Stochastik II (für BSc)

Sommersemester 2021

Universität Freiburg

von Peter Pfaffelhuber

Version: 28. April 2021

Inhaltsverzeichnis

1 Grundlegendes

Es ist nicht übertrieben zu behaupten, dass in der heutigen Welt immer mehr *Daten* jeglicher Art erhoben werden. Diese zu ordnen und aus Daten Schlussfolgerungen zu ziehen ist Aufgabe der Statistik.

Man teilt dabei diese Aufgaben in zwei Gebiete auf. Die *deskriptive Statistik* dient rein der Beschreibung der Daten, etwa durch geeignete Wahl von Statistiken, die die Daten zusammenfassen. Anders ist dies bei der hier behandelten *schließenden* oder *induktiven Statistik*. Die Aufgabe ist hier, mit Hilfe von stochastischen Modellen Aussagen darüber zu treffen, welchen Annahmen den Daten zugrunde liegen könnten. Manchmal sagt man auch, dass man anhand der Daten etwas *lernt*, weshalb man oft auch von statistischem Lernen spricht.

> Angenommen, $X_1, ...X_n \sim \mathbf{P}$. Was können wir anhand von den Realisierungen von $X_1, ..., X_n$ über $\mathbf{P}$ aussagen (oder lernen)?

1.1 Ein Beispiel

Wir beginnen mit einem Beispiel, das sich um die Erfolgswahrscheinlichkeit beim Münzwurf dreht: eine Münze wird 53 mal geworfen. Dabei ist die Wahrscheinlichkeit für *Kopf* (was wir im Folgenden als Erfolg werten wollen) noch unbekannt. Von den 53 Würfen sind 23 ein Erfolg.

Unsere statistischen Überlegungen gehen nun von der Vorstellung aus, dass die 53 Münzwürfe die Realisierung einer Zufallsvariable $X = (X_1, \dots, X_{53})$ sind, wobei $X_1, \dots, X_{53}$ unabhängig und identisch verteilt sind mit

$$X_i = \begin{cases} 1, & \text{falls der } i\text{-te Wurf } \textit{Kopf} \text{ zeigt,} \\ 0, & \text{sonst.} \end{cases}$$

und es gilt

$$\mathbf{P}(X_i = 1) = p.$$

Jetzt ist $X_1 + \dots + X_n$ die Gesamtzahl der Erfolge. Als Summe von n unabhängigen Bernoulliverteilten Zufallsvariablen ist diese Summe $B(n = 53, p)$-verteilt. Wichtig ist, dass zwar $n = 53$ bereits fest steht (schließlich wissen wir ja, dass wir 53 mal die Münze geworfen haben), nicht jedoch p. In dieser Situation gibt es zwei *statistische Probleme*.

- *Schätzproblem:* Wir können versuchen, den Erfolgsparameter p zu schätzen. Entweder werden wir dazu einen aus den Daten (23 Erfolge aus 53 Versuchen) abgeleiteten Wert $\widehat{p}$ angeben (*Punktschätzer*), oder ein aus den Daten abgeleitetes Intervall $[a, b]$, in dem der wahre Parameter p mit hoher Wahrscheinlichkeit liegt (*Intervallschätzer*).

- *Testproblem:* Stellen wir uns vor, der Werfer der Münze behauptet, dass die Münze fair ist, also $p = \frac{1}{2}$ gilt. Dieser Meinung können wir skeptisch gegenüber stehen, da ja nur 23 aus 53 Würfen (etwa 43 %) ein Erfolg waren. Wir können versuchen, die Hypothese $p = \frac{1}{2}$ zu testen. Das bedeutet, dass wir untersuchen, wie gut die Hypothese mit den Daten in Einklang steht.

Die Erfolgswahrscheinlichkeit schätzen: Wir versuchen nun, die Erfolgswahrscheinlichkeit des Münzwurfes zu schätzen. Dies muss auf Grundlage der Daten erfolgen, also basierend auf dem Wissen, dass 23 aus 53 Erfolge zu verzeichnen waren. Ein einfacher Ansatz ist es, zu vermuten, dass die 23 Erfolge aus der 53 Würfen in etwa der Erfolgsquote entspricht. Also ist

$$\widehat{p} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{23}{53} \approx 0.43$$

ein Schätzer für den unbekannten Parameter p. (Im Folgenden werden wir einen Schätzer für einen Parameter $\bullet$ meistens mit $\widehat{\bullet}$ bezeichnen.) Ein anderer Ansatz (die wir als der Maximum-Likelihood-Methode bezeichnen werden) wäre der, $\hat{p}$ so zu setzen, dass die Wahrscheinlichkeit, 23 Erfolge zu erzielen, maximal wird. Es ist hier also $p \mapsto \binom{53}{23} p^{23}(1-p)^{30}$ zu maximieren. Wir berechnen

$$\frac{\partial}{\partial p} p^{23}(1-p)^{30} = p^{22}(1-p)^{29}(23(1-p) - 30p),$$

und es ergibt sich $53\hat{p} = 23$ oder wieder $\hat{p} = \frac{23}{53}$.

In beiden Fällen hängt $\widehat{p}$ von den Daten ab, die wir uns als Realisierung von einer Zufallsvariable gedacht haben. Also ist $\widehat{p}$ auch eine Zufallsvariable. Warum ist der Schätzer $\widehat{p}$ gut? Nehmen wir an, wir wüssten den wahren Parameter p. Dann leistet $\widehat{p}$ zumindest im Mittel das gewünschte: (Wir schreiben hier und im Folgenden $\mathbf{P}_p(\cdot)$ und $\mathbf{E}_p[\cdot]$, wenn wir Wahrscheinlichkeiten und Erwartungswerte unter der Hypothese ausrechnen wollen, dass p der wahre Parameter ist.)

$$\mathbf{E}_p[\widehat{p}] = \frac{1}{n}\mathbf{E}_p[X_1 + \cdots + X_n] = p.$$

Wir sagen auch, der Schätzer $\widehat{p}$ ist erwartungstreu (oder unverzerrt oder unbiased).

Eine weitere wünschenwerte Eigenschaft eines Schätzers ist, dass er immer besser wird, je größer die zu Grunde liegende Datenmenge ist. Eine große Datengrundlage bedeutet in unserem Fall, dass die Münze oft geworfen wurde, also n groß ist. Aus dem schwachen Gesetz großer Zahlen wissen wir, dass

$$\mathbf{P}_p(|\widehat{p} - p| \geq \varepsilon) = \mathbf{P}_p\Big(\Big|\frac{X_1 + \cdots + X_n}{n} - \mathbf{E}_p[X_1]\Big| \geq \varepsilon\Big) \xrightarrow{n\to\infty} 0$$

für alle $\varepsilon > 0$. Die Eigenschaft, dass $\widehat{p}$ mit hoher Wahrscheinlichkeit immer näher am wahren Wert p liegt, wenn mehr Daten zur Verfügung stehen, nennen wir *Konsistenz*.

Stellen wir uns vor, in zwei aufeinander folgenden Experimenten bekommen wir zunächst 23 von 53 Erfolge, und dann 23000 von 53000 Erfolge. In beiden Fällen ist $\widehat{p} = \frac{23}{53}$. Es ist jedoch klar, dass wir dem Wert von $\widehat{p}$ im zweiten Experiment eine viel höhere Bedeutung zumessen und wir uns sicherer sind, dass der wahre Wert in der Nähe von $\widehat{p}$ liegt. Diese Sicherheit können wir mit einem *Intervallschätzer*, also einem Intervall, in dem der wahre

Wert mit hoher Wahrscheinlichkeit liegt, zu Ausdruck bringen. Dazu wählen wir ein (kleines) $\alpha \in (0,1)$, etwa $\alpha = 5\,\%$. Aus dem zentralen Grenzwertsatz folgt, dass es eine nach $N(0,1)$ verteilte Zufallsvariable Z gibt, so dass[1]

$$\mathbf{P}_p\Big(-1.96 \leq \frac{n\widehat{p} - np}{\sqrt{np(1-p)}} \leq 1.96\Big) \approx \mathbf{P}\big(-1.96 \leq Z \leq 1.96\big) \approx 0.95.$$

Weiter ist

$$\begin{aligned} 0.95 &\approx \mathbf{P}_p\Big(-1.96 \leq \frac{\widehat{p} - p}{\sqrt{\widehat{p}(1-\widehat{p})/n}} \leq 1.96\Big) \\ &= \mathbf{P}_p\Big(\widehat{p} - 1.96\sqrt{\widehat{p}(1-\widehat{p})/n} \leq p \leq \widehat{p} + 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}\Big). \end{aligned}$$

Wir haben gerade gezeigt, dass

$$\big[\widehat{p} - 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}; \widehat{p} + 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}\big]$$

ein *Konfidenzintervall* für $\widehat{p}$ zum *Niveau* $95\,\%$ ist. Das bedeutet, dass der wahre Wert mit Wahrscheinlichkeit etwa $95\,\%$ in diesem (zufälligen) Intervall liegt. Haben wir 23 Erfolge in $n = 53$ Versuchen, ist unser Konfidenzintervall also $[0.30, 0.57]$. Haben wir hingegen 23000 Erfolge bei 53000 Versuchen, ist das Konfidenzintervall etwa $[0.430, 0.438]$, also wesentlich kleiner.

Die Erfolgswahrscheinlichkeit testen: Nehmen wir an, der Werfer der Münze behauptet, sie sei fair, also $p = \frac{1}{2}$. Können wir diese Hypothese aufgrund der Daten verwerfen? Zunächst stellen wir fest, dass wir prinzipiell zwei Arten von Fehlern mit unserer Entscheidung machen können. Wenn wir die Hypothese verwerfen, könnte sie doch wahr sein, und wenn wir die Hypothese nicht verwerfen, könnte sie doch falsch sein.

Da wir nicht grundlos dem Werfer der Münze widersprechen wollen, wollen wir die Wahrscheinlichkeit, dass wir die Hypothese ablehnen (wir dem Werfer der Münze widersprechen), obwohl sie wahr ist (die Hypothese des Werfers richtig ist), kontrollieren. Das bedeutet, dass

$$\mathbf{P}_{p=1/2}(\text{Hypothese verwerfen}) \leq \alpha$$

für ein anfangs gewähltes $\alpha \in (0,1)$ sein soll. Klar ist, dass damit die Hypothese umso seltener abgelehnt werden kann, je kleiner α ist. Nun kommen wir zu der Regel, mit der wir die Hypothese ablehnen wollen. In unserem Beispiel haben wir für die Hypothese $p = \frac{1}{2}$ zu wenig (23 von 53) Erfolge. Wir würden die Hypothese ablehnen wollen, wenn

$$\mathbf{P}_{p=1/2}(\text{Daten extremer als tatsächliche Daten}) \leq \alpha. \tag{1.1}$$

Wir wählen (wie oben beim Konfidenzintervall) $\alpha = 5\,\%$. Für Y_n nach $B(n = 53, p = \frac{1}{2})$ verteilt, erwarten wir 26.5 Erfolge. Um die Wahrscheinlichkeit einer Abweichung, die größer

[1] Wir schreiben in dieser einführenden Bemerkung öfter approximative Formeln mittels '$\approx$'. Es sei bemerkt, dass dieses Symbol keine mathematisch exakten, beweisbaren Aussagen trifft. Für Anwendungen sinnvolle Resultate liefert es allerdings allemal.

ist als die der Daten zu berechnen, betrachten wir eine nach $N(0,1)$ verteilte Zufallsvariable Z und berechnen

$$\begin{aligned}\mathbf{P}_{p=1/2}\big(|X_1+\cdots+X_n-26.5|\geq 3.5\big)&\\ =1-\mathbf{P}_{p=1/2}\bigg(-\frac{3.5}{\sqrt{np(1-p)}}<\frac{Y_n-np}{\sqrt{np(1-p)}}<\frac{3.5}{\sqrt{np(1-p)}}\bigg)&\\ \approx 1-\mathbf{P}_{p=1/2}\big(-0.962\leq Z\leq 0.962\big)\approx 33.6\,\%&\end{aligned}$$

Da dieser Wert größer als $\alpha = 5\,\%$ ist, kann die Hypothese nicht verworfen werden, siehe (1.1).

1.2 Statistisches Modell und Entscheidungstheorie

Wir beginnen nun mit der Formalisierung der obigen Situation.

Definition 1.1 (Statistisches Modell)**.** *Seien S, Θ, Θ' Mengen. Ein* statistisches Modell *ist ein Paar $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$, wobei X eine Zufallsvariable mit Zielbereich S ist, bei deren Verteilung noch ein Parameter $\vartheta \in \Theta$ frei, also unbestimmt, ist. Das bedeutet, dass es eine Funktion $\vartheta \mapsto \rho_\vartheta$ gibt mit*[2]

$$\mathbf{P}_\vartheta(X \in da) = \rho_\vartheta(a)da.$$

Die Menge Θ heißt Parameterraum*, die Menge S* Beobachtungsraum*. Jede Zufallsvariable $h(X)$ mit $h : S \to \Theta'$ heißt* Statistik*.*

Beispiel 1.2 (Parametrische und nicht-parametrische Modelle)**.**

1. Sei $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$, $S = \mathbb{R}^n$ und $X = (X_1, ..., X_n)$ unabhängig normalverteilut, d.h. $(X_i)_*\mathbf{P}_{(\mu,\sigma^2)} = N(\mu, \sigma^2)$. Dann heißt $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ das *Normalverteilungsmodell.*

2. Sei $\Theta = [0, 1]$, $S = \{0, 1\}^n$ und $X = (X_1, ..., X_n)$ unabhängig mit $(X_i)_*\mathbf{P}_p = B(1, p)$. Dann heißt $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ das *Binomialmodell.*

3. Ist in Definition 1.1 die Menge Θ endlich-dimensional, also $\Theta \subseteq \mathbb{R}^d$ für ein $d = 1, 2, ...$, so spricht man von einem parametrischen Modell. (Siehe etwa 1. und 2.) Ist Θ hingegen unendlich-dimensional, so spricht man von einem nicht-parametrischen Modell. Ein Beispiel wäre $\Theta = \{F : \mathbb{R} \to [0, 1]$ Verteilungsfunktion$\}$, $S = \mathbb{R}^n$ und $X = (X_1, ..., X_n)$ unabhängig mit
$$\mathbf{P}_F(X_i \leq x) = F(x).$$

[2] Wir wollen im Folgenden die dauernde Unterscheidung zwischen diskreten Zufallsvariablen und Zufallsvariablen mit Dichten durch die Notation $\mathbf{P}(X \in da)$ vermeiden. Ist der Wertebereich S von X diskret und $a \in S$, ist damit
$$\mathbf{P}(X \in da) := \mathbf{P}(X = a)$$
gemeint. Hat X die Dichte $f(a)da$, ist
$$\mathbf{P}(X \in da) := f(a)da.$$

Definition 1.3 (Statistische Fragestellungen)**.** *Zu einer statischen Fragestellung gehört zunächst einmal ein statistisches Modell $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$. Anhand der Daten X muss man nun Schlussfolgerungen über ϑ ziehen. Deshalb gibt es eine (deterministische oder stochastische) Entscheidungsfunktion $h : S \to \aleph$, wobei $\aleph$ auch Entscheidungsraum heißt. Fällt man eine Entscheidung für $\alpha \in \aleph$, was jedoch falsch ist, erhält man einen Verlust $\ell : \Theta \times \aleph \to \mathbb{R}_+$. Die mittlere Verlustfunktion, nämlich die Abbildung $\vartheta \mapsto \mathbf{E}_\vartheta[\ell(\vartheta, h(X))]$, heißt auch Risikofunktion.*

Bemerkung 1.4 (Schätz- und Testprobleme)**.**

1. Ein Punktschätzer ermittelt anhand der Daten X eine Vorstellung über das zugrunde liegende ϑ. Mit anderen Worten will man eine Schätzfunktion $h : S \to \Theta$ finden, die möglichst gut an den wahren Wert ϑ herankommt.

   Hier ist $\aleph = \Theta$, und meist $\ell(\vartheta, \vartheta') = (\vartheta - \vartheta')^2$. Ein durch h gegebene Schätzer hat dann einen Bias (oder eine Verzerrung) von $\mathbf{E}_\vartheta[h(X)] - \vartheta$ und eine Varianz von $\mathbf{V}_\vartheta[h(X)]$. Man berechnet die Risikofunktion leicht durch

   $$\begin{aligned}\mathbf{E}_\vartheta[(h(X) - \vartheta)^2] &= \mathbf{E}_\vartheta[h(X)^2] - \mathbf{E}_\vartheta[h(X)]^2 + \mathbf{E}_\vartheta[h(X)]^2 - 2\theta\mathbf{E}_\vartheta[h(X)] + \theta^2] \\ &= \mathbf{V}_\vartheta[h(X)] + (\mathbf{E}_\vartheta[h(X)] - \vartheta)^2,\end{aligned}$$

   also die Summe aus Varianz und dem quadratischen Bias.

2. Ein Testproblem liegt dann vor, wenn man sich anhand der Datenlage für oder gegen eine Behauptung (Hypothese) über ϑ entscheidet.

   Hier ist[3] $\Theta = \Theta_0 \uplus \Theta_1$, $\aleph = \{\Theta_0, \Theta_1\}$ und

   $$\ell(\vartheta, \Theta_i) = \begin{cases} 0, & \vartheta \in \Theta_i, \\ 1, & \vartheta \notin \Theta_i. \end{cases}$$

   Die Risikofunktion ist dann

   $$\mathbf{E}_\vartheta[1_{\vartheta\notin h(X)}] = \mathbf{P}_\vartheta[\vartheta \notin h(X)],$$

   also gerade der Wahrscheinlichkeit, sich (bei Gültigkeit von ϑ) falsch zu entscheiden.

Bemerkung 1.5 (Regression und Klassifikation)**.** Wir betrachten nun den Fall, dass $X = ((X_1, Y_1), ..., (X_n, Y_n))$ aus Beobachtungspaaren mit Zustandsraum $S = (S_X \times S_Y)^n$ besteht. Dabei nennen wir X auch *Prädiktor* und Y *Ausgang*. (Man denke etwa an X =Blutdruck und Y =restliche Lebenszeit.) Wir nehmen an (d.h. unser statistisches Modell modelliert), dass es ein r gibt mit $Y = r(X) + \varepsilon$ mit einer Zufallsvariable ε mit $\mathbf{E}[\varepsilon] = 0$. Das statistische Modell besteht also aus $(X, (\mathbf{P}_r)_{r\in\Theta})$, wobei Θ die Menge aller möglichen Zusammenhänge zwischen X_i und Y_i ist.

1. Ein Regressionsproblem liegt dann vor, wenn wir für einen neuen Datenpunkt X_{n+1} das zugehörige Y_{n+1} *vorhersagen* wollen. Hier ist also $h(X, X_{n+1})$ die Vorhersage von Y_{n+1}. Die Verlustfunktion ist etwa $(h(X, X_{n+1}) - Y_{n+1})^2$, also gerade der quadratischen Abweichung der Vorhersage vom wahren Wert.

2. Ein Klassifikationsproblem ist ein spezielles Vorhersageproblem im Fall von endlichem S_Y. Hier ist $1_{h(X,X_{n+1})\neq Y_{n+1}}$ der Misklassifikationsfehler.

[3] Wir schreiben $A \uplus B$ für die Vereinigung von A und B, falls $A \cap B = \emptyset$.

Bemerkung 1.6 (Frequentistische und Bayesianische Statistik)**.** Oftmals wird streng zwischen der frequentistischen und der Bayesianischen Statistik unterschieden. Der Hauptunterschied ist der Folgende: Modellparameter (z.B. die Erfolgswahrscheinlichkeit im Münzwurf) sind in der frequentistischen Sichtweise immer deterministisch. In der Bayesianischen Statistik werden sie als Zufallsvariablen modelliert.

Nehmen wir das Beispiel aus Abschnitt 1.1. Vor Beginn der Experimentes wissen wir nichts über den Erfolgsparameter des Münzwurfes. Deshalb nehmen wir an, dass $P \sim U([0,1])$. Führen wir nun 53-mal das Experiment durch, und erlangen hierbei 23 Erfolge, so verändert dies unsere Einschätzung über die Verteilung von P. Es ist $X_1 + \cdots + X_{53} \sim B(53, P)$, wobei $P \sim U([0,1])$ die sogenannte apriori-Verteilung ist. Nach Durchführung des Experimentes ist

$$\mathbf{P}(P \in dp | X_1 + \cdots + X_{53} = 23) = \frac{p^{23}(1-p)^{30}}{\int_0^1 x^{23}(1-x)^{30} dx} dp,$$

was man auch *a posteriori* Verteilung nennt.

2 Einführende Konzepte

2.1 Die multivariate Normalverteilung

Die Normalverteilung hat eine wichtige Stellung in der Statistik. Grund hierfür ist der Zentrale Grenzwertsatz: kommt eine zufällige Größe durch viele unabhängige Einflüsse zu Stande, so ist sie annähernd normalverteilt. Wir stellen nun die mehrdimensionale Normalverteilung vor. Wir erinnern daran, dass ein Wahrscheinlichkeitsmaß $\mathbf{P}$ eine mehrdimensionale Dichte $f : \mathbb{R}^d \to \mathbb{R}$ besitzt, falls

$$\mathbf{P}(B_1 \times \cdots \times B_d) = \int_{B_1} \cdots \int_{B_d} f(x_1, ..., x_d) dx_d \cdots dx_1$$

für alle Intervalle $B_1, ..., B_d$.

Definition 2.1 (Multivariate Normalverteilung)**.** *Sei $\mu \in \mathbb{R}^d$ und $\Sigma \in \mathbb{R}^{d\times d}$ eine symmetrische, positiv definite Matrix. Eine $\mathbb{R}^d$-wertige Zufallsvariable heißt normalverteilt mit Erwartungswert(vektor) μ und Covarianz(matrix) Σ, falls sie die Dichte*[1]

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |det(\Sigma)|}} \exp\Big(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big)$$

besitzt. Wir schreiben dann $X \sim N(\mu, \Sigma)$. Im Falle[2] *$\mu = 0$ und $\Sigma = I_d$ spricht man von einer multivariaten Standard-Normalverteilung.*

Bemerkung 2.2 (Singuläre multivariate Normalverteilungen)**.** Bei der obigen Definition der multivariaten Normalverteilung liegt die Wahrscheinlichkeitsmass auf ganz $\mathbb{R}^d$. Man kann die Definition auch erweitern und erlauben, dass die multivariate Normalverteilung nur auf einem linearen Teilraum von $\mathbb{R}^d$ Wahrscheinlichkeitsmasse legt. In einem solchen Fall wäre Σ nur nicht-negativ definit, d.h. hätte auch verschwindende Eigenwerte. Dann braucht man jedoch maßtheoretische Konstruktionen, um die Verteilung überhaupt hinzuschreiben. Da wir in dieser Vorlesung keine solchen Methoden zur Verfügung haben, begnügen wir uns mit obiger Definition.

