Mathematische Statistik I

Inhaltsverzeichnis

1 Einleitung

Statistik ist die Wissenschaft, die sich mit der Analyse quantitativer Merkmale von großen Grundgesamtheiten befasst. In Lexika findet man, dass der Begriff "Statistik" aus dem Lateinischen und Italienischen stammt und dass dort "statisticum" "den Staat betreffend" bedeutet und dass ein "Statistica" ein Politiker oder ein Staatsmann ist.

Die quantitativen Merkmale der Grundgesamtheit oder Population nennt man Daten. Es gibt verschiedene Aspekte der Datenanalyse, beispielsweise:

1. Die optimale Präsentation der Daten; hierbei muss ein Mittelweg zwischen den Extremen der vollständigen Erhaltung der Information (wie sie etwa die Urliste bietet) und zu großer Vereinfachung, im Extremfall der Zusammenfassung aller Daten in eine Gruppe, gefunden werden. Dies ist die beschreibende oder deskriptive Statistik.

2. Die Untersuchung der Datenqualität. Dieses Gebiet überlappt mit der deskriptiven Statistik und auch mit nicht-mathematischen Disziplinen. Schlechte Daten können beispielsweise durch Messfehler, Schreibfehler, Übertragungsfehler, aber auch durch eine fehlerhafte Versuchskonzeption entstehen. Bekannt sind beispielsweise Untersuchungen über das Sexualverhalten (siehe z. B. `http://www.durex.com/de/gss2005result.pdf`). Die Frage nach der Anzahl verschiedener Sexualpartner im Leben ergab, dass Frauen durchschnittlich sieben verschiedene Sexualpartner in ihrem Leben haben, während es bei Männern zehn sind. Geht man von ungefähr 50 % Männern und 50 % Frauen in einer Population aus, so fragt man sich, wie dies möglich ist.

3. Die explorative Datenanalyse ist ebenfalls mit der deskriptiven Statistik verwandt. Der Anspruch ist hier, mithilfe verschiedener, auch computergestützter Verfahren aus einem vorhandenen Datensatz Hypothesen über diese Daten bzw. das dahinterstehende Modell zu entwickeln.

4. Die schließende oder induktive Statistik geht von einem wahrscheinlichkeitstheoretischen Modell aus, von dem die Daten stammen, das jedoch nicht vollständig bekannt ist. Die induktive Statistik versucht, mithilfe von Schätz- und Testverfahren Aussagen über das Modell zu treffen. Die induktive Statistik ist auch als die mathematische Statistik bekannt.

Die mathematische Statistik ist gewissermaßen invers zur Wahrscheinlichkeitstheorie. Während wir in letzterer ein Modell gegeben haben und Vorhersagen über das Verhalten einer Stichprobe, d. h. einer Familie $X_1, \dots, X_n$ von i.i.d. Zufallsvariablen, die gemäß dieses Modells gezogen werden, treffen wollen, ist die Situation in der Statistik gerade umgekehrt: Hier ist eine Stichprobe, d. h. in der Regel eine Realisierung von i.i.d. Zufallsvariablen, gegeben, und wir wollen auf das zugrunde liegende Modell schließen.

Inhaltlich zerfällt die mathematische Statistik in zwei Gebiete, die parametrische Statistik und die nicht-parametrische Statistik. In der parametrischen Statistik lässt sich das Modell mithilfe eines endlich-dimensionalen Parameters beschreiben, man denke beispielsweise

daran, dass die Daten von einer Poisson-Verteilung zum Parameter $\lambda > 0$ stammen (dies ist dann der Modellparameter) oder aber an eine Stichprobe, die aus einer $\mathcal{N}(\mu, \sigma^2)$-Verteilung gezogen wird, wobei μ und σ^2 unbekannt sind.

In der nicht-parametrischen Statistik lässt sich die Datenquelle nicht durch einen endlichdimensionalen Parameter beschreiben. Man denke zum Beispiel an Situationen, in denen man keine Annahmen über die der Stichprobe zugrunde liegende Verteilung machen kann. Darüber hinaus gibt es noch die sogenannte "semiparametrische" Statistik, auf die hier aber nicht näher eingegangen werden soll.

Offensichtlich ist die nicht-parametrische Statistik weitaus komplexer als die parametrische, daher beginnen wir mit der letzteren. Zunächst aber wollen wir einige Beispiele kennen lernen, die uns davon überzeugen sollen, dass es sich bei statistischen Fragestellungen um Alltags-relevante Probleme handelt.

Beispiel 1.1 *Zur Behandlung einer Krankheit wird eine neue Therapie, sagen wir $T1$, entwickelt. Bei einer Behandlung von 20 Patienten mit $T1$ zeigen 17 einen Erfolg, 3 einen Misserfolg. Die klassische Therapie, $T2$, hat etwa 70 % Heilungschancen.*
Frage: Ist $T1$ besser als $T2$?
Diese Frage lässt sich zunächst so modellieren: Für die $n = 20$ Patienten führen wir Zufallsvariablen $X_1, \ldots, X_n$ ein, die die Werte 1 (für einen Behandlungserfolg) und 0 (für einen Misserfolg) annehmen können. Wir nehmen an, dass die (X_i) i.i.d. sind und

$$\mathbb{P}(X_i = 1) = \vartheta$$

gilt.

Das Schätzproblem besteht nun darin, ϑ aufgrund unserer Beobachtung zu schätzen, d. h. das Testproblem beschäftigt sich mit der Frage, ob wir aufgrund unserer Beobachtung verlässlich sagen können, dass die neue Behandlungsmethode besser ist als die alte, dass also $\vartheta \geq 0.7$ ist. Schließlich gibt es noch eine dritte Fragestellung, die sogenannte Bereichsschätzung von ϑ. Sie besteht darin, bei bekannter Stichprobe x einen möglichst kleinen Bereich $C(x) \subseteq [0, 1]$ anzugeben, in dem sich ϑ mit großer Wahrscheinlichkeit befindet. Hier ist zu betonen, dass die Wahrscheinlichkeit von der zufälligen Beobachtung herrührt. Für jedes feste Intervall C ist natürlich entweder $\vartheta \in C$ oder $\vartheta \notin C$.

Beispiel 1.2 *Bei der Positionsbestimmung per GPS wird die Position im Raum durch Entfernungsbestimmung zu drei Punkten im Raum berechnet. Das konkrete Vorgehen sieht dabei so aus, dass man als diese drei Raumpunkte Satelliten verwendet. Ungefähr 30 Satelliten umkreisen dabei die Erde in ca. 20.000 km Höhe und senden sekündlich Signale zur Erde, die die Zeit des gesendeten Signals und die Position des Satelliten beinhalten. Hierbei kann es durch verschiedene Umstände zu Messfehlern kommen, etwa durch Veränderungen in der Ionosphäre oder durch Uhrenfehler beim Empfänger. Man versucht, diese Fehler auszugleichen, indem man die Signale von mehr als drei Empfängern verwertet und dann mithilfe statistischer Methoden die Position des Empfängers schätzt.*

Beispiel 1.3 *Der Zellstoffwechsel wird durch Proteine gesteuert. Bei DNA-Microarrays wird statt der Proteinaktivität, die schwer zu messen ist, die Aktivität von Genen simultan für 3.000 – 20.000 Gene gemessen. Eine Messung liefert daher einen Datenvektor von der Länge 3.000 – 20.000. Ausgehend von solchen Messungen sollen dann z. B. bei Tumorzellen Vorhersagen gemacht werden bzgl.*

- *Anspruch auf Therapien*
- *Überlebenswahrscheinlichkeit eines Patienten*

etc. Dabei kennt man das Verhalten erkrankter Zellen von anderen Patienten ebenso wie das Verhalten gesunder Zellen.

2 Schätzmethoden

Wir werden in der Folge immer davon ausgehen, dass wir eine Stichprobe $X_1, \ldots, X_n$ gegeben haben. Diese Stichprobe bestehe aus i.i.d. Zufallsvariablen auf einem Raum $(\mathcal{X}, \mathcal{F})$, die wir uns in diesem Kapitel gemäß einer Verteilung $\mathbb{P}_\vartheta$ realisiert vorstellen. ϑ ist dabei ein Element aus einem $\mathbb{R}^d,\ d \geq 1$. Wir nehmen an, dass die Familie der $(\mathbb{P}_\vartheta)_{\vartheta \in \mathbb{R}^d}$ dominiert wird durch ein Maß ν. Die zugehörigen Dichten bezeichnen wir mit f_ϑ, also

$$\frac{d\mathbb{P}_\vartheta}{d\nu} = f_\vartheta.$$

Beispiel 2.1 *a) Die $X_1, \ldots, X_n$ seien i.i.d. Poisson-verteilt zum Parameter $\vartheta > 0$, also $(\mathbb{P}_\vartheta)_{\vartheta>0} = Poi(\vartheta)_{\vartheta>0}$. Hier ist also ν das Zählmaß auf $\mathbb{N} \cup \{0\}$ und*

$$f_\vartheta(k) = \frac{\vartheta^k}{k!} e^{-\vartheta}, \quad k \in \mathbb{N}_0.$$

b) Die $X_1, \ldots, X_n$ seien i.i.d. $\mathcal{N}(\mu, \sigma^2)$-verteilt mit $\mu \in \mathbb{R}$ und $\sigma^2 > 0$. Also ist hier $\vartheta = (\mu, \sigma^2)$ mit $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ und

$$(\mathbb{P}_\vartheta)_{\vartheta \in \mathbb{R} \times \mathbb{R}^+} = (\mathcal{N}(\mu, \sigma^2))_{\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+}.$$

Das dominierende Maß ν ist in diesem Fall das Lebesguemaß $\lambda\!\!\lambda$ und

$$f_\vartheta(x) = \frac{d\mathbb{P}_\vartheta}{d\lambda\!\!\lambda}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

Wir wollen nun verschiedene Schätzmethoden kennenlernen, also Methoden, um einen Schätzer für das unbekannte $\vartheta \in \mathbb{R}^d$ zu finden. Dabei hilft es sicher zunächst zu wissen, was denn ein Schätzer ist.

Definition 2.2 *Es sei $X_1, \ldots, X_n$ eine Stichprobe, die gemäß einer Verteilung $(\mathbb{P}_\vartheta)_{\vartheta \in \mathbb{R}^d}$ gezogen wird. Ein Schätzer für ϑ ist eine Abbildung*

$$\begin{aligned} T : \mathbb{R}^n &\to \mathbb{R}^d \\ (x_1, \ldots, x_n) &\mapsto T(x_1, \ldots, x_n) \end{aligned}$$

die messbar von $X_1, \ldots, X_n$ abhängt. Analog ist ein Schätzer für eine Funktion

$$\begin{aligned} \gamma : \mathbb{R}^d &\to \mathbb{R}^m \\ \vartheta &\mapsto \gamma(\vartheta) \end{aligned}$$

eine Funktion

$$g : \mathbb{R}^n \to \mathbb{R}^m,$$

die messbar von $X_1, \ldots, X_n$ abhängt.

Die Definition eines Schätzers verlangt zunächst einmal nun vernünftigerweise nur, dass man nicht mehr Informationen verwenden darf als man tatsächlich zur Verfügung hat. Sie sagt aber nicht, wie man an einen guten Schätzer kommt und ob der erhaltene Schätzer in einem noch zu spezifizierenden Sinne bestmöglich ist. Damit wollen wir uns in diesem und den folgenden Kapitel befassen.

Zunächst wollen wir drei verschiedene Verfahren kennen lernen, um überhaupt "vernünftige" Schätzer zu konstruieren.

2.1 Der Maximum-Likelihood-Schätzer

Die Maximum-Likelihood-Methode kennen wir schon aus dem Statistikteil der Stochastikvorlesung. Ihre Idee besteht in der Interpretation der Dichte $f_\vartheta(x)$ einer Beobachtung als Wahrscheinlichkeit. Diese Interpretation stammt aus der Situation, in der ν tatsächlich das Zählmaß ist und $f_\vartheta(x)$ dann zwangsläufig die Wahrscheinlichkeit.

Die Idee der Maximum-Likelihood-Methode ist es, den Parameter ϑ so zu schätzen, dass eine gegebene Beobachtung $X_1 = x_1, \ldots, X_n = x_n$ maximale Wahrscheinlichkeit hat. Dies ist daher plausibel, weil die Interpretation von Wahrscheinlichkeit als relative Häufigkeit ja gerade aussagt, dass wahrscheinliche Ergebnisse häufiger auftreten als unwahrscheinliche.

Definition 2.3 *Seien $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen, die gemäß einer Verteilung $\mathbb{P}_\vartheta$ aus einer Familie von Verteilungen $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta\subseteq\mathbb{R}^d}$ gezogen werden. Die Likelihoodfunktion ist dann*

$$L_x(\vartheta) := L_{(x_1,\ldots,x_n)}(\vartheta) := f_\vartheta(x_1)\cdots f_\vartheta(x_n).$$

Hierbei ist ν ein dominierendes Maß für $(\mathbb{P}_\vartheta)_\vartheta$,

$$f_\vartheta = \frac{d\mathbb{P}_\vartheta}{d\nu}$$

und $x = (x_1, \ldots, x_n)$ eine Realisierung der $X_1, \ldots, X_n$. Die logarithmische Likelihoodfunktion oder log-Likelihoodfunktion ist

$$\mathcal{L}_x(\vartheta) = \log L_x(\vartheta).$$

Definition 2.4 *In der Situation von Definition 2.3 ist der Maximum-Likelihood-Schätzer für ϑ jedes $\hat{\vartheta}$ mit*

$$\hat{\vartheta} = \arg\max_\vartheta L_x(\vartheta).$$

Wegen der Monotonie der Logarithmusfunktion ist dies das gleiche wie

$$\hat{\vartheta} = \arg\max_\vartheta \mathcal{L}_x(\vartheta).$$

Ein Schätzer heißt Maximum-Likelihood-Schätzer für $\gamma(\vartheta)$, falls er $\gamma(\hat{\vartheta})$ ist.

Beispiel 2.5 *a) Es seien die $X_1, \dots, X_n$ i.i.d. $Poi(\lambda)$-verteilt mit $\lambda > 0$. In Beispiel 2.1 haben wir festgestellt, dass in dieser Situation die f_λ gegeben sind durch*

$$f_\lambda(x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

Somit ist

$$L_x(\lambda) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! \dots x_n!} e^{-n\lambda}$$

und somit

$$\mathcal{L}_x(\lambda) = \sum_{i=1}^n x_i \log \lambda - n\lambda - \sum_{i=1}^n \log x_i!$$

Um das Maximum zu bestimmen, leiten wir $\mathcal{L}_x(\lambda)$ ab:

$$\frac{d}{d\lambda} \mathcal{L}_x(\lambda) = \sum_{i=1}^n x_i/\lambda - n.$$

Dies ist gleich 0 genau dann, wenn

$$\lambda = \hat{\lambda} := \frac{\sum_{i=1}^n x_i}{n}.$$

Dies ist – wie man leicht nachrechnet – auch ein Maximum. Dieser Schätzer ist auch vernünftig, wenn man bedenkt, dass λ auch der Erwartungswert der $Poi(\lambda)$-Verteilung ist. Tatsächlich gibt es ein kleines Problem, wenn $\hat{\lambda} = 0$ ist, denn dies ist als Parameter nicht zugelassen. Wir erweitern daher das Modell durch

$$\mathbb{P}_0 = \delta_0.$$

b) Schon in der Stochastik haben wir gesehen, dass für den Fall, dass die $X_1, \dots, X_n$ i.i.d. $Ber(p)$, $p \in (0,1)$ sind, der Maximum-Likelihood-Schätzer durch

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$$

gegeben ist.

c) Nun seien die $X_1, \dots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$-verteilt. Es ist also $\nu = \lambda\!\!\lambda$ und

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

Es ist also

$$L_x(\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^n (\frac{x_i-\mu}{\sigma})^2}$$

und

$$\mathcal{L}_x(\mu, \sigma) = -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2} \sum_{i=1}^n (\frac{x_i - \mu}{\sigma})^2.$$

Wir unterscheiden drei Fälle:

(i) μ *unbekannt,* $\sigma > 0$ *bekannt. Dann ist*

$$\begin{aligned} \frac{d}{d\mu}\mathcal{L}_x(\mu,\sigma) &= \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} = 0 \\ &\Leftrightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i. \end{aligned}$$

(ii) $\sigma > 0$ *unbekannt,* μ *bekannt. Dann ist*

$$\frac{d}{d\sigma}\mathcal{L}_x(\mu,\sigma) = -\frac{n}{2}\frac{1}{2\pi\sigma^2}\cdot 4\pi\sigma + \sum \frac{(x_i-\mu)^2}{\sigma^3}.$$

Dies ist Null, wenn

$$\sigma^2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2.$$

(iii) (μ,σ) *unbekannt. Nun ist* $grad(\mathcal{L}_x(\mu,\sigma))$ *gefragt, dies berechnet sich wie oben als*

$$grad\,\mathcal{L}_x(\mu,\sigma) = \begin{pmatrix} \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} \\ -\frac{n}{\sigma} + \sum \frac{(x_i-\mu)^2}{\sigma^3} \end{pmatrix}.$$

Dies ist gleich Null für

$$\mu = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \textit{und} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2.$$

Natürlich ist selbst bei einem vernünftig klingenden Schätzprinzip die Qualität des Schätzers weitestgehend unklar. Wir wollen die Qualität eines Schätzers zunächst in zwei Kriterien ausdrücken.

Definition 2.6 *Ein Schätzer* g *für* $\gamma(\vartheta)$ *heißt erwartungstreu für* $\gamma(\vartheta)$*, falls für alle* $\vartheta \in \Theta$ *gilt*

$$\mathbb{E}_\vartheta[g(X_1,\ldots,X_n)] = \gamma(\vartheta).$$

Beispiel 2.7 *Die Schätzer* $\hat{\lambda}, \hat{p}$ *und* $\hat{\mu}$ *aus Beispiel 2.5 a) – c) sind erwartungstreu, da sie jeweils der Erwartungswert der Zufallsvariablen sind, aus denen sie gebildet werden. All diese Variablen haben die Struktur*

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

und in Situation a) ist $\mathbb{E}X_i = \lambda$*, in b)* $\mathbb{E}X_i = p$ *und in c)* $\mathbb{E}X_i = \mu$*. Der Schätzer* $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ *in Beispiel 2.5 c)(ii) ist erwartungstreu, denn aus der Stochastik wissen*

wir, dass eine Summe von n Quadraten unabhängiger $\mathcal{N}(0,1)$-verteilter Zufallsgrößen χ^2_n-verteilt ist. Daher ist

$$\begin{aligned}\mathbb{E}\hat{\sigma}^2 &= \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2\\ &= \mathbb{E}\frac{\sigma^2}{n}\sum_{i=1}^{n}\left(\frac{X_i-\hat{\mu}}{\sigma}\right)^2\\ &= \frac{\sigma^2}{n}\cdot n=\sigma^2,\end{aligned}$$

denn der Erwartungswert der χ^2_n-Verteilung ist n. Die Größe

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2$$

aus Beispiel 2.5, c) (iii) ist hingegen nicht erwartungstreu. In der Tat gilt einerseits

$$\begin{aligned}\mathbb{E}(\hat{\sigma}^2+\hat{\mu}^2) &= \mathbb{E}(\frac{1}{n}\sum X_i^2-\hat{\mu}^2+\hat{\mu}^2)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i^2\\ &= \mathbb{V}(X_i^2)+(\mathbb{E}X_i)^2=\sigma^2+\mu^2\end{aligned}$$

und andererseits

$$\begin{aligned}\mathbb{E}(\hat{\sigma}^2+\hat{\mu}^2) &= \mathbb{E}(\hat{\sigma}^2)+\mathbb{E}(\hat{\mu}^2)\\ &= \mathbb{E}(\hat{\sigma}^2)+\mathbb{V}(\hat{\mu})+(\mathbb{E}(\hat{\mu}))^2\\ &= \mathbb{E}(\hat{\sigma}^2)+\frac{\sigma^2}{n}+\mu^2.\end{aligned}$$

Bei der zweiten Rechnung haben wir benutzt, dass $\hat{\mu}$ erwartungstreu für μ ist und dass die Varianz von $\hat{\mu}$ sich als

$$\mathbb{V}(\hat{\mu})=\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\cdot n\mathbb{V}(X_1)=\frac{\sigma^2}{n}$$

berechnen lässt. Somit ist

$$\mathbb{E}(\hat{\sigma}^2)=\frac{\sigma^2(n-1)}{n}.$$

Ein zweites Gütekriterium, das im Laufe dieser Vorlesung eine weniger wichtige Rolle spielen wird, da wir uns kaum mit asymptotischen Fragestellungen befassen werden, richtet sich an eine ganze Schätzerfolge. In der Tat haben wir ja in den Beispielen für jedes n eine (einheitliche) Vorschrift, wie die Schätzer $\hat{\lambda}=\hat{\lambda}_n$, $\hat{p}=\hat{p}_n$, etc. zu konstruieren sind. Konvergieren diese Schätzer nun für $n\to\infty$ gegen ihren Schätzwert, so wollen wir sie konsistent nennen.

Definition 2.8 *Es sei $g_n(X)$ ein Schätzer für $\gamma(\vartheta)$ basierend auf einem Stichprobenumfang n. Gilt*

$$\mathbb{P}_\vartheta(|g_n(X) - \gamma(\vartheta)| > \delta) \to 0$$

für $n \to \infty$ und alle $\delta > 0$, so heißt die Folge $(g_n(x))_{n\in\mathbb{N}}$ konsistent.

Beispiel 2.9 *Alle Schätzer aus Beispiel 2.5 sind konsistent. Für $\hat{\lambda}_n$, $\hat{p}_n$ und $\hat{\mu}_n$ folgt dies unmittelbar aus dem schwachen Gesetz der großen Zahlen. Für $\hat{\sigma}_n$ und $\hat{\sigma}_n^2$ nutzt man aus, dass man ihre Verteilung im wesentlichen kennt. So ist z. B. die Varianz einer χ_n^2-Verteilung $2n$ und daher kann man jedes $\delta > 0$ folgendermaßen abschätzen:*

$$\begin{aligned}
\mathbb{P}(|\hat{\sigma}_n^2 - \sigma^2| > \delta) &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 - \sigma^2\right| > \delta\right) \\
&= \mathbb{P}\left(\left|\frac{\sigma^2}{n}\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 - \sigma^2\right| > \delta\right) \\
&\leq \mathbb{V}\left(\frac{\sigma^2}{n}\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2\right)/\delta^2 \\
&= \frac{\sigma^4}{n^2}\frac{2n}{\delta^2} = \frac{2\sigma^4}{n\delta^2},
\end{aligned}$$

was für $n \to \infty$ gegen 0 geht.

2.2 Der Momentenschätzer

Ein weiteres Konzept, um Schätzer zu konzentrieren, setzt beim Begriff der "Erwartungstreue" an. Die Grundüberlegung hierbei ist die, dass viele Verteilungen schon durch ihre Momente bestimmt sind. Weiß man z. B., dass die Zufallsvariable X die Momente

$$\begin{aligned}
&\mathbb{E}X_i^{2n+1} = 0 \quad \text{für alle} \quad n \\
\text{und} \quad &\mathbb{E}X_i^{2n} = (2n-1)(2n-3)\cdots 1\sigma^{2n} \quad \text{für alle} \quad n
\end{aligned}$$

hat, so ist schon bekannt, dass sie $\mathcal{N}(0, \sigma^2)$-verteilt ist. Eine Möglichkeit, den Zentralen Grenzwertsatz zu beweisen, besteht daher auch darin zu zeigen, dass alle Momente von $\frac{1}{\sqrt{n\mathbb{V}X_1}}\sum_{i=1}^n (X_i - \mathbb{E}X_1)$ gegen die Momente der Standard-Normalverteilung konvergieren. Es liegt also nahe, die Parameter einer Verteilung dadurch zu schätzen, dass man ihre Momente schätzt. Nun ist aber ein erwartungstreuer Schätzer für das k-te Moment $\mathbb{E}X^k$ auf Basis einer i.i.d. Stichprobe $X_1, \ldots, X_n$ (die identisch verteilt sind zu X)

$$\hat{M}_k = \frac{1}{n}\sum_{i=1}^n X_i^k.$$

Setzt man nun für ein Modell mit unbekanntem Parameter $\vartheta \in \mathbb{R}^d$

$$\begin{aligned}
\hat{M}_1 &= \mathbb{E}_\vartheta X_1 \\
&\vdots \\
\hat{M}_d &= \mathbb{E}_\vartheta X_1^d,
\end{aligned} \tag{2.1}$$

so erhält man d Gleichungen in den d unbekannten $\vartheta_1, \ldots, \vartheta_d$ (wobei wir $\vartheta = (\vartheta_1, \ldots, \vartheta_d)$ schreiben). Wenn sich diese Gleichungen lösen lassen, so erhält man einen Schätzer.

Definition 2.10 *Haben die Gleichungen 2.1 eine Lösung in (ϑ), so nennt man die Lösung den* Momentenschäter *für ϑ.*

Bemerkung 2.11 *Ein eindeutiger Nachteil der Methode besteht darin, dass die Gleichungen keine Lösung haben müssen.*

Beispiel 2.12 *a) Seien wieder $X_1, \ldots, X_n$ i.i.d. Poisson-verteilt zum Parameter $\lambda > 0$. Da λ eindimensional ist, genügt es, das erste Moment zu betrachten:*

$$\hat{M}_1 = \mathbb{E}_\lambda X = \lambda,$$

also

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\lambda}.$$

Der Momentenschätzer ist also gleich dem Maximum-Likelihood-Schätzer.

b) Sind die $X_1, \ldots, X_n$ i.i.d. Ber(p)-verteilt, sieht man auf gleiche Weise, dass der Momentenschätzer wieder $\hat{p}$ ist.

c) Sind $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$-verteilt und beide Parameter unbekannt, so ist $\vartheta = (\mu, \sigma^2)$ zwei-dimensional. Wir müssen also die beiden Gleichungen

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu \quad \textit{und} \quad \frac{1}{n}\sum X_i^2 = \mathbb{E}X_1^2$$

lösen. Da

$$\mathbb{E}X_1^2 = \mathbb{V}(X_1) + (\mathbb{E}X_1)^2$$

ist, lässt sich dieses Gleichungssystem lösen:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2. \end{aligned}$$

Somit stimmen auch in diesem Fall der Maximum-Likelihood-Schätzer und der Momenten-Schätzer überein.

2.3 Bayes-Schätzer

In den vorherigen Abschnitten haben wir die Konstruktion von Schätzern in Situationen besprochen, bei denen wir als Statistiker keine Vorahnungen und keine Präferenzen für irgendwelche Werte von ϑ, des wahren Parameters, haben. Dies ist oftmals eine realistische Einschätzung. In anderen Situationen hingegen haben wir z. B. aus vorhergehenden Experimenten oder gesundem Menschenverstand sehr wohl eine Präferenz für gewisse ϑ-Werte. Sollen wir beispielsweise die Höhe des Eiffelturms aus verschiedenen Messungen (beispielsweise des Blickwinkels bei gewissem Abstand) schätzen, so scheint uns, selbst wenn wir die Messungen nicht persönlich durchgeführt haben, das Resultat 1,50 m ebenso unplausibel wie 3 500 m. Diesen Überlegungen trägt der Bayes-Schätzer Rechnung. Hierzu definieren wir zunächst

Definition 2.13 *Die* Verlustfunktion *einer Schätzung T eines Parameters ϑ ist eine Funktion*

$$L : \Theta \times \Theta \to \mathbb{R}_0^+.$$

Hierbei ist $\Theta \subseteq \mathbb{R}^d$ der zugrunde liegende Parameterraum.

Beispiel 2.14 *Eine häufige Verlustfunktion bei eindimensionalen Parametern, d. h. $\Theta \subseteq \mathbb{R}$, ist die quadratische Verlustfunktion, d. h. man wählt*

$$L(\vartheta, T(x)) = |\vartheta - T(x)|^2.$$

Wie im Beispiel 2.14 schon angedeutet, hängt ein Schätzer T vernünftigerweise von einer Beobachtung x ab. Somit bekommen wir für jede Beobachtung $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ einen eigenen Wert der Verlustfunktion. Dieser kann für einige Beobachtungen groß sein, für andere klein. Um den Wert eines Schätzers zu ermitteln, müssen wir den Verlust über alle $x \in \mathcal{X}^n$ mischen. Dabei sollte das gewichtete Maß für eine Beobachtung $x \in \mathcal{X}^n$ gerade die Wahrscheinlichkeit $\mathbb{P}_\vartheta(x)$ sein, mit der sie unter dem wahren ϑ auftritt. Wir definieren daher

Definition 2.15 *Das Risiko $\mathcal{R}(\vartheta, T)$ eines Schätzers T für den Parameter ϑ ist der mittlere Verlust L*

$$\mathcal{R}(\vartheta, T) = \mathbb{E}_\vartheta[L(\vartheta, T(X))] = \int L(\vartheta, T(x)) d\mathbb{P}_\vartheta(x).$$

Analog definiert man den Verlust eines Schätzers $g(x)$ für $\gamma(\vartheta)$.

Ein guter Schätzer bezüglich der Verlustfunktion L wird also ein solcher Schätzer sein, der ein kleines Risiko aufweist. Somit könnten wir einen besten Schätzer sofort ausrechnen, wenn wir nur ϑ kennten (und in diesem Fall wäre auch recht klar, was wir als Schätzer nehmen sollten). Der Bayes-Schätzer hilft uns nun aus diesem Teufelskreis, indem er mittels einer Wahrscheinlichkeitsverteilung Gewichte für verschiedene Werte von ϑ einführt.

Definition 2.16 *Sei α eine Wahrscheinlichkeitsverteilung auf Θ. Dann heißt*

$$r(\alpha, T) = \int_\Theta \mathcal{R}(\vartheta, T)\alpha(d\vartheta)$$

das Bayesrisiko *des Schätzers T bzgl. α und des Risikos $\mathcal{R}$ (bzw. der Verlustfunktion L). T heißt* Bayes-Schätzer *von ϑ bzgl. α (und L), falls für alle Schätzer T' von ϑ gilt*

$$r(\alpha, T) \leq r(\alpha, T').$$

Diese unschuldig aussehende Definition hat aufgrund ihrer Interpretation durch manche Statistiker (die sogenannten Baysianer) für einigen Zündstoff in der Statistik gesorgt. Wir wollen dies kurz vorstellen. Das Bayesrisiko ist ja als Doppelintegral

$$r(\alpha, T) = \int_\Theta \int_\mathcal{X} L(\vartheta, T(x))\mathbb{P}_\vartheta(dx)\alpha(d\vartheta)$$

interpretierbar als Erwartungswert des Verlustes L, wenn sowohl ϑ als auch x zufällig gewählt sind und zwar mit gemeinsamer Verteilung $\mathbb{P}_\vartheta(dx)\alpha(d\vartheta)$. α ist dann also die Randverteilung von ϑ und $\mathbb{P}_\vartheta(dx)$ gewissermaßen die bedingte Verteilung von X gegeben ϑ. Man kann sich das ganze also als ein zweistufiges Experiment vorstellen, bei dem man zuerst ϑ gemäß α "zieht" und dann x gemäß $\mathbb{P}_\vartheta(dx)$. Daher heißt α auch die a priori-Verteilung.

Nun lässt sich die Sache gewissermaßen umkehren: Wenn wir x gezogen haben, so verändert diese Information eventuell unsere Informationen über α. Wenn wir annehmen, dass

$$\mathbb{P}_\vartheta(dx) = f_\vartheta(x)\mu(dx)$$

für eine Dichtefunktion $f_\vartheta(\cdot)$ gilt, so folgt mit dem Satz von Bayes:

$$\alpha(d\vartheta|x) = \frac{f_\vartheta(x)\alpha(d\vartheta)}{\int_\Theta f_{\vartheta'}(x)\alpha(d\vartheta')}. \tag{2.2}$$

$\alpha(\cdot|x)$ heißt auch a-posteriori-Verteilung. Sie ist offenbar proportional zum Produkt aus der a-priori-Verteilung $\alpha(d\vartheta)$ und der Likelihood-Funktion $f_\vartheta(x)$ für ϑ bei Beobachtung x. Nennen wir die Randverteilung von X in diesem zweistufigen Experiment Q, so gilt

$$Q(dx) = \int_\Theta \mathbb{P}_\vartheta(dx)\alpha(d\vartheta),$$

haben die Verteilungen $\mathbb{P}_\vartheta$ Dichten bzgl. eines Maßes μ, so auch Q und der Nenner in 2.2 ist gerade die Dichte von Q.

Das Umstrittene an dieser Interpretation von ϑ als Zufallsvariable ist die theoretische Option, dass diese Zufallsvariable in einer Reihe von Experimenten verschiedene Werte annimmt. Zum einen sind verschiedene Experimente nun prinzipiell nur einmal durchführbar, in anderen Fällen ist nicht vorstellbar, dass ϑ verschiedene Werte annimmt: Misst man die Höhe des Eiffelturms, so ist das unbekannte ϑ eben diese Höhe. Und selbst, wenn wir sie nicht kennen, so ist sie fix und es ist nicht denkbar, dass der Eiffelturm bei verschiedenen Messungen seine Höhe jedesmal neu "auswürfelt".

Die Interpretation der a-posteriori-Verteilung und a-priori-Verteilung wird aber sinnvoll, wenn wir α als ein Maß für unsere subjektive (Un-) Kenntnis auffassen. Wenn bei der Messung der Höhe des Eiffelturms α dem Intervall [280,320] eine Wahrscheinlichkeit von 0,95 zumisst, so bedeutet dies eben, dass wir mit sehr großer Wahrscheinlichkeit annehmen, dass der Eiffelturm zwischen 280 und 320 Metern hoch ist. Die a-posteriori-Verteilung beschreibt dann unser Maß für die Lage von ϑ, nachdem wir eine Beobachtung x gemacht haben.

Außer dieser (mehr philosophischen) Diskussion um die Bedeutung von $\alpha(\cdot)$ und $\alpha(\cdot|x)$ gibt es aber auch eine praktische Anwendung der a-posteriori-Verteilung: Sie erlaubt das Auffinden des Bayes-Schätzers durch punktweises Minimieren.

Um dies mathematisch analysieren zu können, benötigen wir eine Nachhilfestunde in Wahrscheinlichkeitstheorie. In den Vorlesungen darüber haben wir schon die bedingte Erwartung kennen gelernt und ein wenig mit dem Begriff der bedingten Dichte gearbeitet. Wir wollen nun für zwei Zufallsvariable

$$X : \Omega \to \mathbb{R}^n,\ Y : \Omega \to \mathbb{R}^m$$

die bedingte Verteilung von X gegeben $Y = y$ berechnen. Wir nehmen zuerst an, dass (X, Y) eine gemeinsame Dichte $f_{X,Y}$ bzgl. des Lebesguemaßes $\lambda\!\!\lambda^{m+n}$ auf $\mathbb{R}^{m+n}$ hat. Dann hat als Konsequenz aus dem Satz von Fubini auch Y eine Lebesguedichte, nämlich

$$f_Y(y) = \int_{\mathbb{R}^n} f_{X,Y}(x, y) d\lambda\!\!\lambda^n(x).$$

Man definiert die bedingte Dichte von X gegeben $Y = y$ als

$$f_{X|Y=y} = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Der Sinn des ganzen erschließt sich, wenn man die Dichte als stetig annimmt und die bedingte Verteilung von X gegeben $\{|Y - y| \le \delta\}$ berechnet und dann δ gegen 0 streben lässt. In diesem Fall ist dann

$$\mathbb{E}[X|Y = y] = \frac{1}{f_Y(y)} \int x f_{X,Y}(x, y) \lambda\!\!\lambda^n(dx)\ \mathbb{P}^Y\text{-f.s.}$$

Dies ergibt

$$\mathbb{E}[X|Y] = \frac{1}{f_Y(Y)} \int x f_{X,Y}(x, Y) \lambda\!\!\lambda^n(dx)\ \mathbb{P}\text{-f.s.}$$

Diese beiden Formeln kennen wir schon aus der Wahrscheinlichkeitstheorie.

Beispiel 2.17 *(X, Y) besitze eine 2-dimensionale Normalverteilung mit Dichte*

$$f(x, y) = \frac{\sqrt{1 - \rho^2}}{2\pi\sigma^2} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2\sigma^2}\right).$$

Somit ist $\mathbb{E}\binom{X}{Y} = 0$ und die Kovarianzmatrix hat die Gestalt

$$\frac{1}{1 - \rho^2} \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix},\ \rho \in (-1, 1),\ \sigma^2 > 0.$$

Damit erhält man:

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)dx = \frac{\sqrt{1-\rho^2}}{2\pi\sigma^2}\exp\left(\frac{-y^2(1-\rho^2)}{2\sigma^2}\right)\int_{-\infty}^{\infty} e^{-\frac{(x-\rho y)^2}{2\sigma^2}}dx \\ &= \frac{\sqrt{1-\rho^2}}{2\pi\sigma^2}\exp\left(\frac{-y^2(1-\rho^2)}{2\sigma^2}\right)\sqrt{2\pi\sigma^2} \\ &= \sqrt{\frac{1-\rho^2}{2\pi\sigma^2}}e^{-\frac{y^2(1-\rho^2)}{2\sigma^2}}. \end{aligned}$$

Dies ist die Dichte einer $\mathcal{N}(0, \frac{\sigma^2}{1-\rho^2})$-Verteilung. Also berechnet sich $\mathbb{E}[X|Y=y]$ als

$$\mathbb{E}[X|Y=y] = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} xe^{-\frac{1}{2}(\frac{x-\rho y}{\sigma})^2}dx = \rho y.$$

Daher ist auch

$$\mathbb{E}[X|Y] = \rho Y \quad \mathbb{P}\text{-}f.s.$$

Wir nutzen die Gelegenheit, um auch die bedingte Verteilung zu definieren.

Definition 2.18 *Sind $(\Omega, \mathcal{A})$, $(\Omega', \mathcal{A}')$ messbare Räume, so ist eine Funktion*

$$K : \Omega \times \mathcal{A}' \to [0, \infty]$$

ein Kern, falls gilt:

- *$K(\omega, \cdot)$ ist ein Maß auf $(\Omega', \mathcal{A}')$ (für alle $\omega \in \Omega$);*
- *$K(\cdot, A')$ ist $\mathcal{A}$-messbar für alle $A' \in \mathcal{A}'$.*

Gilt $K(\omega, \Omega') = 1$ für alle $\omega \in \Omega$, so heißt K stochastisch *oder* Markovsch.

Definition 2.19 *Sei*

$$X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\Omega', \mathcal{A}')$$

eine Zufallsvariable und $\mathcal{F} \subseteq \mathcal{A}$ eine σ-Algebra. Reguläre bedingte Verteilung von X gegeben $\mathcal{F}$ *heißt dann jeder stochastische Kern*

$$\mathbb{P}^{X|\mathcal{F}} : (\Omega, \mathcal{A}') \to [0, 1]$$

derart, dass

$$\omega \mapsto \mathbb{P}^{X|\mathcal{F}}(\omega, A')$$

für jedes $A' \in \mathcal{A}'$ eine Version von $\mathbb{P}(X \in A'|\mathcal{F})$ ist. d. h. für alle $A' \in \mathcal{A}'$, $C \in \mathcal{F}$ gilt

$$\int_C \mathbb{P}^{X|\mathcal{F}}(\omega, \mathcal{A}') = \mathbb{P}(\{X \in A'\} \cap C).$$

Wird $\mathcal{F}$ von einer Zufallsvariablen Y erzeugt, d. h. gilt $\mathcal{F} = \sigma(Y)$, so schreiben wir auch $\mathbb{P}^{X|Y}$ und nennen den Kern reguläre bedingte Verteilung von X gegeben Y.

Kann man

$$\mathbb{P}^{X|Y}(\omega, \mathcal{A}') = K(\cdot, A') \cdot Y(\omega)$$

für einen stochastischen Kern

$$K : (\Omega'', \mathcal{A}') \to [0, 1]$$

von $(\Omega'', \mathcal{A}'')$ nach $(\Omega', \mathcal{A}')$ schreiben, so definieren wir

$$\mathbb{P}^{X|Y}(\omega, A') = K(Y(\omega), A')$$

für alle $\omega \in \Omega$ und $A' \in \mathcal{A}'$. Wir setzen dann

$$\mathbb{P}^{X|Y=y} := K(y, \cdot)$$

und nennen dies die reguläre bedingte Verteilung von X gegeben $Y = y$.

