Mathematische Statistik

von Peter Pfaffelhuber

Version: 10. Februar 2016

WARNUNG: Dieses Skript enthält vermutlich noch viele Fehler. Es wurde teilweise in Eile geschrieben. Für alle Fehler bin ich selbst verantwortlich. Ich hoffe, in Zukunft eine Version mit weniger Fehlern bereit stellen zu können.

Inhaltsverzeichnis

1 Einleitung

Die Mathematische Statistik ist ein eher theoretisches Teilgebiet der Stochastik. Im Gegensatz zur angewandten Statistik, die sich der Methodenentwicklung für Datenanalyse verschrieben hat, geht es in der theoretischen Statistik darum, Eigenschaften solcher Methoden festzustellen, etwa Optimalitätskriterien für Schätzer und Tests. In diesem Kurzskript soll überblicksartig ein kurzer Einblick in dieses Gebiet gegeben werden. Komplettiert wird es (neben den Übungen) durch die Vorstellung statistischer Methoden, die an anderer Stelle im Rahmen dieser Vorlesung zusammengefasst vorgestellt werden.

1.1 Wiederholungen aus der Wahrscheinlichkeitstheorie

Wir setzen Kenntnisse aus den Vorlesungen *Stochastik I*, *Stochastik II* und *Wahrscheinlichkeitstheorie* voraus. Zur Sicherheit jedoch wiederholen wir einige Begriffe, die im Folgenden unerlässlich sein werden. Alle Räume in diesem Skript (z.B. $E, E', E'', \aleph, ...$) seien vollständige und separable metrische Räume, wenn nicht anders angegeben. Im Folgenden sei $(E, \mathcal{A} = \mathcal{B}(E), \mathbb{P})$ ein Wahrscheinlichkeitsraum.

Bemerkung 1.1 (Maß mit Dichte). Das Bildmaß einer Zufallsvariable X, bezeichnet mit $X_*\mathbb{P}$, ist gegeben als $X_*\mathbb{P}(A) = \mathbb{P}(X \in A), A \in \mathcal{B}(E)$. Es hat Dichte p bezüglich λ^n (dem n-dimensionalen Lebesgue-Maß), falls für alle $A \in \mathcal{B}(\mathbb{R})$

$$\mathbb{P}(X \in A) = \int 1_A(x)p(x)\lambda^n(dx).$$

In diesem Fall gilt dann für $f : E \to \mathbb{R}$

$$\mathbb{E}[f(X)] = \int f(x)p(x)\lambda^n(dx),$$

falls eine der beiden Seiten existiert.

Bemerkung 1.2 (Unabhängigkeit, Messbarkeit). Zufallsvariablen X, Y sind (bezüglich $\mathbb{P}$) unabhängig, wenn $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$ für alle $A, B \in \mathcal{B}(E)$. Wir schreiben dann auch $X \perp_{\mathbb{P}} Y$.
Seien X, T Zufallsvariablen. Wir erinnern daran, dass $\sigma(T) \subseteq \mathcal{A}$ die von T erzeugte σ-Algebra ist. Nach einem Satz aus der Wahrscheinlichkeitstheorie ist genau dann X messbar bezüglich $\sigma(T)$ oder T-messbar, falls es eine Funktion g gibt mit $X = g(T)$.

Bemerkung 1.3 (Bedingte Erwartung, bedingte Verteilung). Sei $(E, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, X eine integrierbare Zufallsvariable und $\mathcal{G} \subseteq \mathcal{F}$ eine (Teil-)σ-Algebra. Die bedingte Erwartung von X gegeben $\mathcal{G}$ (bezeichnet mit $\mathbb{E}[X|\mathcal{G}]$) ist die einzige $\mathcal{G}$-messbare Zufallsvariable, für die

$$\mathbb{E}[X, G] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}], G]$$

für alle $G \in \mathcal{G}$ gilt.
Wir erinnern an zwei Eigenschaften der bedingten Erwartung: Für eine weitere σ-Algebra $\mathcal{H} \subseteq \mathcal{G}$ gilt

$$\mathbb{E}[X|\mathcal{H}] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}].$$

Für eine weitere reellwertige Zufallsvariable Y gilt

$$\mathbb{E}\big[X\mathbb{E}[Y|\mathcal{G}]\big] = \mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]Y\big] = \mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\mathbb{E}[Y|\mathcal{G}]\big], \tag{1.1}$$

falls alle Erwartungen existieren.