Lemma 2.3 (Transformationen von multivariaten Normalverteilungen)**.** *Sei $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$ positiv definit, $n \leq d$ mit $b \in \mathbb{R}^n$ und $A \in \mathbb{R}^{n\times d}$ mit vollem Rang. Sei $X \sim N(\mu, \Sigma)$ und $Z = AX + b$. Dann ist $Z \sim N(A\mu + b, A\Sigma A^\top)$.*

Insbesondere gilt[3] *für $A = \Sigma^{-1/2}$ und $b = -A\mu$, dass Z multivariat standard-normalverteilt ist.*

[1] Wir bezeichnen mit $x^\top$ einen Zeilen- und mit x einen Spaltenvektor.

[2] Hier bezeichnet I_d die d-dimensionale Einheitsmatrix.

[3] Ein Ergebnis aus der linearen Algebra besagt, dass man aus einer positiv definiten Matrix Σ eine Wurzel ziehen kann, d.h. es gibt ein A mit $A^2 = \Sigma$.

Beweis. Wir zeigen die Behauptung nur im Fall $n = d$, da der allgemeine Fall etwas mehr Maßtheorie erfordert. Es genügt, den Fall $\mu = 0$ zu betrachten. Der allgemeine Fall folgt dann durch Addition von μ. Wir schreiben mit dem Transformationssatz[4]

$$\begin{aligned}\mathbf{P}(AX \in B_1 \times \cdots \times B_d) &= \int_{A^{-1}(B_1\times\cdots\times B_d)} \frac{1}{\sqrt{(2\pi)^d |\det(\Sigma)|}} \exp\Big(-\frac{1}{2}x^\top A^\top (A\Sigma A^\top)^{-1} Ax\Big) dx \\ &= \int_{A^{-1}(B_1\times\cdots\times B_d)} \frac{\det(A)}{\sqrt{(2\pi)^d |\det(A\Sigma A^\top)|}} \exp\Big(-\frac{1}{2}x^\top A^\top (A\Sigma A^\top)^{-1} Ax\Big) dx \\ &= \int_{B_1\times\cdots\times B_d} \frac{1}{\sqrt{(2\pi)^d |\det(A\Sigma A^\top)|}} \exp\Big(-\frac{1}{2}y^\top (A\Sigma A^\top)^{-1} y\Big) dy,\end{aligned}$$

d.h. $AX \sim N(0, A\Sigma A^\top)$. □

Korollar 2.4. *Sei $X = (X_1, ..., X_d) \sim N(0, I_d)$ und $O \in \mathbb{R}^{d\times d}$ orthogonal. Dann ist $OX \sim N(0, I_d)$.*

Beweis. Nach Lemma 2.3 ist $OX \sim N(0, OI_dO^\top) = N(0, I_d)$. □

2.2 Totalvariationsabstand

In der Statistik müssen wir häufig zwei Verteilungen vergleichen und etwa abschätzen, ob diese ähnlich sind oder nicht. Eine Möglichkeit, dies zu tun, geben wir nun an.

Definition 2.5 (Metrik der Totalvariation)**.** *Seien $\mathbf{P}$ und $\mathbf{Q}$ Wahrscheinlichkeitsmaße (auf derselben σ-Algebra $\mathcal{A}$). Dann ist*

$$||\mathbf{P} - \mathbf{Q}||_{TV} := \sup_{A\in\mathcal{A}} |\mathbf{P}(A) - \mathbf{Q}(A)|$$

der Totalvariationsabstand *von $\mathbf{P}$ und $\mathbf{Q}$.*

Lemma 2.6 (Eine Darstellung des Totalvariationsabstandes)**.** *Seien $\mathbf{P}$ und $\mathbf{Q}$ Wahrscheinlichkeitsverteilungen. Dann gibt es ein A mit $||\mathbf{P} - \mathbf{Q}||_{TV} = (\mathbf{P} - \mathbf{Q})(A)$. Außerdem gilt, falls $\mathbf{P}$ und $\mathbf{Q}$ auf einem diskreten Raum definiert sind,*

$$||\mathbf{P} - \mathbf{Q}||_{TV} = \tfrac{1}{2}\sum_{x\in S} |\mathbf{P}(x) - \mathbf{Q}(x)|.$$

Beweis. Wir setzen $B := \{x : \mathbf{P}(x) \geq \mathbf{Q}(x)\}$. Es gilt

$$\begin{aligned}\sup_{A\in\mathcal{A}} |\mathbf{P}(A) - \mathbf{Q}(A)| &= \sup_{A\in\mathcal{A}} (\mathbf{P}(A) - \mathbf{Q}(A)) = \sup_{A\in\mathcal{A}} \sum_{x\in A} (\mathbf{P}(x) - \mathbf{Q}(x)) \leq \sum_{x\in B} (\mathbf{P}(x) - \mathbf{Q}(x)) \\ &= \mathbf{P}(B) - \mathbf{Q}(B),\end{aligned}$$

woraus die erste Behauptung folgt. Die zweite folgt dann mit

$$\begin{aligned}\sum_{x\in S} |\mathbf{P}(x) - \mathbf{Q}(x)| &= \sum_{x\in B} (\mathbf{P}(x) - \mathbf{Q}(x)) + \sum_{x\in B^c} (\mathbf{Q}(x) - \mathbf{P}(x)) \\ &= (\mathbf{P}(B) - \mathbf{Q}(B)) + (1 - \mathbf{Q}(B) - 1 + \mathbf{P}(B)) = 2(\mathbf{P}(B) - \mathbf{Q}(B)).\end{aligned}$$

□

[4] Dieser besagt, für eine glatte, bijektive Abbildung Φ, dass $\int_{\Phi(\Omega)} f(y)\,\mathrm{d}y = \int_\Omega f(\Phi(x))\,|\det(D\Phi(x))|\,\mathrm{d}x$

Bemerkung 2.7 (Statistische Interpretation)**.** Angenommen, wir haben uns zu entscheiden, ob Daten X nach $\mathbf{P}$ oder nach $\mathbf{Q}$ verteilt sind. Wir wollen hierfür eine Menge A ausmachen, so dass wir uns bei Vorliegen von $X \in A$ für $\mathbf{P}$ und bei $X \notin A$ für $\mathbf{Q}$ entscheiden. Doch wie wählen wir A? Sinnvollerweise wählen wir dies so, dass der Fehler, den wir bei obiger Entscheidungsregel machen, minimal ist. Ein Fehler passiert dabei immer dann wenn X nach $\mathbf{P}$ verteilt ist, aber $X \notin A$ ist, oder wenn X nach $\mathbf{Q}$ verteilt ist, aber $X \in A$ ist. Es gilt also, $\mathbf{P}(A^c) + \mathbf{Q}(A)$ zu minimieren. Es gilt

$$\inf_{A\in\mathcal{A}} (\mathbf{P}(A^c) + \mathbf{Q}(A)) = 1 - \sup_{A\in\mathcal{A}} \mathbf{P}(A) - \mathbf{Q}(A) = 1 - ||\mathbf{P} - \mathbf{Q}||_{TV}.$$

Also legt der Totalvariationsabstand die mögliche Trennschärfe eines Tests $\mathbf{P}$ gegen $\mathbf{Q}$ fest. Je größer der Abstand, desto besser sind die beiden Maße zu trennen.

Beispiel 2.8 (Ziehen mit und ohne Zurücklegen)**.** Zählt man die Anzahl der Erfolge beim Ziehen mit und ohne Zurücklegen, so wird man vermuten, dass bei großer Gesamtkugelzahl das Zurücklegen keine große Rolle spielt. Um diesen Effekt etwas abzuschätzen, sei $\mathbf{P} = Hyp(N, K, n)$ und $\mathbf{Q} = B(n, K/N)$. Wir berechnen für große N und K und $p = K/N$

$$\begin{aligned}
2||\mathbf{P}-\mathbf{Q}||_{TV} &= \sum_{k=0}^{n} \Big| \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} - \binom{n}{k} p^k (1-p)^{n-k} \Big| \\
&= \sum_{k=0}^{n} \binom{n}{k} \Big| \frac{K\cdots(K-k+1)(N-K)\cdots(N-K-n+k+1)}{N\cdots(N-n+1)} - \frac{K^k(N-K)^{n-k}}{N^n} \Big| \\
&\approx \sum_{k=0}^{n} \binom{n}{k} \Big| \frac{K^k\Big(1-K^{-1}\binom{k}{2}\Big)(N-K)^{n-k}\Big(1-(N-K)^{-1}\binom{n-k}{2}\Big)}{N^n\Big(1-N^{-1}\binom{n}{2}\Big)} - \frac{K^k(N-K)^{n-k}}{N^n} \Big| \\
&\approx \sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k} \Big(\frac{1}{K}\binom{k}{2} + \frac{1}{N-K}\binom{n-k}{2} + \frac{1}{N}\binom{n}{2} \Big) \\
&= \frac{1}{2}\Big(\frac{n(n-1)p^2}{K} + \frac{n(n-1)(1-p)^2}{N-K} + \frac{n(n-1)}{N} \Big) = \frac{n(n-1)}{N}.
\end{aligned}$$

Die $\approx$ gelten dabei jeweils ungefähr im Grenzwert großer N. Man sieht also, dass für große N (und moderate n) das Ziehen mit und ohne Zurücklegen im Sinne des Totalvariationsabstandes ähnlich ist.

Beispiel 2.9 (Abstand von Normalverteilungen)**.** Wir berechnen noch den Totalvariations-Abstand von $\mathbf{P} = N(\mu, 1)$ und $\mathbf{Q} = N(\nu, 1)$, d.h. von zwei Normalverteilungen mit gleicher Varianz. Wir behaupten

$$||\mathbf{P} - \mathbf{Q}||_{TV} = 2\Phi\Big(\frac{|\mu - \nu|}{2}\Big) - 1,$$

wobei $x \mapsto \Phi(x)$ die Verteilungsfunktion von $N(0, 1)$ ist.

Zunächst ist für $\mathbf{P}$ mit Dichte f_μ und $\mathbf{Q}$ mit Dichte f_ν immer – wie im Beweis von Lemma 2.6

$$||\mathbf{P} - \mathbf{Q}||_{TV} = \mathbf{P}(f_\mu > f_\nu) - \mathbf{Q}(f_\mu > f_\nu).$$

Nun ist für $\nu > \mu$

$$f_\mu(x) > f_\nu(x) \iff (x-\mu)^2 < (x-\nu)^2 \iff x(\nu-\mu) < \tfrac{1}{2}(\nu^2-\mu^2) \iff 2x < \nu + \mu.$$

Insgesamt folgt also mit $X_*\mathbf{P} = \mathbf{P}$, $Y_*\mathbf{P} = \mathbf{Q}$ und $Z_*\mathbf{P} = N(0,1)$

$$\begin{aligned}||\mathbf{P} - \mathbf{Q}||_{TV} &= \mathbf{P}(2X < \nu + \mu) - \mathbf{P}(2Y < \nu + \mu) = \mathbf{P}(2Z < \nu - \mu) - \mathbf{P}(2Z < \mu - \nu) \\ &= 2\Phi\Big(\frac{|\mu - \nu|}{2}\Big) - 1.\end{aligned}$$

Bereits bei Markov-Ketten haben wir Kopplungen kennengelernt. Diese hängen mit dem Totalvariationsabstand wiefolgt zusammen.

Theorem 2.10. *Seien* $\mathbf{P}$ *und* $\mathbf{Q}$ *Wahrscheinlichkeitsmaße (auf einer* σ*-Algebra* $\mathcal{A}$*). Weiter seien* X, Y *auf einem Wahrscheinlichkeitsraum* $(\Omega, \mathcal{F}, \mathbf{P}')$ *definiert mit* $X_*\mathbf{P}' = \mathbf{P}$ *und* $Y_*\mathbf{P}' = \mathbf{Q}$*. Dann gilt*

$$||\mathbf{P} - \mathbf{Q}||_{TV} \leq \mathbf{P}'(X \neq Y).$$

Bemerkung 2.11. Man kann auch noch zeigen, dass es einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbf{P}')$ mit $X_*\mathbf{P}' = \mathbf{P}$ und $Y_*\mathbf{P}' = \mathbf{Q}$ gibt, auf dem Gleichheit gilt.

Beweis. Wir schreiben, mit dem Erwartungswert $\mathbf{E}'$ bezüglich $\mathbf{P}'$,

$$\begin{aligned}||\mathbf{P} - \mathbf{Q}||_{TV} &= \sup_{A \in \mathcal{A}} |\mathbf{E}'(1_{X \in A} - 1_{Y \in A})| \leq \sup_{A \in \mathcal{A}} \mathbf{E}'(|1_{X \in A} - 1_{Y \in A}|) \\ &= \sup_{A \in \mathcal{A}} \mathbf{E}'(|1_{X \in A} - 1_{Y \in A}|, X \neq Y) \leq \mathbf{E}'(X \neq Y).\end{aligned}$$

□

Beispiel 2.12. Sei $\mathbf{P} = B(n,p)$ und $\mathbf{Q} = B(n,q)$ mit $q > p$. Seien $U_1, ..., U_n \sim U([0,1])$ unabhängig und $X = \sum_{i=1}^n 1_{U_i \leq p}$ sowie $Y = \sum_{i=1}^n 1_{U_i \leq q}$. Nun gilt

$$||\mathbf{P} - \mathbf{Q}||_{TV} \leq \mathbf{P}(X \neq Y) = 1 - (1 - (q-p))^n \leq n(q-p).$$

2.3 Stochastische Ordnung

Wie im letzten Beispiel sei $X \sim \mathbf{P} = B(n,p)$ und $Y \sim \mathbf{Q} = B(n,q)$ mit $q \geq p$. Tendenziell wird also Y größer als X sein, da die Erfolgswahrscheinlichkeit ja höher ist. Basierende auf Daten bedeutet das, dass wir uns bei Vorliegen von vielen Erfolgen eher für das Vorliegen von $\mathbf{Q}$ entscheiden werden, und für $\mathbf{P}$ bei wenigen Erfolgen. Das entsprechende Konzept stellen wir nun vor.

Definition 2.13 (Stochastische Ordnung)**.**

1. *Seien* $\mathbf{P}$ *und* $\mathbf{Q}$ *Wahrscheinlichkeitsmaße auf* $\mathbb{R}$*. Dann sagen wir,* $\mathbf{Q}$ *sei* stochastisch größer *als* $\mathbf{P}$*, falls* $\mathbf{Q}([t;\infty)) \geq \mathbf{P}([t;\infty))$ *für alle* $t \in \mathbb{R}$*.*

2. *Eine Familie* $(\mathbf{P}_\vartheta)_{\vartheta \in \Theta}$ *mit* $\Theta \subseteq \mathbb{R}$ *von Wahrscheinlichkeitsmaßen heißt* stochastisch wachsend in ϑ*, falls* $\mathbf{P}_\vartheta$ *für* $\vartheta \geq \vartheta'$ *stochastisch größer ist als* $\mathbf{P}_{\vartheta'}$*.*

3. *Seien* X, Y *reellwertige Zufallsvariable. Dann heißt* Y *stochastisch größer als* X*, falls* $Y_*\mathbf{P}$ *stochastisch größer ist als* $Y_*\mathbf{P}$*, d.h.* $\mathbf{P}(Y \geq t) \geq \mathbf{P}(X \geq t)$ *für alle* $t \in \mathbb{R}$*.*

Bemerkung 2.14 (Stochastische Ordnung und Verteilungsfunktionen)**.** Seien $F_\mathbf{P}$ und $F_\mathbf{Q}$ die Verteilungsfunktionen von $\mathbf{P}$ und $\mathbf{Q}$. Es ist genau dann $\mathbf{Q}$ stochastisch größer als $\mathbf{P}$, wenn $F_\mathbf{P} \leq F_\mathbf{Q}$.

Lemma 2.15 (Stochastische Ordnung und Kopplung)**.** *Seien* $\mathbf{P}$ *und* $\mathbf{Q}$ *zwei Wahrscheinlichkeitsmaße auf* $\mathbb{R}$*. Dann sind folgende Aussagen äquivalent:*

1. $\mathbf{Q}$ *ist stochastisch größer als* $\mathbf{P}$*.*
2. *Es gibt es einen Wahrscheinlichkeitsraum* $(\Omega, \mathcal{F}, \mathbf{P})$ *und Zufallsvariable* X, Y *mit* $X_*\mathbf{P} = \mathbf{P}, Y_*\mathbf{P} = \mathbf{Q}$ *und* $X \leq Y$*.*

Beweis. 2.⇒1.: Es gilt für alle $t \in \mathbb{R}$

$$\mathbf{Q}([t, \infty)) = \mathbf{P}(Y \geq t) \geq \mathbf{P}(X \geq t) = \mathbf{P}([t, \infty)).$$

1.⇒2.: Seien $F_\mathbf{P}$ und $F_\mathbf{Q}$ die Verteilungsfunktionen von $\mathbf{P}$ und $\mathbf{Q}$. Nach Voraussetzung ist $F_\mathbf{Q} \leq F_\mathbf{P}$. Sei $U_*\mathbf{P} = U([0,1])$, d.h. U ist $U([0,1])$-verteilt. Wir definieren $X := \inf\{t : F_\mathbf{P}(t) \geq U\}$ und $Y := \inf\{t : F_\mathbf{Q}(t) \geq U\}$. Dann gilt

$$\begin{aligned}\mathbf{P}(X \leq t) &= \mathbf{P}(F_\mathbf{P}(t) \geq U) = F_\mathbf{P}(t),\\ \mathbf{P}(Y \leq t) &= \mathbf{P}(F_\mathbf{Q}(t) \geq U) = F_\mathbf{Q}(t),\end{aligned}$$

und wegen $\{t : F_\mathbf{P}(t) \geq U\} \supseteq \{t : F_\mathbf{Q}(t) \geq U\}$ ist auch $X \leq Y$. □

Wir behandeln nun noch eine spezielle Situation, die wir später genauer betrachten werden.

Lemma 2.16. *Sei* X *eine reellwertige Zufallsvariable mit Dichte* f*. Falls* $f(x) = f(-x)$ *für alle* x *und 0 das einzige Maximum von* f *ist, dann ist* $(X + \mu)^2$ *stochastisch wachsend in* μ*.*

Beweis. Sei F die Verteilungsfunktion von X. Für $t \geq 0$ gilt

$$\begin{aligned}\frac{\partial}{\partial\mu}\mathbf{P}((X+\mu)^2 \geq t^2) = \frac{\partial}{\partial\mu}\mathbf{P}(|X+\mu| \geq t) &= \frac{\partial}{\partial\mu}F(-t+\mu) + F(-t-\mu)\\ &= f(t-\mu) - f(t+\mu) \geq 0.\end{aligned}$$

Daraus folgt, dass $\mu \mapsto \mathbf{P}((X+\mu)^2 \geq t^2)$ wachsend in μ ist. □

Definition 2.17 (χ^2-Verteilung)**.** *Seien* $X_1, ..., X_d \sim N(0,1)$ *unabhängig und* $\mu \in \mathbb{R}^d$*. Dann heißt die Verteilung von* $Y = \sum_{i=1}^d (\mu_i + X_i)^2$ *unzentrierte* χ^2*-Verteilung mit* d *Freiheitsgraden und Unzentriertheit* $|\mu|^2$*. Wir schreiben auch* $Y \sim \chi^2_d(|\mu|^2)$*.*

Bemerkung 2.18. In der Tat ist zunächst unklar, ob die Verteilung von Y in obiger Definition nur von $|\mu|^2$, und nicht vom gesamten Vektor μ abhängt. Wir beweisen nun

$$\sum_{i=1}^d (\mu_i + X_i)^2 \sim (|\mu| + X_1)^2 + \sum_{i=2}^d X_i^2.$$

Es sei bemerkt, dass hier die rechte Seite nur von $|\mu|^2$ abhängt.
Es gibt eine orthogonale Matrix O mit $O\mu = |\mu|e_1$. Nun ist $OX \sim N(0, I_d)$, also gilt

$$\sum_{i=1}^d (\mu_i + X_i)^2 = |\mu + X|^2 = |O\mu + OX|^2 \sim ||\mu|e_1 + X| = (|\mu| + X_1)^2 + \sum_{i=2}^d X_i^2.$$

Theorem 2.19. *Die Familie* $(\chi^2_d(|\mu|^2))_{|\mu|^2}$ *ist stochastisch wachsend.*

Beweis. Zunächst behaupten wir folgendes:

> Seien X, Y, Z Zufallsvariable, so dass X, Y und X, Z unabhängig sind. Ist dann Y stochastisch größer als Z, so ist auch $X + Y$ stochastisch größer als $X + Z$. (2.1)

(Für den Beweis hierfür siehe Übung.) Nun ist nach Bemerkung 2.18 $(|\mu| + X_1)^2 + \sum_{i=2}^{d} X_i^2 \sim \chi_d^2(|\mu|^2)$ als Summe zweier unabhängiger Zufallsvariablen. Nach Lemma 2.16 sind die Verteilungen von $(|\mu| + X_1)^2$ stochastisch wachsend in $|\mu|^2$. Nun folgt die Behauptung aus (2.1). □

2.4 Suffizienz

In unserem einführenden Beispiel wollten wir den Erfolgsparameter p eines p-Münzwurfes $X = (X_1, ..., X_{53})$ schätzen, wobei $X_1 + \cdots + X_{53} = 23$ war. Da wir nur Kenntnis dieser Summe hatten, jedoch nicht von den einzelnen Münzwürfen, können wir uns fragen, ob wir den Schätzer von $\hat{p} = 23/53$ verbessern können, wenn wir genauere Kenntnis über die einzelnen Würfe haben. Dies ist nicht der Fall, wie wir sehen werden. Der Grund dafür ist, dass $X_1 + \cdots + X_{53}$ eine für p suffiziente Statistik ist.

Definition 2.20 (Suffiziente Statistik)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\mathbf{P}_\vartheta(X \in da) = \rho_\vartheta(da)$. Eine Zufallsvariable $t(X)$ mit Zielbereich $\widetilde{S}$ heißt suffiziente Statistik für ϑ, falls für alle $y \in \widetilde{S}$*

$$\mathbf{P}_\vartheta(X \in \cdot \mid t(X) \in dy) \text{ nicht von } \vartheta \text{ abhängt.}$$

Beispiel 2.21 (Münzwurf)**.** Sei $X = (X_1, \ldots, X_n) \in \{0, 1\}^n$ ein p-Münzwurf mit noch unbestimmtem p. Der statistische Raum ist also $(X, (\mathbf{P}_p)_{p\in[0,1]})$, wobei

$$\mathbf{P}_p(X = (x_1, \ldots, x_n)) = p^k(1-p)^{n-k}, \text{ falls } k = \sum_{i=1}^{n} x_i.$$

Die Statistik

$$t(X_1, \ldots, X_n) = X_1 + \cdots + X_n$$

ist suffizient für p. Denn es gilt für $k = x_1 + \cdots + x_n$

$$\mathbf{P}_p(X = (x_1, \ldots, x_n) \mid X_1 + \cdots + X_n = k) = \frac{\mathbf{P}_p(X = (x_1, \ldots, x_n))}{\mathbf{P}_p(X_1 + \cdots + X_n = k)} = \frac{p^k(1-p)^{n-k}}{\binom{n}{k} p^k(1-p)^{n-k}} = \frac{1}{\binom{n}{k}},$$

unabhängig von p. Für $k \neq \sum_{i=1}^{n} x_i$ gilt

$$\mathbf{P}_p(X = (x_1, \ldots, x_n) \mid X_1 + \cdots + X_n = k) = 0,$$

ebenfalls unabhängig von p.