Fakt 2.20 *(nicht-trivial)*
Die regulären bedingten Verteilungen von X gegeben Y und X gegeben $Y = y$ existieren in vielen Fällen, insbesondere in allen, in denen wir sie benutzen werden. Für Details kann man fast alle Bücher über Wahrscheinlichkeitstheorie konsultieren.

Obschon die reguläre bedingte Verteilung und Dichte bislang noch wenig vertraut sind, gelten viele der üblichen Formeln, z. B. eine Version der Bayesschen Regel:

$$f(Y|X = x) = \frac{f_{X|Y=y}(x) f_Y(y)}{\int f_{X|Y=y'}(x) f_Y(y') dy'}$$

(wie man durch Nachrechnen verifiziert) und eine Form des Satzes von Fubini. Für integrierbares h gilt nämlich

$$\int \int h(x, y) \mathbb{P}^{X|Y=y}(dx) \mathbb{P}^Y(dy) = \int \int h(x, y) \mathbb{P}^{Y|X=x}(dy) \mathbb{P}^X(x).$$

Ist nun τ eine Zufallsvariable, die Werte aus Θ mit der Verteilung α annimmt, so berechnet man mit dieser Formel für einen Schätzer T von ϑ

$$\begin{aligned} r(\alpha, T) &= \int_\Theta \int_\mathcal{X} L(\vartheta, T(x)) \mathbb{P}_\vartheta(dx) \alpha(d\vartheta) \\ &= \int_\Theta \int_\mathcal{X} L(\vartheta, T(x)) \mathbb{P}^{X|\tau=\vartheta}(dx) \mathbb{P}^\tau(d\vartheta) \\ &= \int_\mathcal{X} \int_\Theta L(\vartheta, T(x)) \mathbb{P}^{\tau|X=x}(d\vartheta) \mathbb{P}^X(dx) \\ &= \int_\mathcal{X} \mathbb{E}[L(\tau, T(x))|X = x] \mathbb{P}^X(dx). \end{aligned}$$

Damit folgt auch der folgende Satz:

Satz 2.21 *Falls für alle $x \in \mathcal{X}$*

$$T(x) = \arg\min_a \mathbb{E}[L(\tau, a)|X = x]$$

existiert, dann ist T ein Bayesschätzer für ϑ bzgl. α und L.

Beweis: Nach Voraussetzung gilt für jeden Schätzer T'

$$\mathbb{E}[L(\tau, T(X))|X = x] \leq \mathbb{E}[L(\tau, T'(X))|X = x].$$

Nun ist aber

$$r(\alpha, T) = \mathbb{E}\mathbb{E}[L(\tau, T(X))|X]$$

die Behauptung. □

Korollar 2.22 *Ist die Verlustfunktion quadratisch, d. h. ist $\Theta \subseteq \mathbb{R}$ und*

$$L(\vartheta, a) = (\vartheta - a)^2,$$

so ist

$$T(x) = \mathbb{E}[\tau|X = x]$$

der Bayesschätzer für ϑ. Ebenso ist

$$\mathbb{E}[\gamma(\tau)|X = x]$$

der Bayesschätzer für $\gamma(\vartheta)$.

Beweis: Aufgrund von Satz 2.21 gewinnt man den Bayesschätzer durch minimieren von

$$a \mapsto \mathbb{E}[(\tau - a)^2|X = x].$$

Dies ist aber (nach dem, was wir aus der Wahrscheinlichkeitstheorie wissen) gerade

$$\mathbb{E}[\tau|X = x].$$

□

Beispiel 2.23 *(Bernoulli-Verteilung)*
Es seien $X = (X_1, \ldots, X_n)$ und die X_i seien i.i.d. Ber(p)-verteilt auf $\{0, 1\}$ mit unbekanntem $p \in (0, 1)$. Wir wählen aufgrund der großen Flexibilität durch Wahl verschiedener Parameter $a, b \in \mathbb{R}^+$ als a priori-Verteilung eine $\beta(a, b)$-Verteilung. Ihre λ-Dichte ist durch

$$g_{a,b}(x) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1 - x)^{b-1} \mathbb{1}_{(0,1)}(x).$$

Für $a = b = 1$ erhält man die Gleichverteilung auf $(0, 1)$. Ferner gilt

$$\mathbb{E}\beta(a, b) = \frac{a}{a + b}$$

und

$$\mathbb{V}(\beta(a, b)) = \frac{ab}{(a + b)^2(a + b + 1)}.$$

(Dies ist eine Übung.)

Um den Bayesschätzer bzgl. der $\beta(a,b)$-Verteilung bei quadratischer Verlustfunktion zu berechnen, müssen wir zunächst die a-posteriori-Verteilung $\mathbb{P}^{\tau|X=x}$ berechnen. Hierfür sei $x \in \{0,1\}^n$ und $\tau \sim \beta(a,b)$-verteilt, $a,b > 0$ seien fest. Nun ist die Verteilung von x bezüglich des Zählmaßes auf $\{0,1\}^n$ absolut stetig und die Dichte ist

$$f_p(x) = p^s(1-p)^{n-s}.$$

Hierbei ist $p \in (0,1)$ und wir haben $s = \sum_{i=1}^n x_i$ gesetzt. Nach Anwendung der oben zitierten Bayesschen Formel gilt

$$f^{\tau|X=x}(\vartheta) = C(a,b,s)\vartheta^{a+s-1}(1-\vartheta)^{b+n-s-1}\mathbb{1}_{(0,1)}(\vartheta)$$

(wobei wir $\vartheta = p$ setzen). Dies ist (als a-posteriori-Verteilung) wieder eine β-Verteilung zu den Parametern $a+s$ und $b+n-s$ und daher ist

$$C(a,b,s) = \frac{\Gamma(a+b+n)}{\Gamma(a+s)\Gamma(b+n-s)}.$$

Mithilfe von Korollar 2.22 erhalten wir somit als Bayesschätzer $\hat{\vartheta}$ für ϑ

$$\hat{\vartheta}(x) = \mathbb{E}[\tau|X=x] = \mathbb{E}\beta(a+s,b+n-s) = \frac{a+s}{a+b+n}.$$

Schreiben wir noch $\bar{x} = \frac{s}{n}$, so erhalten wir

$$\hat{\vartheta}(x) = \left(\frac{a+b}{a+b+n}\right)\frac{a}{a+b} + \frac{n}{a+b+n}\bar{x}.$$

Wir erhalten also als Bayesschätzer ein gewichtetes Mittel aus dem a-priori-Schätzer $\frac{a}{a+b}$ und dem ML-Schätzer $\bar{x}$. Für $n=0$ hat man nur den a-priori-Schätzer, für sehr große Stichproben verschwindet dieser Anteil und es überlebt nur der Maximum-Likelihood-Schätzer.

Beispiel 2.24 *Es sei $X = X_1$ eine Stichprobe aus* einer *Beobachtung. Diese sei $Poi(\lambda)$-verteilt, zu einem unbekannten Parameter $\lambda > 0$. Wir wählen als a-priori-Verteilung α die $\Gamma(\gamma,\eta)$-Verteilung. Diese hat die $\lambda\!\!\lambda$-Dichte*

$$f_{\gamma,\eta}(x) = \frac{1}{\Gamma(\gamma)}\eta^\gamma x^{\gamma-1}e^{-\eta x}\mathbb{1}_{(0,\infty)}(x).$$

Man rechnet nach (dies ist wieder eine Übung), dass für eine gemäß $\Gamma(\gamma,\eta)$-verteilte ZV Y gilt

$$\mathbb{E}Y = \frac{\gamma}{\eta} \quad \textit{und} \quad \mathbb{V}Y = \frac{\gamma}{\eta^2}.$$

Man rechnet für die a-posteriori-Verteilung $\alpha(\lambda|x)$ nach, dass diese wieder eine Dichte bzgl. $\lambda\!\!\lambda$ hat und zwar

$$\frac{\lambda^{\gamma-1}e^{-\eta\lambda}e^{-\lambda}\lambda^x}{z(x)},$$

wobei $z(x)$ passend gewählt ist, um dies zu einer Wahrscheinlichkeitsdichte zu machen. Also ist die a-posteriori-Verteilung wieder eine Gamma-Verteilung zu den Parametern

$$\gamma' = \gamma + x \quad \textit{und} \quad \eta' = \eta + 1.$$

Bei quadratischer Verlustfunktion ist der Bayesschätzer daher

$$T(x) = \mathbb{E}[\tau | X = x] = \frac{\gamma + x}{\eta + 1}.$$

Schreibt man dies als

$$\frac{\gamma + x}{\eta + 1} = \frac{\gamma}{\eta} \frac{\eta}{\eta + 1} + x \frac{1}{\eta + 1},$$

so sieht man wieder, dass der Bayesschätzer eine Kombination aus dem a-priori-Schätzer $\frac{\gamma}{\eta}$ *und dem Maximum-Likelihood-Schätzer* x *ist.*

Eine abschließende kurze Diskussion der Bayesmethode ergibt:

1. Der Vorteil des Bayes-Verfahrens ist seine explizite Form, sein Nachteil seine Abhängigkeit von der a-priori-Verteilung. Verschiedene a-priori-Verteilungen liefern in der Regel verschiedene Bayesschätzer.

2. Auch wenn man der Ansicht der Bayesianer nicht folgt und ϑ als Zufallsvariable interpretiert, lässt sich ein Bayesschätzer benutzen. Satz 2.21 und Korollar 2.22 sind dann einfach nette Tricks zur Bestimmung des Bayesschätzers.

3 Punktschätzungen

Wir haben im zweiten Abschnitt verschiedene Verfahren besprochen, um unbekannte Parameter einer Verteilung zu schätzen. Es ist klar, dass man für den Fall, dass diese verschieden sind, die Qualität dieser Schätzer vergleichen möchte. Hierfür haben wir schon den Begriff des Risikos eingeführt. Darüber hinaus würden wir natürlich am liebsten den "best-möglichen" Schätzer finden, in dem Sinne, dass für dieses T gelte

$$\mathcal{R}(\vartheta, T) \leq \mathcal{R}(\vartheta, T') \quad \text{für alle} \quad \vartheta \in \Theta \quad \text{und alle Schätzer } T'.$$

Das ist aber bei "vernünftigen" Verlustfunktionen L, die nur ein Minimum haben und das bei $a = \vartheta$ liegt (z. B.

$$L(\vartheta, a) = (\vartheta - a)^2$$

oder $L(\vartheta, a) = |\vartheta - a|$), nicht möglich. In diesem Falle müsste ein optimaler Schätzer besser sein als die konstanten Schätzer $T'_\vartheta = \vartheta$ (dies ist für jedes ϑ ein Schätzer, der nicht besonders clever aussieht, weil er die Informationen aus den Beobachtungen komplett ignoriert). Nun ist aber

$$\mathcal{R}(\vartheta, T'_\vartheta) = 0.$$

Somit müsste auch ein bester Schätzer für jedes ϑ Risiko 0 haben. Das ist nur dann möglich, wenn Θ einelementig ist, was für die Statistik eine wenig spannende Situation darstellt.

Man beschränkt sich daher zumeist auf das Auffinden eines besten erwartungstreuen Schätzers, also eines Schätzers T für ϑ mit

$$\mathbb{E}_\vartheta T(X) = \vartheta \quad \text{für alle} \quad \vartheta \in \Theta,$$

so dass für alle erwartungstreuen Schätzer T' von ϑ und alle $\vartheta \in \Theta$ gilt

$$\mathcal{R}(\vartheta, T) \leq \mathcal{R}(\vartheta, T').$$

Wir wollen zwei Methoden kennenlernen, solche Schätzer zu erhalten. Dazu müssen wir zunächst zwei neue Konzepte diskutieren.

3.1 Suffizienz

Wenn wir keine zeitlichen Abhängigkeiten in unseren Daten vermuten können (beispielsweise, wenn wir verschiedene Messungen zur Höhe des Eiffelturms oder zur Lichtgeschwindigkeit anstellen), ist die Reihenfolge unserer Daten offenbar irrelevant. Dies entspricht der häufigen Annahme (wir werden später noch eine Situation kennenlernen, bei der dies anders ist), dass die Daten i.i.d. Zufallsvariablen sind. Wenn aber die Reihenfolge irrelevant ist, so ist jede andere Reihenfolge der Daten genauso gut wie die unsere. Vielleicht lassen sich die Daten ja sogar noch mehr reduzieren. Dabei ist klar, dass im Allgemeinen Informationen verloren gehen, wenn wir die Beobachtungen mit einer nicht-umkehrbaren Transformation

$$S : \mathcal{X} \to \mathcal{Y}$$

transformieren (dabei wollen wir jede messbare Abbildung von $\mathcal{X}$ in ein $\mathcal{Y}$ Statistik nennen). Es gibt jedoch Situationen, in denen $S(X)$ ebenso viele Informationen enthält wie X. In diesem Fall wollen wir S suffizient nennen, d. h. informationserhaltend. Wir beginnen mit der mathematischen Definition und Beispielen und diskutieren sie dann im allgemeinen Rahmen.

Definition 3.1 *Sei $\mathcal{P} = \{\mathbb{P}_\vartheta^X, \vartheta \in \Theta\}$ eine Menge von Wahrscheinlichkeitsverteilungen auf $(\mathcal{X}, \mathcal{A})$ und*

$$S : \mathcal{X} \to \mathcal{Y}$$

eine Statistik. S heißt suffizient für $\mathcal{P}$, falls die bedingte Verteilung

$$\mathbb{P}^{X|S=s} = \mathbb{P}_\vartheta[X \in \cdot | S(X) = s]$$

nicht von ϑ abhängt.

Dahinter steckt die folgende Idee: Wenn S eine nicht-umkehrbare Abbildung ist, so können wir nach Anwendung von S die Beobachtung x nicht mehr rekonstruieren. Wir können aber einen anderen Wert x^* aus dem Urbild $S^{-1}(s)$ zufällig mit der Verteilung

$$\mathbb{P}_\vartheta[X \in \cdot | S(X) = s]$$

ziehen, da diese unabhängig von ϑ ist. Wenn wir x^* verwenden statt x, ändert das die Verteilung des Schätzers nicht.

Beispiel 3.2 *Die $X_1, \ldots, X_n$ seien i.i.d. Ber(p)-verteilt, wobei $p \in (0,1)$ unbekannt ist. Wir wollen zeigen, dass*

$$S = \sum_{i=1}^n X_i$$

suffizient ist für die Familie

$$\mathcal{P} = \{\bigotimes_{i=1}^n Ber(p),\ p \in (0,1)\}.$$

Nun gilt für $X = (X_1, \ldots, X_n)$

$$\mathbb{P}_p^{X|S=s}(\{x\}) = 0,$$

falls $\sum_{i=1}^n x_i \neq s$ (und dies ist unabhängig von p). Falls aber $S(x) = s$ ist, gilt

$$\mathbb{P}_p^{X|S=s}(\{x\}) = \frac{\mathbb{P}_p^X(\{x\})}{\mathbb{P}_p(S=s)} = \frac{p^s(1-p)^{n-s}}{\binom{n}{s}p^s(1-p)^{n-s}} = \frac{1}{\binom{n}{s}},$$

was wiederum unabhängig von p ist. S ist somit suffizient für die Familie $\mathcal{P}$.

Steigen wir noch einmal bei der Diskussion vor dem Beispiel 3.2 ein. Verwendet man dort x^* statt x, so ist die gewählte Aktion x^* nicht nur von x abhängig, sondern auch von einem Zufallsgenerator, der x^* aus der Menge $\{y : S(y) = s\}$ aussucht. Wir haben also eine randomisierte Entscheidung, einen randomisierten Schätzer $T(x^*)$ (da dieser typischerweise verschieden ist von $T(x)$). Immerhin vergrößern wir aber das Risiko nicht:

Satz 3.3 *Sei S eine suffiziente Statistik für $\mathcal{P} = \{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$. Dann existiert für jeden Schätzer*

$$T : \mathcal{X} \to \Theta$$

ein randomisierter Schätzer $\bar{T}$ basierend auf $S(X)$, derart, dass T und $\bar{T}$ das gleiche Risiko haben.

Beweis: Das haben wir in Worten oben bereits beschrieben. Setze

$$\bar{T}(s) = T(x^*),$$

wobei x^* auf $S^{-1}(s)$ zufällig gemäß $\mathbb{P}_\vartheta[X \in \cdot | S(X) = s]$ gezogen werde. Wegen der Suffizienz von S benötigen wir hierfür die Kenntnis von ϑ nicht. Dann gilt:

$$\begin{aligned} \mathcal{R}(\vartheta, \bar{T}) &= \mathbb{E}_\vartheta[L(\vartheta, \bar{T}(S(X)))] \\ &= \mathbb{E}_\vartheta \mathbb{E}_\vartheta[L(\vartheta, T(X))|S(X) = s] \\ &= \mathcal{R}(\vartheta, T), \end{aligned}$$

wobei wir die Definition des Risikos, die bedingte Verteilung und die Glättungseigenschaft der bedingten Erwartung benutzt haben. □

Sind Entscheidungsraum und Verlustfunktion konvex, erhält man sogar ein kleineres Risiko, wenn man sich auf Schätzer beschränkt, die nur von der suffizienten Statistik abhängen. Hierbei kommt man ohne Randomisieren aus, sondern mittelt einfach.

Satz 3.4 *(Rao-Blackwell)*
Es sei $\Theta \subseteq \mathbb{R}^d$ konvex und $L(\vartheta, \cdot)$ konvex für alle $\vartheta \in \Theta$. Ferner sei S eine suffiziente Statistik für $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ und

$$T : \mathcal{X} \to \Theta$$

ein Schätzer mit

$$\mathcal{R}(\vartheta, T) < +\infty \quad \textit{und} \quad \mathbb{E}_\vartheta(|T|) < +\infty$$

für alle $\vartheta \in \Theta$. Setze

$$\bar{T}(s) = \mathbb{E}_\vartheta[T(X)|S(X) = s].$$

Dann ist

$$\mathcal{R}(\vartheta, \bar{T}) \leq \mathcal{R}(\vartheta, T)$$

für alle $\vartheta \in \Theta$. Ist $L(\vartheta, \cdot)$ sogar strikt konvex, so gilt sogar

$$\mathcal{R}(\vartheta, \bar{T}) < \mathcal{R}(\vartheta, T),$$

außer wenn $\bar{T} = T$ $\mathbb{P}_\vartheta$-f.s. gilt.

Bemerkung 3.5 *Wie vorher wird die Suffizienz hier wieder benötigt, damit $\bar{T}$ nicht von ϑ abhängt, also ein gültiger Schätzer ist.*

Beweis: Wegen der Jensenschen Ungleichung folgt

$$\mathbb{E}[L(\vartheta, T(X))|S(X) = s] \geq L(\vartheta, \mathbb{E}[T(X)|S(X) = s]).$$

Bildet man nun auf beiden Seiten den Erwartungswert, so erhält man links $\mathcal{R}(\vartheta, T)$ und rechts $\mathcal{R}(\vartheta, \bar{T})$. Ist L strinkt konvex, so ist die Ungleichung auch strikt, außer es gilt

$$T(X) = \mathbb{E}[T(X)|S(X)] \ \mathbb{P}\text{-f.s.}$$

□

Oftmals ist es ein wenig lästig, die bei der Suffizienz auftretenden bedingten Wahrscheinlichkeiten zu berechnen. Ein handliches Kriterium für Suffizienz liefert der folgende Satz.

Satz 3.6 *(Faktorisierungskriterium von Neyman)*
Es sei $\{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ eine Familie von Wahrscheinlichkeitsmaßen, die durch ein σ-endliches Maß μ dominiert sind. Es sei

$$\frac{d\mathbb{P}_\vartheta}{d\mu} = f_\vartheta.$$

Eine Statistik

$$S : (\mathcal{X}, \mathcal{A}) \to (\mathcal{X}', \mathcal{A}')$$

ist genau dann suffizient für $\{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$, wenn es $\mathcal{A}'$-messbare bzw. $\mathcal{A}$-messbare Funktionen g_ϑ und h gibt, so dass

$$f_\vartheta(x) = g_\vartheta(S(x))h(x)$$

gilt.

Beweis: Wir beweisen den diskreten Fall. Der allgemeine Fall folgt denselben Ideen, ist aber technisch wesentlich aufwändiger (siehe z. B. Alsmeyer: "Mathematische Statistik" oder Lehmann: "Testing statistical hypothesis"). Sei also $\mathcal{X}$ abzählbar und μ das Zählmaß. Es gilt

$$\mathbb{P}_\vartheta[X = x|S = s] = \begin{cases} \frac{\mathbb{P}_\vartheta[X=x]}{\mathbb{P}_\vartheta[S=s]} & \text{falls } S(x) = s \\ 0 & \text{sonst} \end{cases}.$$

Für die Hin-Richtung beachte man, dass die linke Seite aufgrund der Suffizienz von S nicht von ϑ abhängt. Setzen wir also

$$\begin{aligned} g_\vartheta(s) &= \mathbb{P}_\vartheta[S(X) = s] \quad \text{und} \\ h(x) &= \mathbb{P}[X = x|S(X) = s], \end{aligned}$$

so erhalten wir

$$\begin{aligned} g_\vartheta(s) \cdot h(x) &= \mathbb{P}_\vartheta[S(X) = s]\mathbb{P}_\vartheta[X = x|S(X) = s] \\ &= \mathbb{P}_\vartheta[X = x, S(X) = s] \\ &= \mathbb{P}_\vartheta[X = x]. \end{aligned}$$

Für die Rückrichtung geht man von

$$\mathbb{P}_\vartheta[S = s] = g_\vartheta(s) \sum_{x:S(x)=s} h(x)$$

aus, was aus der Voraussetzung folgt. Dies ergibt

$$\mathbb{P}_\vartheta[X = x|S(X) = s] = \frac{h(x)}{\sum_{x':S(x')=S(x)} h(x')},$$

was offenbar von ϑ unabhängig ist. Also ist S suffizient. □

Beispiel 3.7 *Seien $X_1, \ldots, X_n$ i.i.d. gleichverteilt auf $\Theta = (0; \vartheta)$ und ϑ sei unbekannt. Die Familie der $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ ist also gegeben durch*

$$\{\mathbb{P}_\vartheta : \vartheta \in \Theta\} = \{\mathcal{R}^n(0, \vartheta)), \vartheta \in \mathbb{R}^+\}.$$

Wir können ihre Dichten bzgl. des Lebesguemaßes dann schreiben als

$$f_\vartheta(x_1, \ldots, x_n) = \begin{cases} \vartheta^{-n}, & \textit{falls } \max_{i=1,\ldots,n}(x_i) \leq \vartheta \\ 0, & \textit{sonst} \end{cases}.$$

Also ist nach dem Neyman-Kriterium

$$S(X) = \max_{i=1,\ldots,n} X_i$$

eine suffiziente Statistik für $\{\mathbb{P}_\vartheta : \vartheta \in \mathbb{R}^+\}$.

Beispiel 3.8 *Seien $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$-verteilt, also*

$$\{\mathbb{P}_\vartheta : \vartheta \in \Theta\} = \{\mathcal{N}^n(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}.$$

Wählt man als dominierendes Maß das Lebesgue-Maß λ^n, so erhält man als gemeinsame Dichte

$$\begin{aligned} f_{\mu,\sigma^2}(x_1, \ldots, x_n) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-n\frac{1}{2}(\frac{\bar{x}-\mu}{\sigma})^2} e^{-\sum_{i=1}^n \frac{(x_i-\bar{x})^2}{2\sigma^2}} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{n}{2}(\frac{\bar{x}-\mu}{\sigma})^2 - (n-1)\frac{s^2}{2\sigma^2}}, \end{aligned}$$

wobei wir

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i \quad \textit{und} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$$

gesetzt haben. Somit ist die bekannte Statistik $(\bar{x}, s^2)$ suffizient für die Familie der $\{\mathbb{P}_\vartheta$, $\vartheta \in \Theta\}$.

Eine ganze Familie neuer Beispiele gewinnen wir mit der nächsten Definition.

Definition 3.9 *Eine Familie $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ heißt k-parametrige Exponentialfamilie in $Q = (Q_1, \ldots, Q_k)$ und $T = (T_1, \ldots, T_k)$, wenn es ein $\mathcal{P}$-dominierendes Maß μ gibt, so dass*

$$f_\vartheta = \frac{d\mathbb{P}_\vartheta}{d\mu}$$

sich schreiben lässt als

$$f_\vartheta(x) = C(\vartheta) \exp\left(\sum_{i=1}^{k} Q_i(\vartheta) T_i(x)\right) h(x).$$

Hierbei sind die

$$Q_i : \Theta \to \mathbb{R} \quad i = 1, \ldots, k$$

und $h : \mathcal{X} \to \mathbb{R}$ sowie

$$T_i : \mathcal{X} \to \mathbb{R}$$

messbare Abbildungen. Wir sagen darüber hinaus, dass die Exponentialfamilie vollen Rang *besitzt, falls $1, Q_1, \ldots, Q_k$ linear unabhängig auf Θ sind und $1, T_1, \ldots, T_k$ linear unabhängig auf N^c für jede Nullmenge $N \in \mathcal{A}$ ($\mathcal{A}$ die σ-Algebra auf $\mathcal{X}$) sind. Letzteres bedeutet*

$$\begin{aligned} & c_0 + \sum_{j=1}^{k} c_j T_j = 0 \quad \mathcal{P}\text{-f.s.} \\ \Rightarrow \quad & c_0 = \ldots = c_k = 0. \end{aligned}$$

Beispiel 3.10 *Wir sehen, dass bekannte Verteilungsfamilien Exponentialfamilien sind:*

a) Die Familie $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ ist eine 2-parametrige Exponentialfamilie, denn die λ-Dichte ist

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\mu^2}{2\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2} + \frac{\mu}{\sigma^2} x\right).$$

Setzen wir $k = 2$, $C(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\mu^2}{2\sigma^2}}$,

$$\begin{aligned} Q_1(\mu, \sigma^2) &= -\frac{1}{2\sigma^2}, & Q_2(\mu, \sigma^2) &= \frac{\mu}{\sigma^2} \\ T_1(x) &= x^2, & T_2(x) &= x, \end{aligned}$$

so erhält man die gewünschte Form.

b) Die Familie der $B(n, p)$-Verteilungen ist für festes n eine einparametrige Exponentialfamilie bezüglich des Zählmaßes μ auf $\{0, \ldots, n\}$. In der Tat gilt ja für die Dichte $f_p(x)$

$$f_p(x) = \binom{n}{x} p^x (1-p)^{n-x} = (1-p)^n \binom{n}{x} e^{x \log \frac{p}{1-p}}.$$

Wir wählen also

$$\begin{aligned} C(p) &= (1-p)^n \\ Q_1(p) &= \log \frac{p}{1-p} \\ T_1(x) &= x \\ \text{und} \quad h(x) &= \binom{n}{x} \end{aligned}$$

und sehen, dass wir in der Tat die Struktur einer Exponentialfamilie erhalten.

Wir sehen nun, dass in einer k-parametrigen Exponentialfamilie $T = (T_1, \ldots, T_k)$ suffizient ist.

Korollar 3.11 *Es sei $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ eine k-parametrige Exponentialfamilie in $Q = (Q_1, \ldots, Q_k)$ und $T = (T_1, \ldots, T_k)$. Dann ist T suffizient für $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$.*

Beweis: Da $\{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ eine Exponentialfamilie ist, gibt es ein Maß μ, so dass μ-f.s. gilt

$$\frac{d\mathbb{P}_\vartheta}{d\mu} = C(\vartheta) \cdot \exp\left(\sum_{i=1}^{k} Q_i(\vartheta) T_i(x)\right) h(x)$$

(für messbare Funktionen $Q_i, T_i, i = 1, \ldots, k$ und h). Setzt man nun

$$g_\vartheta(T(x)) = C(\vartheta) \exp(< Q(\vartheta) T(x) >),$$

wobei $< \cdot, \cdot >$ das Skalarprodukt in $\mathbb{R}^k$ ist, so ist das Neyman-Kriterium für Suffizienz erfüllt. □

Natürlich ist eine suffiziente Statistik nicht notwendig eindeutig (man kann sie z. B. immer mit Konstanten multiplizieren). Wir sehen in der folgenden Proposition dann auch, dass man aus einer suffizienten Statistik viele andere konstruieren kann.

Proposition 3.12 *Es sei die Familie $\mathcal{P} = \{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ dominiert durch ein σ-endliches Maß μ.*

$$T : \mathcal{X} \to \mathcal{X}'$$

sei suffizient für $\mathcal{P}$. Dann ist jede weitere Statistik

$$S : \mathcal{X} \to \mathcal{X}'',$$

für die sich T in der Form

$$T = k \circ S$$

für eine messbare Funktion

$$k : \mathcal{X}'' \to \mathcal{X}'$$

schreiben lässt, ebenfalls suffizient für $\mathcal{P}$.

Beweis: Wir setzen

$$f_\vartheta := \frac{d\mathbb{P}_\vartheta}{d\mu}.$$

Nach dem Neyman-Kriterium existieren aufgrund der Suffizienz von T messbare Funktionen g_ϑ und h mit

$$f_\vartheta = (g_\vartheta \circ T) \cdot h = (g_\vartheta \circ k \circ S) \cdot h = (\tilde{g}_\vartheta \circ S) \cdot h,$$

wobei $\tilde{g}_\vartheta$ als

$$\tilde{g}_\vartheta := g_\vartheta \circ k$$

definiert ist. Damit folgt die Suffizienz von S wieder aus dem Neyman-Kriterium. □

Beispiel 3.13 *Da wir in Beispiel 3.8 schon gesehen haben, dass*

$$T(x) = \left(\frac{\sum_{i=1}^n x_i}{n}, \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right)$$

suffizient ist für die Familie $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$*, ist auch*

$$S(x) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$$

suffizient für diese Klasse, denn

$$T(x) = \left(\frac{S_1(x)}{n}, \frac{n}{n-1} \left(\frac{1}{n} S_2(x) - \left(\frac{1}{n} S_1(x) \right)^2 \right) \right).$$

Man kann also durch das Anwenden einer suffizienten Statistik eine Datenreduktion erreichen und sogar zeigen, dass ein Schätzer, der auf einer suffizienten Statistik basiert, in der Regel besser ist als ein anderer. Die Frage ist natürlich, wie weit man so eine Datenreduktion treiben kann, ob es eine "einfachste" suffiziente Statistik gibt. Dies wollen wir in der Folge klären.

Definition 3.14 *Gegeben sei eine Familie von Wahrscheinlichkeitsmaßen* $\mathcal{P} = (\mathbb{P}_\vartheta, \vartheta \in \Theta)$ *über einem messbaren Raum* $(\mathcal{X}, \mathcal{A})$*. Eine für* $(\mathbb{P}_\vartheta : \vartheta \in \Theta)$ *suffiziente Statistik* $T^* : \mathcal{X} \to \mathcal{X}'$ *heißt* minimal suffizient*, wenn sie messbar über jeder weiteren suffizienten Statistik faktorisiert, d. h. wenn es für jede weitere suffiziente Statistik* T *eine messbare Funktion* h *gibt, so dass*

$$T^* = h \circ T \quad \mathcal{P}\text{-f.s.}$$

gilt.

Dies ist eine sinnvolle Vereinbarung (insofern Definitionen sinnvoll sein können), als der Übergang von T zu T^* mittels einer messbaren Funktion in der Tat eine Vereinfachung darstellt. Bei der Konstruktion minimal-suffizienter Statistiken konzentrieren wir uns auf Familien $\mathcal{P}$ äquivalenter Maße, z. B. Exponentialfamilien.

Satz 3.15 *Sei $\mathcal{P} = \{\mathbb{P}_j, j = 0, \ldots, n\}$ eine endliche Familie äquivalenter Verteilungen auf $(\mathcal{X}, \mathcal{A})$ mit Dichten $f_0, f_1, \ldots, f_n$ bzgl. eines dominierenden Maßes μ. Dann ist*

$$T(x) = \left(\frac{f_1(x)}{f_0(x)}, \ldots, \frac{f_n(x)}{f_0(x)} \right)$$

eine minimalsuffiziente Statistik für $\mathcal{P}$.

Beweis: Da die $\mathbb{P}_j$ allesamt äquivalent sind, stimmen die Mengen $\{f_j > 0\}$ μ-f.s. überein. Setzt man $\frac{0}{0} := 0$, so ist T auch wohldefiniert. Für jedes $j \in \{1, \ldots, n\}$ gilt

$$\frac{d\mathbb{P}_j}{d\mathbb{P}_0} = \frac{\frac{d\mathbb{P}_j}{d\mu}}{\frac{d\mathbb{P}_0}{d\mu}} = \frac{f_j}{f_0} = \pi_j \circ T \quad \mu\text{-f.s.},$$

wobei π_j die Projektion auf die j-te Koordinate bezeichnet. Somit ist T eine suffiziente Statistik für $\mathcal{P}$. Dies folgt unmittelbar aus dem Neyman-Kriterium, wenn man $\mathbb{P}_0$ als dominierendes Maß wählt. Nach diesem Kriterium existieren für jede weitere suffiziente Statistik S Funktionen $h, g_0, \ldots, g_n$, so dass

$$f_j = (g_j \circ S) \cdot h, \quad \text{also} \quad \frac{f_j}{f_0} = \frac{g_j}{g_0} \circ S$$

gilt. Dies impliziert

$$T = \left(\frac{g_1}{g_0}, \ldots, \frac{g_n}{g_0} \right) \circ S \quad \mu\text{-f.s.}$$

Dies bedeutet T ist minimal suffizient. □

Das folgende Lemma zeigt, dass der vorhergehende Satz auch für beliebige Familien $\mathcal{P}$ seinen Wert hat.

Lemma 3.16 *Sei $\mathcal{P}$ eine Familie äquivalenter Verteilungen und $\mathcal{P}_0 \subseteq \mathcal{P}$ sei eine endliche Teilfamilie. Dann ist jede Statistik, die minimal suffizient für $\mathcal{P}_0$ ist und suffizient für $\mathcal{P}$, auch minimal suffizient für $\mathcal{P}$.*

Beweis: Sei T eine solche Statistik und S eine für $\mathcal{P}$ suffiziente Statistik. Dann ist S auch suffizient für $\mathcal{P}_0$. Da T minimal suffizient für $\mathcal{P}_0$ ist, gibt es eine messbare Funktion h, so dass

$$T = h \circ S \quad \mathcal{P}_0\text{-f.s.}$$

git. Daraus folgt aber auch $T = h \circ S \quad \mathcal{P}$-f.s., denn $\mathcal{P}_0$ und $\mathcal{P}$ sind nach Voraussetzung äquivalent. □

Dies hat besonders für Exponentialfamilien eine interessante Konsequenz.

Satz 3.17 *Sei $\mathcal{P} = \{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ eine k-parametrige Exponentialfamilie in $Q = (Q_1, \ldots, Q_k)$ und $T = (T_1, \ldots, T_k)$. Dann ist T minimal suffizient für $\mathcal{P}$, wenn*

$$\mathcal{Q} := \{(Q_1(\vartheta), \ldots, Q_k(\vartheta)), \vartheta \in \Theta\} \subseteq \mathbb{R}^d$$

innere Punkte besitzt.

Beweis: Nach Proposition 3.12 ist T suffizient. Sei $\mathcal{P}_0 = (\mathbb{P}_\vartheta)_{j=0}^k$ eine Teilfamilie von $\mathcal{P}$. Aus Satz 3.15 folgert man, dass

$$\hat{T}(x) = \left(\sum_{j=1}^k (Q_j(\vartheta_1) - Q_j(\vartheta_0)) T_j(x), \ldots, \sum_{j=1}^k (Q_j(\vartheta_k) - Q_j(\vartheta_0)) T_j(x) \right)$$

minimal suffizient ist für $\mathcal{P}_0$. Nun gilt

$$\hat{T} = \Delta Q \cdot T =: (Q_i(\vartheta_j) - Q_i(\vartheta_0)) \cdot T.$$

Ist ΔQ regulär, d. h. invertierbar, so ist

$$T = (\Delta Q)^{-1} \hat{T},$$

und somit ist auch T minimal suffizient für $\mathcal{P}_0$. Dies impliziert nach Lemma 3.16 auch die Minimalsuffizienz von T für $\mathcal{P}$. Nun lassen sich die $\vartheta_0, \ldots, \vartheta_k$ aber immer so wählen, dass ΔQ regulär ist, denn $\mathcal{Q}$ hat innere Punkte, ist also k-dimensional. □

Beispiel 3.18 *Anhand von Beispiel 3.10 vergewissert man sich schnell, dass die beiden folgenden Beispiele die Voraussetzungen an $\mathcal{Q}$ in Satz 3.17 erfüllen:*

a) Ist $\mathcal{P} = (B(n,p))_{p \in (0,1)}$, dann ist $T(x) = \sum_{i=1}^n x_i$ und $\hat{T} = \frac{1}{n} \sum_{i=1}^n x_i$ minimal suffizient.

b) Sei $\mathcal{P} = \{\mathcal{N}^n(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$. Dann sind

$$\begin{aligned} T(x) &= \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2 \right) \quad \text{und} \\ \hat{T}(x) &= (\bar{x}_n, \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x}_n)^2) \end{aligned}$$

minimal suffizient.

3.2 Vollständigkeit

Für die Diskussion erwartungstreuer Schätzer benötigen wir noch einen weiteren Begriff. Zur Motivation beginnen wir mit der sogenannten "Verteilungsfreiheit".

Definition 3.19 *Es sei $(\mathcal{X}, \mathcal{A})$ ein messbarer Raum und $(\mathbb{P}_\vartheta)_\vartheta$ eine Familie von Wahrscheinlichkeitsmaßen darauf. Eine Statistik*

$$T : \mathcal{X} \to \mathcal{X}'$$

heißt verteilungsfrei*, falls die Verteilung $\mathbb{P}_\vartheta^T$ unabhängig von ϑ ist. Sie heißt* verteilungsfrei 1. Ordnung*, falls der Erwartungswert*

$$\mathbb{E}_\vartheta T$$

nicht mehr von ϑ abhängt.

Offenbar ist Verteilungsfreiheit eine Art Gegenpol zur Suffizient: Eine suffiziente Statistik behält alle für ϑ relevanten Informationen, eine verteilungsfreie Statistik besitzt gar keine Informationen über den unbekannten Parameter. Dennoch kann auch eine minimal suffiziente Statistik T noch verteilungsfreies Material enthalten. Manchmal lassen sich für solche Statistik nicht konstante Funktionen f finden, so dass $f(T)$ verteilungsfrei ist.

Dennoch ist es plausibel, dass eine suffiziente Statistik T nicht mehr weiter verbessert werden kann, wenn es keine nicht-konstante Funktion f gibt, so dass $f(T)$ verteilungsfrei ist. Es stellt sich heraus, dass dies in der Tat wahr ist, wenn man "verteilungsfrei" durch "verteilungsfrei 1. Ordnung" ersetzt.

Dies lässt sich schreiben als

$$\mathbb{E}_\vartheta f(T) = c \quad \mathbb{P}_\vartheta\text{-f.s.} \quad \forall\, \vartheta \in \Theta \Rightarrow f \equiv c \quad \mathbb{P}_\vartheta\text{-f.s.} \quad \forall\, \vartheta \in \Theta.$$

Durch Subtraktion des Erwartungswerts kann man sich auf die konstante Nullfunktion beschränken.

Definition 3.20 *In der Situation von Definition 3.19 heißt eine Statistik $T : \mathcal{X} \to \mathcal{X}'$ vollständig, falls*

$$\mathbb{E}_\vartheta f(T) = 0 \quad \mathbb{P}_\vartheta\text{-f.s. für alle} \quad \vartheta \in \Theta$$

schon impliziert, dass

$$f \equiv 0 \quad \mathbb{P}_\vartheta^{T(X)}\text{-f.s. für alle} \quad \vartheta \in \Theta$$

gilt.

Nun können wir auch die eingangs aufgestellte Vermutung beweisen.