Wir definieren weiterhin $\mathbb{P}(A|\mathcal{H}) := \mathbb{E}[1_A|\mathcal{H}]$ für $A \in \mathcal{F}$. Angemerkt sei, dass es sich bei $A \mapsto \mathbb{P}(A|\mathcal{H})$ zunächst nicht notwendigerweise um ein Wahrscheinlichkeitsmaß handelt, da bedingte Erwartungen zunächst nur $\mathbb{P}$-fast sicher definiert sind, es also immer Ausnahme-Nullmengen geben kann. Ist jedoch E ein vollständiger und separabler metrischer Raum (was wir hier annehmen werden), so existiert (nach einem Satz aus der Wahrscheinlichkeitstheorie) die *reguläre Version der bedingten Verteilung*, d.h. ein stochastischer Kern κ von E nach E, so dass für $\mathbb{P}$-f.a. $\omega \in E$

$$\kappa(\omega, B) = \mathbb{P}(X \in B|\mathcal{G})(\omega).$$

Bemerkung 1.4 (Varianzzerlegung). Neben der bedingten Erwartung kann man auch die bedingte Varianz

$$\mathbb{V}[X|\mathcal{G}] := \mathbb{E}[(X - \mathbb{E}[X|\mathcal{G}])^2|\mathcal{G}] = \mathbb{E}[X^2|\mathcal{G}] - \mathbb{E}[X|\mathcal{G}]^2$$

definieren. Es gilt dann die Varianzzerlegung

$$\mathbb{V}[X] = \mathbb{E}[\mathbb{V}[X|\mathcal{G}]] + \mathbb{V}[\mathbb{E}[X|\mathcal{G}]]. \tag{1.2}$$

Bemerkung 1.5 (Notation). Für $x, y \in \mathbb{R}^n$ ist $x^\top y$ das euklidische Skalarprodukt. Allgemeiner bezeichnen wir für $A \in \mathbb{R}^{m\times n}$ und $x \in \mathbb{R}^n$ die Matrixmultiplikation mit $A\,\underline{x}$ und mit $A^\top$ die Transponierte von A.
Wir bezeichnen das n-dimensionale Lebesgue-Maß (auf $\mathbb{R}^n$) mit λ^n, und zur Vereinfachung der Notation n-dimensionale Zählmaß ebenfalls mit λ^n (also ist in diesem Fall $\lambda^n = \sum_{\underline{x}\in\mathbb{Z}^n} \delta_x$). Das n-dimensionale Produktmaß eines Maßes μ bezeichnen wir im Allgemeinen mit μ^n. Etwa ist für die Standardnormalverteilung $\mathcal{N}(0,1)$ die Verteilung einer unabhängigen Stichprobe gerade $\mathcal{N}(0,1)^n$.
Für die vollständigen, separablen metrische Räume (D, r_D), (E, r_E), ... seien $\mathcal{B}(D), \mathcal{B}(E), ...$ die Borel'schen σ-Algebren.
Hat μ die Dichte f bezüglich ν, so schreiben wir $\mu = f\cdot\nu$. Das Dirac-Maß auf $x \in E$ bezeichnen wir mit δ_x. Das Bildmaß von X unter μ bezeichnen wir mit $X_*\mu$.

1.2 Beispiele

Als theoretische Wissenschaft mit Anwendungsbezügen lebt die Statistik von guten Beispielen, anhand denen man die zu entwickelnde Theorie ausprobieren kann. Auch wenn im Verlauf des Skriptes noch weitere Beispiele auftreten werden, sammeln wir hier drei besonders wichtige, die wir zunächst ohne großen Formalismus vorstellen.