Beispiel 2.22 (Uniformes Modell)**.** Wir betrachten das statistische Modell $(X, (\mathbf{P}_\vartheta)_{\vartheta=[0,\infty)})$, wobei $X = (X_1, ..., X_n)$ ein unabhängiger Vektor ist mit $(X_i)_*\mathbf{P}_\vartheta = U([0, \vartheta])$, d.h. $X_1, ..., X_n$ sind unabhängig und uniform auf $[0, \vartheta]$ verteilt. Wir behaupten nun, dass

$$t(X) := \sup_{i=1,...,n} X_i$$

für ϑ suffizient ist.
Denn: Es gilt etwa für $x < y$

$$\mathbf{P}_\vartheta(X_1 \leq x | t(X) = y) = \frac{x/\vartheta \cdot (n-1)y^{n-2}/\vartheta^{n-1}}{ny^{n-1}/\vartheta^n} = \frac{n-1}{n}\frac{x}{y}$$

unabhängig von ϑ.

Theorem 2.23 (Fisher-Neyman'scher Faktorisierungssatz)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $T = t(X)$ mit $t : S \to S'$. Dann sind äquivalent:*

1. *T ist suffizient,*
2. *Es gibt $g_\theta : S' \to \mathbb{R}$ und $h : S \to \mathbb{R}$, so dass*

$$\mathbf{P}_\vartheta(X \in dx) = g_\vartheta(t(x))h(x)dx.$$

Beweis. Wir beweisen die Aussage nur im diskreten Fall. '2. $\Rightarrow$ 1.': Zunächst gilt $\mathbf{P}_\vartheta(X = x|t(X) = t) = 0$ für $t \neq t(x)$, was unabhängig von ϑ ist. Für $t = t(x)$ hingegen ist unter 2.

$$\mathbf{P}_\vartheta(X = x|t(X) = t) = \frac{g_\vartheta(t(x))h(x)}{\sum_{y:t(y)=t} g_\vartheta(t(y))h(y)} = \frac{g_\vartheta(t(x))h(x)}{g_\vartheta(t(x))\sum_{y:t(y)=t} h(y)} = \frac{h(x)}{\sum_{y:t(y)=t} h(y)},$$

was ebenfalls unabhängig von ϑ ist. Für '1. $\Rightarrow$ 2.' setzen wir

$$g_\vartheta(t) := \mathbf{P}_\vartheta(t(X) = t), \qquad h(x) = \mathbf{P}_\vartheta(X = x|t(X) = t(x)).$$

Dann ist $h(x)$ nach Voraussetzung unabhängig von ϑ und es gilt

$$\mathbf{P}_\vartheta(X = x) = \mathbf{P}_\vartheta(X = x, t(X) = t(x)) = h(x)g_\vartheta(t(x))$$

und die Behauptung ist gezeigt. □

Beispiel 2.24 (Münzwurf)**.** Im Beispiel des Münzwurfs aus Beispiel 2.21 ist $t(X) = X_1 + \cdots + X_n$. Hier ist

$$\mathbf{P}_\vartheta(X_1 = x_1, ..., X_n = x_n) = \vartheta^{t(X)}(1-\vartheta)^{n-t(X)},$$

woraus sich die Darstellung aus Theorem 2.23.2 mit $h = 1$ ergibt.

Beispiel 2.25 (Uniformes Modell)**.** Im uniformen Modell aus Beispiel 2.22 ist $t(X) = \sup_{i=1,...,n} X_i$ und

$$\mathbf{P}_\vartheta(X_1 \in dx_1, ..., X_n \in dx_n) = \frac{1}{\vartheta^n}1_{X_1,...,X_n\in[0,\vartheta)} = \frac{1}{\vartheta^n}1_{\sup_{i=1,...,n} X_i<\vartheta} \cdot 1_{\inf_{i=1,...,n} X_i\geq 0}.$$

Nun folgt die Suffizienz von $t(X)$ mit $h(x) = 1_{\inf_{i=1,...,n} x_i\geq 0}$ aus Theorem 2.23.

2.5 Exponentialfamilien

Viele Verteilungen, etwa die Normal- Poisson-, Binomial- und Exponentialverteilung, haben eine gemeinsame Struktur, die oftmals direkte Rechnungen ermöglicht. Diese Struktur wird in der folgenden Definition formalisiert.

Definition 2.26 (Exponentialfamilie)**.** *Sei $\Theta \subseteq \mathbb{R}^k$. Ein parametrisches statistisches Modell $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ heißt* k-parametrige Exponentialfamilie *(mit c, t, d, h) für*

$$c_1, ..., c_k, d : \Theta \to \mathbb{R}, \qquad t_1, ..., t_k, h : \mathbb{R}^n \to \mathbb{R},$$

falls

$$\mathbf{P}_\vartheta(X \in dx) = h(x) \cdot \exp\Big(\sum_{j=1}^{k} c_j(\vartheta)t_j(x) - d(\vartheta)\Big)dx = h(x) \cdot \exp\big(c(\vartheta)^\top t(x) - d(\vartheta)\big)dx.$$

Gilt insbesondere $c_j(\vartheta) = \vartheta_j$, also

$$p_\vartheta(x) = h(x) \cdot \exp\big(\vartheta^\top t(x) - d(\vartheta)\big), \qquad x \in \mathbb{R}^n,$$

so sagt man, die Exponentialfamilie sei in kanonischer Form*.*

Bemerkung 2.27 (1-parametrige Exponentialfamilie)**.** Für den Spezialfall einer 1-parametrigen Exponentialfamilie gibt es Funktionen c, d, t, h mit

$$p_\vartheta(x) = h(x) \cdot \exp\Big(c(\vartheta)t(x) - d(\vartheta)\Big), \qquad x \in \mathbb{R}^n.$$

So ziemlich alle statistischen Modelle, die auf uns bekannten Verteilungen basieren, sind Exponentialfamilien:

Beispiel 2.28. Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell.

1. Ist $\Theta = [0, 1]$ und $X = (X_1, ..., X_n)$ mit $X_1, ..., X_n$ unabhängig und $(X_i)_*\mathbf{P}_\vartheta = B(1, \vartheta)$, (d.h. wir betrachten das Binomialmodell aus Bemerkung 1.2), so gilt

$$\begin{aligned}\mathbf{P}_\vartheta(X = (x_1, ..., x_n)) &= \vartheta^{\sum_{i=1}^n x_i}(1-\vartheta)^{n-\sum_{i=1}^n x_i} \\ &= \exp\Big(\sum_{i=1}^n x_i \log\vartheta + \Big(n - \sum_{i=1}^n x_i\Big)\log(1-\vartheta)\Big) \\ &= \exp\Big(\log\frac{\vartheta}{1-\vartheta}\sum_{i=1}^n x_i + n\log(1-\vartheta)\Big)\end{aligned}$$

   und damit haben wir es mit einer 1-parametrigen Exponentialfamilie mit

$$c(\vartheta) = \log\frac{\vartheta}{1-\vartheta}, \qquad t(x) = \sum_{i=1}^n x_i, \qquad d(\vartheta = -n\log(1-\vartheta)$$

   zu tun.

2. Ist $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$ und $X_* \mathbf{P}_{(\mu,\sigma^2)} = N(\nu, \sigma^2)$ (d.h. wir betrachten das Normalverteilungsmodell aus Bemerkung 1.2), so ist

$$\begin{aligned}\mathbf{P}_{(\mu,\sigma^2)}(X \in dx) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)dx \\ &= \exp\Big(\frac{\mu}{\sigma^2}x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\Big(\frac{\mu}{\sigma^2} + \log(2\pi\sigma^2)\Big)\Big)dx.\end{aligned}$$

Also ist die Familie der (ein-dimensionalen) Normalverteilungen eine 2-parametrige Exponentialfamilie mit

$$\begin{aligned} c_1(\mu,\sigma^2) &= \frac{\mu}{\sigma^2}, & t_1(x) &= x, \\ c_2(\mu,\sigma^2) &= -\frac{1}{2\sigma^2}, & t_2(x) &= x^2, \\ h(x) &= 1, & d(\mu,\sigma^2) &= -\frac{1}{2}\Big(\frac{\mu}{\sigma^2} + \log(2\pi\sigma^2)\Big). \end{aligned}$$

Diese ist nun allerdings nicht in kanonischer Form.

3. Ist $\Theta = [0, \infty)$ und $X = (X_1, ..., X_n)$ mit $X_1, ..., X_n$ unabhängig und $(X_i)_* \mathbf{P}_\vartheta = U([0, \vartheta))$ (d.h. wir betrachten das uniforme Modell), so handelt es sich nicht um eine Exponentialfamilie.

Bemerkung 2.29 (Stichprobe aus einer Exponentialfamilie)**.** Ist $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ eine k-parametrige Exponentialfamilie mit Funktionen $c_1, ..., c_k, d, t_1, ..., t_k, h$, und sind $X_1, ..., X_n$ unabhängig und identisch nach $\mathbf{P}_\vartheta$ verteilt. Dann ist $((X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$, (d.h. die gemeinsame Verteilung von $X_1, ..., X_n$) ebenfalls eine Exponentialfamilie mit

$$c_1, ..., c_k, nd, \sum_{i=1}^N t_1 \circ \pi_i, ..., \sum_{i=1}^N t_k \circ \pi_i, \prod_{i=1}^N h \circ \pi_i$$

mit der Projektion $\pi_i(x) = x_i$.
Denn: Wir schreiben

$$\mathbf{P}(X_1 \in dx_1, ..., X_n \in dx_n) = h(x_1) \cdots h(x_n) \cdot \exp\Big(\sum_{j=1}^k c_j(\vartheta) \sum_{i=1}^n t_j(x_i) - nd(\vartheta)\Big).$$

Proposition 2.30 (Suffiziente Statistik bei Exponentialfamilien)**.**
Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ eine Exponentialfamilie (mit c, t, d, h). Dann ist die Statistik $T := t(X) = (t_1(X), ..., t_k(X))$ suffizient.

Beweis. Um die Suffizienz von T zu sehen, schreiben wir zunächst

$$\mathbf{P}_\vartheta(X \in dx) = h(x) g_\vartheta(t(x))$$

für

$$g_\vartheta(t(x)) = \exp(c(\vartheta)^\top t(x) - d(\vartheta)).$$

Damit folgt die Aussage aus dem Fisher-Neyman'schen Faktorisierungssatz, Theorem 2.23. □

2.6 Bayes'sche Modelle

Die Formel von Bayes ist wohlbekannt. Auf ihr beruht der große Zweig der Bayesianischen Statistik. Grundlegend ist hier, dass in einem statistischen Modell $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein Vorwissen über die Möglichkeiten besteht, welcher Parameter $\vartheta \in \Theta$ zutrifft. Dies wird in der a-priori-Verteilung zusammengefasst, einer Verteilung auf Θ.

Definition 2.31 (A-priori-Verteilung, a-posteriori-Verteilung)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}$. Eine* a-priori-Verteilung *ist die Verteilung π einer Zufallsvariable Ξ auf Θ. In diesem Fall wird durch*

$$\mathbf{P}(\Xi \in A, X \in B) := \int_A \mathbf{P}(\Xi \in d\vartheta)\mathbf{P}_\vartheta(X \in B)$$

die gemeinsame Verteilung von Ξ und X auf $\Theta \times S$ definiert. Die a-posteriori-*Verteilung ist dann die Verteilung*

$$\pi_x(d\vartheta) := \mathbf{P}(\Xi \in d\vartheta | X = x).$$

Bemerkung 2.32 (A-posteriori-Verteilung bei diskreten und stetigen Modellen)**.**

1. Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\Theta \subseteq \mathbb{Z}^k$ (also diskret), und Ξ eine Θ-wertige Zufallsvariable. Für die a-posteriori Verteilung gilt dann

   $$\mathbf{P}(\Xi = \vartheta | X = x) = \frac{\mathbf{P}(\Xi = \vartheta) \cdot \mathbf{P}_\vartheta(X \in dx)}{\sum_{\eta\in\Theta} \mathbf{P}(\Xi = \eta)\mathbf{P}_\eta(X \in dx)}.$$

   Wir bemerken, dass der Nenner nicht von ϑ abhängt und damit nur eine Normierungskonstante darstellt. Deshalb ist es äquivalent,[5]

   $$\mathbf{P}(\Xi = \vartheta | X = x) \sim_x \mathbf{P}(\Xi = \vartheta) \cdot \mathbf{P}_\vartheta(X \in dx)$$

   zu schreiben.

2. Ist $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\Theta \subseteq \mathbb{R}^k$, und ist Ξ eine Θ-wertige Zufallsvariable und $\Xi \sim \pi$ mit Dichte g, so hat die gemeinsame Verteilung von Ξ und X die Dichte $g(d\vartheta) \cdot \mathbf{P}_\vartheta(X \in dx)$. Für die a-posteriori Verteilung gilt dann

   $$\mathbf{P}(\Xi \in d\vartheta | X = x) = \frac{g(\vartheta) \cdot \mathbf{P}_\vartheta(X \in dx)}{\int g(\eta)\mathbf{P}_\eta(X \in dx)d\eta} d\vartheta,$$

   also

   $$\mathbf{P}(\Xi \in d\vartheta | X = x) \sim_x g(\vartheta) \cdot \mathbf{P}_\vartheta(X \in dx).$$

Es stellt sich heraus, dass die a-priori-Verteilung und die a-posteriori-Verteilung gerade bei Exponentialfamilien einen besonderen Zusammenhang haben. Dies unterstreicht nochmal die gute Handhabbarkeit dieser Verteilungen.

[5] Wir schreiben $a \sim_x b$, falls a/b nur von x abhängt (d.h. a und b sind proportional).

Proposition 2.33 (Konjugierte Familie bei Exponentialfamilien)**.** *Sei* $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ *eine* 1*-parametrige Exponentialfamilie (mit* c, t, d, h*) mit Dichte. Weiter sei* π *die a-priori-Verteilung einer* $\mathbb{R}$*-wertigen Zufallsvariable* Ξ *mit Dichte*

$$\mathbf{P}_{(r,s)}(\Xi \in d\vartheta) = \frac{\exp\Big(c(\vartheta)r - sd(\vartheta)\Big)}{\int \exp\Big(c(\eta)r - sd(\eta)\Big)d\eta} \sim \exp\Big(c(\vartheta)r - sd(\vartheta)\Big).$$

Dann ist die a-posteriori-Verteilung

$$\mathbf{P}(\Xi \in d\vartheta | X = x) = \mathbf{P}_{(r+t(x),s+1)}(\Xi \in d\vartheta).$$

Beweis. Es gilt

$$\begin{aligned}\mathbf{P}(\Xi \in d\vartheta | X = x) \sim_x \mathbf{P}_{(r,s)}(\Xi \in d\vartheta)\mathbf{P}_\vartheta(X \in dx) &\sim_x \exp\Big(c(\vartheta)(t(x)+r) - (s+1)d(\vartheta)\Big)\\ &\sim_x \mathbf{P}_{(r+t(x),s+1)}(\Xi \in d\vartheta).\end{aligned}$$

□

Beispiel 2.34 (A-posteriori-Verteilung bei der Normalverteilung)**.** Wir betrachten das Normalverteilungsmodell bei bekanntem σ^2, d.h. das statistische Modell $(X = (X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ mit $X_1, ..., X_n$ unabhängig und $(X_i)_*\mathbf{P}_\vartheta = N(\vartheta, \sigma^2)$. Also ist

$$\mathbf{P}_\vartheta(X_1 \in dx) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\vartheta)^2}{2\sigma^2}\Big) \sim_x \exp\Big(\frac{\vartheta x}{\sigma^2} - \frac{\vartheta^2}{2\sigma^2}\Big)$$

Nach Bemerkung 2.29 ist damit $(X = (X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ eine Exponentialfamilie mit

$$t(x) = \sum_{i=1}^n x_i, \qquad c(\vartheta) = \vartheta/\sigma^2 \qquad \text{und} \qquad d(\vartheta) = n\vartheta^2/(2\sigma^2).$$

Angenommen, Ξ ist die a-priori-verteilte Zufallsvariable mit Verteilung $\mathcal{N}(a, b^2)$ für $a, b \in \mathbb{R}$. Wie sieht dann die a-posteriori-Verteilung aus? Und ist diese um den wahren Wert ϑ konzentriert?
Um dies zu beantworten, verwenden wir Proposition 2.33. Wir haben für die a-priori-Verteilung

$$\mathbf{P}(\Xi \in d\vartheta) \sim \exp(-(\vartheta - a)^2/2b^2) \sim \exp(-\vartheta^2/2b^2 + \vartheta a/b^2).$$

Setzen wir

$$r = \sigma^2 a/b^2, \qquad s = \sigma^2/(nb^2),$$

so ist die a-priori-Verteilung also

$$\mathbf{P}_{(r,s)}(\Xi \in d\vartheta) \sim \exp\Big(\frac{\vartheta}{\sigma^2}r - s\frac{n\vartheta^2}{2\sigma^2}\Big).$$

Die a-posteriori-Verteilung ergibt sich damit mit dem Mittelwert $\overline{x}$ zu

$$\begin{aligned}\mathbf{P}(\Xi \in d\vartheta | X = x) = \mathbf{P}_{(r+t(x),s+1)}(\Xi \in d\vartheta) &\sim \exp\Big(\frac{\vartheta a}{b^2} + \frac{\vartheta}{\sigma^2}(x_1 + \cdots + x_n) - \frac{n\vartheta^2}{2\sigma^2} - \frac{\vartheta^2}{2b^2}\Big)\\ &\sim \exp\Big(-\frac{(\vartheta - \overline{x})^2}{2\sigma^2/n} - \frac{(\vartheta - a)^2}{2b^2}\Big) = \exp\Big(-\frac{(\vartheta - \alpha)^2}{2\beta^2}\Big)\end{aligned}$$

für

$$\alpha = \frac{\frac{\overline{x}}{\sigma^2/n} + \frac{a}{b^2}}{\frac{1}{\sigma^2/n} + \frac{1}{b^2}} = \frac{1}{\sigma^2/(nb^2) + 1}\overline{x} + \frac{\sigma^2/(nb^2)}{\sigma^2/(nb^2) + 1}a,$$
$$\beta = \frac{1}{n/\sigma^2 + 1/b^2}.$$

Damit ist gezeigt, dass die a-posteriori-Verteilung für große n um $\overline{x}$ konzentriert ist.

3 Schätzprobleme

Der in einem statistischen Modell freie Parameter ϑ (oder ein $h(\vartheta)$) kann aus Daten, d.h. der Realisierung der Zufallsvariable X geschätzt werden. Führt so eine Schätzung auf einen einzigen Wert, sprechen wir von Punktschätzern.

Definition 3.1 (Punktschätzer, unverzerrte und konsistente Schätzer)**.**

1. *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $m : \Theta \to \Theta'$. Jede Statistik $\widehat{m}(X)$ mit $\widehat{m} : S \to \Theta'$ heißt (Punkt-)Schätzer für $m(\vartheta)$. (Wir schreiben im Folgenden auch $m(\mathbf{P}_\vartheta) \equiv m(\vartheta)$.)*

   Der Bias *oder die* Verzerrung *von $\widehat{m}(X)$ ist gegeben als*

   $$b(\vartheta, m, \widehat{m}) := \mathbf{E}_\vartheta[\widehat{m}(X)] - m(\vartheta).$$

   Ist $b(\vartheta, m, \widehat{m}) = 0$ für alle ϑ, so sagt man, $\widehat{m}$ ist ein unverzerrter *(*erwartungstreuer, unbiased*) Schätzer für m.*

2. *Sei $(X^n, (\mathbf{P}^n_\vartheta)_{\vartheta\in\Theta})_{n=1,2,\dots}$ eine Folge statistischer Modelle mit derselben Parametermenge Θ und $\widehat{m}_1(X^1), \widehat{m}_2(X^2), \dots$ eine Folge von Schätzern für $m(\vartheta)$. Die Folge $\widehat{m}_1(X^1), \widehat{m}_2(X^2), \dots$ heißt konsistent, falls*

   $$\mathbf{P}^n_\vartheta[|\widehat{m}_n(X^n) - m(\vartheta)| \geq \varepsilon] \xrightarrow{n\to\infty} 0$$

   für alle $\varepsilon > 0, \vartheta \in \Theta$.

Bemerkung 3.2 (Varianz und mittlere quadratische Abweichung eines Schätzers)**.** Die Varianz eines Schätzers ist gegeben als

$$v(\vartheta, \widehat{m}) := \mathbf{V}_\vartheta[\widehat{m}(X)].$$

Die mittlere quadratische Abweichung ist

$$\mathbf{E}_\vartheta[(\widehat{m}(X) - m(\vartheta))^2].$$

Mit Bemerkung 1.4 sehen wir, dass

$$\mathbf{E}_\vartheta[(\widehat{m}(X) - m(\vartheta))^2] = v(\vartheta, \widehat{m}) + b(\vartheta, m, \widehat{m})^2.$$

Mit der Chebycheff-Ungleichung (Proposition **??**) ist also eine Folge $\widehat{m}_1(X^1), \widehat{m}_2(X^2), \dots$ von Schätzern für $m(\vartheta)$ insbesondere dann konsistent, wenn sie approximativ (d.h. im Grenzwert großer n) unverzerrt ist und

$$v(\vartheta, \widehat{m}_n) \xrightarrow{n\to\infty} 0$$

gilt.

Beispiel 3.3 (Erfolgswahrscheinlichkeit beim Münzwurf)**.** Betrachten wir noch einmal das einführende Beispiel der Schätzung des Erfolgsparameters in einem Münzwurf $X = (X_1, ..., X_n)$ in n Versuchen, d.h. $X_1, ..., X_n$ sind unabhängig und $B(1, p)$-verteilt. Wir betrachten zwei Schätzer, nämlich

$$\widehat{p}(X) := \overline{X}, \qquad \widehat{p}'(X) = X_1.$$

Man beachte, dass beide unverzerrt sind, denn

$$\mathbf{E}_p[\widehat{p}] = \mathbf{E}_p[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}_p[X_i] = \mathbf{E}_p[X_1] = \mathbf{E}_p[\widehat{p}'] = p.$$

Allerdings ist

$$\mathbf{V}_p[\widehat{p}] = \mathbf{V}_p\Big[\frac{X_1 + \cdots + X_n}{n}\Big] = \tfrac{1}{n^2}\sum_{i=1}^{n} \mathbf{V}_p[X_i] = \tfrac{1}{n}\mathbf{V}[X_1] = \tfrac{1}{n}p(1-p),$$
$$\mathbf{V}_p[X_1] = p(1-p),$$

und damit hat $\widehat{p}$ eine kleinere Varianz als $\widehat{p}'$ (was nicht erstaunen sollte).