Satz 3.21 *Es sei $(\mathcal{X}, \mathcal{A})$ ein messbarer Raum und $\{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ eine Familie von Wahrscheinlichkeitsmaßen auf $(\mathcal{X}, \mathcal{A})$. Ist eine Statistik*

$$T : \mathcal{X} \to \mathcal{X}'$$

suffizient und vollständig, so ist sie auch minimal suffizient.

Beweis: Es sei

$$\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$$

und

$$S : (\mathcal{X}, \mathcal{A}) \to (\mathcal{X}'', \mathcal{A}'')$$

eine beliebige suffiziente Statistik. Zu zeigen ist, dass es eine messbare Funktion

$$f : (\mathcal{X}'', \mathcal{A}'') \to (\mathcal{X}', \mathcal{A}')$$

gibt mit $T = f \circ S$. Eine Übung zeigt, dass dies gezeigt ist, falls wir für alle $\mathcal{A} \in \mathcal{A}'$ ein $B \in \mathcal{A}''$ angeben können mit

$$\mathbb{1}_A(T) = \mathbb{1}_B(S) \quad \mathcal{P}\text{-f.s.} \tag{3.1}$$

(3.1) gilt, wenn

$$\mathbb{P}[T \in \mathcal{A}|S] = \mathbb{1}_A(T) \quad \mathcal{P}\text{-f.s.} \tag{3.2}$$

für alle $A \in \mathcal{A}'$. (In der Tat ist ja (3.2) gleichbedeutend mit

$$\mathbb{1}_A(T) \in S^{-1}(\mathcal{A}'') \quad \text{für alle} \quad A \in \mathcal{A}'.$$

Dies wiederum ist gleichbedeutend mit (3.1).) Nun gilt aber (3.2) zumindest, wenn man unter T bedingt, denn

$$\int_{\mathcal{X}} (\mathbb{P}[\mathbb{P}[T \in A|S]|T] - \mathbb{1}_A(T)) d\mathbb{P}_\vartheta = \mathbb{P}_\vartheta(T \in A) - \mathbb{P}_\vartheta(T \in A) = 0 \quad \text{für alle} \quad \vartheta \in \Theta.$$

Da T als vollständig angenommen war, erhalten wir hieraus

$$\mathbb{P}[\mathbb{P}[T \in A|S]|T] = \mathbb{1}_A(T) \quad \mathcal{P}\text{-f.s.}$$

Damit erhalten wir

$$\begin{aligned} 0 &\leq \mathbb{P}[(\mathbb{P}[T \in A|S] - \mathbb{1}_A(T))^2|T] \\ &= \mathbb{P}[\mathbb{P}[T \in A|S]^2|T] - 2\mathbb{1}_A(T)\mathbb{P}[\mathbb{P}[T \in A|S]|T] + \mathbb{1}_A(T)^2 \\ &= \mathbb{P}[\mathbb{P}[T \in A|S]^2|T] - \mathbb{1}_A^2(T) \\ &\leq \mathbb{P}[\mathbb{P}[T \in A|S]|T] - \mathbb{1}_A(T) = 0. \end{aligned}$$

In der letzten Ungleichung haben wir hierbei die Positivität der Differenz ausgenutzt, dass $\mathbb{1}_A(T) = \mathbb{1}_A^2(T)$ gilt und

$$\mathbb{P}[T \in A|S]^2 \leq \mathbb{P}[T \in A|S].$$

Also folgt offenbar

$$\mathbb{P}[T \in A|S] = \mathbb{1}_A(T) \quad \mathcal{P}\text{-f.s.}$$

□

Zu betonen ist, dass die Umkehrung von Satz 3.21 nicht gilt.

Eingangs hatten wir den Begriff der Vollständigkeit über das Fehlen weiterer verteilungsfreier Informationen motiviert. Es ist daher plausibel, dass eine vollständige, suffiziente Statistik von jeder verteilungsfreien Statistik unabhängig ist. Dies ist der Inhalt des folgenden Satzes.

Satz 3.22 *(Basu)*
Es sei $(\mathcal{X}, \mathcal{A})$ ein messbarer Raum und $(\mathbb{P}_\vartheta : \vartheta \in \Theta)$ eine Familie von Wahrscheinlichkeitsmaßen darauf. Es sei

$$T : \mathcal{X} \to \mathcal{X}'$$

eine vollständige, suffiziente Statistik für $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$. Dann ist jede verteilungsfreie Statistik

$$S : (\mathcal{X}, \mathcal{A}) \to (\mathcal{X}'', \mathcal{A}'')$$

unabhängig von T.

Beweis: Da S verteilungsfrei ist, ist $Q := \mathbb{P}_\vartheta^S$ unabhängig von ϑ. Für $A'' \in \mathcal{A}''$ sei nun

$$f_{A''}(t) = \mathbb{P}[S \in A''|T = t].$$

Dann gilt

$$\mathbb{E}_\vartheta[f_{A''}(T(X)) - Q(A'')] = \int \mathbb{P}(S \in A''|T) - Q(A'')d\mathbb{P}_\vartheta = \int (\mathbb{1}_{S\in A''} - Q(A''))d\mathbb{P}_\vartheta = 0$$

für alle $\vartheta \in \Theta,\ A'' \in \mathcal{A}''$. Da T vollständig ist, folgt daraus

$$f_{A''} = \mathbb{P}[S \in A''|T = \cdot] = Q(A'') \quad \mathcal{P}\text{-f.s.}$$

Dies ist aber die behauptete Unabhängigkeit von S und T. □

Dieser Satz hat eine interessante und überraschende Konsequenz:

Beispiel 3.23 *Es seien $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$-verteilt, wobei $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ gilt. Wir haben früher schon gesehen, dass*

$$\hat{T}(x) = \left(\frac{1}{n}\sum_{i=1}^n x_i, \frac{1}{n-1}\sum_{i=1}^n \left(x_i - \frac{1}{n}\sum_{i=1}^n x_i \right)^2 \right)$$

minimal suffizient für die Familie

$$\{\mathcal{N}^n(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$$

ist. Wir wollen nun sehen, dass das Stichprobenmittel $\frac{1}{n}\sum_{i=1}^n x_i =: \bar{x}$ und die Stichprobenvarianz

$$v^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$$

unter jeder der $\mathcal{N}(\mu, \sigma^2)$-Verteilungen unabhängig sind. Dies ist auf den ersten Blick überraschend, da v^2 ja explizit $\bar{x}$ benutzt. Sei dafür $\sigma^2 > 0$. Dann ist $\bar{x}$ suffizient für die Familie

$$\{\mathcal{N}^n(\mu, \sigma^2) : \mu \in \mathbb{R}^+\}.$$

(Dies ist eine Übung.) Weiter ist $\bar{x}$ auch vollständig (das ist die nächste Übung). Können wir zeigen, dass v^2 verteilungsfrei ist für $\{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}\}$ (was naheliegend ist, denn v^2

soll ja σ^2 schätzen und σ^2 ist ja fest für diese Klasse), so können wir mithilfe des Satzes von Basu die Unabhängigkeit von $\bar{x}$ und v^2 folgern. Dazu überlegen wir: Sei

$$Y_j = Y_j(\mu) = X_j - \mu, \quad j = 1, \ldots, n.$$

Dann sind $Y_1, \ldots, Y_n$ unter $\mathcal{N}(\mu, \sigma^2)$ stochastisch unabhängig und $\mathcal{N}(0, \sigma^2)$-verteilt. Wegen der Translationsinvarianz ist $v^2(X) = v^2(Y)$, wobei $X = (X_1, \ldots, X_n)$ und $Y = (Y_1, \ldots, Y_n)$ gesetzt ist. Somit ist

$$\mathbb{P}^{v^2(X)}_{\mu,\sigma^2} = \mathbb{P}^{v^2(Y)}_{\mu,\sigma^2},$$

d. h. die Verteilung ist in der Tat unabhängig von μ. Dies ist aber die Verteilungsfreiheit von v^2.

3.3 Erwartungstreue Schätzer

Nun werden wir versuchen, optimale Schätzer zu konstruieren. Wie am Anfang des Kapitels besprochen beschränken wir uns hierbei auf gleichmäßig beste erwartungstreue Schätzer, also solche Schätzer T, für die

$$\mathcal{R}(\vartheta, T) \leq \mathcal{R}(\vartheta, T')$$

für alle $\vartheta \in \Theta$ und alle Schätzer T' gilt. Bei quadratischer Verlustfunktion ist das Risiko eines erwartungstreuen Schätzers nichts anderes als seine Varianz, denn (wollen wir mit dem Schätzer T die Funktion $\gamma(\vartheta)$ schätzen) es gilt:

$$\mathbb{E}_\vartheta[(T(X) - \gamma(\vartheta))^2] = \mathbb{V}_\vartheta T + (\mathbb{E}_\vartheta T(X) - \gamma(\vartheta))^2$$

und der hintere Teil ist für ein erwartungstreues T gleich null.

Definition 3.24 *Ein Schätzer $T : \mathcal{X} \to \mathbb{R}^m$ heißt* gleichmäßig bester erwartungstreuer Schätzer *(GBES oder UMVU = uniform minimum variance unbiased) für die Parameterfunktion $\gamma(\vartheta)$, falls T erwartungstreu für $\gamma(\vartheta)$ ist, d. h.*

$$\mathbb{E}_\vartheta T = \gamma(\vartheta) \quad \textit{für alle} \quad \vartheta \in \Theta,$$

und falls

$$\mathbb{V}_\vartheta(T) \leq \mathbb{V}_\vartheta(T')$$

für alle $\vartheta \in \Theta$ und alle erwartungstreuen T' gilt.

Auf den ersten Blick scheint Erwartungstreue ein sehr vernünftiges Konzept zu sein. Zum Beispiel schließt es die lästigen konstanten Schätzer aus. Es hat aber auch Schwachpunkte:

- Es gibt nicht immer erwartungstreue Schätzer (das ist eine Übung);

- UMVUs können unzulässig sein, d. h. ist T ein UMVU, so ist es möglich, dass es einen Schätzer S gibt (der dann natürlich nicht erwartungstreu ist) mit

$$\mathcal{R}_\vartheta(S) \leq \mathbb{V}_\vartheta(T)$$

für alle $\vartheta \in \Theta$ und

$$\mathcal{R}_{\vartheta'}(S) < \mathbb{V}_{\vartheta'}(T)$$

für ein $\vartheta' \in \Theta$.

- Erwartungstreue ist nicht invariant unter Parametertransformationen: Ist T erwartungstreu für ϑ, so ist i. a. $\gamma(T)$ nicht erwartungstreu für $\gamma(\vartheta)$.

Schon im Satz von Rao und Blackwell haben wir gesehen, dass wir einen erwartungstreuen Schätzer bei quadratischer Verlustfunktion durch Bedingen auf eine suffiziente Statistik verbessern können. In den nächsten beiden Sätzen zeigen wir, dass ein solcher Schätzer sogar optimal und eindeutig ist, wenn die Statistik zudem noch vollständig ist.

Satz 3.25 *(Lehmann-Scheffé)*
Es sei T eine suffiziente Statistik für die Familie $\{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ und

$$g : \mathcal{X} \to \mathbb{R}^d$$

ein erwartungstreuer Schätzer für $\gamma(\vartheta)$

$$\gamma : \Theta \to \mathbb{R}^d.$$

Definiere

$$g^*(t) = \mathbb{E}[g(X)|T(X) = t]. \tag{3.3}$$

Ist T suffizient, so ist $g^* \circ T$ ein gültiger Schätzer. Ist T vollständig, so ist $g^* \circ T$ ein UMVU-Schätzer.

Beweis: Es sei g ein erwartungstreuer Schätzer für $\gamma(\vartheta)$ und g^* gebildet wie in (3.3). Es sei

$$h : \mathcal{X} \to \mathbb{R}^d$$

ein anderer erwartungstreuer Schätzer für $\gamma(\vartheta)$. Es ist somit zu zeigen, dass

$$\mathbb{V}_\vartheta g^*(T(X)) \leq \mathbb{V}_\vartheta h(X) \tag{3.4}$$

für alle $\vartheta \in \Theta$ gilt. Wenn wir

$$h^*(t) = \mathbb{E}_\vartheta[h(T(X))|T(X) = t]$$

setzen, so ist aufgrund der Glättungseigenschaft der bedingten Erwartung auch h^* erwartungstreu und es gilt aufgrund der Rao-Blackwell-Ungleichung

$$\mathbb{V}_\vartheta h^*(T(X)) \leq \mathbb{V}_\vartheta h(X) \quad \forall\, \vartheta \in \Theta.$$

Wir müssen (3.4) also nur für h^* überprüfen. Da $h^* \circ T$ erwartungstreu ist, folgt nun

$$\mathbb{E}_\vartheta h^* \circ T = \gamma(\vartheta) = \mathbb{E}_\vartheta g^* \circ T,$$

also

$$\mathbb{E}_\vartheta (h^* - g^*) \circ T = 0$$

für alle $\vartheta \in \Theta$. Da T als vollständig vorausgesetzt war, erhalten wir

$$h^* - g^* = 0 \quad \mathbb{P}_\vartheta^T\text{-f.s. für alle} \quad \vartheta \in \Theta,$$

also auch

$$\mathbb{V}_\vartheta\, g^* \circ T(X) = \mathbb{V}_\vartheta\, h^* \circ T(X) \leq \mathbb{V}_\vartheta h(X)$$

für alle $\vartheta \in \Theta$. Also ist g^* ein UMVU. □

Korollar 3.26 *Es sei*

$$T : (\mathcal{X}, \mathcal{A}) \to (\mathcal{X}', \mathcal{A}')$$

eine vollständige, suffiziente Statistik für die Familie $\{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$. *Ist dann* g *ein UMVU-Schätzer für* $\gamma(\vartheta)$, *so ist dieses* $\mathbb{P}_\vartheta$*-f.s. eindeutig für alle* $\vartheta \in \Theta$ *mit* $\mathbb{V}_\vartheta g(X) < +\infty$.

Beweis: Seien h und g zwei UMVU-Schätzer für $\gamma(\vartheta)$. Da man durch Bedingen von h und g auf T höchstens bessere Schätzer erhält, folgt für

$$g^* \circ T = \mathbb{E}[g|T] \quad \text{und} \quad h^* = \mathbb{E}[h|T],$$

dass

$$g = g^* \circ T \quad \text{und} \quad h = h^* \circ T \quad \mathbb{P}_\vartheta\text{-f.s.} \quad \forall\, \vartheta \in \Theta$$

gilt. Nun folgt wie oben

$$\mathbb{E}_\vartheta[(g^* - h^*) \circ T] = 0 \quad \forall\, \vartheta \in \Theta,$$

und aus der Vollständigkeit von T folgt

$$h = g \quad \mathbb{P}_\vartheta\text{-f.s.} \quad \forall\, \vartheta \in \Theta.$$

□

Der Satz von Lehmann-Scheffé hilft uns nun, einen UMVU-Schätzer zu konstruieren. Hierzu können wir entweder

a) intelligent raten und einen erwartungstreuen Schätzer angeben, der nur von einer suffizienten, vollständigen Statistik abhängt;

b) rechnen, indem wir einen beliebigen erwartungstreuen Schätzer auf eine vollständige und suffiziente Statistik bedingen.

Wir betrachten Beispiele.

Beispiel 3.27 *a) Es seien $X_1, \ldots, X_n$ i.i.d. Poi(λ)-verteilt und*

$$\gamma(\lambda) = \mathbb{P}[X_1 = 0] = e^{-\lambda}.$$

Der Schätzer

$$T(x_1, \ldots, x_n) = \mathbb{1}_{\{x_1=0\}}$$

ist erwartungstreu für $\gamma(\lambda)$, denn

$$\mathbb{E}_\lambda T = \mathbb{P}_\lambda[X_1 = 0] = e^{-\lambda} \quad \textit{für alle} \quad \lambda \in \mathbb{R}^+.$$

Wir wissen, dass die Statistik

$$S = \sum_{i=1}^{n} X_i$$

suffizient ist für die Familie

$$\{Poi^n(\lambda) : \lambda \in \mathbb{R}^+\}.$$

Also können wir T durch Bedingen auf S verbessern.

$$\begin{aligned} T'(s) &:= \mathbb{E}[T|S = s] = \mathbb{P}[X_1 = 0|S = s] \\ &= \frac{\mathbb{P}[X_1 = 0, S = s]}{\mathbb{P}(S = s)} = \frac{\mathbb{P}_\lambda(X_1 = 0, S = s)}{\mathbb{P}_\lambda(S = s)} \\ &= \frac{e^{-\lambda}\mathbb{P}_\lambda(\sum_{i=2}^n X_i = s)}{\mathbb{P}_\lambda(\sum_{i=1}^n X_i = s)} = \frac{e^{-\lambda}(e^{-(n-1)\lambda}\lambda^s(n-1)^s/s!}{e^{-n\lambda}\lambda^s n^s/s!} \\ &= (1 - \frac{1}{n})^s. \end{aligned}$$

Hierbei haben wir die Unabhängigkeit der X_i verwendet, sowie die Tatsache, dass $\sum_{i=1}^n X_i \sim Poi(n\lambda)$-verteilt ist. Wenn wir nun noch zeigen können, dass S auch vollständig ist, so ist T' ein UMVU-Schätzer für $\gamma(\lambda)$. Dazu nehmen wir an, dass für eine messbare Funktion

$$f : \mathbb{R} \to \mathbb{R}$$

gilt

$$\mathbb{E}_\lambda[f(S)] = e^{-\lambda n} \sum_{n=0}^{\infty} f(k) \frac{(n\lambda)^k}{k!} = 0$$

für alle $\lambda \in \mathbb{R}^+$. Offenbar ist das eine Potenzreihe in λ. Diese kann nur identisch in λ verschwinden, wenn alle Koeffizienten 0 sind, dies bedeutet, wenn

$$f(k) = 0 \quad \textit{für alle} \quad k \in \mathbb{N}_0$$

gilt. Also ist

$$T' = (1 - \frac{1}{n})^{X_1 + \ldots + X_n}$$

ein UMVU-Schätzer für $\gamma(\lambda) = e^{-\lambda}$.

b) Es sei $X = (X_1, \ldots, X_n)$ ein Vektor mit i.i.d. Komponenten, X_i sei $\hat{\mathbb{P}}_\vartheta$-verteilt für alle i und

$$\mathbb{P}_\vartheta = \bigotimes_{i=1}^{n} \hat{\mathbb{P}}_\vartheta.$$

Ferner sei $\hat{\mathbb{P}}_\vartheta$ eine ein-parametrige Exponentialfamilie in $Q(\vartheta)$ und $\hat{T}(x) = x$, d. h.

$$\frac{d\hat{\mathbb{P}}_\vartheta}{d\nu} = C(\vartheta)e^{Q(\vartheta)\cdot x} \quad \nu\text{-f.s.}$$

für ein dominierendes Maß ν. Da dann

$$\frac{d\mathbb{P}_\vartheta}{d\bigotimes_{i=1}^{n}\nu}(x) = \prod_{j=1}^{n} \frac{d\hat{\mathbb{P}}_\nu}{d\nu} = C^n(\nu)\prod_{j=1}^{n} e^{Q(\vartheta)x_j} = C^n(\vartheta)e^{Q(\vartheta)\sum_{j=1}^{n} x_j}$$

gilt, ist auch die Familie der $\{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ eine einparametrige Exponentialfamilie in Q und

$$T = \sum_{j=1}^{n} x_j =: s_n.$$

Daher ist $T(x) = s_n$ suffizient für $\{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$. Man kann auch zeigen, dass T vollständig ist (Lehmann: "Testing Statistical Hypothesis", Kapitel 4.3), falls die Menge

$$\{Q(\vartheta) : \vartheta \in \Theta\}$$

innere Punkte besitzt. In diesem Fall wissen wir, dass

$$g(x) = \frac{1}{n}T(x) =: \bar{x}_n$$

ein erwartungstreuer Schätzer für $\gamma(\vartheta) = \mathbb{E}_\vartheta X_1$ ist. Da g nur von T abhängt, folgt die Optimalität. Dies lässt sich auf viele Spezialfälle anwenden, z. B. Bernoulli- oder Binomialverteilungen zu unbekanntem $p \in (0,1)$, Poisson-Verteilungen zu unbekanntem $\lambda \in \mathbb{R}^+$ oder $\mathcal{N}(\mu, \sigma^2)$-Verteilungen bei festem σ^2 und unbekanntem μ.

c) Seien $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$-verteilt und $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ seien unbekannt. Diese Verteilungen bilden daher eine zweiparametrige Exponentialfamilie in $Q = (Q_1, Q_2)$ mit

$$Q_1(\mu, \sigma^2) = -\frac{1}{2\sigma^2} \quad \textit{und} \quad Q_2(\mu, \sigma^2) = \frac{\mu}{\sigma^2}$$

und $T(x) = (T_1(x), T_2(x))$ mit

$$T_1(x) = \sum_{i=1}^{n} x^2 \quad \textit{und} \quad T_2(x) = \sum_{i=1}^{n} x.$$

Die Statistik

$$S(x) = \left(\sum_{i=1}^{n} x_1^2,\ \sum_{i=1}^{n} x_1\right)$$

ist daher suffizient für

$$\{\bigotimes_{i=1}^{n}\mathcal{N}(\mu,\sigma^2) : \mu\in\mathbb{R},\sigma^2>0\}.$$

Da $\mathbb{R}\times\mathbb{R}^+$ innere Punkte besitzt, ist S auch vollständig. Schon in der Stochastik haben wir gesehen, dass

$$g_1(x)=\frac{1}{n}\sum_{i=1}^{n}x_i \quad \textit{und} \quad g_2(x)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-g_1(x))^2$$

ein erwartungstreues Schätzerpaar für μ und σ^2 bildet. Da diese nur von S abhängen, folgt ihre Optimalität.

Abschließend zeigen wir noch, dass UMVU-Schätzer allgemein nicht besonders gut sein müssen: In einigen Fällen sind sie noch nicht einmal zulässig. Erstaunlicherweise muss man hierfür nicht auf besonders exotische Beispiele zurückgreifen. Zunächst beweist man:

Lemma 3.28 *Es seien $X_1,\dots,X_n$ i.i.d. Zuvallsvariablen mit $\mathbb{V}(X_1)=\sigma^2$ und*

$$\mathbb{E}[(X_1-\mathbb{E}X_1)^4]=:\mu_4<+\infty.$$

Dann gilt für

$$\begin{aligned}\bar{X}_n : &= \frac{1}{n}\sum_{i=1}^{n}X_i\\ \mathbb{E}(\sum_{k=1}^{n}(X_k-\bar{X}_n)^2)^2 &= \frac{(n-1)^2}{n}\mu_4+\frac{(n-1)(n(n-2)+3)}{n}\sigma^4.\end{aligned}$$

Beweis: Das rechnet man einfach nach (siehe z. B. Alsmeyer: "Mathematische Statistik", S. 56/57). □

Mithilfe dieses Lemmas lässt sich nun das Risiko eines Schätzers für die Varianz berechnen. Genauer seien $X_1,\dots,X_n,\ n\geq 2$ i.i.d. und $\mathcal{N}(\mu,\sigma^2)$-verteilt, wobei $\mu\in\mathbb{R}$ und $\sigma^2>0$ unbekannt sind. Setze

$$\hat{\sigma}^2_{n,c}=\frac{1}{c}\sum_{k=1}^{n}(X_k-\bar{X}_n)^2.$$

Hierbei sei $c>0$. Für $c=n$ erhält man den Maximum-Likelihood-Schätzer (und Momentenschätzer) für σ^2, für $c=n-1$ den UMVU-Schätzer. Erstaunlicherweise sind beide nicht zulässig.

Satz 3.29 *Unter den obigen Voraussetzungen gilt für das Risiko $R(\vartheta,\cdot)$ bei quadratischer Verlustfunktion*

$$\mathcal{R}(\vartheta,\hat{\sigma}^2_{n,c})=\sigma^4\left((n^2-1)\left(\frac{1}{c}-\frac{1}{n+1}\right)^2+\frac{2}{n+1}\right).$$

Es wird minimiert für $c = n + 1$. Es gilt

$$(1 + \frac{2}{n-1})\frac{2\sigma^4}{n+1} = \mathcal{R}(\vartheta, \hat{\sigma}_{n,n-1}) > R(\vartheta, \hat{\sigma}^2_{n,n}) > R(\vartheta, \sigma^2_{n,n+1}) = \frac{2\sigma^4}{n+1}$$

für alle $\vartheta = (\mu, \sigma^2)$.

Beweis: Schätzer und Varianz ändern sich nicht, wenn wir $\mu = 0$ annehmen, also

$$R((\mu, \sigma^2), \hat{\sigma}^2_{n,c}) = R((0, \sigma^2), \sigma^2_{n,c}).$$

Ferner gilt wegen

$$\mathbb{E}_{\mu,\sigma^2}\hat{\sigma}^2_{n,c} = \frac{n-1}{c}\sigma^2$$

(was man aus der Tatsache gewinnt, dass $\sum_{i=1}^n \left(\frac{X_i - \bar{X}_n}{\sigma}\right)^2$ χ^2_{n-1}-verteilt ist, also Erwartungswert $n-1$ hat) für alle $c > 0$

$$R((0, \sigma^2), \sigma^2_{n,c}) = \mathbb{E}_{(0,\sigma^2)}(\hat{\sigma}^2_{n,c} - \sigma^2)^2 = \mathbb{E}_{(0,\sigma^2)}\hat{\sigma}^4_{n,c} - \frac{2(n-1)}{c}\sigma^4 + \sigma^4.$$

Beachtet man, dass

$$\mathbb{E}_{(0,\sigma^2)}\sigma^4_{n,c} = \frac{1}{c^2}\mathbb{E}_{(0,\sigma^2)}\left(\sum_{k=1}^n (X_k - \bar{X}_n)^2\right)^2$$

und

$$\mathbb{E}_{(0,\sigma^2)}X_1^4 = 3\sigma^4,$$

so folgt die Aussage aus dem letzten Lemma. □

3.4 Die Cramér-Rao-Ungleichung

Das Risiko eines erwartungstreuen Schätzers ist bei quadratischer Verlustfunktion durch seine Varianz gegeben. Wir leiten in diesem Abschnitt eine untere Schranke für die Varianz $\mathbb{V}_\vartheta(T)$ eines erwartungstreuen Schätzers T her. Finden wir also einen Schätzer, für den diese untere Schranke angenommen wird, haben wir automatisch einen UMVU-Schätzer gefunden.

Um die folgenden Operationen auch durchführen zu können, benötigen wir ein paar Annahmen. Gegeben sei ein messbarer Raum $(\mathcal{X}, \mathcal{A})$ und eine Familie von Wahrscheinlichkeitsmaßen $\{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ über $(\mathcal{X}, \mathcal{A})$. Wir sagen, dass $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein reguläres statistisches Experiment ist, falls

1. $\Theta \subseteq \mathbb{R}$ ein offenes Intervall ist;
2. $A = \{x : f_\vartheta(x) = \frac{d\mathbb{P}_\vartheta}{d\mu}(x) > 0\}$ nicht von ϑ abhängt (dabei ist μ ein die Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ dominierendes Maß);

3. Für alle $x \in A$ und alle $\vartheta \in \Theta$ existiert

$$f'_\vartheta(x) := \frac{\partial f_\vartheta(x)}{d\vartheta},$$

ist endlich und stetig.

Um die gewünschte Ungleichung abzuleiten, starten wir mit der folgenden Beobachtung: Ist

$$\psi : \Theta \times \mathcal{X} \to \mathbb{R}$$

eine beliebige Funktion mit

$$0 \leq \mathbb{V}_\vartheta[\psi(\vartheta, X)] < +\infty,$$

so folgt aus der Cauchy-Schwarz-Ungleichung

$$\mathbb{V}_\vartheta(T) \geq \frac{Cov_\vartheta(T, \psi(\vartheta, X))^2}{\mathbb{V}_\vartheta(\psi(\vartheta, X))}$$

für jeden Schätzer T. Kann man ψ so wählen, dass

$$Cov_\vartheta(T, \psi(\vartheta, X))$$

unabhängig von T wird, so hat man eine untere Schranke für $\mathbb{V}_\vartheta(T)$ gefunden, also auch für das quadratische Risiko von T. Dies ist – wie wir sehen werden – der Fall, wenn wir

$$\psi(\vartheta, x) = \frac{f_{\vartheta+\Delta}(x) - f_\vartheta(x)}{\Delta \cdot f_\vartheta(x)}$$

für ein $\Delta > 0$ setzen, oder den Grenzwert $\Delta \to 0$ bilden:

$$\psi(\vartheta, x) = \frac{\frac{\partial}{\partial\vartheta} f_\vartheta(x)}{f_\vartheta(x)} = \frac{\partial}{\partial\vartheta} \log f_\vartheta(x).$$

Genauer beweisen wir:

Satz 3.30 *Unter den obigen Regularitätsannahmen gilt für jeden erwartungstreuen Schätzer T der Parameterfunktion*

$$\gamma : \Theta \to \mathbb{R}$$

die Chapman-Robbins-Ungleichung

$$\mathbb{V}_\vartheta(T) \geq \sup_{\Delta>0} \frac{(\gamma(\vartheta+\Delta) - \gamma(\vartheta))^2}{\mathbb{V}_\vartheta(\frac{f_{\vartheta+\Delta}(X) - f_\vartheta(X)}{f_\vartheta(X)})}.$$

Falls auch

$$\frac{1}{\Delta}\frac{f_{\vartheta*1} - f_\vartheta}{\Delta f_\vartheta} \to \frac{\partial}{\partial\vartheta} \log f_\vartheta(x) \quad \textit{für} \quad \Delta \to 0$$

in $L^2(\mathbb{P}_\vartheta)$ konvergiert, so ist $\gamma(\vartheta)$ differenzierbar, es gilt

$$\mathbb{E}_\vartheta[\frac{\partial}{\partial\vartheta} \log f_\vartheta(X)] = 0$$

und die Cramér-Rao-Ungleichung

$$\mathbb{V}_\vartheta(T) \geq \frac{(\gamma'(\vartheta))^2}{I(\vartheta)}.$$

Hierbei ist

$$\begin{aligned} I(\vartheta) &= \mathbb{E}_\vartheta\left[\left(\frac{\partial}{\partial\vartheta}\log f_\vartheta(X)\right)^2\right] \\ &= \mathbb{V}_\vartheta\left[\frac{\partial}{\partial\vartheta}\log f_\vartheta(X)\right] \end{aligned}$$

die Fisher-Information.

Beweis: Da Δ nicht von ϑ abhängt, ist für

$$\psi(\vartheta, x) = \frac{f_{\vartheta+\Delta}(x) - f_\vartheta(x)}{\Delta \cdot f_\vartheta(x)}$$

der Erwartungswert

$$\begin{aligned} \mathbb{E}_\vartheta[\psi(\vartheta, X)] &= \int_{\mathcal{X}} \frac{f_{\vartheta+\Delta}(x) - f_\vartheta(x)}{\Delta f_\vartheta(x)}\, \mathbb{P}_\vartheta(dx) \\ &= \int_{\mathcal{X}} \frac{f_{\vartheta+\Delta}(x) - f_\vartheta(x)}{\Delta f_\vartheta(x)} f_\vartheta(x)\mu(dx) \\ &= \int \frac{f_{\vartheta+\Delta}(x)}{\Delta} - \frac{f_\vartheta(x)}{\Delta}\mu(dx) \\ &= \frac{1}{\Delta} - \frac{1}{\Delta} = 0. \end{aligned}$$

Analog rechnet man nach, dass

$$\begin{aligned} Cov_\vartheta[T, \psi(\vartheta, X)] &= \mathbb{E}_\vartheta[T(X)\frac{f_{\vartheta+\Delta}(X) - f_\vartheta(X)}{\Delta \cdot f_\vartheta(X)}] \\ &= \int \frac{T(x)}{\Delta}(f_{\vartheta+\Delta}(x) - f_\vartheta(x))\mu(dx) \\ &= \frac{\gamma(\vartheta+\Delta) - \gamma(\vartheta)}{\Delta}. \end{aligned}$$

Somit erhalten wir aus der Cauchy-Schwarz-Ungleichung:

$$\mathbb{V}_\vartheta(T) \geq \frac{(\gamma(\vartheta+\Delta) - \gamma(\vartheta))^2}{\mathbb{V}_\vartheta[\frac{f_{\vartheta+\Delta}(X) - f_\vartheta(X)}{f_\vartheta(X)}]}.$$

Da dies für alle $\Delta > 0$ gilt, folgt die Chapman-Robbins-Ungleichung.

Zur Herleitung der Cramér-Rao-Ungleichung benutzt man wieder die Cauchy-Schwarz-Ungleichung. Für jedes $U \in L^2(\mathbb{P}_\vartheta)$ impliziert diese ja für jedes $\Delta > 0$

$$\begin{aligned} &\left(\mathbb{E}_\vartheta\left(U \cdot \frac{f_{\vartheta+\Delta}(X) - f_\vartheta(X)}{\Delta \cdot f_\vartheta(X)}\right) - \mathbb{E}_\vartheta\left(U \cdot \frac{\partial}{\partial\vartheta}\log f_\vartheta(X)\right)\right)^2 \qquad (3.5) \\ \leq\ &\mathbb{E}_\vartheta[U^2]\, \mathbb{E}_\vartheta\left[\left(\frac{f_{\vartheta+\Delta}(X) - f_\vartheta(X)}{\Delta f_\vartheta(X)} - \frac{\partial}{\partial\vartheta}\log f_\vartheta(X)\right)^2\right]. \end{aligned}$$

Nun konvergiert nach Voraussetzung für $\Delta \to 0$

$$\frac{f_{\vartheta+\Delta}(x) - f_\vartheta(x)}{\Delta f_\vartheta(x)} \to \frac{\partial}{\partial\vartheta} \log f_\vartheta(x) \text{ in } L^2(\mathbb{P}_\vartheta).$$

Somit konvergiert die rechte Seite von (3.5) gegen 0, also auch die linke. Setzt man nun $U = 1$, so ist der erste Summand auf der linken Seite von (3.5) gleich 0. Somit folgt

$$\mathbb{E}_\vartheta \left[\frac{\partial}{\partial\vartheta} \log f_\vartheta(X) \right] = 0.$$

Wählt man hingegen $U = T(X)$, so ist der erste Summand auf der linken Seite von (3.5)

$$\frac{\gamma(\vartheta+\Delta) - \gamma(\vartheta)}{\Delta}.$$

Die Konvergenz dieses Ausdrucks für $\Delta \to 0$ ist mithin die Differenzierbarkeit von γ. Darüber hinaus bekommen wir eben aus (3.5)

$$\begin{aligned} \gamma'(\vartheta) &= \mathbb{E}_\vartheta[T(X) \frac{\partial}{\partial\vartheta} \log f_\vartheta(X)] \\ &= Cov_\vartheta(T, \frac{\partial}{\partial\vartheta} \log f_\vartheta(X)]. \end{aligned}$$

Die Cauchy-Schwarz-Ungleichung ergibt also

$$\begin{aligned} \mathbb{V}_\vartheta(T) &\geq \frac{(Cov(T, \frac{\partial}{\partial\vartheta} \log f_\vartheta(X)))^2}{\mathbb{V}(\frac{\partial}{\partial\vartheta} \log f_\vartheta(X))} \\ &= \frac{(\gamma'(\vartheta))^2}{I(\vartheta)}. \end{aligned}$$

□

Bemerkung 3.31 *Die Chapman-Robbins-Schranke ist zwar i. a. schärfer als die Cramér-Rao-Schranke, aber auch schwieriger zu berechnen.*

Lemma 3.32 *a) Sei f_ϑ eine Dichte, die die Bedingungen aus Satz 3.30 erfüllt. Dann sind die Bedingungen auch erfüllt für*

$$f_\vartheta(\vec{x}) = f_\vartheta(x_1) \dots f_\vartheta(x_n)$$

(wobei $\vec{x} = (x_1, \dots, x_n)$ ist) und es gilt

$$I_n(\vartheta) = nI_1(\vartheta).$$

b) Unter stärkeren Regularitätsbedingungen als in Satz 3.30 gilt

$$I(\vartheta) = -\mathbb{E}_\vartheta \left[\frac{\partial^2}{\partial\vartheta^2} \log f_\vartheta(X) \right].$$

Beweis:

a) Dass
$$I_n(\vartheta) = n \cdot I_1(\vartheta)$$
gilt, folgt sofort aus der Produktgestalt von $f_\vartheta(\vec{x})$. Der Rest ist mühsames Rechnen, das wir uns hier sparen wollen.

b) Durch Differenzieren unter dem Integral von
$$\mathbb{E}_\vartheta \left[\frac{\partial}{\partial \vartheta} \log f_\vartheta(X) \right] = 0$$
folgt die Behauptung.

□

Beispiel 3.33 *Es seien $X_1, \ldots, X_n$ i.i.d. Poi(λ)-verteilte Zufallsvariablen. $\lambda > 0$ sei unbekannt. Für $n = 1$ gilt*
$$\log f_\lambda(x) = -\lambda + x \log \lambda - \log x!,$$
wobei f_λ die Zähldichte ist. Also ist
$$\frac{\partial}{\partial \lambda} \log f_\lambda(x) = -1 + \frac{x}{\lambda} \quad \textit{und} \quad \frac{\partial^2}{\partial \lambda^2} \log f_\lambda(x) = \frac{-x}{\lambda^2}.$$
Somit folgt
$$\begin{aligned} I_1(\lambda) &= \mathbb{E}_\lambda \left[-\frac{X}{\lambda^2} \right] \\ &= \sum_{k=1}^{\infty} \frac{k}{\lambda^2} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=1}^{\infty} \frac{\lambda^{k-2}}{(k-1)!} e^{-\lambda} \\ &= \frac{1}{\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = \frac{1}{\lambda}. \end{aligned}$$
Also ist
$$I_n(\lambda) = \frac{n}{\lambda}.$$
Ist nun $\gamma(\lambda) = \lambda$ zu schätzen, so ist
$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$
(der Maximum-Likelihood-Schätzer) und hat die Varianz
$$\mathbb{V}_\lambda(\bar{X}) = \frac{1}{n} \mathbb{V}_\lambda(X_1) = \frac{\lambda}{n} = \frac{1}{I_n(\lambda)}.$$

Also erreicht $\bar{X}$ die Cramér-Rao-Schranke, ist also ein UMVU-Schätzer für λ.

Betrachten wir die Parameterfunktion

$$\gamma(\lambda) = \mathbb{P}(X_1 = 0) = e^{-\lambda},$$

so ist (wie bereits gesehen)

$$T(X) = (1 - \frac{1}{n})^{n\bar{X}}$$

ein UMVU-Schätzer. Es gilt aber

$$\begin{aligned}
\mathbb{V}_\lambda(T) &= \mathbb{E}_\lambda(T^2) - (\mathbb{E}_\lambda(T))^2 \\
&= e^{-n\lambda} \sum_{k=0}^{\infty} (1 - \frac{1}{n})^{2k} \frac{(n\lambda)^k}{k!} - e^{-2\lambda} \\
&= e^{(1-\frac{1}{n})^2 n\lambda} e^{-n\lambda} - e^{-2\lambda} \\
&> e^{-2\lambda} \frac{\lambda}{n} \\
&= \frac{(\gamma'(\lambda))^2}{I_n(\lambda)}.
\end{aligned}$$

Obwohl also T ein UMVU-Schätzer ist, wird die Cramér-Rao-Schranke nicht angenommen.

Interessanterweise steht die Frage, ob ein Schätzer die Cramér-Rao-Schranke annimmt, in engem Zusammenhang zur Frage, ob das zugrundeliegende Modell die Struktur einer Exponentialfamilie besitzt. Genauer gilt

Satz 3.34 *Es sei $(\mathcal{X}, \mathcal{A})$ ein messbarer Raum und die Familie $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ sei regulär auf $(\mathcal{X}, \mathcal{A})$ im Sinne der eingangs gegebenen Definition. Ein erwartungstreuer Schätzer T der Parameterfunktion $\gamma(\vartheta)$ erreicht die Cramér-Rao-Schranke genau dann, wenn zwei differenzierbare Funktionen $c(\vartheta)$ und $d(\vartheta)$ existieren, so dass für das die Familie $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ dominierende Maß μ und eine messbare Funktion h*

$$\frac{d\mathbb{P}_\vartheta}{d\mu}(x) = f_\vartheta(x) = \exp(c(\vartheta)T(x) + d(\vartheta))h(x)$$

und

$$\gamma(\vartheta) = -\frac{d'(\vartheta)}{c'(\vartheta)}$$

gilt.