Beispiel 1.6 (Beispiel *Bern*). Ein Bernoulli-Experiment besteht aus einer (endlich oder unendlich oft) unabhängig wiederholten Durchführung eines Zufallsexperiments, in dem jede Durchführung entweder einen *Erfolg* oder einen *Misserfolg* liefert. Es wird beschrieben durch einen Zufallsvektor $X = (X_1, X_2, ...)$ und eine Wahrscheinlichkeitsverteilung $\mathbb{P}_\theta$, so dass für $x_i \in \{0,1\}, i = 1, 2, ...$

$$\mathbb{P}_\theta((X_1, ..., X_n) = (x_1, ..., x_n)) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \qquad (Bern\ 0)$$

für ein $\theta \in [0,1]$ gilt. Hierbei ist θ die Wahrscheinlichkeit eines Erfolges (in jeder der Durchführungen). Bei einer gegebenen Folge von Erfolgen und Misserfolgen $x_1, ..., x_n$ wären naheliegende Fragen etwa:

- Wie groß ist θ?
- Ist $\theta = \frac{1}{2}$?

Beispiel 1.7 (Beispiel *Norm*). Der zentrale Grenzwertsatz betont die Bedeutung der Normalverteilung. Für statistische Fragestellungen bedeutet dies, dass man – zumindest approximativ – oft eine Stichprobe von Daten $X_1, ..., X_n$ erhebt, die unabhängig und normalverteilt sind. Da man typischerweise die Stichprobe aus der gleichen Grundgesamtheit zieht, sollten hierbei die Erwartungswerte und Varianzen gleich sein. Hier ist also $X = (X_1, ..., X_n)$ und $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$ hat für $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$ die Dichte

$$p_{(\mu,\sigma^2)}(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x_i - \mu)^2}{\sigma} \Big) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(- \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma} \Big) \quad (Norm\ 0)$$

bezüglich des Lebesgue-Maßes in $\mathbb{R}^n$. Bei einer Stichprobe $X_1, ..., X_n$ könnte man etwa fragen:

- Wie groß ist μ, wie groß ist σ^2?
- Ist $\mu = \mu_0$ für einen vorgegebenen Wert μ_0, wenn man weiß, wie groß σ^2 ist?
- Ist $\mu = \mu_0$ für einen vorgegebenen Wert μ_0, wenn man nicht weiß, wie groß σ^2 ist?

Beispiel 1.8 (Beispiel *Unif*). Etwas pathologischer klingt folgendes Beispiel: Wir nehmen an, dass Daten $X = (X_1, ..., X_n)$ unabhängig uniform auf $[0, \theta]$ verteilt sind (für ein $\theta \in \mathbb{R}_+$). Das heißt, dass $\mathbb{P}_\theta$ so ist, dass $X_* \mathbb{P}_\theta = p_\theta \cdot \lambda^n$ mit

$$p_\theta(x) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{0 \le x_i \le \theta} = \frac{1}{\theta^n} \mathbb{1}_{\max_{i=1,...,n} x_i \le \theta}$$

für $x \in \mathbb{R}_+^n$. Hier könnte man etwa fragen:

- Wie groß ist θ?

Für alle Beispiele kann man also sagen, dass *Daten* X erhoben wurden, deren Verteilung von einem Parameter θ abhängen. Dies führt zur ersten Definition.

