldots, x_n; \theta)}{\sup_{\theta \in \Theta_0} L(x_1, \ldots, x_n; \theta)}$$

oder auch

$$\tilde{R}(x_1, \ldots, x_n) := \frac{\sup_{\theta \in \Theta_A \cup \Theta_0} L(x_1, \ldots, x_n; \theta)}{\sup_{\theta \in \Theta_0} L(x_1, \ldots, x_n; \theta)}$$

und wählt als Teststatistik $T_0 := R(X_1, \ldots, X_n)$ bzw. $\tilde{T} := \tilde{R}(X_1, \ldots, X_n)$ mit kritischem Bereich $K_0 := (c_0, \infty)$. Durch Umformen erhält man daraus oft einen äquivalenten, aber einfacheren Test (T, K) von einer leicht anderen Form; siehe Beispiele. Die Konstante c_0 bzw. den Bereich K muss man dabei noch so wählen, dass der Test ein in der Regel a priori gewähltes Signifikanzniveau einhält.

Beispiel: tea tasting lady.

Im Modell $\mathbb{P}_\theta$ sind $X_1, \ldots, X_n$ i.i.d. $\sim Be(\theta)$; die Gewichtsfunktion eines X_i unter $\mathbb{P}_\theta$ ist also $p_X(x_i;\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$, und damit wird die Likelihood-Funktion

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} p_X(x_i;\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Der Likelihood-Quotient ist also

$$\begin{aligned} R(x_1, \ldots, x_n; \theta_0, \theta_A) &= \frac{L(x_1, \ldots, x_n; \theta_A)}{L(x_1, \ldots, x_n; \theta_0)} \\ &= \left(\frac{\theta_A}{\theta_0}\right)^{\sum_{i=1}^n x_i} \left(\frac{1-\theta_A}{1-\theta_0}\right)^{n-\sum_{i=1}^n x_i} \\ &= \left(\frac{\theta_A(1-\theta_0)}{\theta_0(1-\theta_A)}\right)^{\sum_{i=1}^n x_i} \left(\frac{1-\theta_A}{1-\theta_0}\right)^{n}. \end{aligned}$$

Nun ist ja $\theta_0 = \frac{1}{2}$ und $\theta_A > \frac{1}{2}$, also $\theta_0 < \theta_A$. Damit ist

$$\frac{\theta_A(1-\theta_0)}{\theta_0(1-\theta_A)} = \frac{\theta_A - \theta_0\theta_A}{\theta_0 - \theta_0\theta_A} > 1,$$

und damit ist $R(x_1, \ldots, x_n; \theta_0, \theta_A)$ genau dann gross, wenn der Exponent $\sum_{i=1}^n x_i$ gross ist. Statt des komplizierten Quotienten wählen wir als Teststatistik also

$$T := \sum_{i=1}^{n} X_i = S_n,$$

und der kritische Bereich "Quotient gross" hat die äquivalente Form "Summe (= Exponent) gross", also

$$K := (c, \infty).$$

Also liefert hier der Neyman–Pearson-Ansatz genau das Testverfahren, das wir oben schon aufgrund von Plausibilitätsargumenten benutzt haben. ◇

Beispiel: Seien $X_1, \ldots, X_n$ unter $\mathbb{P}_\theta$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ mit **bekannter Varianz** σ^2; der unbekannte Parameter ist hier also $\theta = \mu \in \mathbb{R}$. Die Dichtefunktion von X_i ist

$$f_X(x_i;\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right),$$

die Likelihood-Funktion ist

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_X(x_i;\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right),$$

und der Likelihood-Quotient wird

$$\begin{aligned} R(x_1,\dots,x_n;\theta_0,\theta_A) &= \frac{L(x_1,\dots,x_n;\theta_A)}{L(x_1,\dots,x_n;\theta_0)} \\ &= \exp\left(-\frac{1}{2\sigma^2}\Big(\sum_{i=1}^n (x_i-\theta_A)^2 - \sum_{i=1}^n (x_i-\theta_0)^2 \Big)\right) \\ &= \text{const.}(\sigma,\theta_0,\theta_A) \exp\left(\frac{1}{\sigma^2}(\theta_A-\theta_0) \sum_{i=1}^n x_i \right). \end{aligned}$$