Beweis: Wir erinnern uns, dass der Beweis der Cramér-Rao-Ungleichung auf der Cauchy-Schwarz-Ungleichung

$$\mathbb{V}_\vartheta(T) \geq \frac{Cov(T, \psi(\vartheta, X))^2}{\mathbb{V}_\vartheta(\psi(\vartheta, X))}$$

mit

$$\psi(\vartheta, X) = \frac{\partial}{\partial \vartheta} \log f_\vartheta(X)$$

beruhte. Gleichheit gilt in der Cauchy-Schwarz-Ungleichung, wenn sich beide Seiten nur durch eine affin-lineare Transformation unterscheiden, wenn also $\mathbb{P}_\vartheta$-f.s. gilt

$$\frac{\partial}{\partial \vartheta} \log f_\vartheta(x) = a(\vartheta)T(x) + b(\vartheta)$$

für geeignete $a(\vartheta)$ und $b(\vartheta)$. Dies ist äquivalent zu

$$f_\vartheta(x) = \exp(c(\vartheta)T(x) + d(\vartheta))h(x) \quad \mathbb{P}_\vartheta\text{-f.s.}$$

Wollen wir dies aber auch $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$-fast sicher behaupten, haben wir das Problem, dass die Nullmenge

$$\{x : \frac{\partial}{\partial \vartheta} \log f_\vartheta(x) \neq a(\vartheta)T(x) + b(\vartheta)\}$$

von ϑ abhängt. Wir definieren daher

$$\mathcal{X}^* = \{x : \frac{\partial}{\partial \vartheta} \log f_\vartheta(x) = a(\vartheta)T(x) + b(\vartheta) \quad \forall\, \vartheta \in \Theta\}.$$

Wir betrachten nur den interessanten Fall, dass $\gamma(\vartheta)$ nicht konstant ist. Dann ist auch T nicht konstant. Also gibt es $x, y \in \mathcal{X}$ mit $T(x) \neq T(y)$. Somit lassen sich $a(\vartheta)$ und $b(\vartheta)$ als Lösung eines linearen Gleichungssystems

$$\begin{aligned} \frac{\partial}{\partial \vartheta} \log f_\vartheta(x) &= a(\vartheta)T(x) + b(\vartheta) \\ \frac{\partial}{\partial \vartheta} \log f_\vartheta(y) &= a(\vartheta)T(y) + b(\vartheta) \end{aligned}$$

gewinnen. Da alle beteiligten Größen messbar in ϑ sind, sind $a(\vartheta)$ und $b(\vartheta)$ auch messbar. Da zu den Annahmen der Regularität auch die Stetigkeit von $\frac{\partial}{\partial \vartheta} \log f_\vartheta$ in ϑ zählt, sind auch $a(\cdot)$ und $b(\cdot)$ stetig.

Wegen der paarweisen Äquivalenz der $\mathbb{P}_\vartheta$, die aus (2) der Regularitätsannahmen folgt, erhalten wir für alle $\vartheta, \tau \in \Theta$

$$\mathbb{P}_\vartheta\{x \in \mathcal{X} : \frac{\partial}{\partial \tau} \log f_\tau(x) = a(\tau)T(x) + b(\tau)\} = 1.$$

Sei nun $\Theta^* \subseteq \Theta$ eine beliebige abzählbare, dichte Teilmenge. Dann folgt einerseits

$$\mathbb{P}_\vartheta\{x \in \mathcal{X} : \frac{\partial}{\partial \vartheta} f_\vartheta(x) = a(\vartheta)T(x) + b(\vartheta)\} = 1$$

für alle $\vartheta \in \Theta$ und zum anderen, da alle beteiligten Funktionen stetig sind,

$$\mathcal{X}^* = \{x : \frac{\partial}{\partial \vartheta} \log f_\vartheta(x) = a(\vartheta)T(x) + b(\vartheta) \quad \forall\, \vartheta \in \Theta^*\}.$$

Es gilt

$$\mathbb{P}_\vartheta(\mathcal{X}^*) = 1 \quad \forall\, \vartheta \in \Theta.$$

Auf $\mathcal{X}^*$ aber bekommen wir für jedes feste $\vartheta_0 \in \Theta$

$$f_\vartheta(x) = \exp\left(\left(\int_{\vartheta_0}^{\vartheta} a(t)dt\right) T(x) + \left(\int_{\vartheta_0}^{\vartheta} b(t)dt\right)\right) f_{\vartheta_0}(x).$$

Setzen wir

$$c(\vartheta) = \int_{\vartheta_0}^{\vartheta} a(t)dt \quad \text{und} \quad d(\vartheta) = \int_{\vartheta_0}^{\vartheta} b(t)dt \quad \text{und} \quad f_{\vartheta_0}(x) = h(x),$$

so folgt die eine Richtung.

Ist nun f_ϑ umgekehrt von der Form

$$f_\vartheta(x) = e^{c(\vartheta)T(x)+d(\vartheta)}h(x),$$

so ist

$$\frac{\partial}{\partial\vartheta}\log f_\vartheta(x) = c'(\vartheta)T(x) + d'(\vartheta).$$

Nach Satz 3.30 ist dies null im Erwartungswert, also

$$\mathbb{E}_\vartheta[c'(\vartheta)T(x) + d'(\vartheta)] = 0 \quad \text{für alle} \quad \vartheta \in \Theta.$$

Dies bedeutet

$$\mathbb{E}_\vartheta[T(x)] = -\frac{d'(\vartheta)}{c'(\vartheta)} \quad \text{für alle} \quad \vartheta \in \Theta.$$

Somit ist T erwartungstreu für

$$\gamma(\vartheta) = -\frac{d'(\vartheta)}{c'(\vartheta)}.$$

Da $\frac{\partial}{\partial\vartheta}\log f_\vartheta(x)$ und $T(x)$ affin-linear abhängig sind, nimmt T auch die Cramér-Rao-Schranke an. □

Wir wollen nun noch kurz auf eine mehrdimensionale Erweiterung des Satzes von Cramér und Rao eingehen. Sei nun $\vartheta \in \Theta \subseteq \mathbb{R}^d$, aber noch $\gamma(\vartheta) \in \mathbb{R}$ zu schätzen. Für die Cramér-Rao-Ungleichung wählen wir

$$\psi(\vartheta, x) = \sum_{i=1}^{d} a_i \frac{\partial}{\partial\vartheta_i}\log f_\vartheta(x)$$

mit zunächst beliebigen $a_i \in \mathbb{R}$.

Ähnlich wie in Satz 3.30 erhält man aus der Cauchy-Schwarz-Ungleichung, dass für jeden erwartungstreuen Schätzer T von $\gamma(\vartheta)$ gilt

$$\mathbb{V}_\vartheta[T] \geq \frac{(\sum_{i=1}^{d} a_i \frac{\partial}{\partial\vartheta_i}\gamma(\vartheta))^2}{\sum_{i,j}^{d} a_i a_j (I(\vartheta))_{i,j}}.$$

Hierbei ist $I(\vartheta)$ die sogenannte Fisher-Informationsmatrix, definiert als

$$(I(\vartheta))_{i,j} = \mathbb{E}_\vartheta[\frac{\partial}{\partial\vartheta_i}\log f_\vartheta(X)\frac{\partial}{\partial\vartheta_j}\log f_\vartheta(X)] = -\mathbb{E}_\vartheta[\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}\log f_\vartheta(X)].$$

Die mehrdimensionale Cramér-Rao-Ungleichung erhalten wir, indem wir diese Ungleichung in den a_i optimieren.

Satz 3.35 *Sei $\mathbb{P}_\vartheta(dx) = f_\vartheta(x)\mu(dx)$ für alle $\vartheta \in \Theta \subseteq \mathbb{R}^d$ offen und sei T erwartungstreu für $\gamma(\vartheta) \in \mathbb{R}$. Unter Regularitätsbedingungen gilt*

$$\mathbb{V}_\vartheta[T] \geq \left(\frac{\partial}{\partial\vartheta}\gamma(\vartheta)\right)^T (I(\vartheta))^{-1} \left(\frac{\partial}{\partial\vartheta}\gamma(\vartheta)\right).$$

Beweis: Sei V eine positiv definite $d \times d$-Matrix und $c \in \mathbb{R}^d$. Mithilfe von Lagrange-Multiplikatoren sieht man, dass $a^T c$ maximal unter der Nebenbedingung $a^T V a = 1$ ist, wenn $a = const.V^{-1}c$ gilt. Dies wendet man auf $V = I(\vartheta)$ und $c = \frac{\partial}{\partial\vartheta}\gamma(\vartheta)$ an. □

4 Testtheorie

4.1 Einführung und das Neyman-Pearson-Lemma

Hier nehmen wir einen etwas anderen Standpunkt ein. Es kann passieren, beispielsweise im Fall von n unabhängigen Bernoulli-Variablen zum Parameter $p \in (0,1)$, dass der bestmögliche erwartungstreue Schätzer (in diesem Fall $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$) mit Wahrscheinlichkeit 1 nicht den wahren Wert liefert (z. B., wenn $p \in \mathbb{R}\backslash\mathbb{Q}$ ist). Hier geht es eher darum, Hypothesen über den unbekannten Parameter zu verifizieren oder abzulehnen.

Es sei also $(\mathcal{X}, \mathcal{A})$ ein messbarer Raum und $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen. Auf der Basis einer Stichprobe $X_1, \ldots, X_n$, die i.i.d. gemäß $\mathbb{P}_\vartheta$ gezogen wird, wollen wir entscheiden, ob die Hypothese

$$\vartheta \in H \subseteq \Theta \quad \text{oder die Alternative} \quad \vartheta \in K := \Theta\backslash H$$

vorliegt. Offenbar gibt es zwei Möglichkeiten, einen Fehler zu machen:
Fehler 1. Art: Verwerfe H, wenn H vorliegt;
Fehler 2. Art: Nehme H an, obwohl K vorliegt.
Es wird dabei ein Test gesucht, dessen Wahrscheinlichkeit für einen Fehler 1. Art unterhalb eines gegebenen Signifikanzniveaus $\alpha \in [0,1]$ liegt. Hierbei definieren wir "Test" folgendermaßen:

Definition 4.1 *Jede messbare Funktion*

$$\varphi : \mathcal{X} \to [0,1]$$

heißt Test *für das oben beschriebene Testproblem. Nimmt ein Test nur die Werte 0 und 1 an, so heißt der Test* nicht-randomisiert*, anderenfalls heißt er* randomisiert*. Ähnlich wie bei den Schätzproblemen lässt sich die Klasse der Testprobleme mithilfe einer Verlust- und einer zugehörigen Risikofunktion modellieren.*

Definition 4.2 *Die* Neyman-Pearson-Verlustfunktion *ist die Funktion*

$$L(\vartheta, \gamma) = \begin{cases} \gamma & \vartheta \in H \\ 1-\gamma & \vartheta \in K \end{cases}$$

für alle $\gamma \in [0,1]$. Speziell gilt für nicht-randomisierte Tests

$$L(\vartheta, 0) = \begin{cases} 0 & \vartheta \in H \\ 1 & \vartheta \in K \end{cases} \quad \textit{und} \quad L(\vartheta, 1) = \begin{cases} 1 & \vartheta \in K \\ 0 & \vartheta \in H \end{cases}.$$

Dies ergibt die Risikofunktion

$$\mathcal{R}(\vartheta, \varphi) = \begin{cases} \int \varphi d\mathbb{P}_\vartheta = \mathbb{E}_\vartheta \varphi(X), & \vartheta \in H \\ \int (1-\varphi) d\mathbb{P}_\vartheta = 1 - \mathbb{E}_\vartheta \varphi(X), & \vartheta \in K \end{cases}.$$

Die Neyman-Pearson-Verlustfunktion ist sinnvoll, da wir das Ergebnis $\varphi(x) = \gamma$ des Tests φ so interpretieren wollen, dass φ sich bei Beobachtung von x mit Wahrscheilichkeit γ für K entscheidet. Offenbar ist $\mathbb{E}_\vartheta[\varphi(X)]$ bei dieser Beschreibung eine wichtige Größe.

Definition 4.3 *Die Funktion*

$$\beta_\varphi : \vartheta \mapsto \mathbb{E}_\vartheta[\varphi(X)]$$

nennt man Gütefunktion *des Tests φ.*

Offenbar beschreibt $\beta_\varphi(\vartheta)$ für $\vartheta \in H$ die Wahrscheinlichkeit eines Fehlers 1. Art. Für $\vartheta \in K$ ist $1 - \beta_\varphi(\vartheta)$ die Wahrscheinlichkeit eines Fehlers 2. Art. Wir werden von nun an an Tests zum Niveau $\alpha \in [0, 1]$ interessiert sein, d. h. solchen Tests, die

$$\mathbb{E}_\vartheta \, \varphi(X) \leq \alpha \quad \forall \, \vartheta \in H$$

erfüllen. Für solche Tests wollen wir den Fehler 2. Art minimieren.

Definition 4.4 *φ heißt gleichmäßig bester Test zum Niveau $\alpha \in [0, 1]$, falls er unter allen Tests zum Niveau α den Fehler 2. Art minimiert, d. h. falls*

$$\mathbb{E}_\vartheta \, \varphi(X) = \max_{\psi \in \Phi_\alpha} \mathbb{E}_\vartheta \, \psi(X)$$

für alle $\vartheta \in K$ gilt. Hierbei ist

$$\Phi_\alpha := \{\psi : \mathcal{X} \to [0, 1] | \mathbb{E}_\vartheta \, \psi \leq \alpha \text{ für alle } \vartheta \in H\}$$

die Menge aller Tests zum Niveau α.

Grundlegend für die Konstruktion solcher Tests ist das folgende Resultat, das die Situation im einfachsten Falle klärt, in dem sowohl H als auch K nur aus einem Punkt bestehen.

Satz 4.5 *(Neyman-Pearson-Lemma)*
Es seien $\mathbb{P}_0$ und $\mathbb{P}_1$ zwei Wahrscheinlichkeitsmaße auf $(\mathcal{X}, \mathcal{A})$ mit Dichten f_0 bzw. f_1 bzgl. eines σ-endlichen dominierenden Maßes μ (man kann stets $\mu = \mathbb{P}_1 + \mathbb{P}_2$ wählen). Ferner sei $\alpha \in (0, 1)$. Dann gilt:

a) Ist $\psi \in \Phi_\alpha$ ein Test, der

$$\int \psi \, d\mathbb{P}_0 = \alpha$$

erfüllt und

$$\psi(x) = \begin{cases} 1, & \text{falls } f_1(x) > k \cdot f_0(x) \\ 0, & \text{falls } f_1(x) < k \cdot f_0(x) \end{cases} \tag{4.1}$$

μ-fast sicher für ein $k \in [0, \infty]$, dann gilt

$$\int \psi \, d\mathbb{P}_1 = \max_{\varphi \in \Phi_\alpha} \int \varphi \, d\mathbb{P}_1. \tag{4.2}$$

b) Es gibt einen Test ψ wie unter a) beschrieben.

c) Gilt ψ für (4.2), so existiert ein $k \in [0, \infty]$, so dass (4.1) gilt. Gilt zudem

$$\int \psi \, d\mathbb{P}_1 < 1,$$

so erfüllt ψ auch $\int \psi \, d\mathbb{P}_0 = \alpha$.

Beweis:

a) Sei $\varphi \in \Phi_\alpha$. Es gilt nun

$$\begin{aligned} f_1(x) - k f_0(x) > 0 \quad &\Rightarrow \quad \psi(x) = 1 \Rightarrow \psi(x) - \varphi(x) \geq 0 \quad \text{und} \\ f_1(x) - k f_0(x) < 0 \quad &\Rightarrow \quad \psi(x) = 0 \Rightarrow \psi(x) - \varphi(x) \leq 0. \end{aligned}$$

Also gilt μ-f.s.

$$(\psi(x) - \varphi(x))(f_1(x) - k f_0(x)) \geq 0.$$

Integriert man dies, so ergibt sich

$$\int \psi f_1 d\mu - \int \varphi f_1 d\mu - k(\int \psi f_0 d\mu - \int \varphi f_0 d\mu) \geq 0.$$

Also

$$\int \psi d\mathbb{P}_1 - \int \varphi d\mathbb{P}_1 \geq k(\int \psi d\mathbb{P}_0 - \int \varphi d\mathbb{P}_0).$$

Da

$$\mathbb{E}_{\mathbb{P}_0} \psi = \alpha \quad \text{und} \quad \mathbb{E}_{\mathbb{P}_0} \varphi \leq \alpha,$$

folgt

$$\int \psi d\mathbb{P}_1 - \int \varphi d\mathbb{P}_1 \geq 0,$$

also

$$\int \psi d\mathbb{P}_1 \geq \int \varphi d\mathbb{P}_1.$$

b) Für einen Test φ der Form

$$\varphi(x) = \begin{cases} 1 & f_1(x) > k f_0(x) \\ \gamma & f_1(x) = k f_0(x) \\ 0 & f_1(x) < k f_0(x) \end{cases}$$

müssen wir zeigen, dass wir (γ, k) so finden können, dass er ein vorgegebenes Niveau $\alpha \in (0, 1)$ ausschöpft, d. h. dass

$$\int \varphi d\mathbb{P}_0 = \alpha$$

gilt. Da φ auf der Menge $\{x : f_1(x) > k \cdot f_0(x)\}$ gleich 1 ist, liegt es nahe, k als das $(1-\alpha)$-Quantil von $\frac{f_1}{f_0}$ zu wählen und γ so zu verwenden, dass φ auch das Niveau erreicht, wenn $\frac{f_1}{f_0}$ gerade an der Stelle k springt. Wir setzen also

$$T(x) = \frac{f_1(x)}{f_0(x)},$$

wobei wir Divisionen durch Null stets als ∞ bewerten. Für x mit $f_0(x) > 0$ ergibt sich dann

$$\begin{aligned} f_1(x) > k \cdot f_0(x) &\Leftrightarrow T(x) > k \\ f_1(x) = k \cdot f_0(x) &\Leftrightarrow T(x) = k \quad \text{und} \\ f_1(x) < k \cdot f_0(x) &\Leftrightarrow T(x) < k. \end{aligned}$$

Also folgt

$$\begin{aligned} \int \varphi d\mathbb{P}_0 &= \int_{\{f_0>0\}} \varphi f_0 d\mu \\ &= \int (\mathbb{1}_{\{T>k\}} + \gamma \mathbb{1}_{\{T=k\}}) d\mathbb{P}_0 \\ &= \mathbb{P}_0(T > k) + \gamma \mathbb{P}_0(T = k). \end{aligned}$$

Wir suchen also k und γ so, dass dies gleich α ist. Setze

$$k := \inf\{y > 0 : \mathbb{P}_0(T > y) \leq \alpha) = \inf\{y > 0 : \mathbb{P}_0(T \leq y) > 1 - \alpha\}.$$

k ist kleiner ∞, da $\alpha > 0$ ist. Da

$$y \mapsto \mathbb{P}_0(T > y)$$

rechtsseitig stetig ist, folgt außerdem $\mathbb{P}_0(T > k) \leq \alpha$. Ist zudem $\mathbb{P}_0(T > k) < \alpha$, so gilt

$$\begin{aligned} \mathbb{P}_0(T = k) &= \mathbb{P}_0(T \geq k) - \mathbb{P}_0(T > k) \\ &> \mathbb{P}_0(T \geq k) - \alpha \\ &= \lim_{y \uparrow k} \mathbb{P}_0(T > y) - \alpha \geq 0, \end{aligned}$$

denn angenommen

$$\mathbb{P}_0(T \geq k) < \alpha,$$

so wäre auch

$$\lim_{n \to \infty} \mathbb{P}_0(T > k - \frac{1}{n}) < \alpha,$$

d. h. k wäre auch nicht das Infimum aller $y > 0$ mit

$$\mathbb{P}_0(T > y) < \alpha.$$

Wir können somit setzen:

$$\gamma = \begin{cases} 0, & \text{falls } \mathbb{P}_0(T > k) = \alpha \\ \frac{\alpha - \mathbb{P}_0(T>k)}{\mathbb{P}(T_0=k)}, & \text{falls } \mathbb{P}_0(T > k) < \alpha \end{cases}$$

Man beachte, dass $0 \leq \gamma \leq 1$. Man sieht nun, dass für diese Parameterwahl gerade gilt:

$$\mathbb{E}_{\mathbb{P}_0}\varphi = \mathbb{P}_0[T > k] + \mathbb{P}_0[T = k]\frac{\alpha - \mathbb{P}_0(T > k)}{\mathbb{P}_0(T = k)} = \alpha.$$

c) Sei $\psi \in \Phi_a$ ein Test, für den (4.2) gilt. Sei φ ein Test, der die Gestalt (4.1) hat und

$$\mathbb{E}_{\mathbb{P}_0}\varphi = \alpha$$

erfüllt. Ein solcher existiert nach Teil b). Um nachzuweisen, dass auch ψ die Gestalt (4.1) hat, betrachte die Menge

$$A = \{x : \psi(x) = \varphi(x) \quad \text{oder} \quad f_1(x) = kf_0(x)\}.$$

Wir zeigen, dass $\mu(A^c) = 0$ gilt. Das genügt offenbar, um die Behauptung zu beweisen.

Angenommen, es gelte $\mu(A^c) > 0$. Dann folgt

$$\int (\varphi - \psi)(x)(f_1 - kf_0)(x)d\mu = \int_{A^c} (\varphi - \psi)(x)(f_1(x) - kf_0(x))d\mu > 0.$$

Letzteres ergibt sich, da auf

$$A^c \cap \{x : f_1(x) > kf_0(x)\}$$

gilt

$$(\varphi - \psi)(x)(f_1(x) - kf_0(x)) = (1 - \psi(x))(f_1(x) - kf_0(x)) > 0$$

und analog auf

$$A^c \cap \{x : f_1(x) < kf_0(x)\}$$

gilt

$$(\varphi - \psi)(x)(f_1(x) - kf_0(x)) = -\psi(x)(f_1(x) - kf_0(x)) > 0.$$

Damit erhalten wir

$$\int \varphi d\mathbb{P}_1 - \int \psi d\mathbb{P}_1 > k(\int \varphi d\mathbb{P}_0 - \int \psi d\mathbb{P}_0) = k(\alpha - \int \psi d\mathbb{P}_0) \geq 0.$$

Dies ist ein Widerspruch zur Optimalität von ψ. Nehmen wir nun schließlich an, das obige ψ erfüllte nicht

$$\int \psi d\mathbb{P}_0 = \alpha,$$

sondern

$$\int \psi d\mathbb{P}_0 < \alpha.$$

Dann folgt für die Menge

$$B := \{x : \psi(x) < 1\}$$

$\mathbb{P}_1(B) > 0$. Wir können $\varepsilon > 0$ mit

$$\varepsilon \cdot \mathbb{P}_0(B) \leq \alpha - \int \psi d\mathbb{P}_0$$

wählen. Aber dies impliziert die Existenz eines Tests $\hat{\psi} \in \Phi_\alpha$, der strikt besser ist als ψ: Wir setzen

$$\hat{\psi}(x) = \psi(x)\mathbb{1}_{B^c}(x) + \min\{\psi(x) + \varepsilon, 1\}\mathbb{1}_B(x).$$

In der Tat gilt dann

$$\int \hat{\psi} d\mathbb{P}_1 > \int \psi d\mathbb{P}_1$$

sowie

$$\int \hat{\psi} d\mathbb{P}_0 \leq \int \psi d\mathbb{P}_0 + \varepsilon \mathbb{P}_0(B) \leq \alpha.$$

Also ist $\hat{\psi} \in \Phi_\alpha$ und $\hat{\psi}$ ist strikt besser als ψ.

□

4.2 Zusammengesetzte Hypothesen und Alternativen

Wir wollen uns nun den interessanteren und schwierigeren Fällen zuwenden, bei denen sowohl H als auch K nicht-notwendig einelementige Mengen sind. Schon in der Vorlesung über Stochastik haben wir gesehen: Will man im n-fachen Münzwurf etwa die Hypothese

$$H : p \leq p_0 \quad \text{gegen} \quad K : p > p_0$$

testen, so genügt es, einen Test für

$$H' : p = p_0 \quad \text{gegen} \quad K : p > p_0$$

zu konstruieren. Der Schlüssel hierfür ist einerseits die Intervallstruktur von H und K und andererseits die Monotonie von

$$p \mapsto \mathbb{P}_p(\sum_{i=1}^{n} x_i > t).$$

Ähnliche Überlegungen sind auch in der allgemeineren Situation relevant.

Definition 4.6 *Ein Testproblem heißt einseitiges Testproblem, wenn gilt*

$$\begin{aligned} H &= \{\vartheta \in \Theta : \vartheta \leq \vartheta_0\},\ K = \{\vartheta \in \Theta : \vartheta > \vartheta_0\} \quad \textit{oder} \\ H &= \{\vartheta \in \Theta : \vartheta \geq \vartheta_0\},\ K = \{\vartheta \in \Theta : \vartheta < \vartheta_0\} \end{aligned}$$

für ein $\vartheta_0 \in \Theta$.

Definition 4.7 *Es sei*

$$\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$$

eine durch ein σ-endliches Maß μ dominierte Familie von Wahrscheinlichkeitsverteilungen auf einem messbaren Raum $(\mathcal{X}, \mathcal{A})$.

$$T : \mathcal{X} \to \mathbb{R}$$

sei messbar. $\mathcal{P}$ hat (streng) isotone Dichtequotienten in T, wenn es zu jedem Paar $\vartheta_0, \vartheta_1 \in \Theta$ mit $\vartheta_0 < \vartheta_1$ eine (streng) isotone Funktion

$$H_{\vartheta_0,\vartheta_1} : \mathbb{R} \to [0, \infty]$$

gibt mit

$$\frac{f_{\vartheta_1}}{f_{\vartheta_0}}(x) := \frac{\frac{d\mathbb{P}_{\vartheta_1}}{d\mu}(x)}{\frac{d\mathbb{P}_{\vartheta_0}}{d\mu}(x)} = H_{\vartheta_0,\vartheta_1} \circ T(x) \quad (\mathbb{P}_{\vartheta_0} + \mathbb{P}_{\vartheta_1})\text{-fast sicher.}$$

$\frac{f_{\vartheta_1}}{f_{\vartheta_0}}$ heißt Likelihood- oder Dichtequotient.

Beispiel 4.8 *a)* Bernoulli-Verteilung

Die Dichten der $B(n,p)$-Verteilung bzgl. des Zählmaßes auf $\{0, \ldots, n\}$ sind

$$f_p(i) = \binom{n}{i} p^i (1-p)^{n-i}.$$

Also ist

$$\frac{f_{p_1}(i)}{f_{p_0}(i)} = \left(\frac{p_1}{p_0}\right)^i \left(\frac{1-p_1}{1-p_0}\right)^{n-i};$$

dies hat die Form

$$\frac{f_{p_1}(i)}{f_{p_0}(i)} = C \cdot \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^i,$$

je nachdem, ob $\frac{p_1}{p_0}\frac{1-p_0}{1-p_1}$ größer oder kleiner ist als 1, steigt oder fällt dieser Ausdruck streng monoton in i. Die Familie der Binomialverteilungen hat somit einen isotonen Dichtequotienten in der Statistik $T = Id$ (bzw. $T = -Id$).

b) Normalverteilung $T = Id$

Für die Familie der Normalverteilungen zu fester Varianz σ_0^2

$$\mathcal{P} = \{\mathcal{N}(\mu, \sigma_0^2) : \mu \in \mathbb{R}\}$$

gilt

$$\frac{f_{\mu_1}(x)}{f_{\mu_0}(x)} = \frac{e^{-\frac{1}{2}(\frac{x-\mu_1}{\sigma_0})^2}}{e^{-\frac{1}{2}(\frac{x-\mu_0}{\sigma_0})^2}} = e^{\frac{x(\mu_1-\mu_0)}{\sigma_0^2}} e^{-\frac{(\mu_1^1-\mu_0^2)}{2\sigma_0^2}}.$$

Dies ist, je nach Lage von μ_0 und μ_1, isoton oder antiton in x; wieder liegt also eine Familie mit isotonem Dichtequotienten in

$$T = Id \quad (bzw.\ T = -Id)$$

vor.

c) Exponentialfamilien

Sind die $(\mathbb{P}_\vartheta)\vartheta \in \Theta$ eine Exponentialfamilie bzgl. eines dominierenden Maßes μ in T, d. h. gilt für messbare Q, T und h

$$f_\vartheta(x) = C(\vartheta) \cdot e^{Q(\vartheta)T(x)} h(x),$$

so folgt natürlich für die Konstance C

$$C(\vartheta) = \left[\int_{\mathcal{X}} e^{Q(\vartheta)T(x)} h(x) \mu(dx) \right]^{-1}.$$

$C(\vartheta)$ hängt also nur über Q von ϑ ab. Wir parametrisieren daher um

$$\vartheta \mapsto Q(\vartheta),$$

wobei der neue Parameterraum nun $\mathcal{Q} := Q[\Theta]$ ist. Mit dieser Parametrisierung gilt

$$\frac{f_{Q_1}}{f_{Q_0}}(x) = \frac{C(Q_1)}{C(Q_0)} e^{(Q_1 - Q_0)T(x)},$$

d. h. die Klasse bildet wieder eine Familie mit isotonem Dichtequotienten in der Statistik T. Viele der wichtigsten praktischen Beispiele fallen in diese Klasse.

Für Verteilungsklassen von diesem Typ gilt nun:

Satz 4.9 *Es sei*

$$\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$$

eine Familie von Wahrscheinlichkeitsmaßen mit isotonem Dichtequotienten auf einem messbaren Raum $(\mathcal{X}, \mathcal{A})$. Ferner sei $\alpha \in (0,1)$, $\vartheta_0, \vartheta_1 \in \Theta$, $\vartheta_0 \leq \vartheta_1$ und

$$\begin{aligned} H &= \{\vartheta \in \Theta : \vartheta \leq \vartheta_0\} \neq \emptyset \\ K &= \{\vartheta \in \Theta : \vartheta > \vartheta_1\} \neq \emptyset. \end{aligned}$$

Dann gilt für den Test

$$\varphi^*(x) = \begin{cases} 0 & T(x) < k^* \\ \gamma^* & T(x) = k^* \\ 1 & T(x) > k^* \end{cases}, \tag{4.3}$$

wobei $\gamma^* \in [0,1]$ und $k^* \in \mathbb{R}$ so bestimmt werden, dass

$$\mathbb{P}_{\vartheta_0}(T > k^*) + \gamma^* \mathbb{P}_{\vartheta_0}[T = k^*] = \alpha$$

gilt:

a) φ^* minimiert unter allen Tests φ von H gegen K mit

$$\mathbb{E}_{\vartheta_0} \varphi = \alpha$$

gleichmäßig die Fehlerwahrscheinlichkeiten erster und zweiter Art.

b) φ^* ist ein gleichmäßig bester Test zum Niveau α für H gegen K.

c) $\vartheta \mapsto \mathbb{E}_\vartheta \varphi^*$ ist streng isoton auf $\{\vartheta : 0 < \mathbb{E}_\vartheta \varphi^* < 1\}$.

Beweis:

a) Wir testen zunächst auf dem Niveau α

$$\tilde{H} := \{\vartheta_0\} \quad \text{gegen} \quad \tilde{K} := \{\vartheta'\},$$

wobei $\vartheta' \in K$ beliebig aber fest gewählt ist. Nach dem Neyman-Pearson-Lemma ist hierfür jedes

$$\tilde{\varphi}(x) = \begin{cases} 1, & \text{falls } k f_{\vartheta_0}(x) < f_{\vartheta_1(x)} \\ 0, & \text{falls } k f_{\vartheta_0}(x) > f_{\vartheta_1}(x) \end{cases}$$

mit $\mathbb{E}_{\vartheta_0}\tilde{\varphi} = \alpha$ ein bester Test. Aufgrund der Voraussetzung über den monotonen Dichtequotienten gilt

$$\begin{aligned} H_{\vartheta_0,\vartheta_1}(T(x)) > H_{\vartheta_0,\vartheta_1}(k^*) &\Rightarrow T(x) > k^* \quad \text{und} \\ H_{\vartheta_0,\vartheta_1}(T(x)) < H_{\vartheta_0,\vartheta_1}(k^*) &\Rightarrow T(x) < k^*. \end{aligned}$$

Setzen wir

$$H_{\vartheta_0,\vartheta_1}(k^*) = k,$$

so lässt sich der Test φ^* aus (4.3) als $\tilde{\varphi}$ wählen, denn es gilt

$$\mathbb{E}_{\vartheta_0}\varphi^* = \alpha.$$

Wichtig ist, dass die Festlegung von γ^* und k^* nicht von der Wahl des ϑ' abhängt, sondern nur davon, dass $\vartheta_0 < \vartheta'$ gilt. Somit ist φ^* sogar ein gleichmäßig bester Test für $\tilde{H}$ gegen K unter der Randbedingung

$$\mathbb{E}_{\vartheta_0}\varphi = \alpha,$$

d. h. φ^* minimiert die Fehlerwahrscheinlichkeit zweiter Art.

φ^* minimiert aber auch die Wahrscheinlichkeit für den Fehler erster Art. Um dies einzusehen, führt man die Minimierung von

$$\mathbb{E}_{\vartheta''}\varphi, \ \vartheta'' < \vartheta_0$$

unter der Randbedingung

$$\mathbb{E}_{\vartheta_0}\varphi = \alpha$$

auf das Neyman-Pearson-Lemma zurück. Hierfür setzen wir

$$\psi := 1 - \varphi$$

und bestimmen eine Lösung des Optimierungsproblems

$$\mathbb{E}_{\vartheta_0}\psi = 1 - \alpha, \quad \mathbb{E}_{\vartheta''}\psi \stackrel{!}{=} \max.$$

Für dieses Problem ist nach dem Neyman-Pearson-Lemma $1 - \varphi^*$ ein optimaler Test und zwar unabhängig von $\vartheta'' < \vartheta_0$. Dies aber bedeutet, dass φ^* auch die Fehlerwahrscheinlichkeit erster Art minimiert.

b) φ^* ist nach dem Neyman-Pearson-Lemma auch ein bester Test für $\tilde{H}$ gegen $\tilde{K}$ unter allen Tests φ mit

$$\mathbb{E}_{\vartheta_0}\varphi \leq \alpha.$$

Wegen der Unabhängigkeit von φ^* von ϑ' ist φ^* auch ein gleichmäßig bester Test für $\tilde{H}$ gegen K. Nach dem ersten Schritt gilt für $1-\varphi^*$

$$\mathbb{E}_{\vartheta''}[1-\varphi^*] \geq 1-\alpha = \mathbb{E}_{\vartheta''}[1-\alpha]$$

für alle $\vartheta'' < \vartheta_0$, daher ist φ^* ein Test zum Niveau α für H gegen K, d. h. es gilt

$$\mathbb{E}_{\vartheta''}\varphi^* \leq \alpha \quad \text{für alle} \quad \vartheta'' \leq \vartheta_0.$$

Da weiterhin jeder Test zum Niveau α für H gegen K auch ein Test zum Niveau α für $\tilde{H}$ gegen K ist, ist φ^* gleichmäßig bester Test zum Niveau α für H gegen K.

c) Ergibt sich schließlich wegen $\mathbb{P}_{\vartheta'} \neq \mathbb{P}_{\vartheta''}$ und der Struktur von φ^* (der Test hängt nicht von ϑ' und ϑ'' ab, ist aber für $\vartheta' < \vartheta''$ ein Test wie im Neyman-Pearson-Lemma) aus dem folgenden Korollar zum Neyman-Pearson-Lemma.

□

Korollar 4.10 *In der Situation des Neyman-Pearson-Lemmas gilt für jeden besten Test φ^* zum Niveau $\alpha \in (0,1)$*

$$\mathbb{E}_{\vartheta_1}\varphi^* \geq \alpha.$$

Beweis: Übung. □

Da sich die Rollen von H und K mühelos vertauschen lassen, folgt

Korollar 4.11 *Für einseitige Testprobleme bei Verteilungsklassen mit isotonen Dichtequotienten in T gibt es gleichmäßig beste Tests φ^* zum Niveau $\alpha \in (0;1)$ der Form:*

$$\begin{aligned} \varphi^*(x) &= \mathbb{1}_{(k^*,\infty)}(T(x)) + \gamma^*\mathbb{1}_{\{k^*\}}(T(x)) \quad \textit{bzw.} \\ \varphi^*(x) &= \mathbb{1}_{(-\infty,k^*)}(T(x)) + \gamma^*\mathbb{1}_{\{k^*\}}(T(x)). \end{aligned}$$

Beweis: Das ist offensichtlich. □

Beispiel 4.12 *Sei $\mathcal{X} = \{0,1\}^n$, $\mathcal{A} = \mathcal{P}(\mathcal{X})$ und die Familie $\mathcal{P}$ gegeben durch*

$$\mathcal{P} = \{Ber^n(p),\ p \in [0,1]\}.$$

Weiter seien

$$H = [0,p_0] \quad \textit{und} \quad K = (p_0;1].$$

Dann gilt für jedes $\alpha \in (0;1)$, dass

$$\varphi_n^*(x_1,\ldots,x_n) = \mathbb{1}_{(k_{n,\alpha},\infty^*)}\left(\sum_{i=1}^n x_i\right) + \gamma_{n,\alpha}^* \mathbb{1}_{\{k_{n,\alpha}^*\}}\left(\sum_{i=1}^n x_i\right)$$

mit

$$\mathbb{P}_{p_0}\left(\sum_{i=1}^n X_i > k_{n,\alpha}^*\right) + \gamma_{n,\alpha}^* \mathbb{P}_{p_0}\left(\sum_{i=1}^n X_i = k_{n,\alpha}^*\right) = \alpha$$

ein gleichmäßig bester Test für H gegen K zum Niveau α ist. Die Werte für $k_{n,\alpha}^*$ lassen sich mit dem Computer ermitteln (früher waren sie in Tafelwerken vertafelt). Damit kann man auch $\gamma_{n,\alpha}^*$ bestimmen. Für größere n lässt sich der Satz von de Moivre-Laplace verwenden, für größere n und kleine p auch der Poissonsche Grenzwertsatz.

Das zuletzt diskutierte Problem "Wie lassen sich die Werte k^* und γ^* finden?" ist allgemein für Statistiken T^* schwer zu beantworten. Man kann allerdings verwenden, dass eine isotone Transformation einer monotonen Funktion wieder monoton ist, d. h. man kann versuchen, eine isotone Funktion h zu finden, so dass $h \circ T$ eine bekannte Dichte hat. Dies wird gerechtfertigt durch

Lemma 4.13 *In der Situation von Satz 4.9 sei*

$$h : \mathbb{R} \to \mathbb{R}$$

strikt isoton und

$$\tilde{T} = h \circ T.$$

Sei

$$\tilde{\varphi}^*(x) = \mathbb{1}_{(k^*,\infty)}(\tilde{T}(x)) + \tilde{\gamma}^* \mathbb{1}_{(\{k^*\})}(\tilde{T}(x))$$

mit $\tilde{k}^*$ und $\tilde{\gamma} \in [0,1]$, so dass

$$\mathbb{P}_{\vartheta_0}(\tilde{T} > \tilde{k}^*) + \tilde{\gamma}^* \mathbb{P}_{\vartheta_0}(\tilde{T} = \tilde{k}^*) = \alpha.$$

Dann stimmt $\tilde{\varphi}^*$ mit φ^* aus Satz 4.9 fast sicher überein und ist somit gleichmäßig bester Test für H gegen K zum Niveau α.