3.1 Plugin- und momentenbasierte Schätzer

Wir behandeln nun grundlegende Prinzipien des Schätzens. Diese sind Plugin-Schätzer (Definition 3.4), den Spezialfall von momentenbasierten Schätzern (Definition 3.8). Im nächsten Abschnitt behandeln wir dann Maximum-Likelihood-Schätzer (Definition 3.10).

Definition 3.4 (Empirische Verteilung, Plugin-Schätzer)**.** *Sei $(X = (X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell. Dann heißt die (zufällige, diskrete) Wahrscheinlichkeitsverteilung auf $\{X_1, ..., X_n\}$, gegeben durch*

$$\widehat{\mathbf{P}}_X(A) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i\in A}$$

die empirische Verteilung *von X. (Es hat also $\widehat{\mathbf{P}}_X$ ein Wahrscheinlichkeitsgewicht von $1/n$ auf $X_1, ..., X_n$, d.h. für $Z \sim \widehat{\mathbf{P}}_X$ ist $\widehat{\mathbf{P}}_X(Z = x_i) = 1/n, i = 1, ..., n$.)*

Für $m : \Theta \to \Theta'$ heißt

$$\widehat{m}(X) := m(\widehat{\mathbf{P}}_X)$$

Plugin-Schätzer *für m.*

Beispiel 3.5 (Plugin-Schätzer für Erwartungswert und Varianz)**.** Seien $X_1, ..., X_n$ unter $\mathbf{P}_\vartheta$ identisch verteilt.

1. Wir wollen den Plugin-Schätzer für $m(\vartheta) := \mathbf{E}_\vartheta[X_1]$ angeben. Wir berechnen (und bezeichnen mit $\widehat{\mathbf{E}}_X$ die Erwartung bezüglich $\widehat{\mathbf{P}}_X$ und $Z \sim \widehat{\mathbf{P}}_X$)

$$\widehat{m}(X) = m(\widehat{\mathbf{P}}_X) = \widehat{\mathbf{E}}_X[Z] = \sum_{i=1}^{n} X_i \widehat{\mathbf{P}}_X(Z = X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i =: \overline{X},$$

   d.h. der Plugin-Schätzer ist gerade der Mittelwert der Daten.

2. Nun zum Plugin-Schätzer für $m(\vartheta) := \mathbf{V}_\vartheta[X_1]$ angeben. Wir berechnen (und bezeichnen mit $\widehat{\mathbf{V}}_X$ die Varianz bezüglich $\widehat{\mathbf{P}}_X$)

$$\widehat{m}(X) = \widehat{\mathbf{V}}_X[Z] = \widehat{\mathbf{E}}_X[Z^2] - \widehat{\mathbf{E}}_X[Z]^2 = \Big(\frac{1}{n}\sum_{i=1}^n X_i^2\Big) - \overline{X}^2$$
$$= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 =: \widetilde{s}(X).$$

Die gerade ermittelten Schätzer für Erwartungswert und Varianz einer Verteilung wollen wir nun etwas näher beleuchten.

Proposition 3.6 (Mittelwert und empirische Varianz als Schätzer für Erwartungswert und Varianz)**.** *Sei $(X = (X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, wobei $X_1, ..., X_n$ unter $\mathbf{P}_\vartheta$ reellwertig, unabhängig und identisch verteilt sind. Weiter sei $\mathbf{E}_\vartheta[X_1^4] < \infty$ für alle $\vartheta \in \Theta$.*

1. *Der Mittelwert*
$$\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$$
ist ein unverzerrter, konsistenter Schätzer für $\mathbf{E}_\vartheta[X_1]$.

2. *Die empirische Varianz*
$$s^2(X) := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$$
ist ein unverzerrter, konsistenter Schätzer für $\mathbf{V}_\vartheta[X_1]$.

Bemerkung 3.7. Der Plugin-Schätzer $\widetilde{s}^2(X)$ für $\mathbf{V}_\vartheta[X_1]$ unterscheidet sich vom unverzerrten Schätzer $s^2(X)$ durch den Faktor $(n-1)/n$ Insbesondere gilt also, da $s^2(X)$ unverzerrt ist,

$$b(\vartheta, \mathbf{V}_\cdot[X_1], \widetilde{s}^2) = \mathbf{E}_\vartheta[\widetilde{s}^2(X) - s^2(X)] = -\frac{1}{n}\mathbf{E}_\vartheta[s^2(X)] = -\frac{1}{n}\mathbf{V}_\vartheta[X_1].$$

Da dies für $n \to \infty$ gegen 0 konvergiert, ist immerhin $\widetilde{s}^2(X)$ nach Bemerkung 3.2 konsistent.

Beweis. 1. Zunächst ist

$$\mathbf{E}_\vartheta[\overline{X}] = \tfrac{1}{n}(\mathbf{E}_\vartheta[X_1] + \cdots + \mathbf{E}_\vartheta[X_n]) = \mathbf{E}_\vartheta[X_1],$$

was bereits die Unverzerrtheit von $\overline{X}$ als Schätzer von μ_ϑ zeigt. Für die Konsistenz schreiben wir für $\varepsilon > 0$

$$\mathbf{P}_\vartheta[|\overline{X} - \mathbf{E}_\vartheta[X_1]| \geq \varepsilon] = \mathbf{P}_\vartheta\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mathbf{E}_\vartheta[X_1]\right| \geq \varepsilon\right) \xrightarrow{n\to\infty} 0$$

nach dem schwachen Gesetz der großen Zahlen.
2. Wir schreiben zunächst

$$\mathbf{E}_\vartheta[s^2(X)] = \tfrac{1}{n-1}\sum_{i=1}^n \mathbf{E}_\vartheta[X_i^2 - 2X_i\overline{X} + \overline{X}^2] = \tfrac{n}{n-1}\mathbf{E}_\vartheta[X_1^2 - 2X_1\overline{X} + \overline{X}^2].$$

Nun ist

$$\begin{aligned}\mathbf{E}_\vartheta[X_1^2] &= \mathbf{E}_\vartheta[X_1]^2 + \mathbf{V}_\vartheta[X_1],\\ \mathbf{E}_\vartheta[X_1\overline{X}] &= \mathbf{E}_\vartheta[X_1]^2 + \tfrac{1}{n}\mathbf{V}_\vartheta[X_1],\\ \mathbf{E}_\vartheta[\overline{X}^2] &= \mathbf{E}_\vartheta[X_1\overline{X}],\end{aligned}$$

also

$$\mathbf{E}_\vartheta[s^2(X)] = \tfrac{n}{n-1}\mathbf{E}_\vartheta[X_1^2 - X_1\overline{X}] = \mathbf{V}_\vartheta[X_1],$$

was die Unverzerrtheit bereits zeigt. Die Konsistenz wird in einer Übungsaufgabe nachgeprüft. □

Definition 3.8 (Momentenbasierte Schätzer)**.** *Sei $(X = (X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, so dass $X_1, ..., X_n$ unabhängig und identisch verteilt sind. Weiter sei*

$$m(\mathbf{P}_\vartheta) = h(m_1(\mathbf{P}_\vartheta), ..., m_\ell(\mathbf{P}_\vartheta))$$

mit

$$m_k(\mathbf{P}_\vartheta) := \mathbf{E}_\vartheta[X_1^k]$$

das k-te Moment von $\mathbf{P}_\vartheta$. (Das heißt, die zu schätzende Funktion $m(\mathbf{P}_\vartheta)$ lässt sich als Funktion der ersten ℓ Momente schreiben.) Dann heißt

$$\widehat{m}_k(X) := \frac{1}{n}\sum_{i=1}^n X_i^k$$

auch k-tes empirisches Moment und der Schätzer

$$\widehat{m}(X) = h(\widehat{m}_1(X), ..., \widehat{m}_\ell(X))$$

heißt momentenbasierter Schätzer für m.

Beispiel 3.9 (Schätzung des Parameters einer Poisson-Verteilung)**.** Momentenbasierte Schätzer müssen nicht eindeutig sein, wie folgendes Beispiel zeigt. Seien $X_1, ..., X_n$ unabhängig und identisch verteilt mit $(X_1)_*\mathbf{P} = \mathrm{Poi}(\vartheta)$. Es gilt

$$m(\mathbf{P}_\vartheta) := \vartheta = m_1(\mathbf{P}_\vartheta) = m_2(\mathbf{P}_\vartheta) - m_1(\mathbf{P}_\vartheta)^2.$$

Deshalb ist sowohl

$$\widehat{m}(X) = \widehat{m}_1(X) := \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}$$

als auch

$$\widehat{\widehat{m}}(X) := \widehat{m}_2(X) - \widehat{m}_1(X)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2 = \widetilde{s}^2(X)$$

ein momentenbasierter Schätzer für ϑ.

3.2 Maximum-Likelihood-Schätzer

Das Konzept von Maximum-Likelihood-Schätzern geht davon aus, dass bereits Daten $X = x$ erhoben wurden. Der Maximum-Likelihood-Schätzer ist dann dasjenige ϑ, für das die Wahrscheinlichkeit, die Daten zu erhalten, maximiert wird.

Definition 3.10 (Maximum Likelihood)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell. Die Abbildung*

$$L : \begin{cases} S \times \Theta & \to [0,1] \\ (x,\vartheta) & \mapsto \mathbf{P}_\vartheta(X \in dx) \end{cases}$$

heißt Likelihood-Funktion*. Für eine Abbildung $h : S \mapsto \Theta$ mit*

$$L(x, h(x)) = \max_{\vartheta\in\Theta} L(x,\vartheta)$$

heißt $\widehat{\vartheta}_{ML} = h(X)$ Maximum-Likelihood-Schätzer *von ϑ.*

Bemerkung 3.11 (Interpretation von Maximum-Likelihood-Schätzern)**.** Sei X diskret und $X = x$. Da die Vorstellung die ist, dass die erhobenen Daten Ergebnis eines Zufallsexperiments (d.h. die Realisierung einer Zufallsvariable X) sind, sagt man auch, dass die Daten $X = x$ sind. Ein Maximum-Likelihood-Schätzer ist also ein Parameter ϑ, unter dem die Wahrscheinlichkeit, die Daten $X = x$ zu beobachten – das ist $\mathbf{P}_\vartheta(X = x)$ – maximal ist.

Beispiel 3.12 (Maximum-Likelihood-Schätzer für μ und σ^2 von Normalverteilungen)**.** Wir betrachten den Fall einer unabhängigen, normalverteilten Stichprobe. Sei also $(X, (\mathbf{P}_{(\mu,\sigma^2)})_{(\mu,\sigma^2)\in\Theta})$ mit $\Theta = \mathbb{R} \times \mathbb{R}_+$ so, dass $X_1, \dots, X_n$ unabhängig und identisch nach verteilt sind mit $(X_1)_*\mathbf{P}_{(\mu,\sigma^2)} = N(\mu, \sigma^2)$.

Wir berechnen nun die Maximum-Likelihood-Schätzer für μ und σ^2. Anstatt die Likelihood-Funktion zu maximieren, werden wir dasselbe für deren Logarithmus tun. Wir schreiben

$$\begin{aligned} \log L((X_1,\dots,X_n),(\mu,\sigma^2)) &= \log\left(\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2}\Big)\right) \\ &= -n\log\sigma - \sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2} + C, \end{aligned}$$

wobei C weder von μ noch von σ abhängt. Ableiten nach μ und σ ergibt

$$\begin{aligned} \frac{\partial \log L((X_1,\dots,X_n),(\mu,\sigma^2))}{\partial\mu} &= \sum_{i=1}^n \frac{X_i-\mu}{\sigma^2}, \\ \frac{\partial \log L((X_1,\dots,X_n),(\mu,\sigma^2))}{\partial\sigma} &= -\frac{n}{\sigma} + \sum_{i=1}^n \frac{(X_i-\mu)^2}{\sigma^3}. \end{aligned}$$

Für die Maximum-Likelihood-Schätzer $\widehat{\mu}_{ML}$ und $\widehat{\sigma}^2_{ML}$ gilt notwendigerweise

$$\begin{aligned} \sum_{i=1}^n (X_i - \widehat{\mu}_{ML}) &= 0, \\ \frac{n}{\widehat{\sigma}_{ML}} - \sum_{i=1}^n \frac{(X_i-\widehat{\mu}_{ML})^2}{\widehat{\sigma}^3_{ML}} &= 0. \end{aligned}$$

Die Maximum-Likelihood-Schätzer sind also gegeben durch

$$\widehat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X},$$
$$\widehat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \widetilde{s}^2(X).$$

Insbesondere sehen wir, dass $\overline{X}$ nicht nur erwartungstreu und konsistent (siehe Theorem **??**) ist, sondern auch ein Maximum-Likelihood-Schätzer für μ. Allerdings ist der Maximum-Likelihood-Schätzer für σ^2 nicht erwartungstreu, wie man aus Proposition 3.6 abliest. Immerhin ist $\widehat{\sigma}^2_{ML}$ für große n annähernd erwartungstreu, da $\widehat{\sigma}^2_{ML} - s^2(X) \xrightarrow{n\to\infty} 0$.

Beispiel 3.13 (Maximum-Likelihood-Schätzer des Parameters einer Poisson-Verteilung)**.** Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\geq 0})$ so, dass $X = (X_1, \dots, X_n)$ und $X_1, \dots, X_n$ unabhängig und identisch nach $\mathrm{Poi}(\vartheta)$ verteilt ist. Wir wissen bereits, dass $\mathbf{E}_\vartheta[X_1] = \mathbf{V}_\vartheta[X_1] = \vartheta$. Damit folgt, dass sowohl $\overline{X}$ als auch $s^2(X)$ unverzerrte Schätzer für ϑ sind, siehe Proposition 3.6. Wir berechnen nun den Maximum-Likelihood-Schätzer für ϑ (welcher sich als $\overline{X}$ herausstellen wird). Später, in Beispiel 3.26, werden wir sehen, dass dieser dem Schätzer $s^2(X)$ vorzuziehen ist, da er eine kleinere Risikofunktion besitzt.

Für den Maximum-Likelihood-Schätzer für ϑ berechnen wir zunächst wieder die log-Likelihood-Funktion

$$\log L((X_1, \dots, X_n), \vartheta) = \log \prod_{i=1}^{n} e^{-\vartheta}\frac{\vartheta^{X_i}}{X_i!} = -n\vartheta + \log(\vartheta)\sum_{i=1}^{n} X_i + C,$$

wobei C nicht von ϑ abhängt. Also ist

$$\frac{\partial \log L((X_1, \dots, X_n), \vartheta)}{\partial \vartheta} = -n + \frac{1}{\vartheta}\sum_{i=1}^{n} X_i. \tag{3.1}$$

Die log-Likelihood-Funktion ist also maximal für

$$-n + \frac{1}{\widehat{\vartheta}_{ML}}\sum_{i=1}^{n} X_i = 0, \qquad \widehat{\vartheta}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Damit ist $\widehat{\vartheta}_{ML} = \overline{X}$ der Maximum-Likelihood-Schätzer für ϑ. □

Maximum-Likelihood-Schätzer sind in vielen Fällen konsistent, was sicher eine wünschenswerte Eigenschaft ist. Der nächste Satz diese Konsistent von Maximum-Likelihood-Schätzern in einem einfachen Fall.

Theorem 3.14 (Konsistenz von Maximum-Likelihood-Schätzern)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}$ endlich, so dass $X = (X_1, \dots, X_n)$ unabhängig und identisch verteilt sind und X_1 eine Zufallsvariable mit Zielbereich $\mathbb{R}$ und einer beschränkten Dichte f_ϑ ist. Dann ist die Folge von Maximum-Likelihood-Schätzern von ϑ für $n = 1, 2, \dots$ konsistent.*

Bemerkung 3.15. Der Satz gilt auch unter schwächeren Voraussetzungen. (Etwa sollte Θ kompakt sein und $\vartheta \mapsto L(a, \vartheta)$ stetig.)

Beweis von Theorem 3.14. Zunächst ist für alle ϑ, ϑ' wegen der Jensen'schen Ungleichung, und, da $x \mapsto \log(x)$ konkav ist,

$$\mathbf{E}_{\vartheta'}\Big[\log \frac{f_\vartheta(X)}{f_{\vartheta'}(X)}\Big] \leq \log \mathbf{E}_{\vartheta'}\Big[\frac{f_\vartheta(X)}{f_{\vartheta'}(X)}\Big] = \log \int f_{\vartheta'}(x)\frac{f_\vartheta(x)}{f_{\vartheta'}(x)}dx = \log \int f_\vartheta(x)dx = \log 1 = 0.$$

Der Maximum-Likelihood-Schätzer maximiert die Funktion, die ϑ auf

$$\frac{1}{n}\log L((X_1,\ldots,X_n),\vartheta) - \frac{1}{n}\log L((X_1,\ldots,X_n),\vartheta_0) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{f_\vartheta(X_i)}{f_{\vartheta_0}(X_i)}$$

abbildet. Ist ϑ_0 der wahre Parameter, ist wegen des starken Gesetzes der großen Zahlen

$$\frac{1}{n}\log L((X_1,\ldots,X_n),\vartheta) - \frac{1}{n}\log L((X_1,\ldots,X_n),\vartheta_0) \xrightarrow{n\to\infty} \mathbf{E}_{\vartheta_0}\Big[\log\frac{f_\vartheta(X)}{f_{\vartheta_0}(X)}\Big] \leq 0$$

mit $= 0$ genau dann, wenn $f_\vartheta = f_{\vartheta_0}$. Da Θ diskret ist, konvergiert die Folge der Maximum-Likelihood-Schätzer für ϑ gegen ϑ_0. □

3.3 Optimalitätskriterien von Schätzern

Natürlich wollen wir möglichst gute Schätzer finden. Vorher ist zu klären, in welchen Sinn diese Qualität der Schätzer denn gemeint ist. Wir beschränken uns hier auf den mittleren quadratischen Fehler, also eine quadratische Verlustfunktion.

Definition 3.16 (Mittlerer quadratischer Fehler)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $m : \vartheta \mapsto m(\vartheta)$.*

1. *Für einen Schätzer $h(X)$ von $m(\vartheta)$ ist der* mittlere quadratische Fehler *(oder die* Risikofunktion*) definiert als*

$$R_{h(X)} : \begin{cases} \Theta & \to [0,\infty], \\ \vartheta & \mapsto \mathbf{E}_\vartheta[(h(X)-m(\vartheta))^2]. \end{cases}$$

2. *Sei $m : \Theta \to \Theta'$ und $\mathcal{S} \subseteq \{h : S \to \Theta'\}$ eine Menge von Schätzern. Falls für ein $h \in \mathcal{S}$ gilt, dass für alle $\vartheta \in \Theta$*

$$R_{h(X)}(\vartheta) = \inf_{h\in\mathcal{S}} R_{h(X)}(\vartheta),$$

   so heißt $h(X)$ bester Schätzer in $\mathcal{S}$.

3. *Ist insbesondere $\mathcal{S} = \{h(X) : \mathbf{E}_\vartheta[h(X)] = m(\vartheta), \vartheta \in \Theta\}$ die Menge der unverzerrten Schätzer, so heißt ein bester Schätzer in $\mathcal{S}$ auch* UMVUE *(Uniformly Minimal Variance Unbiased Estimator).*

Mit Hilfe suffizienter Statistiken kann man vorhandene Schätzer besser machen. Dies ist Inhalt folgenden Satzes.

Theorem 3.17 (Satz von Rao-Blackwell)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $m : \Theta \to \Theta'$, $h : S \to \Theta'$ mit $h(X)$ einem Schätzer für $m(\vartheta)$. Ist $t(X)$ suffizient für ϑ, so ist*

$$\widetilde{h}(X) = \mathbf{E}_\vartheta[h(X)|t(X)] \tag{3.2}$$

unabhängig von ϑ und

$$\mathbf{E}_\vartheta[(\widetilde{h}(X)-m(\vartheta))^2] \leq \mathbf{E}_\vartheta[(h(X)-m(\vartheta))^2].$$

Bemerkung 3.18 (Interpretation)**.** Das Theorem definiert eine Funktion $\widetilde{h} : S \to \Theta$. Wichtig ist, dass $\widetilde{h}(X)$ nur von $t(X)$ abhängt. Es gilt nämlich

$$\widetilde{h}(x) = \mathbf{E}_\vartheta[h(X)|t(X) = t(x)]$$

nach der Definition der bedingten Erwartung, und die rechte Seite hängt wegen der Definition der bedingten Erwartung nur von $t(x)$ ab. Die Aussage des Satzes ist, dass der Schätzer $\widetilde{h}(X)$ eine kleinere Risikofunktion hat als $h(X)$.

Beispiel 3.19 (Die Schätzer $\widehat{p}$ und $\widehat{p}'$ beim Münzwurf)**.** Sei wieder $X = (X_1, \ldots, X_n)$ ein p-Münzwurf mit noch unbestimmten p. In Beispiel 3.3 haben wir zwei erwartungstreue Schätzer für p kennen gelernt, nämlich

$$\widehat{p}(X) = \tfrac{1}{n}(X_1 + \cdots + X_n), \qquad \widehat{p}'(X) = X_1$$

und hatten auch festgestellt, dass $\widehat{p}$ eine kleinere Varianz (also Risikofunktion) besitzt als $\widehat{p}'$. Weiter wissen wir aus Beispiel 2.21, dass $t(X) = X_1 + \cdots + X_n$ suffizient für p ist. Wir können also nun den Schätzer $\widehat{p}'$ mit Hilfe des Satzes von Rao-Blackwell verbessern, indem wir $h(X) = \widehat{p}'(X)$ setzen und $\widetilde{h}(X)$ aus (3.2) berechnen. Dies ergibt aus Symmetriegründen

$$\begin{aligned}\widetilde{h}(X) = \mathbf{E}_p[\widehat{p}'(X) \mid t(X)] &= \mathbf{E}_p[X_1 \mid X_1 + \cdots + X_n] = \tfrac{1}{n}\sum_{i=1}^{n} \mathbf{E}_p[X_i \mid X_1 + \cdots + X_n] \\ &= \tfrac{1}{n}\mathbf{E}_p[X_1 + \cdots + X_n \mid X_1 + \cdots + X_n] = \tfrac{1}{n}\big(X_1 + \cdots + X_n\big) \\ &= \widehat{p}(X).\end{aligned}$$

Insbesondere sehen wir hier ein konkretes Beispiel dafür, dass $\widetilde{h}(X)$ eine echt kleinere Risikofunktion besitzt als $h(X)$.