Definition 1.9 (Statistisches Modell). *Sei $(\Omega, \mathcal{A})$ ein Messraum.*

1. *Ein* statistisches Modell (auf E) *ist ein Paar $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$, wobei $X : \Omega \to E$ messbar und $\{\mathbb{P}_\theta : \theta \in \mathcal{P}\}$ eine Familie von Wahrscheinlichkeitsmaßen auf $\mathcal{A}$ ist. Hierbei heißt $\mathcal{P}$ auch der* Parameterraum*.*
2. *Das statistische Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ heißt* identifizierbar*, wenn $\theta = \theta'$ genau dann, wenn $X_* \mathbb{P}_\theta = X_* \mathbb{P}_{\theta'}$ gilt. (Im weiteren Verlauf werden wir immer annehmen dass statistische Modelle diese Eigenschaft haben.)*

3. *Weiter heißt das statistische Modell* $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ regulär*, falls* $E = \mathbb{R}^n$ *für ein* n *und für alle* $\theta \in \mathcal{P}$

$$X_* \mathbb{P}_\theta = p_\theta \cdot \lambda^n$$

für eine geeignete Dichtefunktion p_θ *bzw. eine geeignete Zähldichte* p_θ *gilt. (Wir erinnern daran, dass wir* λ^n *sowohl für das Lebesgue-Maß in* $\mathbb{R}^n$ *oder das Zählmaß auf* $\mathbb{Z}^n$ *verwenden.) Im ersten Fall sprechen wir von einem regulären stetigen Modell, im zweiten Fall von einem regulären, diskreten Modell.*

Bemerkung 1.10 (Parametrische und nicht-paramatrische statistische Modelle). Ist $\mathcal{P} \subseteq \mathbb{R}^k$ für ein k, so spricht man oft von parametrischen Modellen. Die Idee ist, dass $\mathbb{P}_\theta$ von einem Vektor θ von verschiedenen Parametern abhängt, etwa Mittelwert und Varianz einer Normalverteilung. In allen anderen Fällen spricht man von nicht-parametrischen Modellen. Im allgemeinsten Fall ist $\mathcal{P}$ die Menge aller Wahrscheinlichkeitsmaße auf $\mathcal{B}(\mathbb{R}^n)$.

Beispiel 1.11 (Beispiel *Norm*). Im Fall von Beispiel 1.7 setzen wir $\theta = (\mu, \sigma^2)$, also

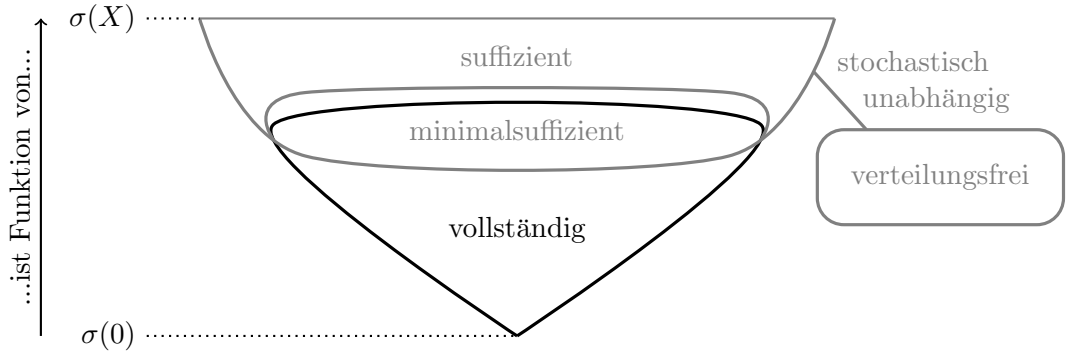
$$(X, \{\mathbb{P}_{\theta=(\mu,\sigma^2)} = \mathcal{N}(\mu, \sigma^2)^n : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}). \tag{Norm 1a}$$

Zu dieser Situation sagen wir auch, dass sowohl μ als auch σ^2 unbekannt sind. In einigen Situation werden wir annehmen, dass wir etwa σ^2 bereits kennen (aber μ nicht). Dann setzen wir für dieses σ^2 das statistische Modell

$$(X, \{\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)^n : \theta \in \mathbb{R}\}). \tag{Norm 1b}$$

2 Grundlegende Konzepte

In diesem und dem nächsten Abschnitt führen wir die Konzepte Suffizienz, Minimalsuffizienz, Vollständigkeit und Verteilungsfreiheit von Statistiken ein. (Eine Statistik ist dabei einfach eine $\sigma(X)$-messbare Zufallsvariable, d.h. für eine Abbildung t gilt $T = t(X)$.) Folgende Grafik illustriert kurz die Zusammenhänge.