Beweis: Dies ist eine einfache Übung. □

Beispiel 4.14 *$X_1,\ldots,X_n$ seien i.i.d. $\mathcal{N}(\mu,\sigma_0^2)$-verteilt mit bekanntem $\sigma_0^2 > 0$. Für $\mu \in \mathbb{R}$ seien die Hypothese*

$$H = (-\infty,\mu_0] \quad \textit{gegen die Alternative} \quad K = (\mu_0,\infty)$$

für ein $\mu_0 \in \mathbb{R}$ zum Niveau $\alpha \in (0,1)$ zu testen. Für die Dichten

$$f_\mu(x_1,\ldots,x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma_0})^2}$$

gilt, dass sie einer Verteilungsklasse mit isotonem Dichtequotienten entstammen. Satz 4.9 liefert die Existenz eines gleichmäßig besten Tests zum Niveau α der Gestalt

$$\varphi^*(x_1,\dots,x_n) = \mathbb{1}_{(k^*,\infty)}(\sum_{i=1}^n x_i) + \gamma^* \mathbb{1}_{\{k^*\}}(\sum_{i=1}^n x_i).$$

Nun kommt man leichter an die Werte einer $\mathcal{N}(0,1)$-Verteilung als an die einer beliebigen $\mathcal{N}(\mu,\sigma^2)$-Verteilung. Nimmt man die (strikt isotone) Transformation

$$h(t) = \sqrt{n}\frac{\frac{t}{n} - \mu_0}{\sigma_0},$$

betrachtet also

$$\sqrt{n}\frac{\frac{\sum_{i=1}^n x_i}{n} - \mu_0}{\sigma_0},$$

so besitzt diese Größe unter $\mathcal{N}(\mu_0,\sigma_0^2)$ eine $\mathcal{N}(0,1)$-Verteilung. Man kann also φ^* wählen als

$$\varphi^*(x_1,\dots,x_n) = \mathbb{1}_{(u_\alpha,\infty)}(\sqrt{n}\frac{\frac{\sum_{i=1}^n x_i}{n} - \mu_0}{\sigma_0}),$$

wobei für u_α gilt

$$\mathbb{P}(X \geq u_\alpha) = \alpha,$$

wobei X eine $\mathcal{N}(0,1)$-verteilte Zufallsvariable ist. Da deren Verteilungsfunktion stetig ist, spielt die Wahl von γ^* keine Rolle.

4.3 Zweiseitige Tests

Wir wollen uns nun zweiseitigen Testproblemen zuwenden, also solchen, bei denen entweder die Alternative (im eindimensionalen Fall) auf beiden Seiten der Hypothese zu finden ist oder umgekehrt die Hypothese auf beiden Seiten der Alternative. Es liegt auf der Hand, dass hierfür die herkömmliche Form des Neyman-Pearson-Lemmas, bei der $H = \{\vartheta_0\}$ gegen $K = \{\vartheta_1\}$ zu testen und dabei eine Nebenbedingung $\mathbb{E}_{\vartheta_0}\varphi = \alpha$ einzuhalten ist, nicht mehr ausreicht. Wir werden dies in einem ersten Schritt verallgemeinern, indem wir mehr als eine Nebenbedingung zulassen.

Satz 4.15 *(Verallgemeinertes Neyman-Pearson-Lemma)*
Es sei μ ein σ-endliches Maß auf einem messbaren Raum $(\mathcal{X},\mathcal{A})$ und $g_1,\dots,g_m,g_{m+1}$ μ-integrierbare Funktionen

$$g_i : \mathcal{X} \to \mathbb{R}.$$

Weiter sei $\alpha = (\alpha_1,\dots,\alpha_m) \in \mathbb{R}^m$. Wir definieren

$$\Phi_{\leq}(\alpha) := \{\varphi \in \Phi : \int \varphi g_i \, d\mu \leq \alpha_i,\ i = 1,\dots,m\}$$

$$\Phi_{=}(\alpha) := \{\varphi \in \Phi : \int \varphi g_i \, d\mu = \alpha_i,\ i = 1,\dots,m\},$$

wobei

$$\Phi := \{\varphi : \mathcal{X} \to [0,1] \text{ } messbar\}$$

ist. Schließlich sei

$$Q_m := \{(\int \varphi g_i d\mu, \ldots, \int \varphi g_m d\mu) : \varphi \in \Phi\}.$$

Dann gilt:

1. *Hinreichende Bedingung*
   Sei φ^* ein Test mit

   (a) *$\varphi^* \in \Phi_=(\alpha)$.*

   (b) *Es gibt $k_1, \ldots, k_m \in \mathbb{R}$ mit*

$$\varphi^*(x) = \begin{cases} 1 & g_{m+1}(x) > \sum_{i=1}^m k_i g_i(x) \\ 0 & g_{m+1}(x) < \sum_{i=1}^m k_i g_i(x) \end{cases} . \tag{4.4}$$

   Dann gilt

$$\int \varphi^* g_{m+1} d\mu = \sup_{\varphi \in \Phi_=(\alpha)} \int \varphi g_{m+1} d\mu. \tag{4.5}$$

   Sind die $k_i \geq 0$ für alle $i = 1, \ldots, m$, so gilt sogar

$$\int \varphi^* g_{m+1} d\mu = \sup\{\int \varphi g_{m+1}\, d\mu : \varphi \in \Phi_\leq(\alpha)\}.$$

2. *Existenz*
   Bildet α einen inneren Punkt von Q_m, so existiert ein φ^* wie unter (a) und (b) unter 1.

3. *Notwendige Bedingung*
   Ist α ein innerer Punkt von Q_m, so ist jeder Test, der (4.5) erfüllt, von der Form (4.4).

Beweis: Wir zeigen nur (1) und verweisen für den Rest auf das Buch "Mathematische Statistik" von Witting oder das gleichnamige Skript von Schmitz.

Der Beweis von (1) folgt den Ideen des Beweises des Neyman-Pearson-Lemmas. φ^* erfülle (4.4) und sei φ ein beliebiger Test. Dann folgt

$$\int (\varphi^* - \varphi)(g_{m+1} - \sum_{i=1}^m k_i g_i) d\mu \geq 0,$$

denn nach Konstruktion von φ^* ist der Integrand μ-fast sicher größer oder gleich 0. Also

$$\int \varphi^* g_{m+1} d\mu - \int \varphi g_{m+1} d\mu \geq \sum_{i=1}^m k_i (\int \varphi^* g_i d\mu - \int \varphi g_i d\mu) \geq 0,$$

falls $\varphi \in \Phi_=(\alpha)$ oder $\varphi \in \Phi_\leq(\alpha)$ und $k_i > 0$ für alle $i = 1, \ldots, m$. □

Wählt man als g_i die Dichten von $\mathbb{P}_{\vartheta_i}$ bzgl. μ, so ergibt sich

Korollar 4.16 *Es seien $\mathbb{P}_{\vartheta_0}, \ldots, \mathbb{P}_{\vartheta_m}$ Wahrscheinlichkeitsmaße über einem messbaren Raum $(\mathcal{X}, \mathcal{A})$, und $\mathbb{P}_{\vartheta_0}$ sei keine Linearkombination von $\mathbb{P}_{\vartheta_1}, \ldots, \mathbb{P}_{\vartheta_m}$. Dann gilt für $\alpha \in (0;1)$: Es existiert ein Test φ mit*

$$\mathbb{E}_{\vartheta_i}[\varphi] = \alpha \quad \textit{für alle} \quad 1 \leq i \leq m \quad \textit{und} \quad \mathbb{E}_{\vartheta_0}[\varphi] > \alpha.$$

Beweis: Wir führen den Beweis per Induktion nach m. Für $m = 1$ ist dies Korollar 4.10. Sei die Aussage für $m-1$ ($m \geq 2$) gezeigt.

Fall I: $\mathbb{P}_{\vartheta_1}, \ldots, \mathbb{P}_{\vartheta_m}$ sind linear abhängig. Dann ist also

$$\mathbb{P}_{\vartheta_m} = \sum_{i=1}^{m-1} \lambda_i \mathbb{P}_{\vartheta_i},$$

wobei die $\lambda_i \in \mathbb{R}$ sind. Da die $\mathbb{P}_{\vartheta_m}$ ein Wahrscheinlichkeitsmaß ist, folgt zudem

$$\sum_{i=1}^{m-1} \lambda_i = 1.$$

Nach Induktionsvoraussetzung existiert ein Test φ mit

$$\mathbb{E}_{\vartheta_i}\varphi = \alpha \quad \text{für alle} \quad i = 1, \ldots, m-1 \quad \text{und} \quad \mathbb{E}_{\vartheta_0}\varphi > \alpha.$$

Somit folgt auch

$$\mathbb{E}_{\vartheta_m}\varphi = \sum_{i=1}^{m-1} \lambda_i \mathbb{E}_{\vartheta_i}\varphi = \alpha \sum_{i=1}^{m-1} \lambda_i = \alpha.$$

Fall II: $\mathbb{P}_{\vartheta_1}, \ldots, \mathbb{P}_{\vartheta_m}$ sind linear unabhängig. Nach Induktionsvoraussetzung existieren zu $k \in \{1, \ldots, m\}$ Tests φ_k und ψ_k mit

$$\begin{aligned} \mathbb{E}_{\vartheta_i}\varphi_k = \alpha \quad & \text{für alle} \quad i \neq 0, k \quad \text{und} \quad \mathbb{E}_{\vartheta_k}\varphi_k > \alpha \quad \text{und} \\ \mathbb{E}_{\vartheta_i}\psi_k = 1-\alpha \quad & \text{für alle} \quad i \neq 0, k \quad \text{und} \quad \mathbb{E}_{\vartheta_k}\psi_k > 1-\alpha. \end{aligned}$$

Wir setzen

$$\varphi_k' := 1 - \psi_k.$$

Dann folgt

$$\mathbb{E}_{\vartheta_i}\varphi_k = \mathbb{E}_{\vartheta_i}\varphi_k' = \alpha \quad \text{für alle} \quad i \neq k \quad \text{und} \quad \mathbb{E}_{\vartheta_k}\varphi_k' < \alpha < \mathbb{E}_{\vartheta_k}\varphi_k.$$

Also ist $\alpha = (\alpha_1, \ldots, \alpha_m)$ ein innerer Punkt von

$$Q_m = \{(\mathbb{E}_{\vartheta_1}\varphi, \ldots, \mathbb{E}_{\vartheta_m}\varphi) : \varphi \in \Phi\}.$$

Angenommen, es gelte für jeden Test φ mit $\mathbb{E}_{\vartheta_i}\varphi = \alpha$ für alle $i = 1, \ldots, m$ auch

$$\mathbb{E}_{\vartheta_0}\varphi \leq \alpha,$$

dann wäre der konstante Test

$$\varphi_\alpha \equiv \alpha$$

ein Test aus $\Phi_=(\alpha)$ mit

$$\mathbb{E}_{\vartheta_0}\varphi_\alpha = \sup_{\varphi\in\Phi_\alpha} \mathbb{E}_{\vartheta_0}\varphi.$$

Sind dann $f_i := \frac{d\mathbb{P}_{\vartheta_i}}{d\mu}$, wobei wir als dominierendes Maß μ

$$\mu = \mathbb{P}_{\vartheta_0} + \ldots + \mathbb{P}_{\vartheta_m}$$

wählen, so können wir aus dem verallgemeinerten Neyman-Pearson-Lemma folgern (man beachte, dass $(\alpha, \ldots, \alpha) \in Q_m$ gilt):

$$\varphi_\alpha(x) = \begin{cases} 1, & \text{falls } f_0 > \sum_{i=1}^m k_i f_i(x) \ \mu\text{-f.s.} \\ 0, & \text{falls } f_0 < \sum_{i=1}^m k_i f_i(x) \ \mu\text{-f.s.} \end{cases}$$

für geeignete $k_i \in \mathbb{R}$. Das aber heißt

$$\mu\left(x : f_0(x) \neq \sum_{i=1}^m k_i f_i(x)\right) = 0.$$

Also folgt

$$\mathbb{P}_{\vartheta_0} = \sum_{i=1}^m k_i \mathbb{P}_{\vartheta_i}$$

im Widerspruch zur Annahme. □

Das soeben bewiesene verallgemeinerte Neyman-Pearson-Lemma und seine Konsequenzen stellen das wichtigste Hilfsmittel bei der Untersuchung zweiseitiger Testprobleme der Form

$$\begin{array}{lcll} H & = & \Theta\backslash(\vartheta_1, \vartheta_2) & \text{gegen} \quad K = (\vartheta_1, \vartheta_2) \\ H & = & [\vartheta_1, \vartheta_2] & \text{gegen} \quad K = \Theta\backslash[\vartheta_1, \vartheta_2] \quad \text{oder} \\ H & = & \{\vartheta_0\} & \text{gegen} \quad K = \Theta\backslash\{\vartheta_0\} \end{array}$$

über einen eindimensionalen Parameter ϑ dar.

Für eine befriedigende Analyse solcher Testprobleme müssen die Maße gewisse Regularitätsannahmen erfüllen. Wir werden daher stets annehmen, dass das zugrunde liegende statistische Experiment sich in Termen eines messbaren Raumes $(\mathcal{X}, \mathcal{A})$ und einer Familie von Wahrscheinlichkeitsmaßen $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$, $\Theta \subseteq \mathbb{R}$, beschreiben lässt. Wir nehmen an, dass die $(\mathbb{P}_\vartheta)$ durch ein σ-endliches Maß μ dominiert werden und bzgl. μ eine Exponentialfamilie bilden, d. h. dass

$$\frac{d\mathbb{P}_\vartheta}{d\mu} = C(\vartheta) e^{Q(\vartheta)T(x)} h(x)$$

gilt. Gehen wir vom Maß μ auf das Maß $h \cdot \mu =: \nu$ über, so können wir annehmen, dass

$$\frac{d\mathbb{P}_\vartheta}{d\nu} = C(\vartheta) e^{Q(\vartheta)T(x)} \tag{4.6}$$

gilt. Liegt nun eine Familie der Form (4.6) vor, so liegt es nahe,

$$Q := Q(\vartheta)$$

als neuen Parameter zu wählen, denn $C(\vartheta)$ hängt wegen

$$\frac{1}{C(\vartheta)} = \int e^{Q(\vartheta)T(x)} d\nu(x)$$

nur über Q von ϑ ab. Wir schreiben die Dichten von nun an in der Form

$$\frac{d\mathbb{P}_Q}{d\nu} = C(Q)e^{QT(x)}$$

und entnehmen den Parameter Q der Menge

$$\mathcal{Q} := \{Q(\vartheta) : \vartheta \in \Theta\}.$$

Als natürlichen Parameterraum der Exponentialfamilie bezeichnet man die Menge $\tilde{\mathcal{Q}}$ aller $Q \in \mathbb{R}$ mit

$$0 < \int e^{QT(x)} d\nu(x) < +\infty.$$

Es gilt stets $\mathcal{Q} \subseteq \tilde{\mathcal{Q}}$.

Satz 4.17 *$\tilde{\mathcal{Q}}$ ist konvex und enthält, falls Q nicht konstant ist, ein nicht-entartetes Intervall.*

Beweis: Es seien $Q_1, Q_2 \in \tilde{\mathcal{Q}},\ \lambda \in (0,1)$. Dann folgt

$$\begin{aligned}
0 &< \int_{\mathcal{X}} e^{(\lambda Q_1 + (1-\lambda)Q_2)T(x)} d\nu(x) \\
&= \int_{\mathcal{X}} e^{\lambda Q_1 T(x)} e^{(1-\lambda)Q_2 T(x)} d\nu(x) \\
&\leq \int_{\mathcal{X}} \left(\max_{i=1,2} (e^{Q_i T(x)}) \right)^{\lambda} \left(\max_{i=1,2} (e^{Q_i T(x)}) \right)^{1-\lambda} d\nu(x) \\
&= \int_{\mathcal{X}} \max_{i=1,2} e^{Q_i T(x)} d\nu(x) \\
&\leq \int_{\mathcal{X}} e^{Q_1 T(x)} + e^{Q_2 T(x)} d\nu(x) < +\infty.
\end{aligned}$$

Also ist $\tilde{\mathcal{Q}}$ konvex. Da außerdem Q als nicht-konstant vorausgesetzt ist (sonst ist das Modell langweilig), d. h. wenn $Q_1 \in \tilde{\mathcal{Q}}$ und $Q_2 \in \tilde{\mathcal{Q}}$ gilt, enthält $\tilde{\mathcal{Q}}$ mindestens das Intervall $[Q_1, Q_2]$. □

Für diese einparametrigen Exponentialfamilien gilt nun

Satz 4.18 *Es sei $\mathcal{P}$ eine einparametrige Exponentialfamilie mit ν-Dichten*

$$f_Q(x) := C(Q)e^{QT(x)},$$

$\tilde{\mathcal{Q}}$ sei dessen natürlicher Parameterraum, φ sei eine beschränkte, $\mathcal{A}$-messbare Funktion und

$$U := \{z = Q + \eta, Q \in \overset{\circ}{\mathcal{Q}}, \eta \in \mathbb{R}\} \subseteq \mathbb{C}.$$

Dann wird durch

$$\beta(z) = \int_{\mathcal{X}} \varphi(x) e^{zT(x)} d\nu(x)$$

eine holomorphe Funktion

$$\beta : U \to \mathbb{C}$$

definiert und es gilt

$$\frac{d\beta(z)}{dz} = \int_{\mathcal{X}} \varphi(x) T(x) e^{zT(x)} d\nu(x),$$

d. h. man kann unter dem Integral differenzieren.

Beweis: Siehe Schmitz, "Mathematische Statistik", Satz 2.4.1. □

In der Anwendung des Satzes ist φ natürlich ein Test.

Satz 4.19 *Es sei $\mathcal{P}$ eine einparametrige Exponentialfamilie. Dann gilt:*

a) Falls $Q \in \overset{\circ}{\mathcal{Q}}$ ist, so existieren Momente $\mathbb{E}_Q T^m$ von beliebiger Ordnung m.

b) Die Gütefunktion eines jeden Tests φ ist im Inneren von $\tilde{Q}$ stetig und beliebig oft differenzierbar. Es gilt

$$\frac{d}{dQ}\mathbb{E}_Q \varphi = \mathbb{E}_Q(\varphi \cdot T) - \mathbb{E}_Q \varphi \mathbb{E}_Q T.$$

Beweis:

a) Es ist

$$\mathbb{E}_Q T^m = C(Q) \int_{T(\mathcal{X})} t^m e^{Qt} d\nu^T(x).$$

Man folgert die Aussage induktiv aus Satz 4.18 (mit $t^{m-1} e^{Qt} d\nu^T(x)$ anstelle von $e^{Qt} d\nu^T(x)$)

$$\mathbb{E}_Q T^m = C(Q) \frac{d^m}{dQ^m} \int_{T(\mathcal{X})} e^{Qt} d\nu^T(x).$$

b) In

$$\beta(Q) = \mathbb{E}_Q \varphi = C(Q) \int_{\mathcal{X}} \varphi(x) e^{QT(x)} d\nu(x)$$

muss

$$C(Q) = [\int e^{QT(x)} d\nu(x)]^{-1}$$

gelten. Insbesondere ist

$$0 < C(Q) < \infty$$

und C ist nach Satz 4.18 differenzierbar. Es folgt

$$\frac{d}{dQ}\mathbb{E}_Q\varphi = \frac{dC(Q)}{dQ}\frac{1}{C(Q)}\mathbb{E}_Q(\varphi) + \mathbb{E}_Q(\varphi T).$$

Setzt man nun $\varphi = 1$, so ergibt sich

$$0 = \frac{d}{dQ}1 = \frac{dC(Q)}{dQ}\frac{1}{C(Q)} \cdot 1 + \mathbb{E}_Q T,$$

also die Behauptung.

□

Wir können nun einen ersten zentralen Satz herleiten.

Satz 4.20 *Unter den bisherigen Bedingungen sei*

$$H = \Theta\backslash(\vartheta_1, \vartheta_2) \quad \textit{gegen} \quad K = (\vartheta_1, \vartheta_2)$$

auf dem Niveau $\alpha \in (0,1)$ zu testen, wobei $\vartheta_1, \vartheta_2 \in \overset{\circ}{\Theta}$ mit $\vartheta_1 < \vartheta_2$ seien. Dann gilt

(i) φ^* sei ein Test mit

a) $\mathbb{E}_{\vartheta_1}\varphi^* = \mathbb{E}_{\vartheta_2}\varphi^* = \alpha$.

b) Es gibt $c_1, c_2 \in \mathbb{R}$ und $\gamma_1, \gamma_2 \in [0,1]$ mit

$$\varphi^*(x) = \begin{cases} 1, & \textit{falls } T(x) \in (c_1, c_2) \\ \gamma_i, & \textit{falls } T(x) = c_i \\ 0, & \textit{falls } T(x) \notin [c_1, c_2] \end{cases}.$$

Dann ist φ^* ein gleichmäßig bester Test zum Niveau α für H gegen K.

(ii) Ein solches φ^* existiert.

Für den Beweis benötigen wir noch ein vorbereitendes

Lemma 4.21 *Seien $b_1 < 0 < b_2$. Dann gilt:*

a) Für $a_1, a_2 > 0$ ist die Menge

$$\{y : a_1 e^{b_1 y} + a_2 e^{b_2 y} < 1\}$$

ein beschränktes, offenes Intervall.

b) Zu $c_1, c_2 \in \mathbb{R}$ mit $c_1 < c_2$ gibt es $a_1, a_2 > 0$ mit

$$(c_1, c_2) = \{y : a_1 e^{b_1 y} + a_2 e^{b_2 y} < 1\}.$$

c) Zu $c \in \mathbb{R}$ gibt es $a_1, a_2 > 0$ derart, dass c die einzige Lösung (in y) von

$$a_1 e^{b_1} y + a_2 e^{b_2 y} = 1$$

ist.

Beweis:

a) Da

$$\lim_{y \to \pm\infty} a_1 e^{b_1 y} + a_2 e^{b_2 y} = \lim_{y \to \pm\infty} g(y) = +\infty$$

gilt, ist

$$\{y : a e^{b_1 y} + a_2 e^{b_2 y < 1}\}$$

beschränkt. Da die beteiligten Funktionen offen sind, ist die Menge offen, und da g strikt konvex ist, ist sie ein Intervall (eventuell allerdings leer).

b) Dies ergibt sich wieder aus der Konvexität von γ, der Tatsache, dass für geeignete a_1, a_2

$$g(0) = a_1 + a_2 < 1$$

ist, und daraus, dass die Nullstellen von

$$g(y) - 1$$

stetig von c_1 und c_2 abhängen.

c) geht sehr ähnlich zu b) und ist eine Übung.

□

Beweis von Satz 4.20

(i) Es sei φ^* ein Test, der (i) a) und b) mit $c_1 \leq c_2$ erfüllt. Es sei $\vartheta' \in (\vartheta_1, \vartheta_2)$. Wegen

$$\vartheta_1 - \vartheta' < 0 < \vartheta_2 - \vartheta'$$

existieren nach Lemma 4.21 Konstanten $a_1, a_2 > 0$ mit

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } a_1 e^{(\vartheta_1 - \vartheta')T(x)} + a_2 e^{(\vartheta_2 - \vartheta')T(x)} < 1 \\ 0, & \text{falls } a_1 e^{(\vartheta_1 - \vartheta')T(x)} + a_2 e^{(\vartheta_2 - \vartheta')T(x)} > 1 \end{cases},$$

d. h. falls wir

$$k_i := a_i \frac{C(\vartheta')}{C(\vartheta_i)} > 0, \quad i = 1, 2$$

setzen,

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } C(\vartheta')e^{\vartheta' T(x)} > \sum_{i=1}^{2} k_i C(\vartheta_i) e^{\vartheta_i T(x)} \\ 0, & \text{falls } C(\vartheta')e^{\vartheta' T(x)} < \sum_{i=1}^{2} k_i C(\vartheta_i) e^{\vartheta_i T(x)} \end{cases}.$$

Aus dem verallgemeinerten Neyman-Pearson-Lemma folgt daher

$$\mathbb{E}_{\vartheta'} \varphi^* \geq \mathbb{E}_{\vartheta'} \varphi$$

für alle $\varphi \in \Phi$ mit $\mathbb{E}_{\vartheta_1} \varphi \leq \alpha$ und $\mathbb{E}_{\vartheta_2} \varphi \leq \alpha$ und somit, da

$$\Phi_\alpha \subseteq \{\varphi \in \Phi : \mathbb{E}_{\vartheta_1} \varphi \leq \alpha \quad \text{und} \quad \mathbb{E}_{\vartheta_2} \varphi \leq \alpha\},$$

auch

$$\mathbb{E}_{\vartheta'} \varphi^* \geq \mathbb{E}_{\vartheta'} \varphi \quad \text{für alle} \quad \varphi \in \Phi_\alpha.$$

Dies gilt für beliebige $\vartheta' \in K$. Können wir also nachweisen, dass $\varphi^* \in \Phi_\alpha$, also dass

$$\mathbb{E}_{\vartheta} \varphi^* \leq \alpha \quad \text{für alle} \quad \vartheta \in H$$

gilt, so sind wir fertig.

Dazu sei $\vartheta' \in H$ und zunächst $\vartheta' < \vartheta_1$. Wieder mithilfe des verallgemeinerten Neyman-Pearson-Lemmas folgern wir: Für einen Test ψ^* mit

$$\mathbb{E}_{\vartheta_i} \psi^* = 1 - \alpha \quad \text{für} \quad i = 1, 2$$

und

$$\psi^* = \begin{cases} 1, & \text{falls } C(\vartheta')e^{\vartheta' T(x)} < \sum_{i=1}^{2} \tilde{k}_i e^{\vartheta_i T(x)} C(\vartheta_i) \\ 0, & \text{falls } C(\vartheta')e^{\vartheta' T(x)} > \sum_{i=1}^{2} \tilde{k}_i e^{\vartheta_i T(x)} C(\vartheta_i) \end{cases}$$

mit geeignet gewählten $\tilde{k}_i$ gilt:

$$\begin{aligned} \mathbb{E}_{\vartheta'} \psi^* &\geq \mathbb{E}_{\vartheta'} \psi \quad \text{für alle} \quad \psi \in \Phi \quad \text{mit} \\ \mathbb{E}_{\vartheta_1} \psi &= \mathbb{E}_{\vartheta_2} \psi = 1 - \alpha. \end{aligned}$$

Diese Form des Tests lässt sich aber erreichen, wenn geeignete $\tilde{a}_1, \tilde{a}_2 > 0$ existieren, so dass

$$\psi^*(x) = \begin{cases} 1, & \text{falls } \tilde{a}_1 e^{(\vartheta' - \vartheta_1)T(x)} + \tilde{a}_2 e^{(\vartheta_2 - \vartheta')T(x)} > 1 \\ 0, & \text{falls } \tilde{a}_1 e^{(\vartheta_1' - \vartheta_1)T(x)} + \tilde{a}_2 e^{(\vartheta_2 - \vartheta')T(x)} < 1 \end{cases}$$

gilt.

Aus Lemma 4.21 folgt, dass für φ^* Konstanten $a_1, a_2 > 0$ existieren mit

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } a_1 e^{(\vartheta' - \vartheta_1)T(x)} + a_2 e^{(\vartheta_2 - \vartheta_1')T(x)} < 1 \\ 0, & \text{falls } a_1 e^{(\vartheta' - \vartheta_1)T(x)} + a_2 e^{(\vartheta_2 - \vartheta')T(x)} > 1. \end{cases}.$$

Somit hat $\psi^* := 1 - \varphi^*$ die gewünschte Eigenschaft. Es gilt daher

$$\mathbb{E}_{\vartheta'}(1 - \varphi^*) \geq \mathbb{E}_{\vartheta'} \psi$$

für alle $\psi \in \Phi$ mit

$$\mathbb{E}_{\vartheta_1} \psi = \mathbb{E}_{\vartheta_2} \psi = 1 - \alpha.$$

Da für den konstanten Test $\varphi_\alpha \equiv \alpha$

$$\mathbb{E}_{\vartheta_1}(1-\varphi_\alpha) = \mathbb{E}_{\vartheta_2}(1-\varphi_\alpha) = 1-\alpha$$

gilt, folgt insbesondere

$$\mathbb{E}_{\vartheta'}\varphi^* \leq \mathbb{E}_{\vartheta'}\varphi_\alpha = \alpha.$$

Für $\vartheta' > \vartheta_2$ geht der Beweis analog. Zusätzlich sehen wir hieraus, dass für alle $\varphi \in \Phi$ mit

$$\mathbb{E}_{\vartheta_1}\varphi = \mathbb{E}_{\vartheta_2}\varphi = \alpha$$

gilt:

$$\begin{array}{rcll} \mathbb{E}_\vartheta\varphi^* & \leq & \mathbb{E}_\vartheta\varphi & \text{für alle} \quad \vartheta \in H \quad \text{und} \\ \mathbb{E}_\vartheta\varphi^* & \geq & \mathbb{E}_\vartheta\varphi & \text{für alle} \quad \vartheta \in K, \end{array}$$

d. h. die Fehlerwahrscheinlichkeiten erster und zweiter Art werden unter der Randbedingung

$$\mathbb{E}_{\vartheta_1}\varphi = \mathbb{E}_{\vartheta_2}\varphi = \alpha$$

durch φ^* gleichmäßig minimiert.

(ii) Sei

$$\mathcal{Q}_2 := \{(\mathbb{E}_{\vartheta_1}\varphi, \mathbb{E}_{\vartheta_2}\varphi) : \varphi \in \Phi\}.$$

Wir bemerken, dass $(\alpha,\alpha) \in \overset{\circ}{\mathcal{Q}}_2$, so dass wir das verallgemeinerte Neyman-Pearson-Lemma anwenden können. Dies folgt aus der Konvexität von $\mathcal{Q}_2$ zusammen mit der Tatsache, dass

$$(\alpha,\alpha), (0,0), (1,1) \in \mathcal{Q}_2,$$

denn die konstanten Tests $\varphi_c \equiv c$ sind in Φ, und außerdem folgt aus Korollar 4.10, dass

$$(\alpha, \alpha+\varepsilon), (\alpha, \alpha-\varepsilon) \in \mathcal{Q}_2.$$

$\tilde{\mathcal{Q}}_2$ enthält also eine Umgebung von (α,α). Für $\vartheta' \in K = (\vartheta_1, \vartheta_2)$ liefert daher das verallgemeinerte Neyman-Pearson-Lemma die Existenz eines Tests ψ^* mit

$$\mathbb{E}_{\vartheta_1}\psi^* = \mathbb{E}_{\vartheta_2}\psi^* = \alpha$$

der Gestalt

$$\psi^*(x) = \begin{cases} 1, & \text{falls } C(\vartheta')e^{\vartheta' T(x)} > \sum_{i=1}^2 k_i C(\vartheta_i) e^{\vartheta_i T(x)} \\ 0, & \text{falls } C(\vartheta')e^{\vartheta' T(x)} < \sum_{i=1}^2 k_i C(\vartheta_i) e^{\vartheta_i T(x)} \end{cases}$$

mit geeigneten $k_i \in \mathbb{R}$. Setzt man $a_i := \frac{k_i C(\vartheta_i)}{C(\vartheta')},\ i = 1,2$, und

$$b_1 := \vartheta_1 - \vartheta' < 0 < \vartheta_2 - \vartheta' := b_2,$$

so ist

$$\psi^*(x) = \begin{cases} 1, & \text{falls } a_1 e^{b_1 T(x)} + a_2 e^{b_2 T(x)} < 1 \\ 0, & \text{falls } a_1 e^{b_1 T(x)} + a_2 e^{b_2 T(x)} > 1 \end{cases}.$$

Nun gilt $a_1, a_2 > 0$. In der Tat: Gälte $a_1 \leq 0$ und $a_2 \leq 0$, so folgte $\psi^* \equiv 1$, also auch

$$\mathbb{E}_{\vartheta_1}\psi^* = 1 \neq \alpha.$$

Gilt hingegen $a_1 > 0,\ a_2 \leq 0$, so ist

$$y \mapsto a_1 e^{b_1 y} + a_2 e^{b_2 y}$$

streng fallend. ψ^* ist also von der Gestalt

$$\psi^* = \begin{cases} 1 & T(x) < c \\ 0 & T(x) > c \end{cases}$$

für ein geeignetes c.
Da aber $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ isotone Dichtequotienten in T hat, folgt nach dem Vorherigen

$$\mathbb{E}_{\vartheta_1}\psi^* > \mathbb{E}_{\vartheta_2}\psi^*$$

im Widerspruch zu

$$\mathbb{E}_{\vartheta_i}\psi^* < \mathbb{E}_{\vartheta_2}\psi^* = \alpha.$$

Ebenso argumentiert man im Falle $a_1 \leq 0$ und $a_2 > 0$. Da aber $a_1, a_2 > 0$ gilt, kann man mithilfe von Lemma 4.21 auf die Existenz von $c_1, c_2 \in \mathbb{R}$ mit $c_1 \leq c_2$ schließen, so dass

$$\psi^*(x) = \begin{cases} 1, & \text{falls } T(x) \in (c_1, c_2) \\ 0, & \text{falls } T(x) \notin [c_1, c_2] \end{cases}$$

Sei nun für $i = 1, 2$

$$\gamma_i := \begin{cases} \frac{1}{\mu(T(x)=c_i)} \int_{\{T(x)=c_i\}} \psi^*(x) d\mu, & \text{falls } \mu(T(x) = c_i) > 0 \\ 0 & \text{sonst} \end{cases}.$$

Dann gilt für alle $\vartheta \in \Theta$

$$\begin{aligned} \int_{\{x:T(x)=c_i\}} \psi^*(x) d\mathbb{P}_\vartheta(x) &= C(\vartheta) \int_{\{x:T(x)=c_i\}} \psi^* e^{\vartheta T(x)} d\mu(x) \\ &= C(\vartheta) e^{\vartheta c_i} \int_{\{x:T(x)=c_i\}} \psi^* d\mu(x) \\ &= C(\vartheta) e^{\vartheta c_i} \gamma_i \, \mu(T(x) = c_i) \\ &= \gamma_i \mathbb{P}_\vartheta(T(x) = c_i). \end{aligned}$$

Definiert man also φ^* durch

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } T(x) \in (c_1, c_2) \\ 0, & \text{falls } T(x) \notin [c_1, c_2) \\ \gamma_i, & \text{falls } T(x) = c_i,\ i = 1, 2 \end{cases},$$

so gilt

$$\mathbb{E}_\vartheta \varphi^* = \mathbb{E}_\vartheta \psi^* \quad \text{für alle} \quad \vartheta \in \Theta$$

und φ^* erfüllt (i) a) und b). □

Abschließend sei noch bemerkt, dass es für realistisch kleine Stichprobenumfänge wenig sinnvoll ist, den oben genannten Test für "nahe beieinander liegende" $\mathbb{P}_{\vartheta_1}$ und $\mathbb{P}_{\vartheta_2}$ durchzuführen. Im Limes "$\vartheta_1 \mapsto \vartheta_2$" testet man

$$H : \vartheta \neq \vartheta_0 \quad \text{gegen} \quad K : \vartheta = \vartheta_0.$$

Da für jeden Test φ die Gütefunktion

$$\vartheta \mapsto \mathbb{E}_\vartheta \varphi$$

stetig ist, folgt aus $\mathbb{E}_\vartheta \varphi \leq \alpha$ für alle $\vartheta \neq \vartheta_0$ auch

$$\mathbb{E}_{\vartheta_0} \varphi \leq \alpha \quad \text{für alle} \quad \varphi \in \Phi.$$

Somit ist der triviale Test $\varphi_\alpha \equiv \alpha$ schon optimal.

Nun wollen wir Tests von

$$\begin{aligned} H : \vartheta \in [\vartheta_1, \vartheta_2] \quad &\text{gegen} \quad K : \vartheta < \vartheta_1 \text{ oder } \vartheta > \vartheta_2 \\ \text{bzw.} \quad H : \vartheta = \vartheta_0 \quad &\text{gegen} \quad K : \vartheta \neq \vartheta_0 \end{aligned}$$

untersuchen. Hierzu muss allerdings zunächst die Klasse der zulässigen Testfunktionen eingeschränkt werden, wie man sich schnell überlegt. Wir haben nämlich gesehen, dass bei Familien mit isotonem Dichtequotienten, also insbesondere Exponentialfamilien, der gleichmäßig beste Test von

$$H_1 = \{\vartheta_0\} \quad \text{gegen} \quad K : \{\vartheta : \vartheta > \vartheta_0\}$$

mit dem gleichmäßig besten Test von

$$H_2 = \{\vartheta : \vartheta \leq \vartheta_0\} \quad \text{gegen} \quad K : \{\vartheta : \vartheta > \vartheta_0\}$$

übereinstimmt. Weiter ist dieser Test φ^* im wesentlichen eindeutig und es gilt

$$\mathbb{E}_\vartheta \varphi^* < \alpha \quad \text{für alle} \quad \vartheta < \vartheta_0.$$

Es kann also keinen Test geben, der für $\vartheta > \vartheta_0$ so gut ist wie φ^* und für $\vartheta < \vartheta_0$ so gut ist wie $\varphi_\alpha \equiv \alpha$. Somit existiert kein gleichmäßig bester Test zum Niveau α für

$$H : \{\vartheta_0\} \quad \text{gegen} \quad K : \{\vartheta : \vartheta \neq \vartheta_0\}.$$

Man betrachtet die folgende vernünftige Einschränkung: Man lässt nur Tests zu, die auf K mindestens die Güte α haben (anderenfalls gäbe es Parameterwerte ϑ, für die $\varphi_\alpha \equiv \alpha$ die größte Güte hätte).

Definition 4.22 *Gegeben sei ein Alternativtestproblem.*

a) $\varphi \in \Phi$ *heißt* unverfälscht *zum Niveau* α*, wenn gilt*

$$\mathbb{E}_\vartheta \varphi \leq \alpha \quad \textit{für alle} \quad \vartheta \in H \quad \underline{\textit{und}} \quad \mathbb{E}_\vartheta \varphi \geq \alpha \quad \textit{für alle} \quad \vartheta \in K.$$

Φ_α^u *sei die Menge aller solcher Tests (es gilt* $\Phi_\alpha^u \neq 0$*, denn* $\varphi_\alpha \in \Phi_\alpha^u$*).*

b) φ^* *heißt gleichmäßig bester, unverfälschter Test zum Niveau* α *für* H *gegen* K*, wenn gilt*

$$\varphi^* \in \Phi_\alpha^u \quad \textit{und} \quad \mathbb{E}_\vartheta \varphi^* = \sup_{\varphi \in \Phi_\alpha^u} \mathbb{E}_\vartheta \varphi$$

für alle $\vartheta \in K$*.*

Bemerkung 4.23 $\varphi^* \in \Phi_\alpha$ *mit*

$$\mathbb{E}_\vartheta \varphi^* = \sup_{\varphi \in \Phi_\alpha^u} \mathbb{E}_\vartheta \varphi \quad \textit{für alle} \quad \vartheta \in K$$

ist auch ein gleichmäßig bester, unverfälschter Test zum Niveau α *für* H *gegen* K*. Dies folgt, da wegen* $\varphi_\alpha \in \Phi_\alpha^u$ *insbesondere gilt*

$$\mathbb{E}_\vartheta \varphi^* \geq \mathbb{E}_\vartheta \varphi_\alpha = \alpha \quad \textit{für alle} \quad \vartheta \in K,$$

also $\varphi^* \in \Phi_\alpha^u$*.*

Wir wollen nun herleiten, dass bei einparametrigen Exponentialfamilien für die oben beschriebene Klasse der zweiseitigen Testprobleme mit einem k, das zwei Zusammenhangskomponenten besitzt, gleichmäßig beste, unverfälschte Tests zum Niveau $\alpha \in (0,1)$ existieren.

Satz 4.24 *Sei* $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ *eine einparametrige Exponentialfamilie bzgl. eines dominierenden Maßes* μ *mit*

$$\frac{d\mathbb{P}_\vartheta}{d\mu} = C(\vartheta) e^{\vartheta T(x)}.$$

Es seien $\vartheta_1, \vartheta_2 \in \Theta$ *mit* $\vartheta_1 < \vartheta_2$ *und*

$$H = \{\vartheta : \vartheta \in (\vartheta_1, \vartheta_2)\}, \; K = \{\vartheta : \vartheta \notin [\vartheta_1, \vartheta_2]\}$$

und $\alpha \in (0,1)$*. Dann gilt*

(i) Ist φ^* *ein Test mit*

a) $\mathbb{E}_{\vartheta_1} \varphi^* = \mathbb{E}_{\vartheta_2} \varphi^* = \alpha$*.*

b) Es gibt $c_1, c_2 \in \mathbb{R}$*,* $\gamma_1, \gamma_2 \in [0,1]$ *mit*

$$\varphi^*(x) = \begin{cases} 1, & \textit{falls } T(x) \notin [c_1, c_2] \\ \gamma_i, & \textit{falls } T(x) = c_i, \; i = 1,2 \\ 0, & \textit{falls } T(x) \in (c_1, c_2) \end{cases}.$$

Dann ist φ^* *ein gleichmäßig bester unverfälschter Test zum Niveau* α *für* H *gegen* K*.*

(ii) Es gibt einen Test der Form wie unter (i) beschrieben.