Beweis von Theorem 3.17. Wir beschränken den Beweis auf den Fall diskreter Zufallsvariablen X. Der Fall von Zufallsvariablen mit Dichten geht analog. Zunächst ist

$$\mathbf{E}_\vartheta[h(X) \mid t(X)] = \sum_{a \in S} h(a)\mathbf{P}_\vartheta(X = a \mid t(X)),$$

was wegen der Suffizienz von $t(X)$ nicht von ϑ abhängt. Wegen (siehe Proposition **??**)

$$\mathbf{E}_\vartheta[\widetilde{h}(X)] = \mathbf{E}_\vartheta[\mathbf{E}_\vartheta[h(X) \mid t(X)]] = \mathbf{E}_\vartheta[h(X)]$$

gilt mit der Varianzformel (Proposition **??**)

$$\begin{aligned}\mathbf{E}_\vartheta[(\widetilde{h}(X) - m(\vartheta))^2] &= \big(\mathbf{E}_\vartheta[\widetilde{h}(X)] - m(\vartheta)\big)^2 + \mathbf{Var}_\vartheta[\mathbf{E}_\vartheta[h(X) \mid t(X)]] \\ &\le \big(\mathbf{E}_\vartheta[h(X)] - m(\vartheta)\big)^2 + \mathbf{Var}_\vartheta[\mathbf{E}_\vartheta[h(X)|t(X)]] + \mathbf{E}_\vartheta[\mathbf{Var}_\vartheta[h(X)|t(X)]] \\ &= \mathbf{E}_\vartheta[(h(X) - m(\vartheta))^2]\end{aligned}$$

und die Behauptung ist gezeigt. □

Bemerkung 3.20 (Satz von Lehmann-Scheffe)**.** Betrachten wir die Klasse erwartungstreuer Schätzer. Der Satz von Rao-Blackwell erlaubt die Verbesserung von Schätzern mit Hilfe suffizienter Statistiken. Wir wissen jedoch nicht, ob die Verbesserung optimal ist, d.h. ob es

sich bei den verbesserten Schätzern um UMVUE-Schätzern handelt. Eine Antwort auf diese Frage gibt der Satz von Lehmann-Scheffe:
Ist die suffiziente Statistik im Satz von Rao-Blackwell vollständig, d.h. für alle g gilt

$$\mathbf{E}_\vartheta[g(t(X))] = 0, \vartheta \in \Theta \Rightarrow \mathbf{P}_\vartheta(g(t(X)) = 0) = 1, \vartheta \in \Theta,$$

so ist $\widetilde{h}(X)$ im Satz von Rao-Blackwell ein UMVUE. Weiter sind die suffizienten Statistiken von Exponentialfamilien vollständig.
Dies bedeutet etwa, dass im Münzwurf-Beispiel der Schätzer $\overline{X}$ ein UMVUE ist.

Nachdem wir nun gesehen haben, wie man Schätzer besser machen kann, wollen wir nun wissen, wie groß die minimale Varianz eines Schätzers denn sein kann. Hierzu benötigen wir ein neues, von der Likelihood-Funktion abgeleitetes, Konzept. Wir beschränken uns dabei auf folgende Situation:

Definition 3.21 (Fisher-Information)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $(x, \vartheta) \mapsto L(x, \vartheta) = \mathbf{P}_\vartheta(X \in dx)$ die Likelihood-Funktion. Folgende Voraussetzungen seien erfüllt:*

1. *$\Theta \subseteq \mathbb{R}$ ist offen;*
2. *$\{x : L(x, \vartheta) > 0\}$ hängt nicht von ϑ ab.*
3. *Die Ableitung $\frac{\partial}{\partial\vartheta} \log L(x, \vartheta)$ existiert und ist endlich;*
4. *Falls X unter $\mathbf{P}_\vartheta$ eine Dichte p_ϑ hat: Ist $t : S \to \mathbb{R}$, so dass $\mathbf{E}_\vartheta[|t(X)|] < \infty$ für alle $\vartheta \in \Theta$, so gilt*
$$\frac{\partial}{\partial\vartheta} \int t(x)p_\vartheta(x)dx = \int t(x)\frac{\partial}{\partial\vartheta}p_\vartheta(x)dx.$$

Dann heißt $\frac{\partial}{\partial\vartheta} \log L(X, \vartheta)$ Score-Funktion *und*

$$\vartheta \mapsto \mathcal{I}(\vartheta) := \mathbf{E}_\vartheta\Big[\Big(\frac{\partial}{\partial\vartheta} \log L(X, \vartheta)\Big)^2\Big]$$

Fisher-Information.

Bemerkung 3.22 (Fisher-Information einer unabhängigen Stichprobe)**.** Ist $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit Likelihood-Funktion $(x, \vartheta) \mapsto L(x, \vartheta)$ und $\mathcal{I}_1(\vartheta)$ die Fisher-Information. Ist dann $X_1, ..., X_n$ unabhängig mit $X_1, ..., X_n \sim X$ unter $\mathbf{P}_\vartheta$, d.h. $X_1, ..., X_n$ sind eine unabhängige Stichprobe der Verteilung $\mathbf{P}_\vartheta$, $\vartheta \in \Theta$. Dann ist

$$\vartheta \mapsto \mathcal{I}(\vartheta) = \mathbf{E}_\vartheta\Big[\Big(\frac{\partial}{\partial\vartheta} \sum_{i=1}^n \log L(X_i, \vartheta)\Big)^2\Big] = \sum_{i=1}^n \mathbf{E}_\vartheta\Big[\Big(\frac{\partial}{\partial\vartheta} \log L(X_i, \vartheta)\Big)^2\Big] = n \cdot \mathcal{I}_1(\vartheta)$$

die Fisher-Information des Modells $((X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$.

Bemerkung 3.23 (Einparametrige Exponentialfamilien erfüllen Voraussetzungen)**.** Ist die Dichte durch eine ein-parametrige Exponentialfamilie mit

$$p_\vartheta(x) = 1_A(x) \exp(c(\vartheta)t(x) - d(\vartheta))$$

und $\frac{\partial}{\partial\vartheta}c(\vartheta) \neq 0$ für alle $\vartheta \in \Theta$ und $\Theta \subseteq \mathbb{R}$ offen, dann sind die Voraussetzungen von Definition 3.21 erfüllt.

Bemerkung 3.24 (Erwartung der Score-Funktion verschwindet)**.** Im Fall einer Dichte schreiben wir

$$\mathcal{I}(\vartheta) = \int \Big(\frac{\partial}{\partial\vartheta}\log p_\vartheta(x)\Big)^2 p_\vartheta(x)dx = \int \frac{\big(\frac{\partial p_\vartheta(x)}{\partial\vartheta}\big)^2}{p_\vartheta(x)}dx$$

und für die Erwartung der Score-Funktion gilt

$$\mathbf{E}_\vartheta\Big[\frac{\partial}{\partial\vartheta}\log p_\vartheta(X)\Big] = \int \frac{\frac{\partial p_\vartheta(x)}{\partial\vartheta}}{p_\vartheta(x)} p_\vartheta(x)dx = \frac{\partial}{\partial\vartheta}\int p_\vartheta(x)dx = 0.$$

Also gilt

$$\mathcal{I}(\vartheta) = \mathbf{V}_\vartheta\Big[\frac{\partial}{\partial\vartheta}\log L(X,\vartheta)\Big]$$

Theorem 3.25 (Cramér-Rao-Schranke)**.** *Sei $(X,(\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\vartheta\mapsto\mathcal{I}(\vartheta)$ die Fisher-Information. Ist $t: S\to\mathbb{R}$, so dass $\mathbf{V}_\vartheta[t(X)]<\infty$ für alle $\vartheta\in\Theta$ und $\Psi(\vartheta):=\mathbf{E}_\vartheta[t(X)]$. Sind die Bedingungen aus Definition 3.21 erfüllt, so ist Ψ differenzierbar und es gilt*

$$\mathbf{V}_\vartheta[t(X)] \geq \frac{(\Psi'(\vartheta))^2}{\mathcal{I}(\vartheta)}, \qquad \vartheta\in\Theta.$$

Ist also insbesondere t ein unverzerrter Schätzer für ϑ, so gilt

$$\mathbf{V}_\vartheta[t(X)] \geq \frac{1}{\mathcal{I}(\vartheta)}, \qquad \vartheta\in\Theta.$$

Gilt Gleichheit, so ist $t(X)$ also ein UMVUE und heißt auch effizient*.*

Beweis. Wir führen den Beweis im Fall kontinuierlicher $\mathbf{P}_\vartheta, \vartheta$, d.h. $X_*\mathbf{P}_\vartheta$ hat die Dichte p_ϑ, $\vartheta\in\Theta$. Zunächst ist

$$\Psi'(\vartheta) = \frac{\partial}{\partial\vartheta}\int t(x)p_\vartheta(x)dx = \int t(x)\frac{\partial}{\partial\vartheta}p_\vartheta(x)dx = \mathbf{E}_\vartheta\Big[\frac{\partial}{\partial\vartheta}t(X)\log p_\vartheta(X)\Big].$$

Daraus folgt, mit Bemerkung 3.24 und der Cauchy-Schwartz-Ungleichung, Proposition **??**

$$\begin{aligned}(\Psi'(\vartheta))^2 &= \Big(\mathbf{E}_\vartheta\Big[t(X)\frac{\partial}{\partial\vartheta}\log p_\vartheta(X)\Big]\Big)^2 = \mathbf{COV}_\vartheta\Big[t(X),\frac{\partial}{\partial\vartheta}\log p_\vartheta(X)\Big]^2\\ &\leq \mathbf{V}_\vartheta[t(X)]\cdot\mathbf{V}_\vartheta\Big[\frac{\partial}{\partial\vartheta}\log p_\vartheta(X)\Big] = \mathbf{V}_\vartheta[t(X)]\cdot\mathcal{I}(\vartheta).\end{aligned}$$

Daraus folgen alle Behauptungen. □

Beispiel 3.26 (Effizienz des Schätzers für den Parameter einer Poisson-Verteilung)**.**
Sei $(X,(\mathbf{P}_\lambda)_{\lambda>0})$ so, dass $X=(X_1,\dots,X_n)$ und $X_1,\dots,X_n$ unabhängig nach $\mathrm{Poi}(\lambda)$ verteilt ist. Wir haben bereits in Beispiel 3.13 berechnet, dass

$$\widehat{\lambda}_{ML} = \tfrac{1}{n}\sum_{i=1}^n X_i$$

der Maximum-Likelihood-Schätzer für λ ist. Wir wollen nun sehen, ob $\widehat{\lambda}_{ML}$ auch effizient ist. Dafür berechnen wir die Fisher-Information (siehe (3.1) für die Ableitung der log-Likelihood-Funktion)

$$\mathcal{I}(\lambda) = \mathbf{E}_\lambda\Big[\Big(\frac{\partial}{\partial\lambda}\log L(X,\lambda)\Big)^2\Big] = \mathbf{E}_\lambda\Big[\Big(-n+\frac{1}{\lambda}\sum_{i=1}^n X_i\Big)^2\Big] = \mathbf{V}_\lambda\Big[\frac{1}{\lambda}\sum_{i=1}^n X_i\Big] = \frac{n}{\lambda}.$$

Der Schätzer $\widehat{\lambda}_{ML}$ ist also effizient, da

$$\mathbf{V}_\lambda[\widehat{\lambda}_{ML}] = \frac{\lambda}{n} = \frac{1}{\mathcal{I}(\lambda)}.$$

4 Testprobleme

4.1 Grundbegriffe

Neben Schätzproblemen sind Testprobleme das wichtigste Thema der induktiven Statistik. In der empirischen Forschung ist es oftmals so, dass aufgrund eines ersten Verständnisses eine Hypothese über die erworbenen Daten aufgestellt werden kann. Das Überprüfen solcher Hypothesen erfolgt dann mittels statistischer Tests. Wie üblich geht man davon aus, dass die Daten die Realisierung einer Zufallsvariable sind.

Wir beginnen mit der Einführung wichtiger Begriffe wie Teststatistik, Nullhypothese, Alternative, Ablehnungsbereich, Signifikanzniveau und p-Wert. Diese sind teilweise aus der Schule bekannt.

Definition 4.1 (Statistischer Test)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, S der Zielbereich von X, und $\Theta_0, \Theta_A \subseteq \Theta$ disjunkt mit $\Theta_0 \cup \Theta_A = \Theta$. $H_0 \vartheta \in \Theta_0$ heißt* Nullhypothese *und $H_A \vartheta \in \Theta_A$ heißt* Alternativhypothese*.*

1. *Die Hypothese $H : \vartheta \in \Theta_H$ (die entweder H_0 oder H_A sein kann) heißt* einfach *wenn $\Theta_H = \{\vartheta^*\}$ für ein $\vartheta^* \in \Theta$. Andernfalls heißt H* zusammengesetzt*.*

2. *Eine Abbildung $\varphi : S \to [0, 1]$ heißt* (randomisierter) statistischer Test *von*

$$H_0 : \vartheta \in \Theta_0 \text{ gegen } H_A : \vartheta \in \Theta_A.$$

Hier ist $\varphi(x)$ die Wahrscheinlichkeit, sich für die Alterantive H_A zu entscheiden (also H_0 abzulehnen), falls die Daten $X = x$ sind. (Man stelle sich also vor, dass die Daten $X = x$ vorliegen, und man anschließend einen $\varphi(x)$-Münzwurf macht und daraus entscheidet, ob man die Null- oder Alternativhypothese annimmt. Insbesondere kann es bei gleichen Daten zu unterschiedlichen Entscheidungen kommen.) Man sagt, der Test hat (Signifikanz-)Niveau $\alpha \in [0, 1]$*, falls*

$$\sup_{\vartheta\in\Theta_0} \mathbf{E}_\vartheta[\varphi(X)] \leq \alpha. \tag{4.1}$$

3. *Der Spezialfall eines* nicht-randomisierten, statistischen Tests *ist es, wenn $\varphi(S) = \{0, 1\}$, man sich also immer gleich für Null- oder Alternativhypothese bei Vorliegen von $X = x$ entscheidet. In diesem Fall ist $Y = t(X)$ eine* Teststatistik *und $C := t(\varphi^{-1}(1)) \subseteq S$ der* kritische Bereich *oder* Ablehnungsbereich *des Test. Insbesondere entscheidet man sich genau dann für die Alternative, falls $\varphi(x) = 1$, d.h. $Y = t(x) \in t(\varphi^{-1}(1)) = C$. Nun sagt man auch, dass das Paar (Y, C) der* statistische Test *von*

$$H_0 : \vartheta \in \Theta_0 \text{ gegen } H_A : \vartheta \in \Theta_A$$

ist. Nun hat der Test (T, C) das (Signifikanz-)Niveau $\alpha \in [0, 1]$*, falls*

$$\sup_{\vartheta\in\Theta_0} \mathbf{P}_\vartheta(Y \in C) = \sup_{\vartheta\in\Theta_0} \mathbf{E}_\vartheta[\varphi(X)] \leq \alpha. \tag{4.2}$$

Falls $Y \in C$, sagt man, dass H_0 abgelehnt (und damit H_A angenommen) ist. Falls $Y \notin C$, sagt man, dass H_0 nicht abgelehnt ist (und H_A abgelehnt ist).

4. *Ist $\Theta = (\underline{\vartheta}, \overline{\vartheta})$ ein Intervall (wobei $\underline{\vartheta} = -\infty$ und $\overline{\vartheta} = \infty$ zugelassen sind und die Intervalle auch abgeschlossen sein können), so heißt der Test* einseitig, *falls $\Theta_0 = (\underline{\vartheta}, \vartheta^*)$ oder $\Theta_0 = (\vartheta^*, \overline{\vartheta})$. Falls $\Theta_0 = (\vartheta^+, \vartheta^*)$ mit $\underline{\vartheta} < \vartheta^+ \leq \vartheta^* < \overline{\vartheta}$, so heißt der Test* zweiseitig.

5. *Gilt*

$$\sup_{\vartheta \in \Theta_0} \mathbf{E}_\vartheta[\varphi(X)] < \alpha,$$

so heißt der Test φ konservativ *zum Niveau α.*

6. *Der Test φ heißt* unverfälscht, *falls*

$$\mathbf{E}_{\vartheta_0}[\varphi(X)] \leq \mathbf{E}_{\vartheta_A}[\varphi(X)]$$

für alle $\vartheta_0 \in \Theta_0, \vartheta_A \in \Theta_A$ gilt.

Bemerkung 4.2 (Interpretation und Fehler eines Tests)**.**

1. Einen (nicht-randomisierten) statistischen Test hat man sich am besten so vorzustellen (siehe auch das nächste Beispiel): die Daten sind gegeben durch die Zufallsvariable X. Diese Daten fasst man durch die meist reellwertige Funktion t zusammen zur Teststatistik $Y = t(X)$. Die Daten können entweder nach $\mathbf{P}_\vartheta$ mit $\vartheta \in \Theta_0$ (d.h. die Nullhypothese ist richtig) oder mit $\vartheta \in \Theta_A$ (d.h. die Alternativhypothese ist richtig) verteilt sein. Ziel ist es, die Nullhypothese genau dann (anhand der Daten X) abzulehnen, wenn H_A richtig ist. Der Ablehnungsbereich C ist so gewählt, dass H_0 genau dann abgelehnt wird, wenn $Y \in C$.

2. Bei randomisierten (und nicht-randomisierten) Tests kommt es immer zu einer Entscheidung für H_0 oder H_A. Dabei können zwei verschiedene Arten von Fehler auftreten:

	H_0 abgelehnt	H_0 nicht abgelehnt
H_0 richtig	Fehler erster Art	richtige Entscheidung
H_0 falsch	richtige Entscheidung	Fehler zweiter Art

Gehen wir zunächst davon aus, dass $\vartheta \in \Theta_0$. Hat der Test ein Niveau α, so wissen wir, dass $\mathbf{E}_\vartheta[\varphi(X)] \leq \alpha$ (bzw. $\mathbf{P}_\vartheta[t(X) \in C] \leq \alpha$). Da H_0 genau mit Wahrscheinlichkeit $\varphi(X)$ verworfen wird (verworfen wird, falls $t(X) \in C$), wissen wir also, dass die Nullhypothese im Mittel höchstens mit Wahrscheinlichkeit α abgelehnt wird, wenn sie zutrifft. Damit hat man also die Wahrscheinlichkeit, die Nullhypothese abzulehnen, falls sie zutrifft, durch α beschränkt. Für $\vartheta \in \Theta_0$ und $X = x$ ist also $\varphi(x)$ die Wahrscheinlichkeit für einen *Fehler erster Art*. (Für $\vartheta \in \Theta_0$ und $t(X) \in C$, die Nullyhpothese also irrtümlicherweise verworfen wird, sprechen wir von einem *Fehler erster Art*).

Geht man davon aus, dass $\vartheta \in \Theta_A$, liegt eine Fehlentscheidung mit Wahrscheinlichkeit $1 - \varphi(X)$ vor (dann vor, wenn $Y \notin C$), die Nullhypothese also nicht abgelehnt wird. In diesem Fall sprechen wir von einem *Fehler zweiter Art*. Das Niveau des Tests liefert keinen Anhaltspunkt dafür, mit welcher Wahrscheinlichkeit ein solcher Fehler auftritt.

3. Auf den ersten Blick besteht eine scheinbare Symmetrie zwischen H_0 und H_A. Schließlich lehnen wir mit Wahrscheinlichkeit $\varphi(X)$ (falls $Y \in C$) H_0 genau dann ab (und nehmen H_A an), und mit Wahrscheinlichkeit $1 - \varphi(X)$ (falls $Y \notin C$) lehnen wir H_0 nicht ab (und lehnen damit H_A ab). Allerdings wird diese Symmetrie durch das Niveau des Tests gebrochen. Weiß man, dass φ ((Y, C)) ein Test zum Niveau α ist, bedeutet das, dass die Wahrscheinlichkeit, die Nullhypothese H_0 abzulehnen, obwohl sie wahr ist, im Mittel höchstens α ist. Mit anderen Worten ist die Wahrscheinlichkeit für einen Fehler erster Art höchstens α. Allerdings hat man keine Kontrolle über den Fehler zweiter Art.

   Wegen dieser Asymmetrie ist in der Praxis die Nullhypothese genau so zu wählen, dass eine Ablehnung der Nullhypothese möglichst sicher auf die Richtigkeit der Alternativhypothese zurückzuführen ist. Wir betrachten das Beispiel 4.5 des Binomialtests, bei dem wir bei $X = 23$ Treffern in $n = 53$ Versuchen eines Münzwurfes testen wollen, ob der Wurf fair gewesen sein kann, d.h. auf die Erfolgswahrscheinlichkeit $p = 1/2$. Zunächst sind wir skeptisch, dass $p = 1/2$ richtig sein kann, da eigentlich zu wenige Erfolge zu verzeichnen waren. Wir legen das Signifikanzniveau α fest (was in der Praxis oft $\alpha = 5\,\%$ ist). Um unsere Vorstellung über die Erfolgswahrscheinlichkeit zu überprüfen, testen wir

$$H_0 : p = 1/2 \qquad \text{gegen} \qquad H_A : p \neq 1/2.$$

   Kommt es nämlich jetzt zu einer Ablehnung von H_0, so wissen wir, dass dies mit Wahrscheinlichkeit höchstens α dann passiert, wenn H_0 wahr ist, die Münze also fair war. Damit können wir uns relativ sicher sein, dass die Ablehnung der Nullhypothese darauf zurückzuführen ist, dass H_A zutrifft. Damit ist unsere Vorstellung, dass die Münze bei einer Ablehnung der Nullhypothese unfair war, höchstwahrscheinlich bestätigt.

4. Ein möglichst großer Ablehnungsbereich bei möglichst kleinem Niveau α ist für jeden Test wünschenswert. Schließlich soll die Nullhypothese in möglichst vielen Fällen abgelehnt werden, ohne dass die Wahrscheinlichkeit, sie irrtümlicherweise abzulehnen, größer als α wird.

5. Die Forderung von unverfälschten Tests ist klar zu verstehen: Da wir H_0 mit Wahrscheinlichkeit $\varphi(X)$ ablehnen, soll zumindest die Wahrscheinlichkeit, dass H_0 abgelehnt wird, unter $\mathbf{P}_{\vartheta_A}$, $\vartheta_A \in \Theta_A$ größer sein als für $\mathbf{P}_{\vartheta_0}$, $\vartheta_0 \in \Theta_0$.

Bemerkung 4.3 (***p*-Werte und alternative Definition eines Tests).** Betrachten wir einen nicht-randomisierten Test (Y, C). Sei $Y = y$, d.h. dass die Teststatistik Y, angewendet auf die echten Daten, ergibt y. Dann heißt der Wert

$$p_y := \sup_{\vartheta \in \Theta_0} \mathbf{P}_\vartheta(Y \text{ extremer als } y)$$

p-Wert des Tests für $Y = y$. Dabei hängt die Bedeutung davon, was 'extremer' heißt davon ab, was genau die Alternative ist. (Dies ist oftmals in konkreten Beispielen einfach zu verstehen, siehe etwa den Binomialtest und den Gauss-Test.) Immer gilt jedoch $p_y \leq p_{y'}$, falls y extremer

als y' ist. Es ist wichtig zu beachten, dass es dadurch einen engen Zusammenhang zwischen dem Niveau α des Tests und dem p-Wert gibt. Ist nämlich (Y, C) ein Test zum Niveau α und

$$C = \{y : y \text{ extremer als } y_0\}$$

für ein y_0, so wird H_0 genau dann abgelehnt, wenn

$$\alpha \geq \sup_{\vartheta \in \Theta_0} \mathbf{P}_\vartheta(Y \text{ extremer als } y_0) = p_{y_0}.$$

Ist $Y = y$ und gilt $p_y \leq p_{y_0}$, so wird H_0 also abgelehnt. Es genügt also, für einen Test zum Niveau α und $Y = y$ den Wert p_y zu bestimmen. Ist $p_y \leq \alpha$, so wird H_0 abgelehnt. Dieses Vorgehen wird bei vielen Statistik-Programmen angewendet, bei denen ausschließlich p-Werte ausgegeben werden. Dabei muss man meist angeben, was genau die Alternative ist (einseitig oder zweiseitig), damit das Programm weiß, in welche Richtungen Abweichungen als extrem zu betrachten sind.