Betrachten wir nun die Hypothese $H_0 : \theta = \theta_0$ und die Alternative $H_A : \theta = \theta_A$. Der obige Quotient wird tendenziell gross, falls der Exponent $(\theta_A - \theta_0)\sum_{i=1}^n x_i$ gross ist. Was das für $\sum_{i=1}^n x_i$ bedeutet, hängt vom Vorzeichen von $\theta_A - \theta_0$ ab. In jedem Fall wählen wir als Teststatistik

$$T' := \sum_{i=1}^n X_i.$$

Ist $\theta_A > \theta_0$, so ist $\theta_A - \theta_0 > 0$, und der Exponent wird dann gross, wenn T' gross ist; hier wählen wir den kritischen Bereich also von der Form $K'_> := (c'_>, \infty)$, d.h. wir lehnen H_0 ab, wenn T' gross ist.

Ist $\theta_A < \theta_0$, so ist $\theta_A - \theta_0 < 0$ und der Exponent gross für T' klein (d.h. negativ). Hier ist also der kritische Bereich von der Form $K'_< := (-\infty, c'_<)$.

In beiden Fällen müssen wir den kritischen Bereich, d.h. hier konkret die Konstanten $c'_>$ bzw. $c'_<$, noch so festlegen, dass der Test ein gewähltes Signifikanzniveau α einhält. Wir wollen also $\mathbb{P}_{\theta_0}[T' \in K'] \le \alpha$ erreichen, und um diese Wahrscheinlichkeit zu berechnen, brauchen wir die Verteilung der Teststatistik T' unter $\mathbb{P}_{\theta_0}$, d.h. unter der Hypothese H_0.

Im vorliegenden Fall ist das einfach. Unter jedem $\mathbb{P}_\theta$ sind die X_i i.i.d. $\sim \mathcal{N}(\theta, \sigma^2)$; also ist die Summe

$$T' = \sum_{i=1}^n X_i \sim \mathcal{N}(n\theta, n\sigma^2) \text{ unter } \mathbb{P}_\theta.$$

Äquivalent ist

$$T = \frac{\overline{X}_n - \theta}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \text{ unter } \mathbb{P}_\theta,$$

und wir können T statt T' als Teststatistik benutzen.

Man beachte, dass T im Modell $\mathbb{P}_\theta$ mit $\theta \in \Theta_0$, d.h. mit $\theta = \theta_0$, also unter der Hypothese H_0, tatsächlich berechenbar ist: die Varianz σ^2 ist nach Annahme bekannt, und der zu testende Erwartungswert θ_0 ist natürlich auch bekannt. (Dasselbe gilt auch für T'; die Verteilung von T unter der Nullhypothese H_0 ist aber einfacher als diejenige von T'.) $\diamond$

Das obige Beispiel zeigt den letzten Schritt, den wir allgemein noch machen müssen. Um den kritischen Bereich K passend zum gewünschten Niveau α festlegen zu können, brauchen wir die Verteilung der Teststatistik T unter der Hypothese H_0, d.h. in jedem Modell $\mathbb{P}_\theta$ mit $\theta \in \Theta_0$. Falls wir diese Verteilung(en) nicht exakt kennen, so ist es wichtig, zumindest eine gute Approximation zu haben.

3.5 Beispiele

In diesem Abschnitt illustrieren wir die obigen Überlegungen durch einige Beispiele. Dabei verzichten wir weitgehend auf Herleitungen und präsentieren nur die Ergebnisse mehr oder weniger in der Form von "Kochrezepten".

Beispiel: Normalverteilung, Test für Erwartungswert bei **bekannter Varianz:** Dieser Test ist unter dem Namen z**-Test** bekannt. Hier sind $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}(\theta, \sigma^2)$ unter $\mathbb{P}_\theta$ mit bekannter Varianz σ^2, und wir wollen die Hypothese $H_0 : \theta = \theta_0$ testen. Mögliche Alternativen H_A sind $\theta > \theta_0$ oder $\theta < \theta_0$ (**einseitig**), oder $\theta \neq \theta_0$ (**zweiseitig**). Welche der Alternativen sinnvoll ist, hängt von der konkreten Fragestellung ab.