Beweis:

(i) φ^* sei wie in a) und b) gefordert. Wir setzen

$$\psi^* := 1 - \varphi^*.$$

Also gilt

$$\mathbb{E}_{\vartheta_1}\psi^* = \mathbb{E}_{\vartheta_2}\psi^* = 1 - \alpha.$$

ψ^* hat die Form wie in Satz 4.20 (i) b). Der Beweis von Satz 4.20 liefert daher: Für $\vartheta \in (\vartheta_1, \vartheta_2)$ gilt

$$\mathbb{E}_\vartheta\psi^* \geq \mathbb{E}_\vartheta\psi$$

für alle $\psi \in \Phi$ mit $\mathbb{E}_{\vartheta_1}\psi = \mathbb{E}_{\vartheta_2}\psi = 1 - \alpha$.

Wählt man für ψ den Test $\psi_{1-\alpha} \equiv 1 - \alpha$, so folgt

$$\mathbb{E}_\vartheta\psi^* \geq 1 - \alpha \quad \text{für alle} \quad \vartheta \in (\vartheta_1, \vartheta_2).$$

Für $\vartheta < \vartheta_1$ bzw. $\vartheta > \vartheta_2$ gilt

$$\mathbb{E}_\vartheta\psi^* \leq \mathbb{E}_\vartheta\psi$$

für alle $\psi \in \Phi$ mit $\mathbb{E}_{\vartheta_1}\psi = \mathbb{E}_{\vartheta_2}\psi = 1-\alpha$. Für $\varphi^* = 1-\psi^*$ ergibt sich daher insgesamt

$$\mathbb{E}_\vartheta\varphi^* \leq \alpha \quad \text{für alle} \quad \vartheta \in H,$$

also ist $\varphi^* \in \Phi_\alpha$. Außerdem gilt

$$\mathbb{E}_\vartheta\varphi^* \geq \mathbb{E}_\vartheta\varphi \quad \text{für alle} \quad \vartheta \in K$$

und alle $\varphi \in \Phi$ mit $\mathbb{E}_{\vartheta_1}\varphi = \mathbb{E}_{\vartheta_2}\varphi = \alpha$. Nach der vorhergehenden Anmerkung bleibt also nur noch zu zeigen, dass für alle $\varphi \in \Phi_\alpha^u$ gilt

$$\mathbb{E}_\vartheta\varphi^* \geq \mathbb{E}_\vartheta\varphi \quad \text{für alle} \quad \vartheta \in K.$$

Dies wiederum ist gezeigt, wenn sich folgendes zeigen lässt:

Behauptung: Für $\varphi \in \Phi_\alpha^u$ gilt

$$\mathbb{E}_{\vartheta_1}\varphi = \mathbb{E}_{\vartheta_2}\varphi = \alpha.$$

Beweis: für $\varphi \in \Phi_\alpha^u$ gilt

$$\mathbb{E}_\vartheta\varphi \leq \alpha \quad \text{für alle} \quad \vartheta \in H \qquad \text{und} \qquad \mathbb{E}_\vartheta\varphi \geq \alpha \quad \text{für alle} \quad \vartheta \in K.$$

Da die ϑ_i innere Punkte sind, folgt die Behauptung aus der Stetigkeit der Gütefunktion. □

(ii) Nach Satz 4.20 (ii) existiert ein Test ψ^* mit

$$\mathbb{E}_{\vartheta_1}\psi^* = \mathbb{E}_{\vartheta_2}\psi^* = 1 - \alpha,$$

der von der Form von Satz 4.20 (i) b) ist. Der Test $\varphi^* = 1 - \psi^*$ ist dann von der gewünschten Gestalt.

□

Wir wollen nun das Testproblem

$$H = \{\vartheta_0\} \quad \text{gegen} \quad K = \{\vartheta : \vartheta \neq \vartheta_0\}$$

behandeln. Hierzu beweisen wir

Lemma 4.25 *Es sei $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ eine einparametrige Exponentialfamilie, $\vartheta_0 \in \overset{\circ}{\Theta}$, $\varphi \in \Phi_\alpha^u$. $\alpha \in (0,1)$ sei das Testniveau für das Testproblem*

$$H = \{\vartheta_0\} \quad \textit{gegen} \quad K = \{\vartheta : \vartheta \neq \vartheta_0\}.$$

Dann gilt

$$\mathbb{E}_{\vartheta_0}\varphi = \alpha \quad \textit{und} \quad \mathbb{E}_{\vartheta_0}(\varphi \cdot T) = \alpha \mathbb{E}_{\vartheta_0} T.$$

Beweis: Wie oben folgt aus der Stetigkeit der Gütefunktion wieder

$$\mathbb{E}_{\vartheta_0}\varphi = \alpha.$$

Außerdem hat die Gütefunktion ein Minimum, nämlich α. Nach Satz 4.19 b) ist die Gütefunktion insbesondere in ϑ_0 differenzierbar und es gilt

$$\frac{d}{d\vartheta}\mathbb{E}_\vartheta\varphi|_{\vartheta=\vartheta_0} = \mathbb{E}_{\vartheta_0}\varphi \cdot T - \mathbb{E}_{\vartheta_0}\varphi \cdot \mathbb{E}_{\vartheta_0} T.$$

Da in ϑ_0 ein Minimum vorliegt und $\mathbb{E}_{\vartheta_0}\varphi = \alpha$ ist, folgt die Behauptung. □

Als Konsequenz sehen wir: Kann man unter allen $\psi \in \Phi$ mit

$$\mathbb{E}_{\vartheta_0}\psi = \alpha \quad \text{und} \quad \mathbb{E}_{\vartheta_0}(\psi \cdot T) = \alpha \mathbb{E}_{\vartheta_0} T$$

einen gleichmäßig besten Test finden, so hat man auch schon einen gleichmäßig besten unverfälschten Test zum Niveau α für

$$H = \{\vartheta_0\} \quad \text{gegen} \quad K = \{\vartheta \in \Theta : \vartheta \neq \vartheta_0\}$$

gefunden, wenn dieser in Φ_α^u liegt.

Satz 4.26 *Es seien $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ eine einparametrige Exponentialfamilie und es sei $\vartheta \in \overset{\circ}{\Theta}$. Zu testen sei*

$$H = \{\vartheta_0\} \quad \textit{gegen} \quad K = \{\vartheta \in \Theta : \vartheta \neq \vartheta_0\}$$

zum Niveau $\alpha \in (0,1)$. Dann gilt:

(i) $\varphi^* \in \Phi$ sei ein Test mit

a) $\mathbb{E}_{\vartheta_0}\varphi^* = \alpha$, $\mathbb{E}_{\vartheta_0}(\varphi^* \cdot T) = \alpha \mathbb{E}_{\vartheta_0} T$.

b) Es gibt $c_1, c_2 \in \mathbb{R}$ und $\gamma_1, \gamma_2 \in [0,1]$, so dass

$$\varphi^*(x) = \begin{cases} 1, & \textit{falls } T(x) \neq [c_1, c_2] \\ \gamma_i, & \textit{falls } T(x) = c_i,\ i = 1,2 \\ 0, & \textit{falls } T(x) \in (c_1, c_2) \end{cases} .$$

Dann ist φ^* ein gleichmäßig bester unverfälschter Test zum Niveau α für H gegen K.

(ii) Es gibt einen Test φ^*, der (i) a) und b) erfüllt.

Analog zum Beweis von Satz 4.20 benötigen wir für den Beweis von Satz 4.26 zunächst ein vorbereitendes Lemma:

Lemma 4.27 *Es sei $b \neq 0$. Dann gilt*

(i) Für alle a_1, a_2 mit $a_2 b > 0$ ist die Menge

$$\{y : a_1 + a_2 y > e^{by}\}$$

ein offenes, beschränktes Intervall.

(ii) Zu $c_1, c_2 \in \mathbb{R}$ existieren $a_1, a_2 \in \mathbb{R}$ mit $a_2 b > 0$, so dass

$$(c_1, c_2) = \{y : a_1 + a_2 y > e^{by}\}.$$

(iii) Zu $c \in \mathbb{R}$ existieren $a_1, a_2 \in \mathbb{R}$ mit $a_2 b > 0$, so dass c die einzige Lösung von

$$a_1 + a_2 y = e^{by}$$

in y ist.

Beweis: Der Beweis verläuft ähnlich zum Beweis von Lemma 4.21. □

Beweis von Satz 4.26

(i) Sei φ^* wie in (i) a) und b). Sei $\vartheta' \in K$, d. h. $\vartheta' - \vartheta_0 \neq 0$. Nach dem vorhergehenden Lemma gibt es $a_1, a_2 \in \mathbb{R}$ mit

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } a_1 + a_2 T(x) < e^{(\vartheta' - \vartheta_0)T(x)} \\ 0, & \text{falls } a_1 + a_2 T(x) > e^{(\vartheta' - \vartheta_0)T(x)} \end{cases} ,$$

d. h.

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } C(\vartheta') e^{\vartheta' T(x)} > k_1 C(\vartheta_0) e^{\vartheta_0 T(x)} + k_2 C(\vartheta_0) T(x) e^{\vartheta_0 T(x)} \\ 0, & \text{falls } C(\vartheta') e^{\vartheta' T(x)} < k_1 C(\vartheta_0) e^{\vartheta_0 T(x)} + k_2 C(\vartheta_0) e^{\vartheta_0 T(x)} T(x) \end{cases} ,$$

wobei $a_i = \frac{k_i C(\vartheta_0)}{C(\vartheta')}$ ist. Wendet man das verallgemeinerte Neyman-Pearson-Lemma auf die (μ-integrierbaren) Funktionen

$$\begin{aligned} g_1(x) &= C(\vartheta_0)e^{\vartheta_0 T(x)}, \\ g_2(x) &= C(\vartheta_0)e^{\vartheta_0 T(x)}T(x), \\ g_3(x) &= C(\vartheta')e^{\vartheta' T(x)} \end{aligned}$$

an, so folgt

$$\mathbb{E}_{\vartheta'}\varphi^* \geq \mathbb{E}_{\vartheta'}\varphi$$

für alle $\varphi \in \Phi$ mit $\mathbb{E}_{\vartheta_0}\varphi = \alpha$, $\mathbb{E}_{\vartheta_0}\varphi \cdot T = \alpha\mathbb{E}_{\vartheta_0}T$. Da $\vartheta' \in K$ beliebig gewählt war, gilt dies für alle $\vartheta' \in K$. Nach Lemma 4.25 ergibt sich also

$$\mathbb{E}_{\vartheta}\varphi^* \geq \mathbb{E}_{\vartheta}\varphi \quad \text{für alle} \quad \varphi \in \Phi_\alpha^u \quad \text{und für alle} \quad \vartheta \in K.$$

Nach Bemerkung 4.23 ist (i) gezeigt.

(ii) Für (ii) benötigen wir zunächst

Behauptung: $(\alpha, \alpha\mathbb{E}_{\vartheta_0}T)$ ist ein innerer Punkt von

$$\tilde{\mathcal{Q}}_2 = \{(\mathbb{E}_{\vartheta_0}\varphi,\, \mathbb{E}_{\vartheta_0}\varphi T) : \varphi \in \Phi\}.$$

Beweis: Der Beweis ähnelt dem Beweis von Satz 4.20, der in den Skripten von Alsmeyer und Schmitz steht. □

Für festes $\vartheta' \in K$ mit $\vartheta' > \vartheta_0$ (der Fall $\vartheta' < \vartheta_0$ geht analog) folgt daher aus dem verallgemeinerten Neyman-Pearson-Lemma die Existenz eines Tests ψ^* mit

$$\mathbb{E}_{\vartheta_0}\psi^* = \alpha \quad \text{und} \quad \mathbb{E}_{\vartheta_0}(\psi^* T) = \alpha\mathbb{E}_{\vartheta_0}T$$

und

$$\psi^*(x) = \begin{cases} 1, & \text{falls } C(\vartheta')e^{\vartheta' T(x)} > (k_1 + k_2 T(x))C(\vartheta_0)e^{\vartheta_0 T(x)} \\ 0, & \text{falls } C(\vartheta')e^{\vartheta' T(x)} < (k_1 + k_2 T(x))C(\vartheta_0)e^{\vartheta_0 T(x)} \end{cases},$$

also

$$\psi^*(x) = \begin{cases} 1, & \text{falls } a_1 + a_2 T(x) < e^{bT(x)} \\ 0, & \text{falls } a_1 + a_2 T(x) > e^{bT(x)} \end{cases}$$

mit geeigneten Konstanten a_1, a_2 und $b > 0$.

Um Lemma 4.27 anwenden zu können, benötigen wir, dass $a_2 b > 0$, also $a_2 > 0$, gilt. Angenommen $a_2 \leq 0$. Dann gilt

$$\psi^*(x) = \begin{cases} 1, & \text{falls } T(x) > k \\ 0, & \text{falls } T(x) < k \end{cases}$$

für ein geeignetes k, d. h.

$$\psi^*(x) = \begin{cases} 1, & \text{falls } T(x)C(\vartheta_0)e^{\vartheta_0 T(x)} > kC(\vartheta_0)e^{\vartheta_0 T(x)} \\ 0, & \text{falls } T(x)C(\vartheta_0)e^{\vartheta_0 T(x)} < kC(\vartheta_0)e^{\vartheta_0 T(x)} \end{cases}.$$

Da außerdem $\mathbb{E}_{\vartheta_0}\psi^* = \alpha$ gilt, folgt aus dem verallgemeinerten Neyman-Pearson-Lemma

$$\int \psi^*(x)T(x)C(\vartheta_0)e^{\vartheta_0 T(x)}d\mu(x) \geq \int \varphi(x)T(x)C(\vartheta_0)e^{\vartheta_0 T(x)}d\mu(x)$$

für alle $\varphi \in \Phi$ mit $\mathbb{E}_{\vartheta_0}\varphi = \alpha$, d. h.

$$\mathbb{E}_{\vartheta_0}(\psi^* T) \geq \mathbb{E}_{\vartheta_0}(\varphi \cdot T)$$

für alle $\varphi \in \Phi$ mit $\mathbb{E}_{\vartheta_0}\varphi = \alpha$. Da aber $(\alpha, \alpha\mathbb{E}_{\vartheta_0}T)$ ein innerer Punkt von $\tilde{\mathcal{Q}}_2$ ist, existiert ein Test $\psi \in \Phi$ mit

$$\mathbb{E}_{\vartheta_0}\psi = \alpha \quad \text{und} \quad \mathbb{E}_{\vartheta_0}\psi T > \alpha\mathbb{E}_{\vartheta_0}T.$$

Insgesamt ergibt sich somit

$$\mathbb{E}_{\vartheta_0}(\psi^* T) \geq \mathbb{E}_{\vartheta_0}(\psi T) > \alpha\, \mathbb{E}_{\vartheta_0}T$$

im Widerspruch zur Wahl von ψ^*. Aus Lemma 4.27 folgt daher die Existenz von $c_1 < c_2$, so dass gilt

$$\psi^*(x) = \begin{cases} 1, & \text{falls } T(x) \notin [c_1, c_2] \\ 0, & \text{falls } T(x) \in (c_1, c_2) \end{cases}.$$

Definiert man schließlich noch die γ_i geeignet als

$$\gamma_i := \begin{cases} \frac{1}{\mu(\{x:T(x)=c_i\})} \int_{\{x:T(x)=c_i\}} \psi^*(x)d\mu(x), & \text{falls } \mu(\{x : T(x) = c_i\}) > 0 \\ 0, & \text{sonst} \end{cases}$$

und

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } T(x) \notin [c_1, c_2] \\ \gamma_i, & \text{falls } T(x) = c_i \\ 0, & \text{falls } T(x) \in (c_1, c_2) \end{cases},$$

so erfüllt φ^* die Bedingung (i) a) und b) und ist somit gleichmäßig bester unverfälschter Test zum Niveau α. □

Für die Werte von $\mathbb{E}_\vartheta\varphi^*$ erhält man:

Lemma 4.28 *In der Situation von Satz 4.26 sei die Verteilung von T unter μ, μ^T, kein 2-Punkt-Maß, φ^* sei ein Test wie in (i) a) und b). Dann gilt*

$$\mathbb{E}_\vartheta\varphi^* > \alpha \quad \textit{für alle} \quad \vartheta \neq \vartheta_0.$$

Beweis: Der Beweis ist eine Übung. □

Ist die Verteilung von $\mathbb{P}^T_{\vartheta_0}$ symmetrisch zu einem Punkt $a \in \mathbb{R}$, d. h. gilt für alle $c \in \mathbb{R}$

$$\mathbb{P}_{\vartheta_0}(x : T(x) - a > c) = \mathbb{P}_{\vartheta_0}(x : T(x) - a < -c),$$

so lassen sich die obigen Konstanten c_i und γ_i leicht bestimmen.

Satz 4.29 *In der Situation von Satz 4.26 sei $\mathbb{P}_{\vartheta_0}^T$ symmetrisch zu $a \in \mathbb{R}$. Es seien $c \in \mathbb{R}^+$ und $\gamma \in [0,1]$, so dass*

$$\mathbb{P}_{\vartheta_0}(x : T(x) - a > c) + \gamma \mathbb{P}_{\vartheta_0}(T(x) - a = c) = \frac{\alpha}{2}$$

gilt und

$$\varphi^*(x) = \begin{cases} 1, & \textit{falls } |T(x) - a| > c \\ \gamma, & \textit{falls } |T(x) - a| = c \\ 0, & \textit{falls } |T(x) - a| < c \end{cases} .$$

Dann ist φ^* ein gleichmäßig bester unverfälschter Test zum Niveau α für H gegen K.

Beweis: Wir zeigen, dass φ^* von der Gestalt ist, die in Satz 4.26 (i) a) und b) angegeben ist. Es gilt

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } T(x) \notin [a - \varepsilon, a + \varepsilon] \\ \gamma, & \text{falls } T(x) = a \pm \varepsilon \\ 0 & \text{sonst} \end{cases}$$

und

$$\mathbb{E}_{\vartheta_0} \varphi^* = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$$

aufgrund der angenommenen Symmetrie. Ebenso gilt aufgrund der Symmetrie von $\mathbb{P}_{\vartheta_0}^T$

$$\mathbb{E}_{\vartheta_0} T = a$$

und

$$\mathbb{E}_{\vartheta_0}(\varphi^*(T - a)) = \int_{\{|T(x)-a|>c\}} (T - a) d\mathbb{P}_{\vartheta_0} + \gamma \int_{\{|T(x)-a|=c\}} (T - a) d\mathbb{P}_{\vartheta_0} = 0.$$

Also folgt insgesamt:

$$\mathbb{E}_{\vartheta_0}(\varphi^* T) = \mathbb{E}_{\vartheta_0}(\varphi^*(T - a)) + a \mathbb{E}_{\vartheta_0} \varphi^* = \alpha a = \alpha \mathbb{E}_{\vartheta_0} T.$$

□

Da hier nur ein Fraktil behandelt werden muss, ist der Aufwand derselbe wie bei einem einseitigen Testproblem.

Beispiel 4.30 *Es sei*

$$\mathcal{X} = \{0,1\}^{20},\ \mathcal{A} = \mathcal{P}(\mathcal{X}) \quad \textit{und} \quad \mathcal{P} = \{\bigotimes_{i=1}^{20} Ber(p) : p \in (0,1)\}.$$

Es sei

$$H = \{\frac{1}{2}\} \quad \textit{und} \quad K = \{p,\ p \neq \frac{1}{2}\}$$

zu testen. Bezüglich des Zählmaßes haben wir eine einparametrige Exponentialfamilie mit

$$C(\vartheta) = (1-p)^{20} \quad \textit{und} \quad Q(\vartheta) = \log \frac{p}{1-p}$$

in

$$T(x) = \sum_{i=1}^{20} X_i.$$

Für $\alpha = 0,1$ ergibt sich z. B. als gleichmäßig bester unverfälschter Test

$$\varphi^*(x) = \begin{cases} 1, & \textit{falls } \left|\sum_{i=1}^{20} X_i - 10\right| > 4 \\ 0,7919, & \textit{falls } \left|\sum_{i=1}^{20} X_i - 10\right| = 4 \\ 0, & \textit{falls } \left|\sum_{i=1}^{20} X_i - 10\right| < 4 \end{cases} .$$

5 Tests im Zusammenhang mit der Normalverteilung

In diesem Kapitel soll eine Reihe von Testsituationen untersucht werden, die in Anwendungssituationen von Statistik häufig vorkommen: Die $X_1, \ldots, X_n$ sind i.i.d. $\mathcal{N}(\mu, \sigma^2)$-verteilt, aber wir kennen zumindest einen der Parameter nicht. Die einfachste Situation (μ ist unbekannt, aber σ bekannt) haben wir schon im Rahmen des letzten Kapitels untersucht. Die anderen Fälle werden wir hier eher beschreibend betrachten. Es lassen sich ähnliche Optimalitätsbetrachtungen anstellen wie in Kapitel 4, wobei man die Klasse der Tests weiter einschränkt. Dies wollen wir uns aber hier ersparen. Zu testen seien also die Hypothesen

$$H : \mu \leq \mu_0 \quad \text{gegen} \quad K : \mu > \mu_0 \tag{5.1}$$

und

$$H : \sigma^2 \leq \sigma_0^2 \quad \text{gegen} \quad K : \sigma^2 > \sigma_0^2, \tag{5.2}$$

in den Fällen, wo es Sinn ergibt. Für den Test unter (5.1) können wir, wie wir im letzten Kapitel schon gesehen haben, die Prüfgröße

$$\tilde{T}(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$$

betrachten. Da wir zu allen praktischen Durchführungen des Tests die Verteilung unserer Prüfgröße kennen müssen, betrachten wir äquivalent

$$T(X_1, \ldots, X_n) = \frac{1}{\sqrt{n}\sigma_0} \sum_{i=1}^{n} (X_i - \mu_0).$$

Ist σ^2 unbekannt, so lässt sich die Varianz durch die empirische Varianz

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

mit

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

schätzen und sodann die Prüfgröße

$$\bar{T}(X_1, \ldots, X_n) = \frac{1}{\tilde{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

betrachten. Testen wir umgekehrt (5.2), so bietet es sich wieder an, als Prüfgröße für σ^2 seinen UMVU-Schätzer

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

bzw. im Falle, dass μ nicht bekannt ist,

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

zu betrachten. Wiederum aus dem Grund, dass sich ihre Verteilung leichter berechnen lässt, betrachten wir äquivalent

$$\begin{aligned} S^2 &= \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i-\mu_0)^2 \\ \text{bzw.} \qquad \bar{S}^2 &= \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i-\bar{X})^2. \end{aligned}$$

Wir verwenden wie im vorherigen Kapitel stets Tests der Struktur

$$\varphi(X_1,\ldots,X_n) = \begin{cases} 1, & \text{falls } \tau(X_1,\ldots,X_n) > c \\ 0, & \text{falls } \tau(X_1,\ldots,X_n) < c \end{cases}.$$

Hierbei ist τ eine der Prüfgrößen $T, \bar{T}, S^2, \bar{S}^2$. Der Wert von c bestimmt sich wieder danach, dass der Test die Fehlerwahrscheinlichkeit α 1. Art einhalten soll.

Definition 5.1 *a) Sind $X_1,\ldots,X_n$ i.i.d. $\mathcal{N}(0,1)$-verteilt, so heißt die Verteilung von*

$$\sum_{i=1}^{n} X_i^2$$

(zentrale) χ_n^2-Verteilung mit n Freiheitsgraden.

b) Sind X und Y unabhängige Zufallsvariablen und ist X $\mathcal{N}(0,1)$-verteilt und Y χ_n^2-verteilt, so heißt die Verteilung von

$$\frac{X}{\sqrt{\frac{Y}{n}}}$$

(zentrale) t_n-Verteilung oder Student-Verteilung mit n Freiheitsgraden.

Satz 5.2 *a) Die χ_n^2-Verteilung hat die Dichte*

$$f_n(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1}e^{-x/2} & x > 0 \\ 0 & \textit{sonst} \end{cases}.$$

Hierbei ist

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

die Γ-Funktion.

b) Die Dichte der t_n-Verteilung ist gegeben durch

$$h_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

Beweis:

a) Beweisen wir per Induktion über n.
$\underline{n=1:}$ Es sei $X \sim \mathcal{N}(0,1)$. Dann gilt

$$\begin{aligned}\mathbb{P}(X_1^2 \leq x) &= \mathbb{P}(-\sqrt{x} \leq X_1 \leq \sqrt{x}) \\ &= 2\int_0^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \\ &= \int_0^x \frac{1}{\sqrt{2\pi}} z^{-1/2} e^{-z/2} dz.\end{aligned}$$

Da $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ ist, beweist dies den Induktionsanfang. Aufgrund der Definition der χ_n^2-Verteilung gilt

$$\begin{aligned}g_n(z) &= g_{n-1} * g_1(z) \\ &= \int_{-\infty}^{\infty} g_{n-1}(x) g_1(z-x) dx \\ &\overset{IV}{=} \int_0^z \frac{1}{2^{-\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} x^{\frac{n-1}{2}-1} e^{-x/2} \frac{1}{\sqrt{2\pi}} (z-x)^{-1/2} e^{-\frac{z-x}{2}} dx.\end{aligned}$$

Setzt man $y = \frac{z}{x}$, so erhält man

$$\begin{aligned}g_n(z) &= \frac{e^{-z/2}}{\sqrt{2\pi} 2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} \int_0^1 z^{\frac{n-1}{2}-1} y^{\frac{n-1}{2}-1} z^{-\frac{1}{2}} (1-y)^{-\frac{1}{2}} z dy \\ &= \frac{z^{\frac{n}{2}-1} e^{-\frac{z}{2}}}{\sqrt{2\pi} 2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} \int_0^1 y^{\frac{n-1}{2}-1} (1-y)^{-\frac{1}{2}} dy \\ &= \frac{z^{\frac{n}{2}-1} e^{-\frac{z}{2}}}{\Gamma(\frac{1}{2}) 2^{n/2} \Gamma(\frac{n-1}{2})} \frac{\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})}{\gamma(\frac{n}{2})} \\ &= \frac{z^{\frac{n}{2}-1} e^{-\frac{z}{2}}}{2^{n/2}\Gamma(\frac{n}{2})},\end{aligned}$$

wobei wir bei der vorletzten Gleichheit die Eigenschaften der β-Funktion ausgenutzt haben.

b) Es sei $X \sim \mathcal{N}(0,1)$ und $Y \sim \chi_n^2$-verteilt. Sei $\lambda > 0$. Dann gilt

$$\mathbb{P}\left(\frac{X}{\sqrt{\frac{Y}{n}}} < \lambda\right) = \mathbb{P}(\sqrt{n}X < \lambda\sqrt{Y}) = \int_0^{\infty} \int_{-\infty}^{\lambda\sqrt{y/n}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} g_n(y) dx\, dy.$$

Wegen $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ ergibt sich mit $\varphi(t) = t\sqrt{y/n}$

$$\mathbb{P}\left(\frac{X}{\sqrt{Y/n}} < \lambda\right) = \int_0^{\infty} \int_{-\infty}^{\lambda} \frac{1}{\sqrt{n} 2^{\frac{n+1}{2}} \Gamma(\frac{n}{2})\Gamma(\frac{1}{2})} e^{-\frac{1}{2}(y + \frac{y+t^2}{n})} y^{\frac{n+1}{2}} dt\, dy.$$

Eine erneute Substitution $\varphi(z) = \frac{2t}{1+\frac{t^2}{n}}$ liefert

$$\begin{aligned}\mathbb{P}\left(\frac{X}{\sqrt{Y/n}} < \lambda\right) &= \int_0^\infty \int_{-\infty}^{\lambda} \frac{1}{\sqrt{n}\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})} e^{-z} z^{\frac{n+1}{2}-1} (1+\frac{t^2}{n})^{-\frac{n+1}{2}} dz\, dt \\ &= \int_{-\infty}^{\lambda} \frac{1}{\sqrt{n}\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})} (1+\frac{t^2}{n})^{-\frac{n+1}{2}} \left(\int_0^\infty e^{-z} z^{\frac{n+1}{2}-1} dz\right) dt.\end{aligned}$$

Die Definition der Γ-Funktion lässt nun das innere Integral als $\Gamma(\frac{n+1}{2})$ erkennen.

□

Um dieses Resultat verwenden zu können, benötigen wir

Satz 5.3 *$X_1, \ldots, X_n$ seien i.i.d. $\mathcal{N}(\mu, \sigma^2)$-verteilte Zufallsvariablen. Setze*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \quad \textit{und} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2.$$

Dann gilt:

(i) $\bar{X}$ *und* S^2 *sind unabhängig;*

(ii) $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$*;*

(iii) $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$*;*

(iv) $\sqrt{n}\frac{\bar{X}-\mu}{S} \sim t_{n-1}$*.*

Zum Beweis benötigen wir

Lemma 5.4 *Seien $Y_1, \ldots, Y_n$ i.i.d. Zufallsvariablen, die allesamt $\mathcal{N}(0,1)$-verteilt sind und sei A eine orthogonale $n \times n$-Matrix. Setze*

$$Z = AY.$$

Dann sind die $Z_1, \ldots, Z_n$ ebenfalls i.i.d. $\mathcal{N}(0,1)$-verteilt.

Beweis: Wir zeigen

$$\mathbb{P}[Z_1 \le z_1, \ldots, Z_n \le z_n] = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_n} \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} dx_n \ldots dx_1.$$

Sei $I = (-\infty, z_1] \times \ldots \times (-\infty, z_n]$. Dann ist

$$\begin{aligned}
\mathbb{P}(Z_1 \leq z_1, \ldots, Z_n \leq z_n) &= \mathbb{P}(Z \in I) \\
&= \mathbb{P}(AY \in I) \\
&= \mathbb{P}(Y \in A^{-1}[I]) \\
&= \int_{A^{-1}[I]} f_Y(x_1, \ldots, x_n) dx_1 \ldots dx_n \\
&= \int_I f_Y(Ay)(det\, A^{-1}) dy_1 \ldots dy_n,
\end{aligned}$$

wobei

$$f_Y(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^n x_i^2/2}$$

die Verteilung von Y ist und wir die Transformationsformel benutzt haben. Da A orthogonal ist, gilt det $A = 1$, also

$$\begin{aligned}
\mathbb{P}[Z_1 \leq z_1, \ldots, Z_n \leq z_n] &= \int_I f_Y(Ay) dy_1 \ldots dy_n \\
&= \int_I f_Y(y) dy_1 \ldots dy_n \\
&= \int_{-\infty}^{z_1} \ldots \int_{-\infty}^{z_n} \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} dx_n \ldots dx_1,
\end{aligned}$$

wobei wir bei der vorletzten Behauptung benutzt haben, dass $f_Y(y)$ nur von der euklidischen Länge von y abhängt und A, A^{-1}, A^T längentreu sind. □

Beweis von Satz 5.3: Da die $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$-verteilt sind, sind die Zufallsvariablen

$$Y_i = \frac{X_i - \mu}{\sigma}$$

i.i.d. $\mathcal{N}(0, 1)$-verteilt. Wir wählen (z. B. nach dem Gram-Schmidtschen Orthogonalisierungsverfahren) die Matrix A, deren erste Zeile gleich

$$\left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right) = v^t$$

ist und setzen

$$Z = AY.$$

Nach dem vorhergehenden Lemma sind die Koordinaten $Z_1, \ldots, Z_n$ von Z i.i.d. und $\mathcal{N}(0, 1)$-verteilt. Wir betrachten

$$\sum_{i=1}^n Z_i^2 = \|Z^T Z\| = \|(AY)^T AY\| = \|Y^T Y\| = \sum_{i=1}^n Y_i^2,$$

da A orthogonal ist. Weiter ist

$$\begin{aligned}
\sqrt{n}\bar{X} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \sigma Y_i + \mu \\
&= \sigma \cdot v^t Y + \sqrt{n}\mu = \sigma \cdot Z_1 + \sqrt{n}\mu
\end{aligned}$$

sowie

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n}(X_i-\bar{X})^2 = \sigma^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2\\ &= \sigma^2\left(\sum_{i=1}^{n}Y_i^2 - n(\bar{Y})^2\right)\\ &= \sigma^2\left(\sum_{i=1}^{n}Y_i^2 - \underbrace{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Y_i\right)^2}_{v^tY=Z_1}\right)\\ &= \sigma^2\left(\sum_{i=1}^{n}Z_i^2 - Z_1^2\right) = \sigma^2\left(\sum_{i=2}^{n}Z_i^2\right),\end{aligned}$$

wobei die vorletzte Gleichheit folgt, da

$$\|Z\|^2 = \|AY\|^2$$

ist. Nun folgt die Behauptung leicht:

(i) Da $Z_1,\ldots,Z_n$ unabhängig sind, sind auch

$$\sqrt{n}\bar{X} = \sigma Z_1 + \sqrt{n}\mu$$

und

$$S^2 = \frac{\sigma^2}{n-1}\sum_{i=2}^{n}Z_i^2$$

unabhängig.

(ii) Da Z_1 $\mathcal{N}(0,1)$-verteilt ist, ist

$$\bar{X} = \frac{\sigma}{\sqrt{n}}Z_1 + \mu$$

$\mathcal{N}(\mu,\sigma^2/n)$-verteilt.

(iii) Da $Z_2,\ldots,Z_n$ unabhängig $\mathcal{N}(0,1)$-verteilt sind, ist

$$\frac{n-1}{\sigma^2}S^2 = \sum_{i=2}^{n}Z_i^2$$

χ^2_{n-1}-verteilt.

(iv)

$$\sqrt{n}\frac{\bar{X}-\mu}{S} = \frac{\sigma\cdot Z_1}{\sqrt{\frac{\sigma^2}{n-1}\sum_{i=2}^{n}Z_i^2}} = \frac{Z_1}{\sqrt{\frac{1}{n-1}\sum_{i=2}^{n}Z_i^2}}.$$

Dies ist somit t_{n-1}-verteilt. □

Die obigen Überlegungen führen auf die folgenden Tests für normalverteilte Zufallsvariablen:

a) Testen bei bekannter Varianz $\sigma^2 = \sigma_0^2$

$$H_0 : \mu \leq \mu_0 \quad \text{gegen} \quad K : \mu > \mu_0$$

mittels

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1, & \text{falls } \frac{\sqrt{n}}{\sigma_0}(\bar{X} - \mu_0) > u_{1-\alpha} \\ 0 & \text{sonst} \end{cases},$$

wobei $u_{1-\alpha}$ das $1-\alpha$-Fraktil von $\mathcal{N}(0,1)$ ist. Dies ist der einseitige Gauß-Test.

b) Testen bei unbekannter Varianz σ^2

$$H_0 : \mu \leq \mu_0 \quad \text{gegen} \quad K : \mu > \mu_0$$

mittels

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1, & \text{falls } \sqrt{n}\frac{\bar{X} - \mu_0}{S} > t_{n-1,1-\alpha} \\ 0 & \text{sonst} \end{cases},$$

wobei $t_{n-1,1-\alpha}$ das $1-\alpha$-Fraktil der t_{n-1}-Verteilung ist. Dies ist der einseitige Studentsche t-Test.

c) Testen bei bekanntem μ

$$H_0 : \sigma^2 \leq \sigma_0^2 \quad \text{gegen} \quad H_1 : \sigma^2 > \sigma_0^2$$

mittels

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1, & \text{falls } \frac{1}{\sigma_0^2}\sum_{i=1}^n (X_i - \mu)^2 > \chi^2_{n,\alpha} \\ 0 & \text{sonst} \end{cases},$$

wobei $\chi^2_{n,1-\alpha}$ das $1-\alpha$-Fraktil der χ^2_n-Verteilung ist. Dies ist der einseitige χ^2-Test bei bekanntem μ.

d) Testen bei unbekanntem μ

$$H_0 : \sigma^2 \leq \sigma_0^2 \quad \text{gegen} \quad H_1 : \sigma^2 > \sigma_0^2$$

mittels

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1, & \text{falls } \frac{1}{\sigma_0^2}\sum_{i=1}^n (X_i - \bar{X})^2 > \chi^2_{n-1,1-\alpha} \\ 0 & \text{sonst} \end{cases},$$

wobei $\chi^2_{n-1,1-\alpha}$ das $1-\alpha$-Fraktil der χ^2_{n-1}-Verteilung ist. Dies ist der einseitige χ^2-Test bei unbekanntem μ.

Bemerkung 5.5 *a) Der einseitige Gauß-Test ist ein gleichmäßig bester Test zum Niveau α für das obige Testproblem, wie wir in Kapitel 4 gesehen haben.*

b) Der einseitige t-Test von Student ist ein Test zum Niveau α, denn:
Für $\mu_1 \leq \mu_0$ gilt:

$$\mathbb{P}_{\mu_1}\left[\sqrt{n}\frac{\bar{X}-\mu_0}{S} > t_{n-1,1-\alpha}\right] \leq \mathbb{P}_{\mu_1}\left[\sqrt{n}\frac{\bar{X}-\mu_1}{S} > t_{n-1,1-\alpha}\right] = \alpha,$$

wobei die Ungleichung folgt, da wir das Ereignis "vergrößert" haben, die letzte Gleichheit, da die normalisierte Zufallsvariable unter $\mu = \mu_1$ t_{n-1}-verteilt ist.

c) Ähnlich zeigt man, dass die χ^2-Tests aus c) und d) Tests zum Niveau α sind.

d) Man kann zeigen, dass der t-Test aus b) und die χ^2-Tests aus c) und d) ähnliche Optimalitätseigenschaften haben wie der Gauß-Test unter a).

Bemerkung 5.6 *a) Möchte man in Satz 5.4 a)*

$$H_0 : \mu \geq \mu_0 \quad \textit{gegen} \quad H_1 : \mu < \mu_0$$

testen, so ersetze man in der Definition des Tests

$$\frac{\sqrt{n}}{\sigma_0}(\bar{X}-\mu_0) > u_{1-\alpha}$$

durch

$$\frac{\sqrt{n}}{\sigma_0}(\bar{X}-\mu_0) < u_{\alpha},$$

wobei u_α das α-Fraktil der $\mathcal{N}(0,1)$-Verteilung ist. Analog geht man in b) – d) vor.

b) Zweiseitige Tests: *Möchte man in a)*

$$H_0 : \mu = \mu_0 \quad \textit{gegen} \quad H_1 : \mu \neq \mu_0$$

testen, so ersetze man in der Definition des Tests

$$\frac{\sqrt{n}}{\sigma_0}(\bar{X}-\mu_0) > u_{1-\alpha}$$

durch

$$\left|\frac{\sqrt{n}}{\sigma_0}(\bar{X}-\mu_0)\right| > u_{1-\frac{\alpha}{2}}.$$

Analog geht man in b) vor.
In c) und d) verwendet man

$$\begin{aligned} \chi^2_{n,\frac{\alpha}{2}} &\leq \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \leq \chi^2_{n,1-\frac{\alpha}{2}} \qquad \textit{bzw.} \\ \chi^2_{n-1,\frac{\alpha}{2}} &\leq \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \leq \chi^2_{n-1,1-\frac{\alpha}{2}} \end{aligned}$$

als Ablehnungsbereich von H_1.

c) Zweistichproben-Probleme

Gegeben seien nun zwei Stichproben

$$X_1, \ldots, X_n \quad \textit{und} \quad Y_1, \ldots, Y_m$$

zweier Normalverteilungen mit unbekanntem Erwartungswert μ_X bzw. μ_Y und gleicher (bekannter oder unbekannter) Varianz σ_0^2 bzw. σ^2. Getestet werden soll

$$H : \mu_X = \mu_Y \quad \textit{gegen} \quad K : \mu_X \neq \mu_Y.$$

Solche Tests sind in der Praxis beim Vergleich zweier Produkte (Medikamente, Schuhsohlen, ...) wichtig. Ist die Varianz σ_0^2 bekannt, so schätzen wir μ_X bzw. μ_Y durch

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \textit{bzw.} \quad \bar{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j.$$

Wir betrachten als Testgröße $|Z|$ mit

$$Z = \sqrt{\frac{n \cdot m}{n+m}}\frac{\bar{X} - \bar{Y}}{\sigma_0}.$$

Z ist als Linearkombination normalverteilter Zufallsvariablen unter H wieder normalverteilt und zwar mit Erwartungswert

$$\begin{aligned} \mathbb{E}Z &= \sqrt{\frac{n \cdot m}{n+m}}\frac{1}{\sigma_0}(\mathbb{E}\bar{X} - \mathbb{E}\bar{Y}) \\ &= \sqrt{\frac{n \cdot m}{n+m}}\frac{1}{\sigma_0}(\mu_X - \mu_Y) \\ &= 0 \end{aligned}$$

und Varianz

$$\begin{aligned} \mathbb{V}(Z) &= \frac{n \cdot m}{n+m}\frac{1}{\sigma_0^2}(\mathbb{V}(\bar{X}) + \mathbb{V}(\bar{Y})) \\ &= \frac{n \cdot m}{n+m}\frac{1}{\sigma_0^2}\left(\frac{\sigma_0^2}{n} + \frac{\sigma_0^2}{m}\right) \\ &= 1. \end{aligned}$$

Es liegt also nahe, H abzulehnen, falls

$$|Z| > u_{1-\frac{\alpha}{2}}$$

ist, wobei $u_{\alpha/2}$ wieder das $\alpha/2$-Quartil der Normalverteilung ist. Dies ist der zweiseitige Gaußtest für 2 Stichproben.