Wir kommen nun zunächst zu zwei konkreten Beispielen für Tests. Den ersten haben wir auch schon in unserem Eingangsbeispiel in Abschnitt 1.1 kennen gelernt.

Proposition 4.4 (Nicht-randomisierter Binomialtest)**.** *Sei $\alpha \in [0,1], n \in \mathbb{N}$ und $(X, (\mathbf{P}_p)_{p\in[0,1]})$ ein statistisches Modell, so dass X unter $\mathbf{P}_p$ nach $B(n,p)$ verteilt ist.*

(a) Ist $\Theta_0 = p^*$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(X, \{0,\dots,k\} \cup \{l,\dots,n\})$ ein unverfälschter Test zum Niveau α, falls

$$\mathbf{P}_{p^*}(X \leq k) \leq \alpha/2, \qquad \mathbf{P}_{p^*}(X \geq l) \leq \alpha/2.$$

(b) Ist $\Theta_0 = [0, p^*]$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(X, \{k,\dots,n\})$ ein unverfälschter Test zum Niveau α, falls

$$\mathbf{P}_{p^*}(X \geq k) \leq \alpha.$$

(c) Ist $\Theta_0 = [p^*, 1]$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(X, \{0,\dots,k\})$ ein unverfälschter Test zum Niveau α, falls

$$\mathbf{P}_{p^*}(X \leq k) \leq \alpha.$$

Beweis. Wir beweisen nur (c), die anderen beiden Aussagen folgen analog. Klar ist, dass der Test unverfälscht ist. Es ist außerdem

$$\sup_{p\in\Theta_0} \mathbf{P}_p(X \in \{0,\dots,k\}) = \mathbf{P}_{p^*}(X \leq k) \leq \alpha$$

nach Voraussetzung. Also folgt bereits die Aussage. □

Beispiel 4.5 (Binomialtest)**.** Sei $\alpha = 5\,\%$, $n = 53$ und $(X, (\mathbf{P}_p)_{p\in[0,1]})$ wie in der Proposition. Wir wollen nun

$$H_0 : p = 1/2 \text{ gegen } H_A : p \neq 1/2$$

testen, wenn wir in 53 Versuchen 23 Erfolge erzielt haben. Nach Proposition 4.4 ist der kritische Bereich von der Form $\{0,\dots,k\} \cup \{l,\dots,53\}$. Es ist

$$\mathbf{P}_{p=1/2}(X \leq 18) + \mathbf{P}_{p=1/2}(X \geq 35) \approx 2.70\,\%.$$

Da $18 < 23 < 35$, liegt 23 nicht im Ablehnungsbereich von H_0. Damit kann die Nullhypothese aufgrund der Daten ($X = 23$) nicht abgelehnt werden. Auf dasselbe Ergebnis kommt man mit Hilfe des p-Wertes. Es ist

$$\mathbf{P}_{p=1/2}(X \text{ extremer als } 23) = \mathbf{P}_{p=1/2}(X \leq 23) + \mathbf{P}_{p=1/2}(X \geq 30) \approx 41.01\,\%.$$

Da dieser Wert größer als $\alpha = 5\,\%$ ist, kann man die Nullhypothese nicht ablehnen.

Beispiel 4.6 (Randomisierter Binomialtest)**.** Betrachten wir den Binomialtest von $H_0 : p \in [0, p^*]$ gegen $H_A : p \in (p^*, 1]$. Im nicht-randomisierten Binomial-Test ist zwar das Signifikanz-Niveau α, jedoch nimmt $\mathbf{P}_{p^*}(X \geq l)$ für $l = 0, ..., n$ nur endlich viele Werte an. Mit anderen Worten wird in vielen Fällen das Signifikanzniveau nicht voll ausgeschöpft, und der Test ist konservativ. Dies wird im *randomisierten Binomial-Test* verbessert:
Für gegebenes $\alpha \in (0, 1)$ sei l minimal mit $\mathbf{P}_{p^*}(X \geq l) \leq \alpha$. Dann definieren wir den randomisierten Test

$$\varphi(x) = \begin{cases} 1, & x = l, ..., n, \\ \frac{\alpha - \mathbf{P}_{p^*}(X \geq l)}{\mathbf{P}_{p^*}(X = l-1)}, & x = l-1, \\ 0, & x = 0, ..., l-2. \end{cases}$$

Dann gilt

$$\mathbf{E}_{p^*}[\varphi(X)] = \mathbf{P}_{p^*}[X \geq l] + \frac{\alpha - \mathbf{P}_{p^*}(X \geq l)}{\mathbf{P}_{p^*}(X = l-1)} \mathbf{P}_{p^*}(X = l-1) = \alpha.$$

Insbesondere ist φ ein (nicht-konservativer) Test zum Niveau α mit einem größeren (wenn auch randomisierten) Ablehnungsbereich.

Proposition 4.7 (Gauss-Test)**.** *Sei $\alpha \in [0, 1], \sigma^2 \in \mathbb{R}_+, \mu^* \in \mathbb{R}$ und $(X = (X_1, \ldots, X_n), (\mathbf{P}_\mu)_{\mu \in \mathbb{R}})$ ein statistisches Modell, so dass $X_1, \ldots, X_n$ unter $\mathbf{P}_\mu$ unabhängig und nach $N(\mu, \sigma^2)$ verteilt sind. Weiter sei*

$$Z := \frac{\overline{X} - \mu^*}{\sqrt{\sigma^2/n}}.$$

und z_p für $p \in [0, 1]$ das p-Quantil von $N(0, 1)$.

(a) Ist $\Theta_0 = \{\mu^*\}$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(Z, (-\infty, z_{\alpha/2}) \cup (z_{1-\alpha/2}, \infty))$ ein unverfälschter Test zum Niveau α.

(b) Ist $\Theta_0 = (-\infty, \mu^*]$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(Z, [z_{1-\alpha}, \infty))$ ein unverfälschter Test zum Niveau α.

(c) Ist $\Theta_0 = [\mu^*, \infty)$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(Z, (-\infty, z_\alpha])$ ein unverfälschter Test zum Niveau α.

Beweis. Wieder beweisen wir nur (c). Es ist klar, dass der Test unverfälscht ist. Wir wissen, dass unter $\mathbf{P}_{\mu^*}$ die Zufallsvariable Z nach $N(0, 1)$ verteilt ist. Damit gilt

$$\sup_{\mu \geq \mu^*} \mathbf{P}_\mu(Z \leq z_\alpha) = \mathbf{P}_{\mu^*}(Z \leq z_\alpha) = \alpha,$$

woraus die Behauptung sofort folgt. □

4.2 Intervallschätzer und Tests

Wir kommen nochmal zurück zu Schätzproblemen. Bisher hatten wir nur *Punktschätzer* für (Funktionen des) Parameter(s) betrachtet, uns jedoch – bis auf Abschätzungen über die Varianz des Schätzers – weniger über die mögliche Streuung Gedanken gemacht. Intervallschätzer unterscheiden sich von Punktschätzern vor allem dadurch, dass das Ergebnis nicht ein einziger Wert ist, sondern ein Intervall, in dem der wahre Wert des Parameters (der Funktion des Parameters) mit einer vorgegebenen Wahrscheinlichkeit liegt.

Definition 4.8. *Sei $\alpha \in [0,1]$ und $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell sowie $m : \Theta \to \Theta' \subseteq \mathbb{R}$. Jedes (von X abhängige) Intervall $(\underline{t}(X), \overline{t}(X))$ mit*

$$\mathbf{P}_\vartheta(m(\vartheta) \in (\underline{t}(X), \overline{t}(X))) \geq 1 - \alpha$$

heißt Konfidenzintervall *für $m(\vartheta)$ zum Konfidenzniveau $1-\alpha$.*

Beispiel 4.9 (Konfidenzintervall im Normalverteilungsmodell)**.** Im Normalverteilungsmodell $(X = (X_1, ..., X_n), (\mathbf{P}_\mu = N(\mu, \sigma^2))_{\mu\in\mathbb{R}})$ bei gegebener Varianz suchen wir ein Konfidenzintervall für μ zum Signifikanzniveau $1-\alpha$. Dies ist gegeben als

$$(\overline{X} - z_{1-\alpha/2}\sqrt{\sigma^2/n}, \overline{X} + z_{1-\alpha/2}\sqrt{\sigma^2/n})$$

(wobei z_x das x-Quantil von $N(0,1)$ ist), denn (mit $Z \sim N(0,1)$)

$$\begin{aligned}
&\mathbf{P}_\mu\Big(\mu \in (\overline{X} - z_{1-\alpha/2}\sqrt{\sigma^2/n}, \overline{X} + z_{1-\alpha/2}\sqrt{\sigma^2/n})\Big) \\
&\quad = \mathbf{P}_\mu\Big(\frac{|\overline{X} - \mu|}{\sqrt{\sigma^2/n}} \leq z_{1-\alpha/2}\Big) = \mathbf{P}(|Z| \leq z_{1-\alpha/2}) = \mathbf{P}(Z < z_{1-\alpha/2}) - \mathbf{P}(Z < -z_{1-\alpha/2}) \\
&\quad = 1 - \alpha/2 - \alpha/2 = 1 - \alpha.
\end{aligned}$$

Es gibt eine Dualität zwischen Konfidenzintervallen und (nicht-randomisierten) Tests. Das bedeutet, dass man oftmals aus Konfidenzintervallen zum Niveau $1-\alpha$ einen Test zum Niveau α herstellen kann und umgekehrt.

Proposition 4.10 (Konfidenzintervalle und Tests)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $m : \Theta \to \Theta' \subseteq \mathbb{R}$ und $I(X) := (\underline{t}(X), \overline{t}(X))$ ein (zufälliges) Intervall. Dann sind äquivalent:*

1. *Das Intervall $I(X)$ ist ein Konfidenzintervall für $m(\vartheta)$ zum Niveau $1-\alpha$.*
2. *Für jedes $\vartheta^* \in \Theta$ ist*
$$\varphi_{\vartheta^*}(X) := 1_{m(\vartheta^*)\notin I(X)}$$
ein (nicht-randomisierter) Test für $H_0 : \vartheta \in m^{-1}(m(\vartheta^*))$ gegen $H_A : \vartheta \notin m^{-1}(m(\vartheta^*))$ zum Niveau α.

Beweis. 1.⇒2.: Gilt H_0, so ist $\vartheta \in m^{-1}(m(\vartheta^*))$ (also $m(\vartheta) = m(\vartheta^*)$). Daraus berechnen wir

$$\mathbf{E}_\vartheta[\varphi_{\vartheta^*}(X)] = \mathbf{P}_\vartheta(m(\vartheta^*) \notin (\underline{t}(X), \overline{t}(X)) = \mathbf{P}_\vartheta(m(\vartheta) \notin (\underline{t}(X), \overline{t}(X)) \leq \alpha.$$

2.⇒1.: Hier ist für jedes $\vartheta^* \in \Theta$

$$\mathbf{P}_{\vartheta^*}(m(\vartheta^*) \in (\underline{t}(X), \overline{t}(X)) = 1 - \mathbf{E}_{\vartheta^*}[\varphi_\vartheta(X)] \geq 1 - \alpha.$$

□

Beispiel 4.11 (Konfidenzintervall und Test im Normalverteilungsmodell)**.** Für das Normalverteilungsmodell bei bekannter Varianz aus Beispiel 4.9 hatten wir das Konfidenzintervall

$$I(X) = \left(\overline{X} - z_{1-\alpha/2}\sqrt{\sigma^2/n}, \overline{X} + z_{1-\alpha/2}\sqrt{\sigma^2/n}\right)$$

bestimmt. Wir bemerken, dass

$$\mu \in I(X) \iff Z := \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \in (z_{\alpha/2}, z_{1-\alpha/2}).$$

In der Tat ist nach Proposition 4.7(a) $(Z, (-\infty, z_{\alpha/2}) \cup (z_{1-\alpha/2}, \infty))$ ein (unverfälschter) Test von $H_0 : \Theta_0 = \{\mu\}$ gegen $H_A : \Theta_A = \mathbb{R} \setminus \{\mu\}$ zum Niveau α.

4.3 Optimale Tests

Wie bei Schätzern will man auch bei Tests Optimalität erreichen. Zunächst muss also wieder geklärt werden, was man darunter versteht.

Definition 4.12 (Gütefunktion, Macht eines Tests, Optimaler Test)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und φ ein statistischer Test von $H_0 : \vartheta \in \Theta_0$ gegen $H_A : \vartheta \in \Theta_A$ zum Niveau $\alpha \in [0, 1]$.*

1. *Die Funktion*
$$g_\varphi : \begin{cases} \Theta & \to [0,1] \\ \vartheta & \mapsto \mathbf{E}_\vartheta[\varphi(X)] \end{cases}$$
heißt Gütefunktion *von φ. Für $\vartheta \in \Theta_A$ heißt $g_\varphi(\vartheta)$ auch die* Macht *des Tests φ in ϑ.*

2. *Ist φ' ein weiterer Test von H_0 gegen H_A zum Niveau α, so heißt φ* gleichmäßig besser *als φ', falls*
$$g_\varphi(\vartheta) \geq g_{\varphi'}(\vartheta)$$
für alle $\vartheta \in \Theta_A$ gilt.

3. *Ein Test φ zum Signifikanzniveau α heißt UMP (uniformly most powerful) für $H_0 : \vartheta \in \Theta_0$ gegen $H_A : \vartheta \in \Theta_A$, falls φ gleichmäßig besser ist als jeder Test φ' von H_0 gegen H_A zum Niveau α, d.h.*
$$g_\varphi(\vartheta) \geq g_{\varphi'}(\vartheta), \qquad \vartheta \in \Theta_A.$$

Bemerkung 4.13 (Idealer Test, Fehler zweiter Art)**.** 1. Ein idealer Test (Y, T) (den es in realen Situationen nie gibt) hätte die Eigenschaft, dass $Y \in C$ nur für $\vartheta \in \Theta_A$ möglich ist und weiter, dass $\mathbf{P}_\vartheta(Y \in C) = 1$ für $\vartheta \in \Theta_A$. Das bedeutet, dass $g_Y(\vartheta) = 1_{\vartheta\in\Theta_A}$ die Gütefunktion eines idealen Tests ist. Für einen solchen Test wäre die Macht für alle $\vartheta \in \Theta_A$ gleich 1.

2. Für $\vartheta \in \Theta_A$ ist $1 - g_\varphi(\vartheta)$ (also 1-*Macht des Tests bei* ϑ) die Wahrscheinlichkeit für einen Fehler zweiter Art.

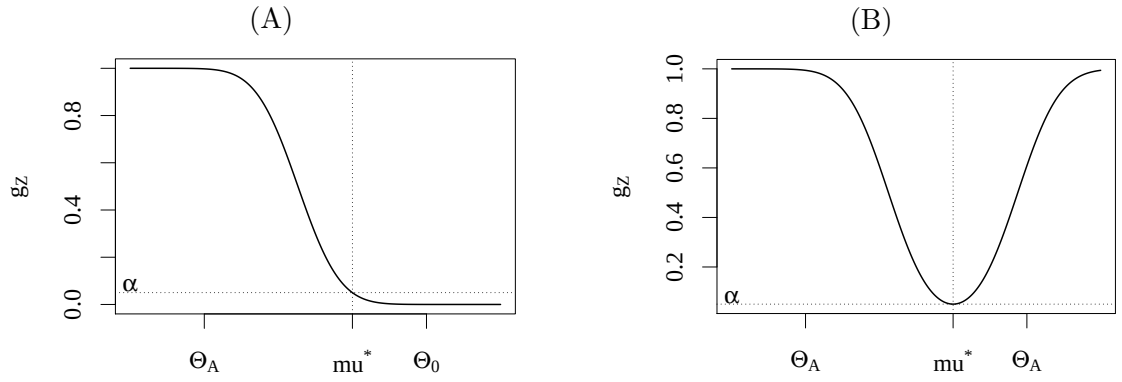

Abbildung 4.1: Die Gütefunktionen für den einseitigen (A) und den zweiseitigen (B) Gauss-Test aus Beispiel 4.14.

Beispiel 4.14 (Gauss-Test)**.** Betrachten wir die Situation des Gauss-Tests φ aus Proposition 4.7. Sei $\widetilde{Z}$ eine (unter allen $\mathbf{P}_\mu$) nach $N(0,1)$ verteilte Zufallsvariable. Hier gilt im Fall des einseitigen Tests (c)

$$\begin{aligned} g_\varphi(\mu) &= \mathbf{P}_\mu(Z \le z_\alpha) = \mathbf{P}_\mu\Big(\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}} + \frac{\mu-\mu^*}{\sqrt{\sigma^2/n}} \le z_\alpha\Big) = \mathbf{P}\Big(\widetilde{Z} \le z_\alpha + \frac{\mu^*-\mu}{\sqrt{\sigma^2/n}}\Big) \\ &= \Phi\Big(z_\alpha + \frac{\mu^*-\mu}{\sqrt{\sigma^2/n}}\Big), \end{aligned}$$

wobei $\Phi(.)$ die Verteilungsfunktion von $N(0,1)$ ist. Analog ist für den zweiseitigen Test (a)

$$\begin{aligned} g_\varphi(\mu) &= \mathbf{P}_\mu(Z \le z_{\alpha/2}) + \mathbf{P}_\mu(Z \ge z_{1-\alpha/2}) \\ &= \mathbf{P}\Big(\widetilde{Z} \le z_{\alpha/2} + \frac{\mu^*-\mu}{\sqrt{\sigma^2/n}}\Big) + \mathbf{P}\Big(Z \ge z_{1-\alpha/2} + \frac{\mu^*-\mu}{\sqrt{\sigma^2/n}}\Big) \\ &= 1 - \Phi\Big(z_{1-\alpha/2} + \frac{\mu^*-\mu}{\sqrt{\sigma^2/n}}\Big) + \Phi\Big(z_{\alpha/2} + \frac{\mu^*-\mu}{\sqrt{\sigma^2/n}}\Big). \end{aligned}$$

Diese zwei Gütefunktionen sind in Abbildung 4.1 dargestellt.

Die Optimalität eines Tests kann man zunächst am besten zeigen, wenn sowohl H_0 als auch H_A einfache Hypothesen sind. Hier hilft das Neyman-Pearson-Lemma 4.17, wo gezeigt wird, dass *Likelihood-Quotienten-Tests* optimal sind.

Definition 4.15 (Likelihood-Quotienten-Test)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\Theta = \Theta_0 \cup \Theta_A$, $\Theta_0 = \{\vartheta_0\}, \Theta_A = \{\vartheta_A\}$ und $\mathbf{P}_\vartheta(X \in dx) = p_\vartheta(x)dx$. Dann heißt*

$$L(x, \vartheta_0, \vartheta_A) := \frac{p_{\vartheta_A}(x)}{p_{\vartheta_0}(x)}$$

Likelihood-Quotient *und $\varphi : S \to [0,1]$* Likelihood-Quotienten-Test *von H_0 gegen H_A, falls für ein $k \in [0,\infty]$ und $\gamma : S \to [0,1]$*

$$\varphi(x) = \begin{cases} 1, & L(x,\vartheta_0,\vartheta_A) > k, \\ \gamma(x), & L(x,\vartheta_0,\vartheta_A) = k, \\ 0, & L(x,\vartheta_0,\vartheta_A) < k. \end{cases}$$

Beispiel 4.16 (Likelihood-Quotienten-Test im Binomialmodell)**.** Wir wollen im Binomialmodell $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ (d.h. $X_*\mathbf{P}_\vartheta = B(n,\vartheta)$) mit $\Theta = \{p, q\}$ einen Likelihood-Quotiententest für $H_0 : \vartheta = p$ gegen $H_A : \vartheta = q$ (mit $q > p$) aufstellen. Wir berechnen

$$L(x, p, q) = \frac{q^x(1-q)^{n-x}}{p^x(1-p)^{n-x}}.$$

Da $q/p > 1$, sind die Abbildungen $x \mapsto (q/p)^x$ und $x \mapsto \frac{(1-q)^{n-x}}{(1-p)^{n-x}}$ und damit $x \mapsto L(x, p, q)$ monoton wachsend. Damit ist jeder Test φ mit

$$\varphi(x) = \begin{cases} 1, & x = \ell+1, ..., n, \\ \gamma, & x = \ell, \\ 0, & x = 0, ..., \ell-1 \end{cases}$$

ein Likelihood-Quotiententest für $k = L(\ell, p, q)$ zum Niveau

$$\mathbf{E}_p[\varphi(X)] = \gamma\mathbf{P}_p[X = \ell] + \mathbf{P}(X \in \{\ell+1, ..., n\}].$$

Insbesondere ist der randomisierte Test aus Beispiel 4.6 ein Likelihood-Quotienten-Test.

Theorem 4.17 (Neyman-Pearson-Lemma)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\Theta = \Theta_0 \cup \Theta_A$, $\Theta_0 = \{\vartheta_0\}, \Theta_A = \{\vartheta_A\}$ und $\mathbf{P}_\vartheta(X \in dx) = p_\vartheta(x)dx$. Sei φ ein Likelihood-Quotienten-Test (mit k und γ) und $\varphi' : S \to [0,1]$ ein weiterer Test mit $g_{\varphi'}(\vartheta_0) \le g_\varphi(\vartheta_0)$ (d.h. φ' ist Test zum selben Niveau wie φ). Dann gilt*

$$g_\varphi(\vartheta_A) \ge g_{\varphi'}(\vartheta_A),$$

d.h. φ ist UMP-Test zum Niveau $g_\varphi(\vartheta_0)$.