Die Teststatistik hier ist in jedem Fall (siehe letztes Beispiel im Abschnitt 3.4)

$$T := \frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \text{unter } \mathbb{P}_{\theta_0}.$$

Der kritische Bereich K ist von der Form $(c_>, \infty)$ für den einseitigen Test gegen die Alternative $H_A : \theta > \theta_0$, bzw. $(-\infty, c_<)$, bzw. $(-\infty, -c_{\neq}) \cup (c_{\neq}, +\infty)$. Im zweiseitigen Fall verwirft man H_0 also zugunsten der Alternative $H_A : \theta \neq \theta_0$, falls $|T| > c_{\neq}$ ist.

Die Konstanten $c_>$, $c_<$, $c_{\neq}$ bestimmt man zum gewählten Niveau mit Hilfe der Verteilung von T unter $\mathbb{P}_{\theta_0}$. Zum Beispiel liefert die Bedingung

$$\alpha = \mathbb{P}_{\theta_0}[T \in K_>] = \mathbb{P}_{\theta_0}[T > c_>] = 1 - \mathbb{P}_{\theta_0}[T \leq c_>] = 1 - \Phi(c_>),$$

dass $c_> = \Phi^{-1}(1-\alpha) =: z_{1-\alpha}$ das sogenannte $(1-\alpha)$**-Quantil** der $\mathcal{N}(0,1)$-Verteilung sein muss; für $\theta > \theta_0$ verwirft man also H_0, falls

$$\overline{X}_n > \theta_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

ist. Analog ist $c_< = z_\alpha = -z_{1-\alpha}$ und $c_{\neq} = z_{1-\frac{\alpha}{2}}$, wobei wir die Symmetrie der $\mathcal{N}(0,1)$-Verteilung ausnutzen; es gilt nämlich

$$\alpha = \mathbb{P}_{\theta_0}[T < c_<] = \mathbb{P}_{\theta_0}[T > -c_<] \text{für } -c_< = z_{1-\alpha}$$

und

$$\alpha = \mathbb{P}_{\theta_0}[T \in K_{\neq}] = \mathbb{P}_{\theta_0}[T < -c_{\neq}] + \mathbb{P}_{\theta_0}[T > c_{\neq}] = \Phi(-c_{\neq}) + 1 - \Phi(c_{\neq}) = 2\big(1 - \Phi(c_{\neq})\big).$$

Die benötigten Quantile findet man in einer Tabelle für die Standard-Normalverteilung; dort ist die Funktion $z \mapsto \Phi(z)$ tabelliert, so dass man in der Tabelle bei den Ergebnissen den Wert $1-\alpha$ suchen und dann zurück zu $\Phi^{-1}(1-\alpha) = z_{1-\alpha}$ gehen kann. Alternativ findet man Quantile der Standard-Normalverteilung oft auch in Tabellen der t-Verteilung für Anzahl der Freiheitsgrade $n = \infty$.

Beispiel: Strausseneier

Die Australier Mr. Smith und Dr. Thurston streiten sich über das Durchschnittsgewicht von **Strausseneiern**. Beide sind damit einverstanden, das Gewicht approximativ als normalverteilt aufzufassen; Mr. Smith behauptet aber, das mittlere Gewicht sei 1100g, während Dr. Thurston darauf besteht, dass die Eier schwerer seien, und zwar im Schnitt 1200g. Um ihren Streit beilegen zu können, reisen die beiden nach Afrika, um in der Savanne Strausseneier zu suchen. Weil diese aber meistens gut versteckt sind, finden sie nur acht, und zwar mit folgenden Gewichten (in g): 1090, 1150, 1170, 1080, 1210, 1230, 1180, 1140.

Dr. Thurston schlägt nun vor, Mr. Smiths Behauptung als Hypothese $\mu = \mu_0 = 1100$ gegen seine Alternative $\mu > 1100$ (oder auch $\mu = 1200$) auf dem 5%-Niveau zu testen. Die Varianz σ^2 ist beiden bekannt; sie beträgt (in g) $\sigma = 55$. Also berechnet Dr. Thurston

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = 1156.25$$

und sucht in der Tabelle $z_{1-\alpha} = z_{0.95} = 1.645$. Damit ist

$$T_{\mathrm{Th}}(\omega) = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} = \sqrt{8} \times \frac{1156.25 - 1100}{55} = 2.89.$$

Wegen $T_{\mathrm{Th}}(\omega) > z_{0.95}$ wird also die Hypothese $\mu = 1100$ auf dem 5%-Niveau verworfen.