Die wichtigsten Sätze des Abschnittes sind der Satz von Bahadur, Theorem 2.16, der besagt, dass suffiziente, vollständige Statistiken minimalsuffizient sind. Weiter besagt der Satz von Basu, Theorem 2.19, dass verteilungsfreie und suffiziente Statistiken unabhängig sind.

2.1 Suffizienz

Suffiziente Statistiken sind solche, die alle (zu statistischen Zwecken) nötigen Informationen über die Daten enthält.

Definition 2.1 (Suffiziente Statistik). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell auf E und $T = t(X)$ für $t : E \to E'$ für einen vollständigen metrischen Raum (E', r') messbar. Dann heißt T* suffizient*, falls unter $\mathbb{P}_\theta$ eine reguläre Version der bedingten Verteilung von X gegeben T existiert, die nicht von θ abhängt. Insbesondere hängt also für $A \in \mathcal{B}(E)$ die bedingte Erwartung $\mathbb{P}_\theta(X \in A|T)$ nicht von θ ab.*

In jedem statistischen Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ist X suffizient. Weitere Beispiele für Suffizienz lassen sich mit Hilfe des Fisher-Neyman'schen Faktorisierungssatzes ausmachen, siehe Theorem 2.5.

Bemerkung 2.2 (Umformulierung).

1. Im Falle eines regulären, diskreten statistischen Modells bedeutet die bedingte Unabhängigkeit von X gegeben $T = t(X)$ (unter $\mathbb{P}_\theta$) gerade, dass (für $t = t(x)$)

$$\mathbb{P}_\theta(X = x|T = t) = \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta(t(X) = t)} = \frac{\mathbb{P}_\theta(X = x)}{\sum_{y:t(y)=t} \mathbb{P}_\theta(X = y)} = \frac{p_\theta(x)}{\sum_{y:t(y)=t} p_\theta(y)}$$

   nicht von θ abhängt. (Im Fall $t \neq t(x)$ ist natürlich $\mathbb{P}_\theta(X = x|t(X) = t) = 0$.)

2. Die reguläre Version der bedingten Verteilung von X gegeben T ist (unter $\mathbb{P}_\theta$) der $\mathbb{P}_\theta$-fast sicher eindeutige stochastische Kern $\kappa_{X,T}$ (von E nach E'), so dass

$$\kappa_{X,T,\theta}(\omega, A) = \mathbb{P}_\theta(X \in A|T)(\omega)$$

   $\mathbb{P}_\theta$-f.s. gilt; siehe auch Bemerkung 1.3. Diese existiert für vollständige und separable metrische Räume (E, r_E) nach einem Satz aus der Wahrscheinlichkeitstheorie. Diese reguläre Version der bedingten Verteilung ist dabei durch die Gleichung

$$\mathbb{E}_\theta[\mathbb{P}_\theta(X \in A|T), X \in B] = \mathbb{P}_\theta(X \in A \cap B)$$

   für $A \in \mathcal{B}(E)$ und $B \in \sigma(T) = t^{-1}(\mathcal{B}(E'))$ eindeutig definiert. Genauer ist $\mathbb{P}_\theta(X \in A|T)$ die $\mathbb{P}_\theta$-f.s. eindeutige, T-messbare Zufallsvariable, die diese Gleichung erfüllt.

Beispiel 2.3 (Beispiel *Bern*). Da das Beispiel 1.6 diskret ist, können wir hier nachrechnen, dass $T := t(X) := \sum_{i=1}^n X_i$ suffizient ist.
Beweis siehe Übung.

Zunächst erscheint es schwierig, suffiziente Statistiken auszumachen. Allerdings geben wir mit Theorem 2.5 eine einfache Charakterisierung von suffizienten Statistiken in Termen der Dichte von X. Erst einmal benötigen wir jedoch ein Lemma.