Ist dagegen auch σ^2 unbekannt, so schätzen wir analog zum zweiseitigen t-Test zunächst die Varianz durch die sogenannte gepoolte Stichprobenvarianz*:*

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2}{n+m-2}.$$

Tatsächlich ist S^2 erwartungstreu für σ^2, denn

$$\mathbb{E}S^2 = \frac{1}{n+m-2}((m-1)\mathbb{V}X_1 + (m-1)\mathbb{V}(Y_1)) = \frac{1}{n+m-2}(n-1+m-1)\sigma^2 = \sigma^2.$$

Als Testgröße $|Z|$ verwenden wir diesmal den Betrag von

$$Z = \sqrt{\frac{n \cdot m}{n+m}} \frac{\bar{X} - \bar{Y}}{\sqrt{S^2}}.$$

Man rechnet nach, dass Z unter H t-verteilt ist mit $n+m-2$ Freiheitsgraden. Daher lehnt man H ab, falls

$$|Z| > t_{\alpha/2, n+m-2}$$

ist, wobei $t_{1-\frac{\alpha}{2}, n+m-1}$ wieder das $1-\frac{\alpha}{2}$-Fraktil der t_{n+m-2}-Verteilung ist. Dies ist der zweiseitige t-Test für zwei Stichproben.

Bemerkung 5.7 Qualitätsprüfung
Es soll überprüft werden, ob bei Mineralwasserflaschen die richtige Füllmenge erreicht wird. Es werden $n = 100$ Flaschen getestet, dabei beobachtet man eine durchschnittliche Füllmenge von $\bar{X} = 0,71$ Litern bei einer empirischen Varianz von $S^2 = 0,003$. Der Sollwert beträgt 0,7 Liter. Wir testen

$$H : \mu \leq 0,7 \quad gegen \quad K : \mu > 0,7$$

auf dem Niveau 5 %. Wegen

$$\sqrt{n}\left(\frac{\bar{X} - \mu_0}{S}\right) = \sqrt{100} \cdot \frac{0,71 - 0,7}{\sqrt{0,003}} \approx 1,83$$

und

$$t_{n-1,1-\alpha} = t_{99,0,95} \approx 1,66$$

kann die Hypothese auf dem Niveau 5 % verworfen werden.

Interessant ist, dass man bei der Auffassung als zweiseitiges Testproblem die Hypothese

$$H : \mu = 0,7 \quad gegen \quad K : \mu \neq 0,7$$

die Hypothese auf dem 5 %-Niveau beibehalten muss, denn stets ist noch

$$\left|\sqrt{n}\frac{\bar{X} - \mu_0}{S}\right| \approx 1,83,$$

aber

$$t_{n-1,1-\frac{\alpha}{2}} = t_{99,0,975} \approx 1,98.$$

Dies ist gewissermaßen paradox, man behält eine schwächere Hypothese bei, erklärt sich aber daraus, dass der Ablehnungsbereich in Richtung $\mu > 0,7$ schrumpft, da nun auch Werte mit $\mu > 0,7$ zu einer Ablehnung führen.

Beispiel 5.8 "Marktforschung"
Im vergangenen Jahr betrug der Wert eines "Warenkorbs" im Durchschnitt 312 Euro. Kaufen wir heute den gleichen Warenkorb in n Kaufhäusern ein, so bezahlen wir dafür $X_1, \ldots, X_n$ Euro. Kann man daraus schließen, dass der Preis des Warenkorbs gestiegen ist?

Als Zahlenbeispiel nehmen wir $n = 40,\ \bar{X} = 315$ *und* $S^2 = 120$ *an und testen*

$$H : \mu \leq 312 \quad \textit{gegen} \quad K : \mu > 312$$

auf dem Niveau $\alpha = 0,05$. *Wegen*

$$t_{n-1,1-\alpha} = t_{39,0,95} \approx 1,69$$

und

$$\sqrt{n}\frac{\bar{X} - \mu_0}{S} = \sqrt{40}\frac{315 - 312}{\sqrt{120}} \approx 1,73$$

lehnen wir H ab. Der Warenkorb ist also teurer geworden.

Beispiel 5.9 "Mietspiegel"
Die Westfälischen Nachrichten bieten $n = 10$ *Vierzimmerwohnungen zu Quadratmeterpreisen 7,52, 6,90, 9,05, 6,60, 7,97, 8,29, 7,48, 10,12, 7,47, 7,45 an. Darüber hinaus gibt es* $m = 5$ *Fünf- oder Sechszimmerwohnungen zu Quadratmeterpreisen von 6,92, 8,94, 9,31, 7,33 und 8,13 (Kaltmiete in Euro pro Quadratmeter). Kann man schließen, dass sich der Quadratmeterpreis zwischen Vier- und Fünf- oder Sechszimmerwohnungen unterscheidet?*

Es sind $\bar{x} = 7,89$ *und* $\bar{Y} = 8,13$ *die Durchschnitts-Quadratmeterpreise. Wir testen somit*

$$H_0 : \mu_X = \mu_Y \quad \textit{gegen} \quad K : \mu_X \neq \mu_Y$$

unter der Annahme, dass $\sigma_X^2 = \sigma_Y^2$ *sowie der Annahme, dass alle beteiligten Daten normalverteilt sind. Das Niveau sei* $\alpha = 5\ \%$. *Es ist* $n = 10,\ m = 5$ *und*

$$S^2 = \frac{1}{13}(9,65 + 4,15) \approx 1,06.$$

Damit ist

$$\left|\sqrt{\frac{n \cdot m}{n - m}}\frac{\bar{X} - \bar{Y}}{\sqrt{S^2}}\right| \approx 0,4$$

und wegen

$$t_{n+m-2,1-\frac{\alpha}{2}} = t_{13;0,975} \approx 2,2$$

kann die Hypothese nicht verworfen werden.

6 Lineare Regression

Wie im eben besprochenen Zweistichproben-Problem haben wir bei einfachen Regressionen zwei Datensätze $(x_1, \ldots, x_n) \in \mathbb{R}^n$ und $(y_1, \ldots, y_n) \in \mathbb{R}^n$ gegeben, die stochastisch modelliert werden sollen. Wir fassen diese zu Paaren

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

als Realisierungen von Zufallsvektoren $(X_1, Y_1), \ldots, (X_n, Y_n)$ auf, die typischerweise nicht identisch verteilt sind. Darüber hinaus deuten wir die Zufallsvariablen $Y_1, \ldots, Y_n$ als Zielvariablen und nehmen an, dass sie folgendermaßen von den Ausgangsvariablen $X_1, \ldots, X_n$ abhängen

$$Y_i = \varphi(X_i) + \varepsilon \quad \text{für alle} \quad i = 1, \ldots, n, \tag{6.1}$$

wobei

- $\varphi : \mathbb{R} \to \mathbb{R}$ eine beliebige, messbare Regressionsfunktion ist und
- $\varepsilon_1, \ldots, \varepsilon_n$ reellwertige Zufallsvariablen sind, die sogenannte Störgrößen, durch die z. B. Messfehler modelliert werden.

Bemerkung 6.1 *a) Ein wichtiger Spezialfall ist der, dass $\varphi : \mathbb{R} \to \mathbb{R}$ eine lineare Funktion ist, die sogenannte* Regressionsgerade. *Es gibt dann also $\alpha, \beta \in \mathbb{R}$, so dass*

$$\varphi(x) = \alpha + \beta x \quad \textit{für alle} \quad x \in \mathbb{R}.$$

Hierbei heißt α Regressionskonstante *und β* Regressionskoeffizient.

b) In diesem Fall sind α, β unbekannte Modellparameter, die aus den Beobachtungen $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ geschätzt werden sollen.

Bei einem solchen Problem erhebt sich die Frage, wodurch sich ein guter Schätzer auszeichnet. Wir wollen hier die Standardmethode vorstellen, die sogenannte Methode der kleinsten Quadrate. Die Idee hierbei ist die, dass wir versuchen, Schätzer $\hat{\alpha}$ und $\hat{\beta}$ für α und β so zu bestimmen, dass der mittlere quadratische Fehler

$$e(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

für $(\alpha, \beta) = (\hat{\alpha}, \hat{\beta})$ minimal wird. Hierzu setzen wir $n \geq 2$ voraus und dass die Reihe der x_i nicht konstant ist.

Satz 6.2 *Der Kleinste-Quadrate-Schätzer (KQS) für (α, β) ist das Paar $(\hat{\alpha}, \hat{\beta})$ mit*

$$\hat{\beta} = \frac{s_{xy}^2}{s_{xx}^2} \quad \textit{und} \quad \hat{\alpha} = \bar{y}_n - \hat{\beta} \bar{x}_n.$$

Hierbei sind $\bar{x}_n$ bzw. $\bar{y}_n$ definiert durch

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \textit{und} \quad \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

also die Stichprobenmittelwerte. Desweiteren sind

$$\begin{aligned} s_{xx}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \\ s_{xy}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) \quad \textit{und} \\ s_{yy}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2, \end{aligned}$$

also die Stichprobenvarianzen bzw. -kovarianzen.

Beweis: Differenziert man $e(\alpha, \beta)$ bei festem β nach α, so sieht man, dass

$$\alpha = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta x_i) = \bar{y}_n - \beta\bar{x}_n$$

stets $e(\alpha, \beta)$ minimiert. D. h. für jedes feste β ist

$$\sum_{i=1}^{n}((y_i - \beta x_i) - (\bar{y}_n - \beta\bar{x}_n))^2 = \sum_{i=1}^{n}((y_i - \bar{y}_n) - \beta(x_i - \bar{x})^2 = (n-1)(s_{yy}^2 - 2\beta s_{xy}^2 + \beta^2 s_{xx}^2)$$

der kleinste Wert des mittleren quadratischen Fehlers. Differenziert man dies nach β, so ergibt sich

$$\beta = \frac{s_{xy}^2}{s_{xx}^2}.$$

□

Bemerkung 6.3 *$e(\alpha, \beta)$ misst den* vertikalen *Abstand zwischen (x_i, y_i) und $(x_i, \varphi(x_i))$ (mit $\varphi(x) = \alpha + \beta x$) an den Stellen $x_1, \ldots, x_n$. Anstelle dessen ließe sich auch der horizontale Abstand messen. Dies entspricht im wesentlichen einer Vertauschung von x und y und führt zur Lösung*

$$\hat{\beta}'(x, y) = \frac{s_{xy}^2}{s_{yy}^2} \quad \textit{und} \quad \hat{\alpha}'(x, y) = \bar{x}_n - \hat{\beta}'\bar{y}_n$$

zur Schätzung der (inversen) Regressionsgeraden

$$\varphi'(y) = x = \alpha' + \beta' y.$$

Da es üblich ist, y als Funktionswert aufzufassen, ergibt dies

$$y = \frac{\alpha'}{\beta'} + \frac{1}{\beta'}x.$$

Zu beachten wäre, dass im allgemeinen

$$\frac{-\hat{\alpha}'}{\beta'} \neq \hat{\alpha} \quad \textit{und} \quad (\hat{\beta}')^{-1} \neq \hat{\beta}$$

gilt.

Beispiel und Übung 6.4 *Im Weinbau werden die Erträge nach der Lese in Tonnen pro 100 m^2 gemessen (t/m^2). Es ist bekannt, dass der Jahresertrag bereits im Juli ziemlich gut aus der mittleren Anzahl von Beeren pro Traube, der sogenannten Clusterzahl, vorhergesagt werden kann. Das folgende Beispiel soll dies illustrieren. Dabei sei der Jahresertrag die Zielvariable, die Clusterzahl die Ausgangsvariable. Gemessen werden die folgenden Größen, wobei die Daten des Jahres 1972 fehlen, weil in diesem Jahr das untersuchte Weinanbaugebiet von einem Wirbelsturm heimgesucht wurde.*

Jahr	*Ertrag*	*Clusterzahl*
1971	*5,6*	*116,37*
1973	*3,2*	*82,77*
1974	*4,5*	*110,68*
1975	*4,2*	*97,50*
1976	*5,2*	*115,88*
1977	*2,7*	*80,19*
1978	*4,8*	*125,24*
1979	*4,9*	*116,15*
1980	*4,7*	*117,36*
1981	*4,1*	*93,31*
1982	*4,4*	*107,46*
1983	*5,4*	*122,30*

- *Zeichnen Sie ein Streudiagramm der Daten.*
- *Bestimmen Sie die Schätzer $\hat{\alpha}$ und $\hat{\beta}$ sowie $\hat{\alpha}'$ und $\hat{\beta}'$ und zeichnen Sie die Regressionsgerade in das Streudiagramm.*
- *1984 werden 100 Beeren pro Traube gezählt. Prognostizieren Sie mit Hilfe der Regressionsgerade*

  $$y = \hat{\alpha} + \hat{\beta}x$$

  den zu erwartenden Jahresertrag.

Bislang wurden keine spezifischen Modellannahmen über die Störgrößen $\varepsilon_1, \ldots, \varepsilon_n$ benötigt. Umgekehrt konnten auch keine Güteeigenschaften der $\hat{\alpha}$ und $\hat{\beta}$ hergeleitet werden, außer

dass eben der mittlere quadratische Fehler $e(\alpha, \beta)$ minimiert wird. Wir wollen von nun an zusätzlich voraussetzen, dass die $\varepsilon_1, \ldots, \varepsilon_n$ paarweise unkorreliert sind und dass

$$\mathbb{E}\varepsilon_i = 0 \quad \text{und} \quad \mathbb{V}\varepsilon_i = \sigma^2$$

für jedes $i = 1, \ldots, n$ ist, und $\sigma^2 > 0$ von i unabhängig und im allgemeinen unbekannt ist. Wir nehmen des Weiteren an, dass die Ausgangsvariablen deterministisch seien, d. h. wir wissen, dass

$$X_1 = x_1, \ldots, X_n = x_n$$

ist, und die (x_i) seien bekannt. Außerdem sei $n \geq 2$ und die (x_i) seien nicht konstant. Für die Zielvariablen $Y_1, \ldots, Y_n$ gelte für alle $i = 1, \ldots, n$

$$Y_i = \alpha + \beta x_i + \varepsilon_i.$$

Somit ist

$$\mathbb{E}Y_i = \alpha + \beta x_i \quad \text{und} \quad \mathbb{V}Y_i = \sigma^2.$$

Wir wollen nun α und β mit einem linearen Schätzer aus den $(y_1, \ldots, y_n)$ schätzen.

Definition 6.5 *Ein* linearer *Schätzer ist eine Linearkombination*

$$L(Y_1, \ldots, Y_n) = \sum_{i=1}^{n} d_i Y_i$$

für feste Konstanten $d_1, \ldots, d_n \in \mathbb{R}$.

Satz 6.6 *Der (lineare) Schätzer*

$$\hat{\beta} = d_1 Y_1 + \ldots + d_n Y_n$$

ist genau dann erwartungstreu für β, wenn

$$\sum_{i=1}^{n} d_i = 0 \quad \textit{und} \quad \sum_{i=1}^{n} d_i x_i = 1.$$

Beweis: $\hat{\beta}$ ist erwartungstreu genau dann, wenn gilt:

$$\mathbb{E}\hat{\beta} = \sum_{i=1}^{n} d_i \mathbb{E}Y_i = \beta.$$

dies ist gleichbedeutend mit

$$\beta = \sum_{i=1}^{n} d_i \mathbb{E}Y_i = \sum_{i=1}^{n} d_i(\alpha + \beta x_i) = \alpha(\sum_{i=1}^{n} d_i) + \beta(\sum_{i=1}^{n} d_i x_i).$$

Das impliziert die Behauptung. □

Analog zum UMVU-Schätzer in Kapitel 3, also demjenigen erwartungstreuen Schätzer, der die Varianz minimiert, suchen wir nun den besten linearen erwartungstreuen Schätzer, also einen Schätzer, so dass es keinen linearen erwartungstreuen Schätzer mit kleinerer Varianz gibt; diesen neuen wir einen BLUE (= best linear unbiased estimator).

Satz 6.7 *Der lineare Schätzer*

$$\hat{\beta} = \sum_{i=1}^{n} d_i Y_i$$

ist genau dann ein BLUE-Schätzer für β, wenn für alle $i = 1, \ldots, n$ gilt

$$d_i = \frac{x_i - \bar{x}_n}{(n-1)s_{xx}^2}.$$

Beweis: Da sowohl die $(\varepsilon_i)_{i=1}^n$ als auch $(Y_i)_{i=1}^n$ unkorreliert sind, ergibt sich

$$\mathbb{V}(\sum_{i=1}^{n} d_i Y_i) = \sum_{i=1}^{n} d_i^2 \mathbb{V}(Y_i) = \sigma^2 \sum_{i=1}^{n} d_i^2,$$

für beliebige $d_1, \ldots, d_n \in \mathbb{R}$. Ein BLUE-Schätzer muss also erfüllen:

$$\sum_{i=1}^{n} d_i = 0, \quad \sum_{i=1}^{n} d_i x_i = 1 \quad \text{und} \quad \sum_{i=1}^{n} d_i^2 \stackrel{!}{=} \text{ minimal.}$$

Somit folgt

$$\frac{(\sum_{i=1}^n d_i x_i)^2}{\sum_{i=1}^n d_i^2} = \frac{1}{\sum_{i=1}^n d_i^2}, \tag{6.2}$$

$\sum_{i=1}^n d_i^2$ ist also genau dann minimal, wenn die linke Seite von (6.2) maximal ist. Da außerdem

$$\bar{d}_n := \frac{1}{n} \sum_{i=1}^{n} d_i = 0$$

gilt, folgt

$$\begin{aligned} \frac{(\sum_{i=1}^n d_i x_i)^2}{\sum_{i=1}^n d_i^2} &= \frac{(\sum_{i=1}^n (d_i - \bar{d}_n)(x_i - \bar{x}_n))^2}{\sum_{i=1}^n (d_i - \bar{d}_n)^2} \\ &= n \sum_{i=1}^{n} \frac{(x_i - \bar{x}_n)^2}{n} \left(\frac{\sum_{i=1}^n \frac{1}{n}(d_i - \bar{d}_n)(x_i - \bar{x}_n)}{\sqrt{\sum_{i=1}^n \frac{(d_i - \bar{d}_n)^2}{n} \sum_{i=1}^n \frac{(x_i - \bar{x}_n)}{n}}} \right)^2. \end{aligned}$$

Der Ausdruck in der Klammer lässt sich als Korrelationskoeffizient der Zufallsvariablen $D, X : \Omega \to \mathbb{R}$ mit

$$D(i) = d_i, \quad X(i) = x$$

und $\Omega = \{1, \ldots, n\}$, $\mathbb{P}(\{i\}) = \frac{1}{n}$ auffassen. Der Ausdruck ist also daher genau dann maximal, wenn D und X linear abhängig sind, also wenn

$$d_i = a x_i + b \tag{6.3}$$

für alle $i = 1, \ldots, n$ und geeignete $a, b \in \mathbb{R}$ gilt. Wegen Satz 6.6 gilt

$$\sum_{i=1}^{n} (a x_i + b) = 0 \quad \text{und} \quad \sum_{i=1}^{n} (a x_i + b) x_i = 1.$$

Hieraus folgt, dass

$$b = -a\bar{x}_n \quad \text{und} \quad a = \frac{1}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

Setzen wir dies in (6.3) ein, so ergibt sich die Behauptung. □

Bemerkung 6.8 *Der in Satz 6.7 hergeleitete BLUE-Schätzer*

$$\hat{\beta} = \sum_{i=1}^n \frac{x_i - \bar{x}_n}{(n-1)s_{xx}^2} Y_i = \sum_{i=1}^n \frac{(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{(n-1)s_{xx}^2} = \frac{s_{xY}^2}{s_{xx}^2} \tag{6.4}$$

für β stimmt mit dem KQS-Schätzer aus Satz 6.2 überein. Aus dem Beweis von Satz 6.7 ist ersichtlich, dass die Varianz von $\hat{\beta}$ in (6.4) gegeben ist durch

$$\mathbb{V}\hat{\beta} = \sigma^2 \sum_{i=1}^n d_i^2 = \frac{\sigma^2}{(n-1)s_{xx}^2} = \frac{\sigma^2}{\sum_{i=1}^{n-1}(x_i - \bar{x}_n)^2}.$$

Zusätzlich wollen wir nun annehmen, dass die Störgrößen $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. normalverteilt sind. Somit ist

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \text{und} \quad Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$$

für alle $i = 1, \ldots, n$. Wegen der Unabhängigkeit der (ε_i) sind auch die (Y_i) unabhängig. Betrachten wir für festes $(x_1, \ldots, x_n)$ die Log-Likelihoodfunktion der unabhängigen Zufallsgrößen $Y_1, \ldots, Y_n$

$$\log L(y_1, \ldots, y_n; a, \beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2}{2\sigma^2}.$$

Für jedes $\sigma^2 > 0$ und jeden Vektor $(y_1, \ldots, y_n)$ nimmt die logarithmische Likelihoodfunktion $\log L$ als Funktion von (α, β) ihr Maximum für denjenigen Vektor $(\hat{\alpha}, \hat{\beta})$ an, der den Ausdruck

$$\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

minimiert. Dies ist das Minimierungsproblem aus Satz 6.2. Die Lösung lautet

$$\hat{\beta} = \frac{s_{xy}^2}{s_{xx}^2}, \quad \hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n.$$

Wir sehen also:

Satz 6.9 *Sind die $\varepsilon_1, \ldots, \varepsilon_n$ $\mathcal{N}(0, \sigma^2)$-verteilt und unabhängig, so stimmt der ML-Schätzer mit dem KQS-Schätzer für (α, β) aus Satz 6.2 überein.*

Bemerkung 6.10 *Weil $(\hat{\alpha}, \hat{\beta})$ die Loglikelihood-Funktion für jedes $\sigma^2 > 0$ maximiert, ergibt sich der ML-Schätzer $\hat{\sigma}^2$ für σ^2 als Maximum von*

$$\log L(y_1, \ldots, y_n; \hat{\alpha}, \hat{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{2\sigma^2}.$$

Ähnlich wie im Fall von unabhängigen und identisch *verteilten Stichprobenvariablen ergibt sich die Lösung dieses Maximierungsproblems durch 2-faches Differenzieren nach* σ^2

$$\hat{\sigma}(y_1, \ldots, y_n) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

Wie im identisch verteilten Fall ist der Schätzer nicht erwartungstreu. Dies wollen wir genauer untersuchen.

Wir setzen

$$\hat{\varepsilon}_i = Y_i - \hat{\alpha} - \hat{\beta} x_i, \quad i = 1, \ldots, n.$$

Offenbar ist

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

Für die Bestimmung von $\mathbb{E}\hat{\sigma}^2$ genügt es also, $\mathbb{E}\hat{\varepsilon}_i^2$ zu bestimmen. Hierfür zeigen wir

Lemma 6.11 *Seien* $Y_1, \ldots, Y_n$ *unkorrelierte Zufallsvariablen mit* $\mathbb{E}(Y_i^2) < +\infty$ *und* $\mathbb{V}Y_i = \sigma^2$ *für jedes* $i = 1, \ldots, n$. *Für beliebige* $c_1, \ldots, c_n \in \mathbb{R},\ d_1, \ldots, d_n \in \mathbb{R}$ *gilt dann*

$$Cov(\sum_{i=1}^{n} c_i Y_i,\ \sum_{j=1}^{n} d_j Y_j) = \sigma^2 \sum_{i=1}^{n} d_i c_i.$$

Beweis: Das ergibt sich durch einfaches Nachrechnen. □

Somit können wir $\mathbb{E}\hat{\varepsilon}_i$ und $\mathbb{V}\hat{\varepsilon}_i$ berechnen:

Satz 6.12 *Für alle* $i = 1, \ldots, n$ *gilt*

$$\mathbb{E}\hat{\varepsilon}_i = 0$$

und

$$\mathbb{V}(\hat{\varepsilon}_i) = \mathbb{E}\hat{\varepsilon}_i^2 = \sigma^2 \left(\frac{n-2}{n} + \frac{1}{(n-1)s_{xx}^2} \left(\frac{1}{n} \sum_{j=1}^{n} x_j^2 + x_i^2 - 2(x_i - \bar{x}_n)^2 - 2x_i \bar{x}_n \right) \right).$$

Beweis: Es gilt $\mathbb{E}\varepsilon_i = 0$, also $\mathbb{E}Y_i = \alpha + \beta x_i$. Außerdem sind $\hat{\alpha}$ und $\hat{\beta}$ erwartungstreu für α und β. Daher folgt

$$\mathbb{E}\hat{\varepsilon}_i = \mathbb{E}(Y_i - \hat{\alpha} + \hat{\beta} x_i) = \alpha + \beta x_i - (\alpha + \beta x_i) = 0.$$

Außerdem gilt:

$$\mathbb{V}\hat{\varepsilon}_i = \mathbb{V}Y_i + \mathbb{V}\hat{\alpha} + x_i^2 \mathbb{V}\hat{\beta} - 2Cov(Y_i, \hat{\alpha}) - 2Cov(Y_i, \hat{\beta}) + 2Cov(\hat{\alpha}, \hat{\beta}).$$

Aus dem Vorherigen ergibt sich

$$\begin{aligned} Cov(Y_i, \hat{\alpha}) &= \sigma^2 \left(\frac{1}{n} - \frac{(x_i - \bar{x}_n)\bar{x}_n}{(n-1)s_{xx}^2} \right) \\ Cov(Y_i, \hat{\beta}) &= \sigma^2 \left(\frac{x_i - \bar{x}_n}{(n-1)s_{xx}^2} \right) \\ Cov(\hat{\alpha}, \hat{\beta}) &= -\frac{\sigma^2 \bar{x}_n}{(n-1)s_{xx}^2}. \end{aligned}$$

Ähnlich berechnet man

$$\mathbb{V}\hat{\alpha} = \frac{\sigma^2}{n(n-1)s_{xx}^2} \sum_{i=1}^{n} x_i^2.$$

Dies ergibt die Behauptung. □

Korollar 6.13 *Für $\hat{\sigma}^2$ gilt*

$$\mathbb{E}\hat{\sigma}^2 = \frac{n-2}{n}\sigma^2.$$

Beweis: Aus dem vorhergehenden Satz folgt:

$$\begin{aligned} \mathbb{E}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\hat{\varepsilon}_i^2 \\ &= \frac{\sigma^2}{n}\sum_{i=1}^{n} \left(\frac{n-2}{n} + \frac{1}{(n-1)s_{xx}^2} \left(\frac{1}{n}\sum_{j=1}^{n} x_j^2 + x_i^2 - 2(x_i - \bar{x}_n)^2 - 2x_i\bar{x}_n \right) \right) \\ &= \sigma^2 \left(\frac{n-2}{n} + \frac{1}{n(n-1)s_{xx}^2} \left(\sum_{j=1}^{n} x_j^2 + \sum_{i=1}^{n} x_i^2 - 2(n-1)s_{xx}^2 - \frac{2}{n}\left(\sum_{i=1}^{n} x_i \right)^2 \right) \right). \end{aligned}$$

Da

$$\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2 = (n-1)s_{xx}^2,$$

folgt die Behauptung. □

Bemerkung 6.14 *Aufgrund der mangelnden Erwartungstreue ist es üblich, anstelle des ML-Schätzers $\hat{\sigma}^2$ den (erwartungstreuen) Schätzer S^2 für σ^2 zu verwenden:*

$$S^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

(Hierbei sei $n > 2$.) Um Hypothesen über α, β oder σ^2 testen zu können, benötigen wir die Verteilung der Zufallsvariablen $\hat{\alpha}$, $\hat{\beta}$ und S^2.

Hierzu zeigen wir

Lemma 6.15 *Es seien $U,\ V$ unabhängige Zufallsvariablen mit $V \sim \chi^2_m$ und $U+V \sim \chi^2_n,\ u,m \in \mathbb{N}$ mit $m<n$. Dann gilt*

$$U \sim \chi^2_{n-m}.$$

Beweis: Seien $\varphi_U,\ \varphi_V,\ \varphi_{U+V}$ die charakteristischen Funktionen der Zufallsvariablen von $U,\ V$ bzw. $U+V$. Wegen der Unabhängigkeit von U und V ist

$$\varphi_{U+V}(t) = \varphi_U(t)\varphi_V(t) \quad \text{für alle} \quad t \in \mathbb{R}.$$

Nun berechnet sich die charakteristische Funktion einer χ^2_n-verteilten Zufallsvariablen X als

$$\frac{1}{(1-2it)^{n/2}}.$$

(Dies ist eine Übung.) Also ergibt sich

$$\varphi_U(t) = \frac{\varphi_{U+V}(t)}{\varphi_V(t)} = \frac{1}{(1-2it)^{\frac{n-m}{2}}}.$$

Dies ergibt

$$U \sim \chi^2_{n-m}.$$

□

Lemma 6.16 *$Y_1,\dots,Y_n$ seien unabhängig und $Y_i \sim \mathcal{N}(\mu_i,\sigma_i^2)$ für $i=1,\dots,n$. Für beliebige $a_{ij}, b_{ik} \in \mathbb{R}$ $(j=1,\dots,l,\ k=1,\dots,m)$ seien die Zufallsvariablen $U_1,\dots,U_l$ und $V_1,\dots,V_m$ gegeben durch*

$$\begin{aligned} U_j &= \sum_{i=1}^n a_{ij}Y_i \quad \textit{für alle} \quad j=1,\dots,l \\ \textit{und} \quad V_k &= \sum_{i=1}^n b_{ik}Y_i \quad \textit{für alle} \quad k=1,\dots,m. \end{aligned}$$

Dann gilt:

1. *Die Zufallsvariablen U_j und V_k sind normalverteilt mit*

$$U_j \sim \mathcal{N}\left(\sum_{i=1}^n a_{ij}\mu_i,\ \sum_{i=1}^n a_{ij}^2\sigma_i^2\right)$$

und

$$V_k \sim \mathcal{N}\left(\sum_{i=1}^n b_{ik}\mu_i,\ \sum_{i=1}^n b_{ik}^2\sigma_i^2\right),$$

wobei

$$Cov(U_j,V_k) = \sum_{i=1}^n a_{ij}b_{ik}\sigma_i^2.$$

2. *U_j und V_k sind unabhängig genau dann, wenn*

$$Cov(U_j, V_k) = 0.$$

3. *Die Zufallsvektoren $(U_1, \ldots, U_l)$ und $(V_1, \ldots, V_m)$ sind genau dann unabhängig, wenn die Komponenten U_j und V_k für beliebige $j = 1, \ldots, l$ und $k = 1, \ldots, m$ unabhängig sind.*

Beweis: Die Normalverteilung für U_j und V_k ist klar. Ihre Kovarianz berechnet sich nach Lemma 6.11. Teil 2 ist eine bekannte Tatsache für normalverteilte Zufallsvariablen. Teilaussage 3 ergibt sich aus der Definition von Unabhängigkeit von Zufallsvariablen. □

Satz 6.17 *1. Für das Regressionsmodell dieses Kapitels gilt*

$$\begin{aligned} \hat{\alpha} &\sim \mathcal{N}\left(\alpha, \frac{\sigma^2}{n(n-1)s_{xx}^2} \sum_{i=1}^{n} x_i^2\right), \\ \hat{\beta} &\sim \mathcal{N}\left(\beta, \frac{\sigma^2}{(n-1)s_{xx}^2}\right), \end{aligned}$$

wobei

$$Cov(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \bar{x}_n}{(n-1)s_{xx}^2}.$$

2. *Die Zufallsvariablen $(\hat{\alpha}, \hat{\beta})$ und S^2 sind unabhängig und es gilt*

$$\frac{n-2}{\sigma^2} S^2 \sim \chi_{n-2}^2. \tag{6.5}$$

Beweis: Lemma 6.16 ergibt, dass $\hat{\alpha}$ und $\hat{\beta}$ normalverteilt sind. Die Erwartungstreue dieses Schätzers haben wir bereits gezeigt. Ebenso haben wir ihre Varianzen bestimmt. Die Unabhängigkeit von $(\hat{\alpha}, \hat{\beta})$ und S^2 ergibt sich folgendermaßen: $\hat{\varepsilon}_i$ lässt sich umschreiben als

$$\hat{\varepsilon}_i = \sum_{j=1}^{n} (S_{ij} - (a_j + d_j x_i)) Y_j,$$

wobei wieder

$$d_i = \frac{x_i - \bar{x}_n}{(n-1)s_{xx}^2}$$

und

$$c_i = \frac{1}{n} - \frac{\bar{x}_n (x_i - \bar{x}_n)}{(n-1)s_{xx}^2}$$

und

$$S_{ij} = \begin{cases} 1, & \text{falls } i = j \\ 0, & \text{falls } i \neq j \end{cases}.$$

Aus Lemma 6.11 berechnen wir für jedes $i = 1, \ldots, n$

$$\begin{aligned} Cov(\hat{\varepsilon}_i, \hat{\alpha}) &= Cov\left(\sum_{j=1}^{n}(\delta_{ij} - (c_j + d_j x_i))Y_j, \sum_{k=1}^{n} c_k Y_k\right) \\ &= \sigma^2\left(\sum_{j=1}^{n}(\delta_{ij} - (c_j + d_j x_i))c_j\right) \\ &= \sigma^2\left(c_i - \sum_{j=1}^{n} c_j^2 - x_i \sum_{j=1}^{n} c_j d_j\right) \\ &= 0. \end{aligned}$$

Dabei ergibt sich die letzte Gleichheit aus den Gleichungen für c_i und d_i, denn hieraus folgt, dass

$$\begin{aligned} \sum_{j=1}^{n} c_j^2 &= \frac{1}{n} - \frac{\bar{x}_n(x_i - \bar{x}_n)}{(n-1)s_{xx}^2}, \\ d_i &= \frac{x_i - \bar{x}_n}{(n-1)s_{xx}^2} \end{aligned}$$

für alle $i = 1, \ldots, n$. Ebenso leitet man aus Lemma 6.11 ab, dass

$$Cov(\hat{\varepsilon}_i, \hat{\beta}) = 0$$

für jedes $i = 1, \ldots, n$ gilt. Aus den Teilaussagen 2 und 3 von Lemma 6.16 folgt nun, dass die Zufallsvektoren $(\hat{\alpha}, \hat{\beta})$ und $(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)$ unabhängig sind. Also sind auch die Zufallsvektoren $(\hat{\alpha}, \hat{\beta})$ von S^2 unabhängig.

Es bleibt noch (6.5) zu zeigen. Da $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ unter der Transformation

$$x_i \mapsto x_i' = x_i - \bar{x}_n \quad \text{für alle} \quad i = 1, \ldots, n$$

unverändert bleibt, können wir voraussetzen, dass $\bar{x}_n = 0$ gilt. Somit sind c_i und d_i von der Form

$$c_i = \frac{1}{n} \quad \text{und} \quad d_i = \frac{x_i}{\sum_{j=1}^{n} x_j^2}. \tag{6.6}$$

Aus dem bisher Gesagten ergibt sich somit

$$\begin{aligned} (n-2)S^2 &= \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \\ &= \sum_{i=1}^{n}(Y_i - \alpha - \beta x_i + (\alpha - \hat{\alpha}) + (\beta - \hat{\beta})x_i)^2 \\ &= \sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2 - n(\hat{\alpha} - \alpha)^2 - \sum_{j=1}^{n} x_j^2(\hat{\beta} - \beta)^2, \end{aligned}$$

wobei die letzte Gleichheit durch Ausmultiplizieren und Einsetzen von (6.6) in die Definitionsgleichung

$$\hat{\alpha} = c_1 Y_1 + \ldots + c_n Y_n$$

und

$$\hat{\beta} = d_1 Y_1 + \ldots + d_n Y_n$$

von $\hat{\alpha}$ und $\hat{\beta}$ folgt, wenn man $n\bar{x}_n = \sum_{i=1}^n x_i = 0$ bedenkt. Mit anderen Worten: Es gilt

$$(n-2)S^2 + Z^2 = \sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2,$$

wobei

$$Z^2 = n(\hat{\alpha} - \alpha)^2 + \sum_{j=1}^n x_j^2 (\hat{\beta} - \beta)^2$$

und die Zufallsvariablen

$$Y_i' = Y_i - \alpha - \beta x_i \quad \text{für jedes} \quad i = 1, \ldots, n$$

unabhängig sind und identisch $\mathcal{N}(0, \sigma^2)$-verteilt. Somit ist aufgrund der Definition der χ_n^2-Verteilung

$$\frac{(n-2)S^2 + Z^2}{\sigma^2} \sim \chi_n^2.$$

Weil bereits gezeigt wurde, dass $(\hat{\alpha}, \hat{\beta})$ und S^2 unabhängig sind, sind somit auch die Zufallsvariablen $(n-2)S^2$ und Z^2 unabhängig. Außerdem gilt

$$Z^2 = Z_1^2 + Z_2^2,$$

wobei aus dem Vorhergehenden folgt, dass die Zufallsvariablen

$$Z_1 = \sqrt{n}(\hat{\alpha} - \alpha) \quad \text{und} \quad Z_2 = \sqrt{\sum_{j=1}^n x_j^2 (\hat{\beta} - \beta)}$$

unabhängig und identisch $\mathcal{N}(0, \sigma^2)$-verteilt sind. Aus der Definition der χ_2^2-Verteilung ergibt sich nun, dass Z^2/σ^2 eine χ_2^2-verteilte Zufallsvariable ist. Die Gültigkeit von (6.5) folgt somit aus Lemma 6.15. □

Für das hier besprochene einfache Regressionsmodell wollen wir nun unter der Normalverteilungsannahme für die Störgröße Hypothesen über die Regressionskonstante und den Regressionskoeffizienten testen. Hierfür seien $\hat{\alpha}, \hat{\beta}$ und S^2 definiert wie bisher, d. h.

$$\hat{\beta} = \frac{s_{xY}^2}{s_{xx}^2}, \quad \hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n \quad \text{und} \quad S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

Aus den Verteilungs- und Unabhängigkeitseigenschaften aus Satz 6.17 und der Definition der t-Verteilung ergibt sich, dass

$$\frac{\hat{\alpha} - \alpha}{S\sqrt{\sum_{i=1}^n x_i^2 / (n(n-1)s_{xx}^2)}} \sim t_{n-2} \quad \text{und} \quad \frac{\hat{\beta} - \beta}{S(\sqrt{(n-1)s_{xx}^2}} \sim t_{n-2}.$$

Beim Test der Hypothese

$$H : \alpha = \alpha_0 \quad \text{gegen} \quad K : \alpha \neq \alpha_0$$

zum Niveau $\gamma \in (0,1)$ wird die Hullhypothese H abgelehnt, falls

$$\frac{|\hat{\alpha} - \alpha_0|}{S - \sqrt{(\sum_{i=1}^n x_i^2)/n(n-1)s_{xx}^2}} > t_{n-2,1-\gamma/2},$$

wobei $t_{n-2,1-\gamma/2}$ das $\gamma/2$-Quantil der t_{n-2}-Verteilung ist.