Mr. Smith kommt sich bei diesem Vorgehen benachteiligt vor und macht deshalb den Gegenvorschlag, doch besser Dr. Thurstons Behauptung als Hypothese $\mu = \mu_1 = 1200$ gegen seine Alternative $\mu < 1200$ (oder auch $\mu = 1100$) zu testen. Er berechnet deshalb

$$T_{\mathrm{Sm}}(\omega) = \frac{\overline{x}_n - \mu_1}{\sigma/\sqrt{n}} = \sqrt{8} \times \frac{1156.25 - 1200}{55} = -2.25.$$

Wegen $z_\alpha = z_{0.05} = -z_{0.95} = -1.645$ ist $T_{\mathrm{Sm}}(\omega) < z_{0.05}$; also wird auch die Hypothese $\mu = 1200$ auf dem 5%-Niveau verworfen.

Dieses Beispiel illustriert sehr schön die Bedeutung der Wahl von Hypothese und Alternative und auch ihre asymmetrische Behandlung. Mit dem ersten Test würde man Dr. Thurston Recht geben, mit dem zweiten hingegen Mr. Smith — und das bei völlig identischen Daten. ◇

Beispiel: Normalverteilung, Test für Erwartungswert bei unbekannter Varianz: Dieser Test ist unter dem Namen t-**Test** bekannt. Hier sind $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ unter $\mathbb{P}_{\vec{\theta}}$, wobei $\vec{\theta} = (\mu, \sigma^2)$ und insbesondere die Varianz σ^2 unbekannt ist. Wir wollen wieder die Hypothese $\mu = \mu_0$ testen.

Genaugenommen ist das eine zusammengesetzte Hypothese, weil der Parameter $\vec{\theta}$ aus den zwei Komponenten μ und σ^2 besteht. Explizit wäre also

$$\Theta_0 = \{\mu_0\} \times (0, \infty) = \{\vec{\theta} = (\mu, \sigma^2) : \mu = \mu_0\}.$$

Die Teststatistik ist hier

$$T := \frac{\overline{X}_n - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \text{unter } \mathbb{P}_{\theta_0};$$

wir ersetzen also die unbekannte Varianz durch den Schätzer

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

für σ^2 und nutzen die Verteilungsaussagen aus Satz ?? aus.

Der kritische Bereich hat (je nach Alternative) eine der drei Formen aus dem letzten Beispiel; die kritischen Werte hier sind $c_> = t_{n-1,1-\alpha}$, bzw. $c_< = t_{n-1,\alpha} = -t_{n-1,1-\alpha}$, bzw. $c_{\neq} = t_{n-1,1-\frac{\alpha}{2}}$. Hier bezeichnen wir mit $t_{m,\gamma}$ das sogenannte γ-Quantil einer t_m-Verteilung, d.h. denjenigen Wert $t_{m,\gamma}$, für den gilt $P[X \leq t_{m,\gamma}] = \gamma$ für X t-verteilt mit m Freiheitsgraden, d.h. $X \sim t_m$. Diese Werte findet man in Tabellen. ◇

Beispiel: Strausseneier

Mr. Smith und Dr. Thurston fragen sich, ob sie bei ihrem ersten Versuch vielleicht eine falsche Information über die Varianz von Strausseneiern benutzt haben. Sie beschliessen deshalb, ihre Tests nochmals ohne die Annahme einer bekannten Varianz durchzuführen, und kommen damit zu einem t-Test.

Dr. Thurston beharrt immer noch darauf, die Hypothese $\mu = 1100$ gegen die Alternative $\mu > 1100$ auf dem 5%-Niveau zu testen. Weil die Varianz σ^2 nun aber unbekannt ist, berechnet er

$$s^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x}_n)^2\right) = 2798.21, \quad \text{also } s = 52.90; \qquad t_{7,0.95} = 1.895.$$

Damit erhält er als Wert für die Teststatistik

$$\tilde{T}_{\mathrm{Th}}(\omega) = \frac{\overline{x}_n - \mu_0}{s/\sqrt{n}} = \sqrt{8} \times \frac{1156.25 - 1100}{52.90} = 3.008.$$

Wegen $\tilde{T}_{\mathrm{Th}}(\omega) > t_{7,0.95}$ wird also die Hypothese $\mu = 1100$ auf dem 5%-Niveau wieder verworfen.