Lemma 2.4 (Vorbereitung des Fisher-Neyman'schen Faktorisierungssatzes). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein reguläres, stetiges statistisches Modell mit $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$, sowie $t : E \to E'$ messbar und $T = t(X)$. Dann gilt:*

1. *Es gibt $c_1, c_2, ... \geq 0$ mit $\sum_{i=1}^\infty c_i = 1$ und $\theta_1, \theta_2, ... \in \mathcal{P}$, so dass $\mathbb{P}_\theta \ll \nu^* := \sum_{i=1}^\infty c_i \mathbb{P}_{\theta_i}$ für alle $\theta \in \mathcal{P}$.*

2. *Ist T suffizient und $A \in \mathcal{B}(E)$, so gilt $\nu^*(X \in A|T) \overset{\nu^*-f.s.}{=} \mathbb{P}_\theta(X \in A|T)$ für alle $\theta \in \mathcal{P}$.*

3. *Gilt $p_\theta(x) = h(x)g_\theta(t(x))$ für ($\mathbb{P}_\theta$-f.a.) $x \in E, \theta \in \mathcal{P}$, dann gibt es eine T-messbare Version von $\frac{d\mathbb{P}_\theta}{d\nu^*}$, nämlich*
$$\frac{d\mathbb{P}_\theta}{d\nu^*} = \frac{g_\theta(t(x))}{\sum_{i=1}^\infty c_i g_{\theta_i}(t(x))}.$$

4. *Falls $\frac{d\mathbb{P}_\theta}{d\nu^*}$ für alle $\theta \in \mathcal{P}$ nur von $t(x)$ abhängt, so ist $T = t(X)$ suffizient.*

Beweis. 1. Sei $P = p \cdot \lambda^n$ ein Wahrscheinlichkeitsmaß auf $\mathbb{R}^n$ mit $p > 0$ (etwa eine n-dimensionale, nicht-entartete Normalverteilung). Sei weiter

$$\mathcal{Q} = \Big\{ \sum_{i=1}^\infty c_i \mathbb{P}_{\theta_i} \text{ für } \theta_1, \theta_2, ... \in \mathcal{P} \text{ und } c_1, c_2, ... \geq 0 \text{ mit } \sum_{i=1}^\infty c_i = 1 \Big\}$$

die Menge der Konvexkombinationen aus $\{\mathbb{P}_\theta : \theta \in \mathcal{P}\}$. Wir setzen

$$\mathcal{C} := \Big\{ C \in \mathcal{B}(E) : \exists Q \in \mathcal{Q} : Q(C) > 0 \text{ und } \frac{dQ}{dP}|_C > 0 \Big\}.$$

Dann ist $\mathcal{C}$ nicht leer und $0 < \sup_{C \in \mathcal{C}} P(C) =: c \leq 1$. Wähle nun $C_1, C_2, ... \in \mathcal{C}$ mit $\sup_i P(C_i) = c$ und $D := \bigcup_i C_i$ sowie $Q_1, Q_2, ... \in \mathcal{Q}$ so, dass $Q_n(C_n) > 0$ und $\frac{dQ_n}{dP}|_{C_n} > 0$. Wir wählen
$$\nu^* = \sum_{i=1}^\infty 2^{-i} Q_i \in \mathcal{Q}.$$

Dann gilt sowohl $\nu^*(D) > 0$ als auch $d\nu^*/dP|_D = \sum_{i=1}^\infty 2^{-i} dQ_n/dP|_D > 0$ und damit $D \in \mathcal{C}$. Wir müssen zeigen, dass für $\theta \in \mathcal{P}$ aus $\nu^*(A) = 0$ folgt, dass $\mathbb{P}_\theta(A) = 0$. Wir setzen $C := \{d\mathbb{P}_\theta/dP > 0\}$ und schreiben

$$\begin{aligned}\mathbb{P}_\theta(A) &= \mathbb{P}_\theta(A \cap D) + \mathbb{P}_\theta(A \cap D^c \cap C^c) + \mathbb{P}_\theta(A \cap D^c \cap C) \\ &\leq \int_{A \cap D} \frac{d\mathbb{P}_\theta}{dP} \cdot \Big(\frac{d\nu^*}{dP}\Big)^{-1} d\nu^* + \mathbb{P}_\theta(C^c) + \mathbb{P}_\theta(C \cap D^c).\end{aligned}$$