Analog testet man

$$H : \beta = \beta_0 \quad \text{gegen} \quad K : \beta \neq \beta_0$$

zum Niveau $\gamma \in (0,1)$. H wird abgelehnt, falls

$$\frac{|\hat{\beta} - \beta|}{S/\sqrt{(n-1)s_{xx}^2}} > t_{n-2,1-\gamma/2}.$$

Bemerkung 6.18 *Von besonderem Interesse ist der Test*

$$H : \beta = 0 \quad \textit{gegen} \quad K : \beta \neq 0$$

(auf dem Niveau γ). Hierbei wird H abgelehnt, falls

$$\frac{|\hat{\beta}|}{S/\sqrt{(n-1)s_{xx}^2}} > t_{n-2,1-\gamma/2}.$$

Beispiel 6.19 *Eine Speditionsfirma will anhand von 10 zufällig ausgewählten LkW-Lieferungen untersuchen, ob ein bzw. welcher Zusammenhang zwischen der Länge des Transportweges (in km) und der Lieferzeit (in Tagen) von der Abholbereitstellung bis zum Eintreffen der Lieferung beim Empfänger besteht. Es werden die folgenden Daten erhoben:*

Nr. der Lieferung	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Weglänge (km)	*825*	*215*	*1070*	*550*	*480*	*920*	*1350*	*325*	*670*	*1215*
Lieferzeit (Tage)	*3,5*	*1,0*	*4,0*	*2,0*	*1,0*	*3,0*	*4,5*	*1,5*	*3,0*	*5,0*

Hierbei wird die Weglänge als Ausgangsvariable und die Lieferzeit als Zielvariable aufgefasst und wir unterstellen einen linearen Zusammenhang.

Die Schätzer für Regressionskoeffizient β und Regressionskonstante α ergeben sich aus diesen Daten als

$$\hat{\beta} = \frac{s_{xy}^2}{s_{xx}^2} = 0,0036, \quad \hat{\alpha} = \bar{y}_{10} - \hat{\beta}\bar{x}_{10} = 0,11.$$

Somit hat die Regressionsgerade die Gestalt

$$\hat{y} = 0,11 + 0,0036x.$$

Beachten wir hieraus die (geschätzten) Störgrößen $\hat{\varepsilon}_i$, so erhalten wir

Nr. der Lieferung	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
beobachtete Lieferzeit	*3,5*	*1,0*	*4,0*	*2,0*	*1,0*	*3,0*	*4,5*	*1,5*	*3,0*	*5,0*
geschätzte Lieferzeit	*3,08*	*0,88*	*3,96*	*2,09*	*1,84*	*3,42*	*4,97*	*1,28*	*2,52*	*4,48*
$\hat{\varepsilon}_i$	*0,42*	*0,12*	*0,04*	*-0,01*	*-0,84*	*-0,42*	*- 0,47*	*0,22*	*0,48*	*0,52*

Somit erhalten wir als Varianzschätzer

$$S^2 = \frac{1}{8}\sum_{i=1}^{6} \hat{\varepsilon}_i^2 \approx 0,48^2.$$

Wir überprüfen nun die Hypothese, dass überhaupt kein signifikanter Zusammenhang zwischen Lieferzeit und Weglänge besteht (d. h. $\beta = 0$ ist) gegen ihre Alternative:

$$H : \beta = 0 \qquad K : \beta \neq 0$$

auf dem Niveau $\alpha = 0,05$. Wir berechnen

$$\bar{x}_{10} = 762,\ \sum_{i=1}^{10} x_i^2 = 7104300 \quad \textit{und} \quad \sqrt{\sum_{i=1}^{4} x_i^2 - 10\bar{x}_{10}^2} = 1139,24.$$

Somit erhalten wir

$$\frac{|\hat{\beta}|}{S/\sqrt{\sum_{i=1}^{10} x_i^2 - 10\bar{x}_{10}^2}} = \frac{0,0036}{0,48/1139,42} = \frac{0,0036}{0,0004} = 9,00.$$

Da

$$t_{8,0,975} = 2,306$$

ist, lehnen wir H ab und vermuten einen Zusammenhang zwischen Lieferzeit und Weglänge.

7 Tests von Verteilungsannahmen

In diesem Kapitel lösen wir uns erstmals von der parametrischen Annahme der ersten Kapitel dieses Skripts. Wieder sei eine Stichprobe $X_1, \ldots, X_n$ reellwertiger i.i.d. Zufallsvariablen gegeben. Bislang haben wir stets angenommen, dass die Verteilung von X_1 zu einer Familie von Wahrscheilichkeitsmaßen

$$\mathcal{P} = \{\mathbb{P}_\vartheta,\ \vartheta \in \Theta \subseteq \mathbb{R}^m\}$$

gehört, wobei Θ oder einige seiner Komponenten unbekannt sind. Diese Situation ist insofern befriedigend, als dass man die Optimalität gewisser Verfahren nachweisen kann. Der Nachteil liegt aber auch auf der Hand: Eine Annahme, dass die Verteilung der X_i der Klasse $\mathcal{P}$ entstammt, ist oftmals eine Annahme, die sich nur mit genauer Kenntnis der Situation, der die Daten entstammen, zu rechtfertigen ist (und manchmal ist diese Annahme überhaupt nicht zu rechtfertigen). In der Folge diskutieren wir daher Tests, die Hypothesen über die Verteilung testen. Solche Tests heißen in der Literatur "Anpassungstest". Wir lernen dabei zunächst einen Test kennen, der die Hypothese einer bestimmten Verteilung überprüft, danach beschäftigen wir uns mit Tests auf Verteilungsklassen.

7.1 Der Kolmogorov-Smirnov-Test

Es soll hier eine Hypothese der Form $H : P = P_0$ getestet werden. Hierbei ist P_0 eine feste, bekannte Verteilung. Die Idee ist es, dabei die "echte" Verteilungsfunktion

$$F_0(t) = \mathbb{P}_0(X_1 \leq t)$$

der X_i mit den sogenannten "empirischen Verteilungsfunktionen"

$$\hat{F}_n(t, x_1, \ldots, x_n) = \frac{1}{n}\{i : x_i \leq t\}$$

zu vergleichen. Dies ist natürlich nur dann sinnvoll, wenn man zunächst weiß, dass für großes n und "typische" x_i F_0 und $\hat{F}_n$ nahe beieinander liegen.

Dies ist der Inhalt des Satzes von Glivenko und Cantelli: Wir bereiten ihn zunächst vor.

Satz 7.1 *Für jedes $x \in \mathbb{R}$ gilt:*

a) Die Zufallsvariable $n\hat{F}_n(x)$ (als Zufallsvariable der $x_1, \ldots, x_n$) ist $\mathcal{B}(n, F(x))$-verteilt, d. h. Binomial-verteilt zu den Parametern n und $p = F(x)$.

b) Es gilt

$$\mathbb{E}\hat{F}_n(x) = F(x), \quad \mathbb{V}\hat{F}_n(x) = \frac{F(x)(1 - F(x))}{n}.$$

c) Für fast alle Realisierungen der x_i gilt

$$\lim_{n\to\infty} \hat{F}_n(x) = F(x).$$

d) Für alle x mit $0 < F(x) < 1$ gilt

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\frac{\hat{F}_n(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \leq y\right) = \int_{-\infty}^{y} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt.$$

Beweis:

a) Man kann die x_i als Realisierungen von i.i.d. Zufallsvariablen X_i auffassen. Mit X_i bezeichnen wir den Indikator, ob die Realisierung von X_i in die Zählung bei $\hat{F}_n(x)$ eingeht oder nicht. Dann ist

$$\begin{aligned} \mathbb{P}(X_i = 1) &= \mathbb{P}(X_i \leq x) = F(x) \\ \text{und} \quad \mathbb{P}(X_i = 0) &= \mathbb{P}(X > x) = 1 - F(x). \end{aligned}$$

Somit ist

$$n \cdot \hat{F}_n(x) = \sum_{i=1}^{n} Y_i$$

$B(n, F(x))$-verteilt.

b) Folgt sofort aus a).

c) Das folgt aus a) und dem Starken Gesetz der Großen Zahlen.

d) Das folgt aus a) und dem Satz von de Moivre-Laplace.

□

Satz 7.1 c) zeigt also schon die punktweise fast sichere Konvergenz von $\hat{F}_n$ gegen F. Wir sind allerdings an einer schärferen Konvergenzart interessiert. Dazu definieren wir:

$$D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)|.$$

Dies ist der sogenannte Kolmogorov-Abstand von $\hat{F}_n$ zu F.

Da F_n eine Treppenfunktion und F monoton und rechtsseitig stetig ist, gilt

$$D_n = \max_{i\in\{1,\ldots,n\}} \max\left\{\left|\frac{i-1}{n} - F(X_{(i)} - 0)\right|, \left|\frac{i}{n} - F(X_{(i)})\right|\right\}$$

bzw.

$$D_n = \max_{i\in\{1,\ldots,n\}} \max\left\{F(X_{(i)} - 0) - \frac{i-1}{n}, \frac{i}{n} - F(X_{(i)})\right\}.$$

Hierbei ist $X_{(i)}$ die i-te Ordnungsstatistik der $X_1, \ldots, X_n$, d. h.

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}.$$

D_n ist somit der maximale Schätzfehler, wenn wir F durch $\hat{F}_n$ schätzen wollen. Es gilt

Satz 7.2 *(Glivenko-Cantelli)*
Es gilt

$$\mathbb{P}(\lim_{n\to\infty} D_n = 0) = 1.$$

Beweis: Wir beginnen mit dem Fall, dass F stetig ist. Zu $m \in \mathbb{N}$ gibt es dann

$$z_0 = -\infty < z_1 < \ldots < z_{m-1} < z_m = +\infty$$

und

$$F(z_0) = 0,\ F(z_1) = \frac{1}{m}, \ldots, F_(z_k) = \frac{k}{m}, \ldots, F(z_{m-1}) = \frac{m-1}{m},\ F(z_m) = 1.$$

Setzen wir $\varepsilon = \frac{1}{m}$, so ergibt sich hieraus für jedes

$$z \in [z_k, z_{k+1}) :$$

$$\hat{F}_n(z) - F(z) \le \hat{F}_n(z_{k+1}) - F(z_k) = \hat{F}_n(z_{k+1}) - F(z_{k+1}) + \varepsilon \tag{7.1}$$

und

$$\hat{F}_n(z) - F(z) \ge \hat{F}_n(z_k) - F(z_{k+1}) = \hat{F}_n(z_k) - F(z_k) - \varepsilon. \tag{7.2}$$

Für $m \in \mathbb{N}$ und $k \in \{0, \ldots, m\}$ sei

$$A_{m,k} = \left\{ w : \hat{F}_n(z_k, w) \underset{n\to\infty}{\longrightarrow} F(z_k) \right\}.$$

Aus Satz 7.1 c) ergibt sich

$$\mathbb{P}(A_{m,k}) = 1 \quad \text{für alle} \quad m, k$$

und daher auch für $A_m := \bigcap_{k=0}^{m} A_{m,k}$

$$\mathbb{P}(A_m) = 1.$$

Für jedes $w \in A_m$ gibt es nun ein $n(w) \in \mathbb{N}$, so dass

$$|\hat{F}_n(z_k, w) - F(z_k)| < \varepsilon$$

für jedes $m \ge n(w)$ und für jedes $k \in \{0, 1, \ldots, m\}$. Hieraus und aus (7.1) und (7.2) folgt, dass

$$\sup_{z\in\mathbb{R}} |\hat{F}_n(z, w) - F(z)| < 2\varepsilon \tag{7.3}$$

für jedes $w \in A_m$ und für jedes $n \ge n(w)$. Also gibt es für jedes

$$w \in A = \bigcap_{m=1}^{\infty} A_m = \bigcap_{m=1}^{\infty} \bigcap_{k=0}^{m} A_{m,k}$$

und für jedes $\varepsilon > 0$ eine natürliche Zahl $n(w, \varepsilon) \in \mathbb{N}$, so dass (7.3) für jedes $n \ge n(w, \varepsilon)$ gilt. Weiter ist natürlich $\mathbb{P}(A) = 1$. Da $\varepsilon > 0$ beliebig klein werden kann, folgt die Behauptung für den Fall, dass F stetig ist.

Für beliebige F gehen wir ähnlich vor. Wir wählen nun für $m \in \mathbb{N}$, $\varepsilon = \frac{1}{m}$ reelle Zahlen

$$z_0 = -\infty < z_1 \ldots < z_{m-1} < z_m = +\infty$$

mit

$$F(z_{k+1} - 0) - F(z_k) \leq \varepsilon.$$

Somit gilt für alle $z \in [z_{k-1}, z_{k+1})$

$$\hat{F}_n(z) - F(z) \leq \hat{F}_n(z_{k+1} - 0) - F(z_{k+1} - 0) + \varepsilon$$

und

$$\hat{F}_n(z) - F(z) \geq \hat{F}_n(z_k) - F(z_k) - \varepsilon.$$

Definieren wir nun

$$A'_{m,k} = \left\{ \omega \in \Omega : \lim_{n \to \infty} a\hat{F}_n(z_k - 0, \omega) = F(z_k - 0) \right\},$$

so folgt wie oben

$$\mathbb{P}(A'_{m,k}) = 1.$$

Mit

$$A'_m = \bigcap_{k=0}^{m} (A_{m,k} \cap A'_{m,k})$$

folgt $\mathbb{P}(A'_m) = 1$ und für $A = \bigcap_{m=1}^{\infty} A'_m$

$$\mathbb{P}(A) = 1.$$

Somit folgt der Beweis wie Fall stetiger F. □

Wir wollen nun die Verteilung des maximalen Schätzfehlers D_n analysieren. Hierzu nennen wir $I \subseteq \mathbb{R}$ Konstanzbereich von F, falls I ein Intervall ist, $\mathbb{P}(X_1 \in I) = 0$ gilt und kein Intervall $J \supseteq I$ existiert, für das auch $\mathbb{P}(X_1 \in J) = 0$ gilt. Wir zeigen nun, dass im Falle stetiger Verteilungsfunktionen F der Kolmogorov-Abstand D_n verteilungsfrei ist, d. h. nicht von der Form von F abhängt.

Satz 7.3 *Für jede stetige Verteilungsfunktion $F : \mathbb{R} \to [0,1]$ gilt:*

$$D_n \stackrel{d}{=} \sup_y |\hat{G}_n(y) - y|,$$

wobei $\hat{G}_n$ die empirische Verteilungsfunktion einer beliebigen Stichprobe ist, die aus n unabhängigen und auf [0,1] gleichverteilten Variablen $Y_1, \ldots, Y_n$ besteht.

Beweis: Sei B die Vereinigung aller Konstanzbereiche von F. Dann gilt mit Wahrscheinlichkeit 1

$$D_n = \sup_{x \in B^c} |\hat{F}_n(x) - F(x)|.$$

Außerdem gilt

$$\{X_i \leq x\} = \{F(X_i) \leq F(x)\} \quad \text{für alle} \quad x \in B^c. \tag{7.4}$$

Wir setzen

$$Y_i = F(X_i) \quad \text{für jedes} \quad i = 1, \ldots, n.$$

Die $(Y_i)_{i=1}^n$ sind unabhängig und identisch verteilt. Weil F stetig ist, gibt es für jedes $y \in (0,1)$ ein $x_y \in \mathbb{R}$, so dass

$$x_y = \inf\{x' : F(x') = y\} \in B^c.$$

Folglich gilt für jedes $y \in (0,1)$

$$\mathbb{P}(Y_i \leq y) = \mathbb{P}(F(X_i) \leq F(x_y)) = \mathbb{P}(X_i \leq x_y) = F(x_y) = y,$$

wobei die zweite Gleichheit aus (7.4) folgt. Die Zufallsvariablen sind also unabhängig und auf [0,1] gleichverteilt. Wegen (7.4) gilt somit, dass $\hat{F}_n(x) = \hat{G}_n(F(x))$ für jedes $x \in B^c$. Hieraus folgt zusammen mit der Eingangsbemerkung

$$\begin{aligned} D_n &= \sup_{x \in B^c} |\hat{F}_n(x) - F(x)| \\ &= \sup_{x \in B^c} |\hat{G}_n(F(x)) - F(x)| \\ &= \sup_{x \in \mathbb{R}} |\hat{G}_n(F(x)) - F(x)| \\ &= \sup_{y \in [0,1]} |\hat{G}_n(y) - y|, \end{aligned}$$

wobei in der letzten Gleichheit erneut die Stetigkeitsvoraussetzung an F ausgenutzt wurde. □

Um nun die Hypothese

$$H : \mathbb{P} = \mathbb{P}_0 \quad \text{bzw.} \quad H : F = F_0$$

zu testen, verwenden wir die Teststatistik

$$T_n(x_1, \ldots, x_n) = \sqrt{n} \sup_{t \in \mathbb{R}} |\hat{F}_n(t; x_1, \ldots, x_n) - F_0(t)|.$$

Dieses ist die sogenannte Kolmogorov-Smirnov-Teststatistik. Sie hängt nicht von $\mathbb{P}_0$ ab, falls F_0 stetig ist. Sei also $s_{n,1-\alpha}$ das $(1-\alpha)$-Quantil der Verteilung von $T_n(X_1, \ldots, X_n)$ unter einer beliebigen stetigen Verteilungsfunktion F_0.

Der Kolmogorov-Smirnov-Test verwirft

$$H : \mathbb{P} = \mathbb{P}_0 \quad \text{gegen} \quad K : \mathbb{P} = \mathbb{P}_0$$

zum Niveau α, wenn

$$T_n(x_1, \ldots, x_n) > s_{n,1-\alpha}.$$

Bemerkung 7.4 *a) Die Quantile $s_{n,1-\alpha}$ lassen sich z. B. durch Simulationen (sogenannte Monte-Carlo-Simulationen) bestimmen. Hierfür verwendet man dann für F_0 die Gleichverteilung auf [0,1].*

b) Setzt man F_0 nicht als stetig voraus, so liefert das Testverfahren einen Test, dessen Niveau kleiner als α sein kann.

c) Wenn jedoch das Quantil $s'_{n,1-\alpha}$ von $T_n(X_1, \ldots, X_n)$ unter F_0 beispielsweise durch Simulationen bestimmt werden kann, so ist stets, also auch bei unstetigem F_0, der beschriebene Test ein Test zum Niveau α.

Will man die Quantile der Teststatistik nicht durch Simulation nähern, so kann man für große n versuchen, sie durch eine bekannte Verteilung zu approximieren. Wir stellen hierfür zunächst einige Hilfsmittel bereit.

Lemma 7.5 *Sei $m \in \mathbb{N}$ und seien $Z, Z_1, Z_2, \ldots : \Omega \to \mathbb{R}^m$ beliebige Zufallsvariablen mit den charakteristischen Funktionen φ_{z_n} und φ_z. Es gilt $Z_n \to Z$ in Verteilung genau dann, wenn*

$$\lim_{k\to\infty} \varphi_{z_n}(t) = \varphi_z(t) \quad \textit{für alle} \quad t \in \mathbb{R}^m$$

gilt.

Die eindimensionale Version dieses Satzes haben wir schon in der Wahrscheinlichkeitstheorie I bewiesen. Daher ersparen wir uns hier den Beweis. Außerdem benötigen wir die folgende mehrdimensionale Version des Zentralen Grenzwertsatzes, der aus Lemma 7.5 und dem 1-dimensionalen CLT folgt (auch ohne Beweis):

Satz 7.6 *Sei $m \in \mathbb{N}$ und $Z_1, Z_2, \ldots$ eine Folge von i.i.d. $\mathbb{R}^m$-wertigen Zufallsvariablen mit Erwartungswertvektor $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix}$ und Kovarianzmatrix K. Dann gilt*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{Z_1 + \ldots + Z_n - n\mu}{\sqrt{n}} \leq x \right) = \Phi_K(x)$$

für alle $x \in \mathbb{R}^m$. Hierbei ist $\Phi_K(x)$ die Verteilungsfunktion der n-dimensionalen Normalverteilung mit Erwartungswertvektor 0 und Kovarianzmatrix K.

Mithilfe dieses Satzes lässt sich nun eine Näherungsformel der Verteilungsfunktion von $T_n(X_1, \ldots, X_n)$ herleiten:

Satz 7.7 *Die Verteilungsfunktion $F_0 : \mathbb{R} \to [0,1]$ sei stetig. Unter der Hypothese*

$$H : \mathbb{P} = \mathbb{P}_0$$

gilt dann

$$\lim_{n\to\infty} \mathbb{P}(T_n(X_1, \ldots, X_n) \leq x) = K(x) \quad \textit{für alle} \quad x \in \mathbb{R},$$

wobei $K : \mathbb{R} \to [0,1]$ die Verteilungsfunktion der sogenannten Kolmogorov-Verteilung ist. Für diese gilt

$$K(x) = \begin{cases} 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2x^2) & \textit{für } x > 0 \\ 0 & \textit{für } x \leq 0 \end{cases}. \tag{7.5}$$

Beweisskizze: (Ausführlicher findet sich der Beweis im Buch von A. van der Vaart und J. Wellner (1996)).
Da die Verteilung von $T_n(X_1, \ldots, X_n) := T_n$ nicht von F_0 abhängt, können wir o. B. d. A. annehmen, dass F_0 die Verteilungsfunktion der Gleichverteilung auf [0,1] ist, also ist

$$F_0(t) = t \quad \text{für alle} \quad t \in [0,1].$$

Wir bezeichnen

$$B_n(t) = \sqrt{n}\left(\hat{F}_n(t; X_1, \ldots, X_n) - F_0(t)\right)$$

für alle $t \in [0,1]$. Die Familie der $\{B_n(t),\ t \in [0,1]\}$ ist ein stochastischer Prozess, der empirischer Prozess heißt. Für beliebige $t_1, \ldots, t_m \in [0,1]$ gilt dann

$$\sqrt{n}(B_n(t_1), \ldots, B_n(t_m)) = \sum_{i=1}^{n} (Y_i(t_1) - t_1, \ldots, Y_i(t_m) - t_m),$$

wobei

$$Y_i(t_j) = \begin{cases} 1, & \text{wenn } X_i \le t_j \\ 0, & \text{wenn } X_i > t_j \end{cases}.$$

Aus Satz 7.6 folgt

$$(B_n(t_1), \ldots, B_n(t_m)) \xrightarrow{d} (B(t_1), \ldots, B(t_m)),$$

wobei die $(B(t_1), \ldots, B(t_m))$ $\mathcal{N}(0, K)$-verteilt sind. Der Erwartungswert der Y_i ist nämlich t_i. Ihre Kovarianzmatrix K berechnet sich als

$$K = (\kappa_{ij}^2)$$

mit

$$\kappa_{ij}^2 = \min\{t_i, t_j\} - t_i t_j.$$

Hieraus ergibt sich

$$\max_{i=1,\ldots,m} \sqrt{n}\left|\hat{F}_n(t_i; X_1, \ldots, X_n) - F_0(t_i)\right| \xrightarrow{d} \max_{i=1,\ldots,m} |B(t_i)|.$$

Die Verteilungen des Zufallsvektors $(B(t_1), \ldots, B(t_m))$ sind die endlich-dimensionalen Verteilungen des sogenannten Brownschen Brückenprozesses $(B(t),\ t \in [0,1])$. Hierbei ist $B(t)$ definiert als

$$B(t) = X(t) - tX(1),$$

wobei $(X(t),\ t \in [0,1])$ eine Standard-Brownsche Bewegung ist. Mithilfe eines Straffheitsarguments wie im Satz von Donsker (oder eines Invarianzprinzips) zeigt man, dass sogar

$$(B_n(t),\ t \in [0,1]) \to (B(t),\ t \in [0,1])$$

bzw.

$$\max_{t\in[0,1]} \sqrt{n}\left|\hat{F}_n(t; X_1, \ldots, X_n) - F_0(t)\right| \xrightarrow{d} \max_{t\in[0,1]} |B(t)|$$

gilt. Außerdem kann man zeigen, dass die Verteilungsfunktion des Maximums $\max_{t\in[0,1]} |B(t)|$ der Brownschen Brücke durch (7.5) gegeben ist. Dies ist eine Übung. □

Bemerkung 7.8 *Wegen Satz 7.7 wird bei großem Stichprobenumfang (Faustregel: $n > 40$) die Hypothese*

$$H : F = F_0$$

abgelehnt, falls

$$T_n(x_1, \ldots, x_n) > \xi_{1-\alpha},$$

wobei $\xi_{1-\alpha}$ das $(1-\alpha)$-Quantil der in (7.5) definierten Kolmogorov-Verteilung ist, d. h. $\xi_{1-\alpha}$ löst

$$K(\xi_{1-\alpha}) = 1 - \alpha.$$

Wir untersuchen nun einige Güte-Eigenschaften des Kolmogorov-Smirnov-Tests.

Satz 7.9 *Die Verteilungsfunktion $F_0 : \mathbb{R} \to [0,1]$ sei stetig. Dann ist der Kolmogorov-Smirnov-Test punktweise konsistent für jede Verteilungsfunktion $F \neq F_0$ der Stichprobenvariablen, d. h. es gilt*

$$\lim_{n\to\infty} \mathbb{P}_F(T_n(X_1, \ldots, X_n) > s_{n,1-\alpha}) = 1.$$

Beweis: Aus dem Satz von Glivenko-Cantelli wissen wir, dass

$$\mathbb{P}_{F_0}(\lim_{n\to\infty} \sup_{t\in\mathbb{R}} |\hat{F}_n(t; X_1, \ldots, X_n) - F_0(t)| = 0) = 1,$$

d. h.

$$\mathbb{P}_F(\lim_{n\to\infty} \sup_{t\in\mathbb{R}} |\hat{F}_n(t, X_1, \ldots, X_n) - F_0(t)| > 0) = 1$$

für alle $F \neq F_0$ gilt. Also gilt mit Wahrscheinlichkeit 1

$$T_n(X_1, \ldots, X_n) \to \infty \quad \text{unter} \quad F \neq F_0.$$

Weiter gilt

$$s_{n,1-\alpha} \to \xi_{1-\alpha} < +\infty \quad \text{für} \quad n \to \infty,$$

wobei $\xi_{1-\alpha}$ das $(1-\alpha)$-Quantil der Kolmogorov-Verteilung ist. Also folgt

$$T_n(X_1, \ldots, X_n) - (s_{n,1-\alpha} - \xi_{1-\alpha}) \xrightarrow{f.s.} \infty,$$

also

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}_F(T_n(X_1, \ldots, X_n) > s_{n,1-\alpha}) &= \lim_{n\to\infty} \mathbb{P}_F(T_n - (s_{n,1-\alpha} - \xi_{1-\alpha} > \xi_{1-\alpha}) \\ &= \lim_{n\to\infty} \mathbb{P}_F(T_n > \xi_{1-\alpha}) = 1. \end{aligned}$$

□

Bemerkung 7.10 *Man kann in Verschärfung von Satz 7.9 sogar die* gleichmäßige *Konsistenz des Kolmogorov-Smirnov-Tests zeigen, d. h. man kann zeigen, dass, falls der* Kolmogorov-Abstand

$$d_K(\Delta_n, F_0) = \inf_{F \in \Delta_n} \sup_{t \in \mathbb{R}} |F(t) - F_0(t)|$$

zwischen der Familie Δ_n der alternativen Verteilungsfunktion und der Verteilungsfunktion F_0 nicht zu schnell gegen 0 konvergiert, gilt:

$$\lim_{n \to \infty} \inf_{F \in \Delta_n} \mathbb{P}_F(T_n(X_1, \ldots, X_n) > s_{n,1-\alpha}) = 1.$$

Umgekehrt kann man zeigen, dass für "kleine Kolmogorov-Abstände", d. h. falls für eine Folge von Verteilungsfunktionen (F_n)

$$\lim_{n \to \infty} \sqrt{n} d_K(F_n, F_0) = 0$$

gilt, auch

$$\limsup \mathbb{P}_{F_n}(T_n(X_1, \ldots, X_n) > s_{n,1-\alpha}) \leq \alpha$$

gilt. Die asymptotische Macht des Kolmogorov-Smirnov-Tests wird also beliebig klein.

7.2 Der χ^2-Anpassungstest

Wir betrachten nun einen asymptotischen Anpassungstest, wobei eine Testgröße betrachtet wird, die bei großem Stichprobenumfang näherungsweise χ^2-verteilt ist. Dabei wird jedoch im allgemeinen nicht die Hypothese

$$H : \mathbb{P} = \mathbb{P}_0 \quad \text{gegen} \quad K : \mathbb{P} \neq \mathbb{P}_0 \tag{7.6}$$

betrachtet, denn wir "vergröbern" das Modell der Zufallsstichprobe $(X_1, \ldots, X_n)$ durch Klassenbildung.

Für eine natürliche Zahl r zerlegen wir den Wertebereich der Zufallsvariablen $X_1, \ldots, X_m$ in r Klassen $(a_1, b_1], \ldots, (a_r, b_r]$ mit

$$-\infty \leq a_1 < b_1 = a_2 < b_2 = \ldots < \ldots = a_r < b_r \leq +\infty.$$

Anstelle der Stichprobe $X_1, \ldots, X_n$ betrachten wir die "Klassenstärke" $Z_1, \ldots, Z_r$, die Zufallsvariablen

$$Z_j = \{i : 1 \leq i \leq n : a_j < X_i \leq b_j\},$$

$j = 1, \ldots, r$. Offenbar gilt

Satz 7.11 *Der Zufallsvektor $(Z_1, \ldots, Z_r)$ ist multinomial-verteilt zu den Parametern n und $p = (p_1, \ldots, p_r)$ mit*

$$p_j = \mathbb{P}(a_j < X_1 \leq b_j)$$

für alle $j = 1, \ldots, r$, d. h.

$$\mathbb{P}(Z_1 = k_1, \ldots, Z_r = k_r) = \frac{n!}{k_1! \ldots k_r!} \cdot p_k^{k_1} \ldots p_r^{k_r}.$$

Bemerkung 7.12 *a) Wir bezeichnen die Multinomialverteilung mit den Parametern $n \geq 1$ und p mit $M_r(n,p)$, für $r = 2$ haben wir eine Binomial-Verteilung $B(n,p)$ mit $p = p_1$ und $1 - p = p_2$.*

b) Anstelle des Testproblems (7.6) prüfen wir die Hypothese

$$H : p = p_0 \quad \textit{gegen} \quad K : p \neq p_0$$

für einen vorgegebenen Vektor

$$p_0 = (p_{0_1}, \ldots, p_{0_r}) \quad \textit{mit} \quad \sum_{i=1}^{r-1} p_{0_i} < 1.$$

Dies bedeutet inhaltlich, dass wir die Familie Δ der insgesamt in Betracht gezogenen Verteilungen der Stichprobenvariablen $X_1, \ldots, X_n$ in die Teilmengen

$$\Delta_0 = \{Q : \mathbb{P}_Q(a_j < X_1 \leq b_j) = p_{0_j}, \textit{ für alle } j\} \quad \textit{bzw.} \quad \Delta_1 = \Delta \backslash \Delta_0$$

zerlegen.

Zu diesem Zweck betrachten wir die Stichprobenfunktion

$$T_n : \mathbb{R}^n \to [0, \infty)$$

mit

$$T_n(x_1, \ldots, x_n) = \sum_{j=1}^{r} \frac{1}{np_{0_j}} (Z_j(x_1, \ldots, x_n) - np_{0_j})^2, \tag{7.7}$$

wobei $Z_j(x_1, \ldots, x_n)$ die Anzahl derjenigen Stichprobenwerte $x_1, \ldots, x_n$ bzeichnet, die im Interall $(a_j, b_j]$ liegen.

Unter

$$H : p = p_0$$

gilt

$$\mathbb{E} Z_j(X_1, \ldots, X_n) = np_{0_j} \quad \text{für jedes} \quad j \in \{1, \ldots, r\}.$$

Es ist daher sinnvoll H abzulehnen, wenn $T_n(x_1, \ldots, x_n)$ signifikant größer als 0 ist. Um zu entscheiden, was "signifikant größer" bedeutet, müssen wir wissen, wie T_n in (7.7) verteilt ist. Hierzu zeigen wir, dass $T_n(X_1, \ldots, X_n)$ in Verteilung gegen die χ^2_{r-1}-Verteilung konvergiert, wenn $n \to \infty$ gilt. Dies ist die Grundlage des von Pearson eingeführten χ^2-Anpassungstests.

Satz 7.13 *Für jedes $\mathbb{P} \in \Delta_0$ gilt*

$$\mathbb{P}(T_n(X_1, \ldots, X_n) > \chi^2_{r-1,1-\alpha}) \to \alpha$$

für alle $\alpha \in (0,1)$, wenn $n \to \infty$ strebt. Hierbei ist $\chi^2_{r-1,1-\alpha}$ das $(1-\alpha)$-Quantil der χ^2_{r-1}-Verteilung.

Beweis: Wir haben schon gesehen, dass Z_n $M(n,p)$-verteilt ist, wobei $p = (p_{0_1}, \ldots, p_{0_r})$ und

$$p_{0_j} = \mathbb{P}_Q(a_j < X_1 \leq b_j).$$

Somit kann man für beliebige $i, j \in \{1, \ldots, r\}$ folgern:

$$\begin{aligned} &\mathbb{E}_Q Z_{n_i} = np_{0_i} \qquad \text{und} \\ &Cov(Z_{n_i}, Z_{n_j}) = \begin{cases} -np_{0_i}p_{0_j}, & \text{wenn } i \neq j \\ np_{0_i}(1-p_{0_i}), & \text{wenn } i = j \end{cases}. \end{aligned}$$

Außerdem gilt

$$Z_{n_j} = \sum_{i=1}^{n} \mathbb{1}_{\{a_j < X_i \leq b_j\}},$$

d. h. Z_n ist eine Summe von n unabhängigen identisch verteilten Zufallsvariablen. Schreiben wir

$$Z'_n = \left(\frac{Z_{n_1}}{\sqrt{n}} - \sqrt{n}p_{0_1}, \ldots, \frac{Z_{n,r-1}}{\sqrt{n}} - \sqrt{n}p_{0,r-1} \right)$$

(die letzte Koordinate von Z_n spielt eine besondere Rolle, da sie sich zwangsläufig aus den anderen ergibt), so folgt mit dem Zentralen Grenzwertsatz Satz 7.7:

$$Z'_n \to Z' \sim \mathcal{N}(0, K).$$

Hierbei ist Z' eine $(r-1)$-dimensionale Zufallsvariable, die einer $(r-1)$-dimensionalen Normalverteilung mit Erwartungsverteilungsvektor 0 und Kovarianzmatrix K mit $K = (\kappa^2_{ij})_{i,j=1}^{r-1}$

$$\kappa^2_{ij} = \begin{cases} -p_{0_i}p_{0_j}, & \text{falls } i \neq j \\ p_{0_i}(1-p_{0_i}), & \text{falls } i = j \end{cases}$$

genügt. Man sieht, dass K invertierbar ist und dass für $A = K^{-1}$ gilt: $A = (a_{ij})_{i,j=1}^{r-1}$

$$a_{ij} = \begin{cases} \frac{1}{p_{0_r}}, & \text{wenn } i \neq j \\ \frac{1}{p_{0_i}} + \frac{1}{p_{0_r}}, & \text{wenn } i = j \end{cases}$$

(nachrechnen).

Da lineare Transformationen stetig sind und Normalverteilungen erhalten, ergibt sich somit aus dem bisher Gesagten

$$A^{1/2} Z'_n \to \mathcal{N}(0, I_{r-1}),$$

wobei I_{r-1} die $(r-1) \times (r-1)$-Einheitsmatrix ist. Somit ist

$$(A^{1/2} Z'_n)^t (A^{1/2} Z'_n)$$

asymptotisch für große n eine Summe von $r-1$ Quadraten von i.i.d. $\mathcal{N}(0,1)$-verteilten Zufallsvariablen, also

$$(A^{1/2} Z'_n)^t (A^{1/2} Z'_n) \xrightarrow{d} \chi^2_{r-1}.$$

Nun ist aber

$$\begin{aligned}
(A^{1/2}Z_n')^t(A^{1/2}Z_n') &= (Z_n')AZ_n' \\
&= n\sum_{j=1}^{r-1}\frac{1}{p_{0_j}}\left(\frac{Z_{nj}}{n}-p_{0_j}\right)^2+\frac{n}{p_{0_r}}\sum_{i=1}^{r-1}\sum_{j=1}^{r-1}\left(\frac{Z_{n_i}}{n}-p_{0_i}\right)\left(\frac{Z_{n_j}}{n}-p_{0_j}\right) \\
&= n\sum_{j=1}^{r-1}\frac{1}{p_{0_j}}\left(\frac{Z_{n_j}}{n}-p_{0_j}\right)^2+\frac{n}{p_{0_r}}\left(\sum_{j=1}^{r-1}\left(\frac{Z_{n_j}}{n}-p_{0_j}\right)\right)^2 \\
&= n\sum_{j=1}^{r-1}\frac{1}{p_{0_j}}\left(\frac{Z_{n_j}}{n}-p_{0_j}\right)^2+\frac{n}{p_{0_r}}\left(\frac{Z_{nr}}{n}-p_{0_r}\right)^2,
\end{aligned}$$

denn offenbar gilt

$$\sum_{j=1}^{r-1}Z_{n_j}=n-Z_{nr}\quad\text{und}\quad\sum_{j=1}^{r-1}p_{0_j}=1-p_{0_r}.$$

Somit ist

$$(A^{1/2}Z_n')^t(A^{1/2}Z_n')=T_n(X_1,\ldots,X_n).$$

Dies impliziert die Behauptung. □

Bemerkung 7.14 *Bei der praktischen Durchführung des χ^2-Anpassungstests zur Prüfung der Hypothese*

$$H: p=p_0$$

ist zunächst die Testgröße $T_n(x_1,\ldots,x_n)$ zu berechnen. Bei hinreichend großem n wird H abgelehnt, wenn

$$T_n(x_1,\ldots,x_n)>\chi^2_{r-1,1-\alpha},$$

wobei $\chi^2_{r-1,1-\alpha}$ das $(1-\alpha)$-Quantil der χ^2_{r-1}-Verteilung ist. Eine "Faustregel" dafür, dass n hinreichend groß ist, ist die Gültigkeit der Ungleichung

$$np_{0,j}\geq a\quad\textit{für alle}\quad j\in\{1,\ldots,r\}$$

und eine Konstante $a>0$. Über die Größe von a gibt es verschiedene Auffassungen in der Literatur, die zwischen $a=2$ und $a=10$ variieren.

Um die Güte des beschriebenen Tests zu diskutieren, zeigen wir den folgenden Satz, der die punktweise Konsistenz des χ^2-Anpassungstests zeigt.

Satz 7.15 *Der χ^2-Anpassungstest ist punktweise konsistent gegen jeden Vektor $p=(p_1,\ldots,p_{r-1})$ mit $p\neq p_0$, d. h. es gilt:*

$$\lim_{n\to\infty}\mathbb{P}_p(T_n(X_1,\ldots,X_n)>\chi^2_{r-1,1-\alpha})=1.$$

Beweis: Ist $p \neq p_0$, so gibt es zumindest ein $j \in \{1, \ldots, r-1\}$ mit

$$p_j \neq p_{0_j}.$$

Das Starke Gesetz der großen Zahlen impliziert, dass für jedes j gilt

$$\frac{Z_{nj}}{n} \to p_j \quad \text{für} \quad n \to \infty \quad \text{und} \quad \mathbb{P}_p\text{-f.s.}$$

Zusammen ergibt dies, dass unter $\mathbb{P}_p$ gilt

$$T_n(X_1, \ldots, X_n) \geq n \left(\frac{Z_{nj}}{n} - p_{0_i} \right)^2 \to \infty$$

$\mathbb{P}_p$-f.s. Dies zeigt den Satz. □