Nicht überraschend ist Mr. Smith mit diesem Vorgehen immer noch nicht einverstanden und will lieber Dr. Thurstons Behauptung als Hypothese $\mu = \mu_1 = 1200$ gegen seine Alternative $\mu < 1200$ (oder auch $\mu = 1100$) testen. Er berechnet deshalb

$$\tilde{T}_{\mathrm{Sm}}(\omega) = \frac{\overline{x}_n - \mu_1}{s/\sqrt{n}} = \sqrt{8} \times \frac{1156.25 - 1200}{52.90} = -2.339.$$

Wegen $t_{n-1,\alpha} = t_{7,0.05} = -t_{7,0.95} = -1.895$ ist $T_{\mathrm{Sm}}(\omega) < -t_{7,0.95}$; also wird auch die Hypothese $\mu = 1200$ auf dem 5%-Niveau verworfen — und damit sind die beiden wieder gleich weit wie vorher. ◇

Die obigen zwei Tests heissen auch **Einstichproben-Tests**, weil man nur Daten aus einer Stichprobe hat. Bei **Zweistichproben-Tests** geht man aus von Zufallsvariable n $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$, die **unter $\mathbb{P}_\theta$ alle unabhängig** sind; zudem sind die X_i und die Y_j unter $\mathbb{P}_\theta$ jeweils für sich betrachtet i.i.d.

Beispiel: Gepaarter Zweistichproben-Test bei **Normalverteilung:** Hier sind $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu_X, \sigma^2)$ und $Y_1, \ldots, Y_n$ i.i.d. $\sim \mathcal{N}(\mu_Y, \sigma^2)$ unter $\mathbb{P}_\theta$; insbesondere ist $m = n$ und die Varianz σ^2 bei beiden Stichproben dieselbe. Eine solche Situation tritt auf, wenn z.B. eine Gruppe von Personen zwei verschiedene Dinge ausprobiert, so dass man eine natürliche Paarbildung zwischen den X_i und Y_i hat.

In dieser Situation kann man Tests über den Vergleich von μ_X und μ_Y auf den Fall nur einer Stichprobe zurückführen; die Differenzen $Z_i := X_i - Y_i$ sind nämlich unter $\mathbb{P}_\theta$ i.i.d. $\sim \mathcal{N}(\mu_X - \mu_Y, 2\sigma^2)$. Damit kann man die bisherigen Tests in leicht angepasster Form benutzen, sowohl für bekannte wie für unbekannte Varianz σ^2. Die resultierenden Tests heissen dann nicht überraschend **gepaarter Zweistichproben-z-Test** (bei bekanntem σ^2) bzw. **gepaarter Zweistichproben-t-Test** (bei unbekanntem σ^2). ◇

Bemerkung 3.8. *Wir haben oben angenommen, dass X_i und Y_i unabhängig sind. Allgemeiner kann man annehmen, dass die Paare (X_i, Y_i), $i = 1, \ldots, n$, unter $\mathbb{P}_\theta$ unabhängig sind mit einer zweidimensionalen Normalverteilung mit Erwartungswerten μ_X, μ_Y, bekannten gleichen Varianzen σ^2 und bekannter Korrelation $\varrho \in (-1, +1)$. (Der Fall $\varrho = 0$ ent-*

spricht Unabhängigkeit.) Dann sind die $Z_i = X_i - Y_i$ unter $\mathbb{P}_\theta$ i.i.d. $\sim \mathcal{N}(\mu_X - \mu_Y, 2(1-\varrho)\sigma^2)$, und man kann wie oben die bisherigen Tests benutzen.

Beispiel: Ungepaarter Zweistichproben-Test bei **Normalverteilung:** Hier sind unter $\mathbb{P}_\theta$ $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu_X, \sigma^2)$ und $Y_1, \ldots, Y_m$ i.i.d. $\sim \mathcal{N}(\mu_Y, \sigma^2)$, wobei die Varianz in beiden Fällen dieselbe ist, aber $m \neq n$ sein kann. Will man einen Vergleich über μ_X und μ_Y hier testen, so kann man nicht mehr paarweise Differenzen bilden. Diesen Test muss man auch benutzen, falls zufällig $m = n$ ist, aber die Daten nicht natürlich gepaart sind. Wir nehmen immer noch an, dass $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ unabhängig sind.