Beweis. Sei zunächst $k \in [0, \infty)$. Es gilt für alle $x \in S$

$$\varphi'(x)(p_{\vartheta_A}(x) - kp_{\vartheta_0}(x)) \le \varphi(x)(p_{\vartheta_A}(x) - kp_{\vartheta_0}(x)).$$

In der Tat: Gilt $p_{\vartheta_A}(x) - kp_{\vartheta_0}(x) > 0$, so ist $\varphi(x) = 1$ und die Behauptung folgt aus $\varphi'(x) \le 1$. Gilt $p_{\vartheta_A}(x) - kp_{\vartheta_0}(x) = 0$, so ist die Aussage trivial, und ist $p_{\vartheta_A}(x) - kp_{\vartheta_0}(x) < 0$, so ist $\varphi(x) = 0$ und die Aussage folgt aus $\varphi'(x) \ge 0$. Nun folgt

$$\begin{aligned}\mathbf{E}_{\vartheta_A}[\varphi'(X)] - k\mathbf{E}_{\vartheta_0}[\varphi'(X)] &= \int \varphi'(x)(p_{\vartheta_A}(x) - kp_{\vartheta_0}(x))dx \\ &\le \int \varphi(x)(p_{\vartheta_A}(x) - kp_{\vartheta_0}(x))dx = \mathbf{E}_{\vartheta_A}[\varphi(X)] - k\mathbf{E}_{\vartheta_0}[\varphi(X)],\end{aligned}$$

also

$$g_{\varphi'}(\vartheta_A) - g_\varphi(\vartheta_A) \le k(g_{\varphi'}(\vartheta_0) - g_\varphi(\vartheta_0)) \le 0.$$

Daraus folgt die Behauptung. Der Fall $k = \infty$ wird in einer Übungsaufgabe behandelt. □

Beispiel 4.18 (Likelihood-Quotienten-Test im Normalverteilungsmodell)**.** Sei $\sigma^2 > 0$ bekannt. Wir betrachten das Normalverteilungsmodell $(X = (X_1, ..., X_n), (\mathbf{P}_\mu)_{\mu\in\mathbb{R}})$. Zunächst schränkten wir unseren Parameterbereich ein und setzen $\Theta_0 = \{\nu_0\}$ und $\Theta_A = \{\nu_A\}$ mit $\nu_A > \nu_0$. Es ist

$$2\sigma^2 \log L(x, \nu_0, \nu_A) = \sum_{i=1}^{n}(x_i - \nu_0)^2 - \sum_{i=1}^{n}(x_i - \nu_A)^2 = n\nu_0^2 - n\nu_A^2 + 2(\nu_A - \nu_0)\sum_{i=1}^{n} x_i.$$

Da monotone wachsende Funktionen von L ebenfalls zu optimalen Tests führen, ist

$$Y = \frac{\sum_{i=1}^{n} X_i - \nu_0}{\sqrt{\sigma^2 n}}$$

mit $Y_*\mathbf{P}_{\nu_0} = N(0, 1)$ eine Teststatistik und für $\alpha \in (0, 1)$ ist mit dem Ablehnungsbereich $C = [z_{1-\alpha}, \infty)$ der Test (Y, C) UMP zum Niveau α für die Hypothese $H_0 : \mu = \nu_0$ gegen $\mu = \nu_A$.

Bislang können wir optimale Tests nur für einfache Null- und Alternativhypothesen mit Hilfe des Neyman-Pearson-Lemmas angeben. Die Übertragung auf zusammengesetzte Hypothesen gelingt uns zumindest im Falle monotoner Dichtequotienten.

Definition 4.19 (Monotoner Dichtequotient)**.** *Ein statistisches Modell* $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ *mit* $\Theta \subseteq \mathbb{R}$ *hat einen* monotonen Dichtequotienten *bezüglich* $t : S \to \mathbb{R}$*, falls für alle* $\vartheta < \vartheta'$ *der Likelihood-Quotient*

$$x \mapsto \frac{p_{\vartheta'}(x)}{p_\vartheta(x)}$$

eine streng monotone wachsende Funktion in $t(x)$ *ist.*

Beispiel 4.20 (Ein-parametrige Exponentialfamilie)**.** Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ eine ein-parametrige Exponentialfamilie mit

$$\mathbf{P}_\vartheta(X \in dx) = p_\vartheta(x)dx = h(x)\exp\Big(c(\vartheta)t(x) + d(\vartheta)\Big)dx,$$

so hat diese genau einen monotonen Dichtequotienten bezüglich t, falls c streng monoton wachsend ist.
Denn: Für den Likelihood-Quotient ist

$$\log\frac{p_{\vartheta'}(x)}{p_\vartheta(x)} = (c(\vartheta') - c(\vartheta))t(x) + d(\vartheta') - d(\vartheta).$$

Offenbar ist dies für $\vartheta < \vartheta'$ genau dann eine streng monoton wachsende Funktion in $t(x)$, wenn c streng monoton wächst.

Theorem 4.21 (Neyman-Pearson-Lemma unter monotonen Dichtequotienten)**.** *Sei* $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ *ein statistisches Modell mit* $\Theta \subseteq \mathbb{R}$ *und*

$$\Theta_0 = \Theta \cap (-\infty, \theta_0], \qquad \Theta_A = \Theta \cap (\theta_0, \infty),$$

$\mathbf{P}_\vartheta(X \in dx) = p_\vartheta(x)dx$ *und einem monotonen Dichtequotienten bezüglich* $t : S \to \mathbb{R}$*. Gilt für ein* k *und ein* γ

$$\varphi(x) = \begin{cases} 1, & t(x) > k, \\ \gamma, & t(x) = k, \\ 0, & t(x) < k, \end{cases}$$

so ist φ ein Likelihood-Quotienten-Test (und damit ein UMP-Test) für jedes Paar $H_0 : \Theta_0 = \{\vartheta\}$ mit $\vartheta \in \Theta_0$ gegen $H_A : \Theta_A = \{\vartheta'\}$ mit $\vartheta' \in \Theta_A$. Weiter ist φ ein UMP-Test für $H_0 : \vartheta \in \Theta_0$ gegen $H_A : \vartheta \in \Theta_A$ zum Niveau $\mathbf{P}_{\vartheta_0}(t(X) > k) + \gamma \mathbf{P}_{\vartheta_0}(t(X) = k)$.

Beweis. Zunächst ist zu zeigen, dass $t(x) > k$ genau dann gilt, wenn $L(x, \vartheta, \vartheta') > \widetilde{k}$ für ein passendes $\widetilde{k}$. Da $(\mathbf{P}_\vartheta)_{\vartheta\in\Theta}$ einen monotonen Dichtequotienten bezüglich t besitzt, gibt es ein streng monotones $f_{\vartheta,\vartheta'}$ mit $\frac{p_{\vartheta'}(x)}{p_\vartheta(x)} = f_{\vartheta,\vartheta'}(t(x))$. Insbesondere ist $t(x) > k$ genau dann, wenn $\widetilde{k} := f_{\vartheta,\vartheta'}(k) < \frac{p_{\vartheta'}(x)}{p_\vartheta(x)} = L(x, \vartheta, \vartheta')$.

Weiter betrachten wir nun die Gütefunktion

$$\vartheta \mapsto g_\varphi(\vartheta) = \mathbf{P}_\vartheta(t(X) > k) + \gamma \mathbf{P}_\vartheta(t(X) = k).$$

Nun ist nach Definition φ ein UMP-Test für $H_0 : \vartheta \in \Theta_0$ gegen $H_A : \vartheta \in \{\vartheta'\}$ für jedes $\vartheta' \in \Theta_A$ mit Signifikanzniveau

$$\sup_{\vartheta\in\Theta_0} \mathbf{E}_\vartheta[\varphi(X)] = \sup_{\vartheta\in\Theta_0} g_\varphi(\vartheta).$$

Daraus folgt dann aber auch $g_\varphi(\vartheta) \geq g_{\varphi'}(\vartheta)$ für alle $\vartheta' \in \Theta_A$. Mit anderen Worten ist φ ein UMP-Test für $H_0 : \vartheta \in \Theta_0$ gegen $H_A : \vartheta \in \Theta_A$. □

Definition 4.22 (Verallgemeinerte Likelihood-Quotienten-Tests)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\mathbf{P}_\vartheta(X \in dx) = p_\vartheta(x)dx$ und $\Theta = \Theta_0 \biguplus \Theta_A$. Der* verallgemeinerte Likelihood-Quotient *ist gegeben durch*

$$L(x, \Theta_A, \Theta_0) := \frac{\sup_{\theta\in\Theta_A} p_\vartheta(x)}{\sup_{\theta\in\Theta_0} p_\vartheta(x)}.$$

Gilt für ein k und ein $\gamma : S \to [0, 1]$ für einen Test φ, dass

$$\varphi(x) = \begin{cases} 1, & L(x, \Theta_A, \Theta_0) > k, \\ \gamma(x), & L(x, \Theta_A, \Theta_0) = k, \\ 0, & L(x, \Theta_A, \Theta_0) < k, \end{cases}$$

so heißt φ verallgemeinerter Likelihood-Quotienten-Test (mit k und γ).

Bemerkung 4.23. In der Praxis berechnet man nicht L, sondern

$$\lambda(x) = \max\{L(x, \Theta_0, \Theta_A), 1\} = \frac{\sup_{\theta\in\Theta} p_\vartheta(x)}{\sup_{\theta\in\Theta_0} p_\vartheta(x)},$$

also eine monotone Funktion von L. Die Suprema erhält man aus den Maximum-Likelihood-Schätzern in Θ_0 und in $\Theta \supseteq \Theta_0$.

5 Einige statistische Tests

Es gibt sehr viele statistische Tests. Wir geben hier nun eine Auswahl wichtiger Tests, inklusive der dazugehörigen Teststatistiken und deren Verteilungen an. Wir konzentrieren uns dabei zunächst auf Tests in Normalverteilungsmodellen (etwa auf Gleichheit von Erwartungswerten oder Varianzen). Anschließend gehen wir auf verteilungsfreie Verfahren ein, d.h. dass das zugrunde liegende statistische Modell nicht-parametrisch (also unendlich-dimensional) ist. Schließlich beschäftigen wir uns mit linearen Modellen.

5.1 Aus der Normalverteilung abgeleitete Verteilungen

Definition 5.1 (Die $\boldsymbol{\chi^2}$- $\boldsymbol{t}$- und $\boldsymbol{F}$-Verteilung)**.** *Seien $X, X_1, X_2, ...$ unabhängige, nach $N(0,1)$ verteilte Zufallsvariablen.*

1. *Die Verteilung der Zufallsvariable*

$$X_1^2 + \cdots + X_n^2$$

heißt (zentrierte) χ^2-Verteilung mit n Freiheitsgraden *und wird mit $\chi^2(n)$ bezeichnet.*

2. *Die Verteilung von*

$$\frac{X}{\sqrt{(X_1^2 + \cdots + X_n^2)/n}}$$

heißt t-Verteilung mit n Freiheitsgraden *und wird mit $t(n)$ bezeichnet.*

3. *Sei $Y \sim \chi^2(m)$ und $Z \sim \chi^2(n)$. Dann heißt die Verteilung von*

$$\frac{Y/m}{Z/n}$$

F-Verteilung mit m und n Freiheitsgraden *und wird mit $F(m,n)$ bezeichnet.*

Bemerkung 5.2. Von χ^2-, t- und F-Verteilung können jeweils Dichten angegeben werden. Dies ist für uns im Folgenden allerdings nicht interessant, so dass wir nur die Ergebnisse angeben:
Die $\chi^2(n)$-Verteilung hat die Dichte[1]

$$f_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} 1_{x>0}.$$

[1]Wir erinnern an die Definition der Γ-Funktion

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

Sie hat Erwartung n und Varianz $2n$. Die $t(n)$-Verteilung hat die Dichte

$$g_n(x)\frac{1}{\sqrt{n\pi}} \cdot \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \cdot \frac{1}{(1+x^2/n)^{(n+1)/2}}.$$

Der Erwartungswert existiert für $n \geq 2$ und ist dann 0. Die Varianz existiert für $n \geq 3$ und ist dann $n/(n-2)$. Die $F(m,n)$-Verteilung hat die Dichte

$$h_{m,n}(x) = m^{m/2}n^{n/2}\frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)}\frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}}1_{x>0}.$$

Der Erwartungswert existiert für $n > 2$ und ist dann $n/(n-2)$, die Varianz existiert für $n > 4$ und ist dann $2n^2(m+n-2)/(m(n-2)^2(n-4))$.

Wir kommen nun zu einem Satz, den wir später noch oft brauchen werden. Wir erinnern an die Begriffe des Mittelwertes und der empirischen Varianz aus Proposition 3.6.

Theorem 5.3 (Satz von Fisher)**.** *Sei $n > 1$, $X = (X_1, .., X_n)$ unabhängig nach $N(\mu, \sigma^2)$ verteilt, $\overline{X}$ der Mittelwert und $s^2(X)$ die empirische Varianz. Dann ist*

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

und

$$(n-1)s^2(X)/\sigma^2 \sim \chi^2(n-1).$$

Außerdem sind $\overline{X}$ und $s^2(X)$ stochastisch unabhängig und

$$T := \frac{\overline{X}-\mu}{\sqrt{s^2(X)/n}} \sim t(n-1). \tag{5.1}$$

Bemerkung 5.4 (Interpretation)**.** Stellen wir uns vor, es liegt uns eine Stichprobe $X = (X_1, \ldots, X_n)$ vor, wobei wir davon ausgehen dürfen, dass $X_1, \ldots, X_n$ unabhängig sind und X_i normalverteilt sind. Wollen wir etwa die Verteilung des Mittelwertes berechnen, wissen wir, dass

$$\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$$

nach $N(0,1)$ verteilt ist. Eine gängige Situation ist nun die, dass wir eine Vermutung haben, wie groß μ ist, jedoch nicht wissen, wie groß σ^2 ist. Es liegt nun nahe, σ^2 in der letzten Formel durch $s^2(X)$ zu ersetzen, wie es in T aus (5.1) geschehen ist. Durch das Ersetzen der festen Größe σ^2 durch die Zufallsvariable $s^2(X)$ verändert sich natürlich die Verteilung. Der Satz von Fisher besagt nun, dass die Verteilung der Standardisierung von $\overline{X}$, wenn man σ^2 durch $s^2(X)$ ersetzt, nach $t(n-1)$ verteilt ist. Bemerkenswert ist dabei, dass die Verteilung von T nicht mehr von σ^2 abhängt.

Beweis von Theorem 5.3. Wir setzen $Z = (Z_1, \ldots, Z_n)$ mit $Z_i := (X_i - \mu)/\sigma$. Wir wissen bereits, dass die Z_i unabhängig und nach $N(0,1)$ verteilt sind. Weiter ist

$$\overline{Z} = \tfrac{1}{n}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma} = (\overline{X}-\mu)/\sigma$$

und

$$(n-1)s^2(Z) = \sum_{i=1}^{n} \Big(\frac{X_i - \mu - \overline{X} + \mu}{\sigma} \Big)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 = (n-1)s^2(X)/\sigma^2.$$

Damit ist

$$T = \frac{\overline{X} - \mu}{\sqrt{s^2(X)/n}} = \frac{\sigma \cdot \overline{Z}}{\sqrt{\sigma^2 s^2(Z)/n}} = \frac{\overline{Z}}{\sqrt{s^2(Z)/n}}$$

und damit genügt es, die Behauptungen für den Vektor Z zu zeigen.

Wir wählen nun eine orthogonale Matrix $O = (o_{ij})_{1 \le i,j \le n}$ mit

$$a_{11} = \cdots = a_{1n} = \frac{1}{\sqrt{n}}$$

und setzen $W = OZ$. Wir verwenden im Beweis die aus Korollar 2.4 bekannte Tatsache, dass W ein Vektor unabhängiger, nach $N(0,1)$ verteilter Zufallsvariablen ist. Weiter ist

$$\begin{aligned} W_1 &= a_{11}Z_1 + \cdots + a_{1n}Z_n = \sqrt{n} \cdot \overline{Z}, \\ (n-1)s^2(Z) &= \sum_{i=1}^{n} (Z_i - \overline{Z})^2 = \sum_{i=1}^{n} Z_i^2 - n\overline{Z}^2 \\ &= \sum_{i=1}^{n} Z_i^2 - W_1^2 = \sum_{i=2}^{n} W_i^2 \end{aligned}$$

da O eine orthogonale Matrix ist und damit

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} (OZ)_i^2 = \sum_{i=1}^{n} W_i^2.$$

Insbesondere haben wir eben gezeigt, dass $(n-1)s^2(X)$ die Summe von $n-1$ Quadraten unabhängiger, nach $N(0,1)$ verteilten Zufallsvariablen ist und damit $\chi^2(n-1)$ verteilt. Da weiter W_1 von $(W_2, \ldots, W_n)$ unabhängig ist, folgt auch, dass $\overline{X}$ von $s^2(X)$ unabhängig ist.

Es bleibt nun zu zeigen, dass T nach $t(n-1)$ verteilt ist. Wir schreiben

$$T = \frac{\overline{Z}}{\sqrt{s^2(Z)/n}} = \frac{\sqrt{n}\overline{Z}}{\sqrt{s^2(Z)}} = \frac{W_1}{\sqrt{(W_2^2 + \cdots + W_n^2)/(n-1)}}.$$

Da $W_1, \ldots, W_n$ unabhängig und nach $N(0,1)$ verteilt sind, folgt, dass T nach $t(n-1)$ verteilt ist. □

5.2 Parametertests bei normalverteilten Daten

Der in Proposition 4.7 besprochene Gauss-Test fällt bereits in die Klasse der Parametertests. Unter diesem Stichwort versteht man statistische Tests, die eine Stichprobe daraufhin testen, ob Parameter der zugrunde liegenden Verteilung gewisse Werte annehmen. Wir werden in diesem Kapitel solche Parametertests für normalverteilten Stichproben kennenlernen, nämlich: den einfachen t-Test, der testet, ob der Erwartungswert einer normalverteilten Stichprobe einen bestimmten Wert annimmt (Proposition 5.5); den doppelten t-Test, der testet, ob die Erwartungswerte von zwei unverbundenen Stichproben identisch sind (Proposition 5.9); den F-Test, der testet, ob die Varianzen zweier normalverteilter Stichproben gleich sind.

Proposition 5.5 (Einfacher t-Test)**.**
Sei $\alpha \in [0,1]$, $\mu^* \in \mathbb{R}$ und $(X = (X_1, \dots, X_n), (\mathbf{P}_{\mu,\sigma^2})_{\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_+})$ ein statistisches Modell, so dass $X_1, \dots, X_n$ unter $\mathbf{P}_{\mu,\sigma^2}$ unabhängig und nach $N(\mu, \sigma^2)$ verteilt sind. Weiter sei

$$T := \frac{\overline{X} - \mu^*}{\sqrt{s^2(X)/n}}.$$

und $t_{n-1,p}$ für $p \in [0,1]$ das p-Quantil von $t(n-1)$.

(a) Ist $\Theta_0 = \{\mu^*\} \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(T, (-\infty, t_{n-1,\alpha/2}) \cup (t_{n-1,1-\alpha/2}, \infty))$ ein unverfälschter Test zum Niveau α.

(b) Ist $\Theta_0 = (-\infty, \mu^*] \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(T, [t_{n-1,1-\alpha}, \infty))$ ein unverfälschter Test zum Niveau α.

(c) Ist $\Theta_0 = [\mu^*, \infty) \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(T, (-\infty, t_{n-1,\alpha}])$ ein unverfälschter Test zum Niveau α.

Beweis. Wir beweisen wir nur (c), da die anderen beiden Aussagen analog folgen. Genau wie im Gauss-Test ist klar, dass der Test unverfälscht ist. Aus dem Satz von Fisher, Theorem 5.3, folgt dass T unter $\mathbf{P}_{\mu^*,\sigma^2}$ nach $t(n-1)$ verteilt ist. Damit gilt

$$\sup_{\mu\geq\mu^*} \mathbf{P}_{\mu,\sigma^2}(T \leq t_{n-1,\alpha}) = \mathbf{P}_{\mu^*,\sigma^2}(T \leq t_{n-1,\alpha}) = \alpha,$$

woraus die Behauptung sofort folgt. □

Bemerkung 5.6 (Vergleich von Gauss-Test und einfachem t-Test)**.** Sowohl der Gauss-Test aus Proposition 4.7, als auch der einfache t-Test basieren auf unabhängigen, normalverteilten Stichproben. Der entscheidende Unterschied der beiden Tests besteht darin, dass beim Gauss-Test die Varianz der zugrunde liegenden Normalverteilung bekannt sein muss, und beim t-Test nicht. Dies sieht man etwa daran, dass beim Gauss-Test die Nullhypothese nur aus einem Bereich für μ, nicht jedoch für σ^2 besteht. Außerdem kann man ja nur bei Kenntnis von σ^2 die Teststatistik Z aus Proposition 4.7 berechnen. Beim t-Test ersetzt man σ^2 durch die empirische Varianz $s^2(X)$, und erhält die Teststatistik T.

Bemerkung 5.7 (Gütefunktion des einfachen t-Tests und Stichprobengröße)**.** Wir betrachten die Situation aus Proposition 5.5(b) mit $\mu^* = 0$ und $\alpha = 5\,\%$. Durch das Signifikanzniveau α wird die Wahrscheinlichkeit für einen Fehler erster Art kontrolliert. Die Wahrscheinlichkeit für einen Fehler zweiter Art wird durch die Gütefunktion angegeben. Ist nämlich $\mu > 0$, so ist

$$\mathbf{P}_{\mu,\sigma^2}(\text{Fehler zweiter Art}) = \mathbf{P}_{\mu,\sigma^2}(T \notin C) = 1 - g_T(\mu).$$

Wir fragen nun, wie gut wir diesen Fehler zweiter Art kontrollieren können, wenn $\mu > 0$ gegeben ist. Die Gütefunktion ist gegeben als

$$\begin{aligned} g_T(\mu) &= \mathbf{P}_{\mu,\sigma^2}(T \in C) = \mathbf{P}_{\mu,\sigma^2}\big(T \geq t_{n-1,1-\alpha}\big) \\ &= \mathbf{P}_{\mu,\sigma^2}\Big(\widetilde{T} + \frac{\mu}{\sqrt{s^2(X)/n}} \geq t_{n-1,1-\alpha}\Big) \end{aligned}$$

für die unter $\mathbf{P}_{\mu,\sigma^2}$ nach $t(n-1)$ verteilte Zufallsgröße

$$\widetilde{T} = \frac{\overline{X} - \mu}{\sqrt{s^2(X)/n}}.$$

Ist n groß, so ist $\widetilde{T}$ etwa nach $N(0,1)$ verteilt und $s^2(X) \approx \sigma^2$. Damit ist für eine nach $N(0,1)$ verteilte Zufallsvariable Z

$$\mathbf{P}_{\mu,\sigma^2}(\text{Fehler zweiter Art}) \approx \mathbf{P}\Big(Z + \frac{\mu}{\sqrt{\sigma^2/n}} \le z_{1-\alpha}\Big).$$

Wollen wir etwa erreichen, dass die Wahrscheinlichkeit für einen Fehler zweiter Art nicht größer als 5 % ist, bedeutet das, dass

$$\frac{\mu}{\sqrt{\sigma^2/n}} \ge 3.29 \tag{5.2}$$

gelten muss, denn

$$\mathbf{P}\Big(Z + 3.29 \le 1.645\Big) = \mathbf{P}\Big(Z \le -1.645\Big) = 5\,\%.$$

Man beachte, dass (5.2) eine Bedingung für die Stichprobengröße liefert. Will man etwa $\mu = 0.1$ (bei $\sigma^2 = 1$) noch mit einem Fehler zweiter Art von 5 % ablehnen können, so muss

$$\sqrt{n} \ge 3.29/0.1 = 32.9, \qquad n \ge 1083$$

gewählt werden.