a) Ist σ^2 **bekannt**, so ist die Teststatistik

$$T := \frac{(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathcal{N}(0,1)$$

unter jedem $\mathbb{P}_\theta$. Dabei ist σ nach Annahme bekannt, und $\mu_X - \mu_Y$ muss sich aus der gewünschten Hypothese H_0 als bekannt ergeben. Die kritischen Werte für den Verwerfungsbereich sind wie oben geeignete Quantile der $\mathcal{N}(0,1)$-Verteilung, je nach Alternative. Das ist der **ungepaarte Zweistichproben-z-Test**.

b) Ist σ^2 **unbekannt**, so brauchen wir zuerst die beiden empirischen Varianzen

$$S_X^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$
$$S_Y^2 := \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2.$$

Mit

$$\begin{aligned} S^2 &:= \frac{1}{m+n-2}\big((n-1)S_X^2 + (m-1)S_Y^2\big) \\ &= \frac{1}{m+n-2}\left(\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2\right) \end{aligned}$$

ist dann die Teststatistik

$$T := \frac{(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

unter jedem $\mathbb{P}_\theta$. Der Rest geht dann analog wie oben. Dieser Test heisst **ungepaarter Zweistichproben-t-Test**. ◇

Die meisten bisherigen Beispiele für Tests gehen von der Annahme normalverteilter Stichproben aus; diese Situation ist sehr angenehm, weil man dann die Verteilung der Teststatistik einfach in expliziter Form hat. Diese Tests sind sehr gut, falls man tatsächlich Normalverteilungen hat; ist das aber nicht der Fall, so verlieren sie sehr schnell einen grossen Teil ihrer Macht. Deshalb ist es gut, auch alternative Tests zu kennen, die von weniger spezifischen Annahmen ausgehen.

3.6 Der p-Wert

Sei $X_1, \dots, X_n$ eine Stichprobe vom Umfang n. Wir wollen eine Hypothese $H_0 : \theta = \theta_0$ gegen eine Alternativhypothese $H_A : \theta \in \Theta_A$ testen.

Definition 3.9 (Geordnete Testsammlung)**.** *Sei T eine Teststatistik. Eine Familie von Tests $(T, (K_t)_{t\geq 0})$ heisst* ***geordert bzgl. T*** *falls $K_t \subset \mathbb{R}$ und*

$$s \leq t \implies K_s \supset K_t$$

gilt.

Typische Beispiele sind $K_t = (t, \infty)$ (rechtsseitiger Test), $K_t = (-\infty, -t)$ (linksseitiger Test) der $K_t = (-\infty, -t) \cup (t, \infty)$ (beidseitiger Test).

Definition 3.10. *Sei $H_0 : \theta = \theta_0$ eine einfache Nullhypothese. Sei $(T, K_t)_{t\geq 0}$ eine geordnete Familie von Test. Der* ***p-Wert*** *ist definiert als Zufallsvariable*

$$\text{p-Wert} = G(T),$$

wobei $G : \mathbb{R}_+ \to [0, 1]$ mittels $G(t) = \mathbb{P}_{\theta_0}[T \in K_t]$ definiert ist.

Anmerkungen:

- Der P-Wert ist als Funktion einer Teststatistik T selbst eine Zufallsvariable.
- Der P-Wert hängt direkt von den anfänglichen Beobachtungen $X_1, \dots, X_n$ ab. Somit wird das Wiederholen des Test auch einen neuen (zufälligen) P-Wert generieren.
- Der p-Wert liegt stets in $[0, 1]$. Sei T stetig ist und $K_t = (t, \infty)$, dann kann gezeigt werden, dass der P-Wert unter $\mathbb{P}_{\theta_0}$ auf $[0, 1]$ gleichverteilt ist.

- Der P-Wert liefert uns die Information, welche Tests in unserer Familie (T, K_t), $t \geq 0$ die Nullhypothese H_0 ablehnen würden.

  Für einen P-Wert mit Wert p gilt, dass alle Tests mit Signifikanzniveau $\alpha > p$ die Nullhypothese H_0 verwerfen würden und alle Tests mit Signifikanzniveau $\alpha \leq p$ die Nullhypothese H_0 nicht verwerfen würden.

- Der P-Wert ist nur von der Nullhypothese abhängig. Die Alternativhypothese spielt keine Rolle in der Definition des P-Werts.

Intuition: Sei $K_t = (t, \infty)$ und nehme an, dass $T \sim \text{Exp}(1)$ unter $\mathbb{P}_{\theta_0}$ gilt. Somit gilt nach obiger Definition $G(t) = e^{-t}$. Für grosse t ist dies gerade die Wahrscheinlichkeit, dass die Teststatistik ünnatürlich"gross ist.

Nehmen wir also an, dass wir aus gegebenen Daten $T(\omega) = 1000$ erhalten. Unter der Nullhypothese H_0, ist die Wahrscheinlichkeit für solch einen grossen Wert sehr gering. Dies legt das Verwerfen von H_0 nahe. Der P-Wert signalisiert somit das Folgende: Falls T ungewöhnlich grosse Werte (unter P_{θ_0}) annimmt, dann ist der p-Wert $G(T)$ sehr klein. Nehmen wir zum Beispiel $p = 0.01$, dann würden alle Tests Signifikanzniveaus strikt grosser als 0.01 **bereits** die Nullhyphothese verwerfen. Zusammenfassend kann man sagen,

p-Wert ist klein $\implies$ H_0 wird wahrscheinlich verworfen.

Umgekehrt muss man mit der Interpretation des p-Wertes mehr als bei einem Test aufpassen; beispielsweise ist die “Aussage”

“der p-Wert ist die Wahrscheinlichkeit, dass die Hypothese richtig ist”

völlig falsch, denn der p-Wert ist eine Zufallsvariable, während die Hypothese mit Sicherheit entweder richtig oder falsch ist (wir wissen einfach nicht, welches von beidem der Fall ist). Zudem kann man mit unüberlegtem oder systematischem Wiederholen von Experimenten den p-Wert deutlich verfälschen.

Ein Vorteil des p-Wertes ist, dass viele Statistik-Pakete direkt einen p-Wert berechnen. Zudem hat man etwas mehr Information als nur die reine Testentscheidung — der p-Wert gibt eine Indikation dafür, wie weit der Wert der Teststatistik im kritischen Bereich liegt oder nicht.

Beispiel:

Wir werfen eine Münze 100 Mal und beobachten dabei 60 Mal Kopf. Ist diese Münze fair?

Unser Modell ist, dass $X_1, \dots, X_n$ unter $\mathbb{P}_\theta$ i.i.d. $\sim Be(\theta)$ mit $\theta \in \Theta = [0,1]$ sind. Als Hypothese wählen wir dann $H_0 : \theta = \frac{1}{2}$, also $\Theta_0 = \{\frac{1}{2}\}$, und die Alternative ist $H_A : \theta \neq \frac{1}{2}$, also $\Theta_A = [0,1] \backslash \{\frac{1}{2}\}$. Die (zufällige) Anzahl der Erfolge in den n Versuchen ist $S_n = \sum_{i=1}^n X_i$, und es gilt $S_n \sim Bin(n, \theta)$ unter $\mathbb{P}_\theta$.

Um vernünftig rechnen zu können, approximieren wir die Binomialverteilung gemäss dem zentralen Grenzwertsatz durch eine Normalverteilung: es gilt für jedes $\theta \in \Theta$, dass

$$S_n \overset{\text{Approx}}{\approx} \mathcal{N}\big(n\theta, n\theta(1-\theta)\big) \text{unter } \mathbb{P}_\theta$$

und damit

$$T' := \frac{S_n - n\theta}{\sqrt{n\theta(1-\theta)}} \overset{\text{Approx}}{\approx} \mathcal{N}(0,1) \text{unter } \mathbb{P}_\theta.$$

Für $\theta = \theta_0 = \frac{1}{2}$ ist also

$$T = \frac{S_n - n/2}{\sqrt{n/4}} = \frac{2S_n - n}{\sqrt{n}} \overset{\text{Approx}}{\approx} \mathcal{N}(0,1) \text{unter } \mathbb{P}_{\theta_0}.$$