Korollar 5.8 (Gepaarter ***t***-Test)**.**
Sei $\alpha \in [0,1]$ und $((X,Y) = (X_1,\dots,X_n,Y_1,\dots,Y_n)), (\mathbf{P}_{\mu,\sigma^2})_{\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_+})$ ein statistisches Modell, so dass $Y - X := (Y_1 - X_1,\dots,Y_n - X_n)$ unter $\mathbf{P}_{\mu,\sigma^2}$ unabhängig und nach $N(\mu,\sigma^2)$ verteilt sind. Weiter sei

$$T := \frac{\overline{Y} - \overline{X}}{\sqrt{s^2(Y-X)/n}}.$$

und $t_{n,p}$ für $p \in [0,1]$ das p-Quantil von $t(n)$.

(a) Ist $\Theta_0 = \{0\} \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(T, (-\infty, t_{n-1,\alpha/2}) \cup (t_{n-1,1-\alpha/2}, \infty))$ ein unverfälschter Test zum Niveau α.

(b) Ist $\Theta_0 = (-\infty, 0] \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(T, [t_{n-1,1-\alpha}, \infty))$ ein unverfälschter Test zum Niveau α.

(c) Ist $\Theta_0 = [\mu^*, \infty) \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(T, (-\infty, t_{n-1,\alpha}])$ ein unverfälschter Test zum Niveau α.

Beweis. Man wendet einfach Proposition 5.5 auf den Vektor $Y - X$ an. □

Proposition 5.9 (Doppelter ***t***-Test)**.**
Sei $\alpha \in [0,1]$ und $((X,Y) = (X_1,\dots,X_m,Y_1,\dots,Y_n)), (\mathbf{P}_{\mu_X,\mu_Y,,\sigma^2})_{\mu_X,\mu_Y\in\mathbb{R},\sigma^2\in\mathbb{R}_+})$ ein statistisches Modell, so dass $X_1,\dots,X_m,Y_1,\dots,Y_n$ unter $\mathbf{P}_{\mu_X,\mu_Y,\sigma^2}$ unabhängig sind, sowie $X_1,\dots,X_m$ nach $N(\mu_X,\sigma^2)$ und $Y_1,\dots,Y_n$ nach $N(\mu_Y,\sigma^2)$ verteilt sind. Weiter sei

$$T := \frac{\overline{Y} - \overline{X}}{\sqrt{s^2(X,Y)(m+n)/(mn)}}$$

mit

$$s^2(X,Y) := \frac{1}{m+n-2}\Big(\sum_{i=1}^{m}(X_i-\overline{X})^2+\sum_{i=1}^{n}(Y_i-\overline{Y})^2\Big), \tag{5.3}$$

und $t_{n,p}$ *für* $p\in[0,1]$ *das* p*-Quantil von* $t(n)$*.*

(a) Ist $\Theta_0=\{(\mu_X,\mu_Y):\mu_X=\mu_Y\}\times\mathbb{R}_+$*,* $\Theta_A=\Theta\setminus\Theta_0$*, so ist* $(T,(-\infty,t_{n-1,\alpha/2})\cup(t_{m+n-2,1-\alpha/2},\infty))$ *ein unverfälschter Test zum Niveau* α*.*

(b) Ist $\Theta_0=\{(\mu_X,\mu_Y):\mu_Y\le\mu_X\}\times\mathbb{R}_+$*,* $\Theta_A=\Theta\setminus\Theta_0$*, so ist* $(T,[t_{m+n-2,1-\alpha},\infty))$ *ein unverfälschter Test zum Niveau* α*.*

(c) Ist $\Theta_0=\{(\mu_X,\mu_Y):\mu_Y\ge\mu_X\}\times\mathbb{R}_+$*,* $\Theta_A=\Theta\setminus\Theta_0$*, so ist* $(T,(-\infty,t_{m+n-2,\alpha}])$ *ein unverfälschter Test zum Niveau* α*.*

Beweis. Wir müssen zunächst zeigen, dass T nach $t(m+n-2)$ verteilt ist. Ohne Einschränkung der Allgemeinheit sei $\sigma^2=1$. Da $X_1,\dots,X_m,Y_1,\dots,Y_n$ unabhängig sind, sind $s^2(X)$ und $s^2(Y)$ unabhängige Zufallsvariablen. Aus dem Satz von Fisher folgt, dass $\overline{X},\overline{Y},s^2(X),s^2(Y)$ unabhängig sind, $\overline{Y}-\overline{X}$ unter $\mathbf{P}_{0,\sigma^2}$ nach $N(0,\frac{1}{m}+\frac{1}{n})$ und $(m-1)s^2(X)$ nach $\chi^2(m-1)$ und $(n-1)s^2(Y)$ nach $\chi^2(n-1)$ verteilt ist. Damit ist

$$T=\frac{\frac{1}{\sqrt{1/m+1/n}}(\overline{Y}-\overline{X})}{\sqrt{((m-1)s^2(X)+(n-1)s^2(Y))/(m+n-2)}}$$

nach $t(m+n-2)$ verteilt. Der Rest des Beweises folgt wie beim einfachen t-Test, Proposition 5.5. □

Beispiel 5.10 (Geburtsgewichte)**.** In einer Kölner Klinik wurden im Jahr 1985 $m=269$ Mädchen und $n=288$ Jungen geboren. (Die einzelnen Geburtsgewichte sind also $X_1,\dots,X_{269}$ und $Y_1,\dots,Y_{288}$.) Das Durchschnittsgewicht und die empirische Varianz der Mädchen in Gramm war $\bar{X}=3050$ und $s^2(X)=211600$, das der Jungen $\bar{Y}=3300$ und $s^2(Y)=220900$. Es soll zum Signifikanzniveau $\alpha=0.01$ getestet werden, ob Jungen und Mädchen das gleiche erwartete Geburtsgewicht haben (Fall (a) in Proposition 5.9). Wir berechnen

$$s^2(X,Y)=\frac{1}{269+288-2}(268\cdot s^2(X)+287\cdot s^2(Y))=216409$$

und

$$T=\frac{3300-3050}{\sqrt{216409\cdot(269+288)/(269\cdot288)}}=6.338>2.585=t_{555,0.995}.$$

Also können wir die Nullhypothese $\mu_X=\mu_Y$ auf dem Signifikanzniveau $\alpha=0.01$ ablehnen. Unter der (sehr plausiblen) Annahme der Normalverteilung und der Annahme der Varianzhomogenität kann man zum Niveau $\alpha=0.01$ schließen, dass Jungen im Mittel schwerer sind als Mädchen.

Im doppelten t-Test ist die Annahme der Gleichheit der Varianzen wichtig. Sind die Varianzen nicht gleich, ist nämlich die Teststatistik T unter H_0 nicht $t(m+n-2)$-verteilt. Um die Gleichheit der Varianzen zu überprüfen, gibt es wiederum einen statistischen Test, den F-Test.

Proposition 5.11 (F-Test auf identische Varianzen)**.**
Sei $\alpha \in [0,1]$ und $((X,Y) = (X_1,\dots,X_m,Y_1,\dots,Y_n)), (\mathbf{P}_{\mu_X,\mu_Y,,\sigma_X^2,\sigma_Y^2})_{\mu_X,\mu_Y\in\mathbb{R},\sigma_X^2,\sigma_Y^2\in\mathbb{R}_+})$ ein statistisches Modell, so dass $X_1,\dots,X_m,Y_1,\dots,Y_n$ unter $\mathbf{P}_{\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2}$ unabhängig sind, sowie $X_1,\dots,X_m \sim N(\mu_X,\sigma_X^2)$ und $Y_1,\dots,Y_n \sim N(\mu_Y,\sigma_Y^2)$. Weiter sei

$$F := \frac{s^2(Y)}{s^2(X)},$$

wobei $s^2(X)$ und $s^2(Y)$ die empirischen Varianzen von X und Y sind. Weiter sei $F_{m,n,p}$ für $p \in [0,1]$ das p-Quantil von $F(m,n)$.

(a) Ist $\Theta_0 = \mathbb{R}^2 \times \{(\sigma_X^2,\sigma_Y^2) : \sigma_X^2 = \sigma_Y^2\} \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(F, (-\infty, F_{m-1,n-1,\alpha/2}) \cup (F_{m-1,n-1,1-\alpha/2}, \infty))$ ein unverfälschter Test zum Niveau α.

(b) Ist $\Theta_0 = \mathbb{R}^2 \times \{(\sigma_X^2,\sigma_Y^2) : \sigma_Y^2 \le \sigma_X^2\}$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(F, [F_{m-1,n-1,1-\alpha}, \infty))$ ein unverfälschter Test zum Niveau α.

(c) Ist $\Theta_0 = \mathbb{R}^2 \times \{(\sigma_X^2,\sigma_Y^2) : \sigma_Y^2 \ge \sigma_X^2\}$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(F, [0, F_{m-1,n-1,\alpha})$ ein unverfälschter Test zum Niveau α.

Beweis. Zunächst sind $(m-1)s^2(X)/\sigma_X^2 \sim \chi^2(m-1)$ und $(n-1)s^2(Y)/\sigma_Y^2 \sim \chi^2(n-1)$ unabhängig. für $\sigma_X^2 = \sigma_Y^2$ ist also $F \sim F(m-1,n-1)$. Weiter ist $F_*\mathbf{P}_{\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2}$ stochastisch wachsend in σ_Y^2 und fallend in σ_X^2. Der Rest des Beweises folgt wie beim einfachen t-Test, Proposition 5.5. □

Beispiel 5.12 (Geburtsgewichte)**.** Wir betrachten nochmal das Beispiel 5.10 und testen auf Gleichheit der Varianzen zum Signifikanzniveau 0.01. Mit $m = 269, n = 288$ und $s^2(X) = 211600, s^2(Y) = 22099$ ist $F = 1.044$. Nach der Hypothese $\sigma_X^2 = \sigma_Y^2$ ist diese Teststatistik nach $F(268,287)$ verteilt. Es ist $F_{268,287,0.005} = 0.733$ und $F_{268,287,0.995} = 1.363$. Damit lässt sich die Nullhypothese gleicher Varianzen nicht ablehnen.

5.3 Anpassungstests

Die bisher behandelten Tests basierten alle auf normalverteilten Stichproben. Etwa können wir mit den zuletzt behandelten t-Tests überprüfen, ob der Mittelwert einer Stichprobe vermutlich einen bestimmten Wert annimmt. Auffällig ist, dass diese t-Tests immer nur auf den Erwartungswert der zu Grunde liegenden Verteilung testen. Dies ist bei den χ^2-Tests, die wir in diesem Abschnitt kennen lernen, anders. Beim χ^2-Anpassungstest (Theorem 5.14) wird überprüft, ob die gesamte Verteilung einer Stichprobe (und nicht nur der Erwartungswert) eine bestimmte Form hat.

Beispiel 5.13 (Mendel's Experimente)**.** Der Naturforscher Gregor Mendel kreuzte rot- und weißblühende Erbsenpflanzen. In der ersten Nachkommensgeneration erhielt er ausschließlich rosa-blühende Pflanzen. Kreuzte er diese weiter, erhielt er in der nächsten Generation rot- rosa- und weißblühende Pflanzen. Seine Theorie sagte voraus, dass diese Farben in der zweiten Nachkommensgeneration im Verhältnis 1:2:1 auftreten sollten.

Ziel des χ^2-Anpassungstests in diesem Beispiel ist es, zu überprüfen, ob das Farbverhältnis von 1:2:1 von einer Stichprobe vermutlich eingehalten wird. Betrachten wir dazu $n = 400$ Pflanzen der zweiten Generation, von denen wir annehmen, dass deren Farbe Realisierung

von unabhängigen Experimenten ist. Sei $\mathcal{I} = \{\text{rot, rosa, weiß}\}$ und $X = (X_1, \dots, X_n)$ ein Vektor von unabhängigen, $\mathcal{I}$-wertigen Zufallsvariablen. Falls die Theorie von Mendel stimmt, gilt für alle i

$$\mathbf{P}(X_i = \text{rot}) = \tfrac{1}{4}, \qquad \mathbf{P}(X_i = \text{rosa}) = \tfrac{1}{2}, \qquad \mathbf{P}(X_i = \text{weiß}) = \tfrac{1}{4}.$$

Gleichbedeutend ist es, dass die Verteilungsgewichte der Verteilung von X_i durch den Vektor $\underline{\pi} = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$ gegeben sind. Die Stichprobe liefert $S_{\text{weiß}} = 115$ weiße Blüten, $S_{\text{rosa}} = 171$ rosa Blüten und $S_{\text{rot}} = 114$ rote Blüten. Kann aufgrund dieser Beobachtung die Nullhypothese über das 1:2:1-Farbverhältnis abgelehnt werden?

Theorem 5.14 ($\boldsymbol{\chi^2}$-Anpassungstest)**.** *Sei $\alpha \in [0,1]$, $\mathcal{I}$ eine endliche Menge mit $|\mathcal{I}| = r$ und*

$$\Theta = \{\underline{p} = (p_i)_{i\in\mathcal{I}} \text{ Verteilungsgewichte einer Verteilung auf } \mathcal{I}\}.$$

Weiter sei $(X = (X_1, \dots, X_n), (\mathbf{P}_{\underline{p}})_{\underline{p}\in\Theta})$ ein statistisches Modell, so dass $X_1, \dots, X_n$ unter $\mathbf{P}_{\underline{p}}$ unabhängig und nach $\underline{p}$ verteilt sind (d.h. $\mathbf{P}(X_k = i) = p_i$). Setze für $i \in \mathcal{I}$

$$S_i := |\{k : X_k = i\}|.$$

Weiter sei $\chi^2_{m,p}$ für $p \in [0,1]$ das p-Quantil von $\chi^2(m)$ und $\Theta_0 = \{\underline{\pi}\}$, $\Theta_A = \Theta \setminus \Theta_0$ für ein $\underline{\pi} \in \Theta$. Für

$$\chi^2_n := \sum_{i\in\mathcal{I}} \frac{(S_i - n\pi_i)^2}{n\pi_i}.$$

ist $(\chi^2_n, (\chi^2_{r-1,1-\alpha}, \infty))$ im Grenzwert $n \to \infty$ ein Test zum Niveau α.

Bemerkung 5.15 (Approximativer Test)**.** Der χ^2-Anpassungstest ist ein approximativer Test für große Stichproben. Unter H_0 gilt nämlich für jedes x (also insbesondere für $x = \chi^2_{r-1,\alpha}$)

$$\lim_{n\to\infty} \mathbf{P}_{\underline{\pi}}(\chi^2_n > x) = \mathbf{P}(X > x)$$

für eine nach $\chi^2(r-1)$ verteilte Zufallsgröße X. In der Praxis ist die Stichprobengröße natürlich nie beliebig groß. Deswegen hat man Faustregeln für die Anwendbarkeit des χ^2-Anpassungstests aufgestellt. Man verwendet den χ^2-Test dann, wenn entweder

$$\begin{aligned} &n\pi_i \geq 3 \text{ für alle } i \in \mathcal{I} \qquad \text{oder} \\ &r \geq 10 \quad \text{und} \quad n\pi_i \geq 1 \text{ für alle } i \in \mathcal{I} \end{aligned} \tag{5.4}$$

gilt.

Beweisskizze von Theorem 5.14. Wir werden nur einige heuristische Bemerkungen anstelle eines kompletten Beweises des Theorems machen. Zunächst bemerken wir, dass die Teststatistik genau dann klein ist, wenn alle S_i nahe an $n\pi_i$ sind. Unter H_0 ist schließlich auch $\mathbf{E}[S_i] = n\pi_i$. Also misst die Teststatistik quadratische Abweichungen von $(S_i)_{i\in\mathcal{I}}$ von den erwarteten Anzahlen $(n\pi_i)_{i\in\mathcal{I}}$. Sind diese quadratischen Abweichungen zu groß, wird H_0 abgelehnt, da ja $(\chi^2_{r-1,1-\alpha}, \infty)$ der Ablehnungsbereich ist.

Wir zeigen nun noch, dass im speziellen Fall $r = 2$ die Teststatistik χ^2_n approximativ $\chi^2(1)$ verteilt ist. Sei also $\mathcal{I} = \{1, 2\}$. Dann ist S_1 nach $B(n, \pi_1)$ verteilt, $S_2 - n\pi_2 = (n - S_1 - n(1 -$

$\pi_1) = -(S_1 - n\pi_1)$, $\frac{1}{\pi_1} + \frac{1}{\pi_2} = \frac{1}{\pi_1\pi_2} = \frac{1}{\pi_1(1-\pi_1)}$ und, wegen dem zentralen Grenzwertsatz, für jedes $x > 0$, für eine nach $N(0,1)$ verteilte Zufallsvariable Z

$$\begin{aligned}\lim_{n\to\infty} \mathbf{P}_{\underline{\pi}}\left(\frac{(S_1 - n\pi_1)^2}{n\pi_1} + \frac{(S_2 - n\pi_2)^2}{n\pi_2} \le x\right) &= \lim_{n\to\infty} \mathbf{P}_{\underline{\pi}}\left(\frac{(S_1 - n\pi_1)^2}{n\pi_1(1-\pi_1)} \le x\right) \\ &= \lim_{n\to\infty} \mathbf{P}_{\underline{\pi}}\left(-\sqrt{x} \le \frac{S_1 - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}} \le \sqrt{x}\right) \\ &= \mathbf{P}(-\sqrt{x} \le Z \le \sqrt{x}) = \mathbf{P}(Z^2 \le x).\end{aligned}$$

Da Z^2 nach $\chi^2(1)$ verteilt ist, haben wir gezeigt, dass approximativ χ^2_n für große n nach $\chi^2(1)$ verteilt ist. □

Beispiel 5.16 (Mendel's Experiment)**.** Wir führen nun den χ^2-Anpassungstest für das Mendel'sche Experiment aus Beispiel 5.13 mit $\alpha = 0.05$ durch. Hier ist $n = 400$,

$$H_0 : \underline{p} = \underline{\pi} := \left(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}\right).$$

Außerdem liefert die Stichprobe $S_{\text{weiß}} = 115$, $S_{\text{rosa}} = 171$ und $S_{\text{rot}} = 114$. Die erste Bedingung aus (5.4) ist sicher erfüllt, so dass wir den χ^2-Test anwenden können. Wir berechnen

$$\chi^2_{400} = \frac{(115-100)^2}{100} + \frac{(171-200)^2}{200} + \frac{(114-100)^2}{100} \approx 8.46 > 5.99 = \chi^2_{2,0.95},$$

H_0 wird also abgelehnt. (Nun darf man nach den biologischen Ursachen für die Abweichung von den Mendel'schen Verhältnissen forschen: z.B. könnte Selektion gegen Heterozygote im Spiel sein, oder Wechselwirkung des Gens für die Blütenfarbe mit anderen Teilen des Genoms.)

Bemerkung 5.17 (Erweiterung des $\boldsymbol{\chi^2}$-Anpassungstests für unbekannte Parameter)**.** Nicht immer ist die Verteilung, nach der die Daten verteilt sein sollen, so klar wie im Beispiel 5.13. Oftmals will man wissen, ob die Daten einer Verteilungsklasse (etwa die der Poisson-Verteilungen mit Parameter $\lambda > 0$) angehören, die noch einen oder mehr Parameter besitzt. In diesem Fall muss man zunächst die Parameter aus den Daten schätzen und kann erst anschließend den χ^2-Anpassungstest durchführen. In einer solchen Situation ist klar, dass schon durch das Schätzen der Parameter eine Anpassung der Verteilung an die Daten vollzogen wird. Den χ^2-Anpassungstest kann man jedoch nachwievor durchführen, indem man die Anzahl der Freiheitsgrade der χ^2-Verteilung reduziert, falls die verwendeten Schätzer Maximum-Likelihood-Schätzer sind. Man geht also folgendermaßen vor:

1. Schätzung der l fehlenden Parameter der Verteilung mittels Maximum-Likelihood, basierend auf $X_1, \ldots, X_n$.
2. Die Teststatistik χ^2 aus Theorem 5.14 ist dann approximativ (unter den Annahmen (5.4)) nach
   $\chi^2(r - l - 1)$ verteilt.

Also ist dann $(\chi^2_n, (\chi^2_{r-l-1,1-\alpha}, \infty))$ für große n ein Test zum Niveau α.

Beispiel 5.18 (Hufschlagtote)**.** Wir betrachten das klassische Beispiel von Hufschlagtoten der preussischen Armee. Hierbei wurden 14 Regimenter über 20 Jahre beobachtet, und für jedes Regiment und jedes Jahr die Zahl der Hufschlagtoten aufgezeichnet. Folgende Häufigkeiten wurden dabei beobachtet:

Anzahl der Todesfälle	0	1	2	3	4
Häufigkeit	144	91	32	11	2

Es liegt nahe zu vermuten, dass die Anzahl der Hufschlagtoten in jedem Jahr eine Poisson-verteilte Zufallsvariable ist. Um dies zu überprüfen, testen wir mit $\alpha = 0.01$ mittels eines χ^2-Anpassungstests.

Zunächst berechnen wir den Maximum-Likelihood-Schätzer für λ. Dieser ist nach Beispiel 3.13 gegeben durch

$$\widehat{\lambda} = \frac{1}{280}\big(0 \cdot 144 + 1 \cdot 91 + 2 \cdot 32 + 3 \cdot 11 + 4 \cdot 2\big) = 0.7.$$

Damit ergeben sich folgende erwartete Größen:

Anzahl der Todesfälle	0	1	2	3	4
Häufigkeit	139.04	97.33	34.07	7.95	1.39

Da wir fordern, dass alle erwarteten Anzahlen mindestens 5 sind, müssen wir die Fälle mit vielen Hufschlagtoten gruppieren:

Anzahl der Todesfälle	0	1	2	3 oder mehr
Häufigkeit	139.04	97.33	34.07	9.56

Wir setzen also $\mathcal{I} = \{0, 1, 2, 3 \text{ oder mehr}\}$,

$$H_0 : \underline{p} \in \Big\{ e^{-\lambda}\Big(\frac{\lambda^0}{0!}, \frac{\lambda^1}{1!}, \frac{\lambda^2}{2!}, \dots\Big) \text{ für ein } \lambda > 0\Big\}.$$

Damit ergibt sich die Teststatistik

$$\begin{aligned}\chi^2_{280} &= \frac{(144 - 139.04)^2}{139.04} + \frac{(91 - 97.33)^2}{97.33} + \frac{(32 - 34.07)^2}{34.07} + \frac{(11 - 7.95)^2}{7.95} + \frac{(2 - 1.61)^2}{1.61} \\ &= 1.95 \leq 9.21 = \chi^2_{2,0.99}.\end{aligned}$$

Damit können wir die Nullhypothese Poisson-verteilter Daten auf dem Signifikanzniveau $\alpha = 0.01$ nicht ablehnen.