Für einen Test wählen wir nun den kritischen Bereich von der symmetrischen Form $K := (-\infty, -c) \cup (+c, +\infty)$, d.h. wir verwerfen H_0 für $|T| > c$; dabei wählen wir K symmetrisch, weil die approximative Verteilung von T unter $\mathbb{P}_{\theta_0}$ symmetrisch ist. Wollen wir die Bedingung $\mathbb{P}_{\theta_0}[T \in K] \leq \alpha$ für ein gegebenes Niveau α möglichst scharf (approximativ) erfüllen, so fordern wir Gleichheit; dann brauchen wir also wegen der approximativen Symmetrie der Verteilung von T unter $\mathbb{P}_{\theta_0}$

$$\alpha = \mathbb{P}_{\theta_0}[T \in K] = \mathbb{P}_{\theta_0}[|T| > c] \approx 2\mathbb{P}_{\theta_0}[T > c] \approx 2\big(1 - \Phi(c)\big),$$

und daraus ergibt sich

$$c \approx \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = z_{1-\frac{\alpha}{2}} = 2.576 \text{für } \alpha = 1\%.$$

Wir verwerfen also die Hypothese H_0 einer fairen Münze für $|T| > c$, und zurückübersetzt passiert das für $S_n > 62.88$ oder $S_n < 37.12$. Also glauben wir auf dem 1%-Niveau nicht an eine faire Münze, sofern wir mindestens 63 oder höchstens 37 Erfolge beobachten.

Den realisierten approximativen p-Wert berechnen wir hier als

$$\text{p-Wert}(\omega) = \mathbb{P}_{\theta_0}[|T| > t_0]\big|_{t_0 = T(\omega)} \approx 2\mathbb{P}_{\theta_0}[T > t_0]\big|_{t_0 = T(\omega)} \approx 2\big(1 - \Phi(t_0)\big)\big|_{t_0 = T(\omega)}.$$

Wegen

$$T(\omega) = \frac{2S_n(\omega) - n}{\sqrt{n}} = \frac{120 - 100}{10} = 2$$

ist also

$$\text{p-Wert}(\omega) \approx 2\big(1 - \Phi(2)\big) = 2(1 - 0.97725) = 0.0455.$$

Bei 60 Erfolgen verwerfen wir die Hypothese einer fairen Münze also auf dem 5%-Niveau, aber (wie oben gesehen) nicht auf dem 1%-Niveau. ◇

3.6.1 Zusammenfassung zu Tests

Zum Abschluss dieses Abschnitts fassen wir das allgemeine Vorgehen bei Tests noch einmal kurz zusammen. Man hat die folgenden 5 Schritte:

1) Wahl des Modells.

2) Formulierung von Hypothese und Alternative.

3) Bestimmung der Teststatistik T und der Form des kritischen Bereichs K; das kann aus einer Herleitung via verallgemeinerten LQ-Test oder direkt aus einem Statistik-Buch stammen.

4) Festlegung des Niveaus α liefert (die Grenze für) den kritischen Bereich K; dazu braucht man die Verteilung von T unter $\mathbb{P}_\vartheta$ für alle $\vartheta \in \Theta_0$ (exakt oder approximativ).

5) Berechnen der Teststatistik $T(\omega)$ aus den Daten; ist $T(\omega) \in K$, so wird die Hypothese abgelehnt, andernfalls wird die Hypothese nicht verworfen.

5′) Berechnen von Teststatistik $T(\omega)$ und entsprechendem realisiertem p-Wert(ω) aus den Daten; ist letzterer $\leq \alpha$, so wird die Hypothese abgelehnt, andernfalls nicht.

Literaturverzeichnis

[Bronstein et al.] I. N. Bronstein, K. A. Semendjajew, G. Musiol, H. Mühlig, "Taschenbuch der Mathematik", 4. Auflage, Harri Deutsch (1999)

[Krengel] U. Krengel, "Einführung in die Wahrscheinlichkeitstheorie und Statistik", 8. Auflage, Vieweg (2005)

[LSW21] J. Lengler, A. Steger, and E. Welzl, **Algorithmen und Wahrscheinlichkeit**, 2021.

[Lehn, Wegmann] J. Lehn, H. Wegmann, "Einführung in die Statistik", 4. Auflage, Teubner (2004)

[Rice] J. A. Rice, "Mathematical Statistics and Data Analysis", second edition, Duxbury Press (1995)

[Sch10] M. Schweizer, **Wahrscheinlichkeit und Statistik**, 2010.

[Stahel] W. A. Stahel, "Statistische Datenanalyse. Eine Einführung für Naturwissenschaftler", 2. Auflage, Vieweg (1999)

[Williams] D. Williams, "Weighing the Odds. A Course in Probability and Statistics", Cambridge University Press (2001)