Die ersten beiden Summanden verschwinden. Angenommen, $\mathbb{P}_\theta(C \cap D^c) > 0$, dann wäre $C \cup D \in \mathcal{C}$ und $P(C \cup D) = P(D) + P(C \cap D^c) > P(D)$ im Widerspruch zur Maximalität von $P(D)$. Damit folgt $\mathbb{P}_\theta(A) = 0$.
2. Sei T suffizient. Für $A \in \mathcal{B}(E)$ hängt $\mathbb{P}_\theta(X \in A|T)$ ($\mathbb{P}_\theta$-f.s.) nicht von θ ab, es gibt also eine Funktion ψ mit $\mathbb{P}_\theta(X \in A|T) \overset{\mathbb{P}_\theta-\text{fs}}{=} \psi(A, T)$ für alle θ. Deshalb gilt auch

$$\nu^*(X \in A|T) = \sum_{i=1}^\infty c_i \mathbb{P}_{\theta_i}(X \in A|T) \overset{\nu^*-\text{fs}}{=} \psi(A, T) \overset{\mathbb{P}_\theta-\text{fs}}{=} \mathbb{P}_\theta(X \in A|T).$$

Die Behauptung folgt nun auch ν^*-f.s., wenn man ψ auf $\mathbb{P}_\theta$-Nullmengen geeignet definiert.
3. Es gilt
$$\frac{d\nu^*}{d\lambda^n} = \sum_{i=1}^\infty c_i \frac{d\mathbb{P}_{\theta_i}}{d\lambda^n} = \sum_{i=1}^\infty c_i h(x) g_{\theta_i}(t(x))$$

und weiter, da $\mathbb{P}_\theta \ll \nu^*$ für alle $\theta \in \mathcal{P}$ (d.h. falls $\nu^*(A) = 0$, so ist zwar $d\nu^*/d\lambda^n$ auf A nicht invertierbar, aber es gilt $d\mathbb{P}_\theta/d\lambda^n = 0$) und

$$\frac{d\mathbb{P}_\theta}{d\nu^*} = \frac{d\mathbb{P}_\theta}{d\lambda^n}\Big(\frac{d\nu^*}{d\lambda^n}\Big)^{-1} = \frac{g_\theta(t(x))}{\sum_{i=1}^\infty c_i g_{\theta_i}(t(x))}.$$

Damit ist $\frac{d\mathbb{P}_\theta}{d\nu^*}$ eine Funktion von $t(x)$ und die Behauptung folgt.
4. Für $A \in \mathcal{B}$, $B \in \sigma(T)$ zeigen wir

$$\mathbb{E}_\theta[\mathbb{P}_\theta(X \in A|T), X \in B] = \mathbb{E}_\theta[\mathbb{P}_{\nu^*}(X \in A|T), X \in B].$$

Dann ist nämlich $\mathbb{P}_\theta(X \in A|T)$ ($\mathbb{P}_\theta$-fs) unabhängig von θ und T ist suffizient. Wir schreiben, da $\frac{d\mathbb{P}_\theta}{d\nu^*}$ nach Voraussetzung messbar bezüglich $\sigma(T)$ ist,

$$\begin{aligned}\mathbb{E}_\theta[\mathbb{P}_\theta(X \in A|T), X \in B] &= \mathbb{E}_\theta[1_{X\in A}1_{X\in B}] = \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}, X \in A \cap B\Big] \\ &= \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}\mathbb{P}_{\nu^*}(X \in A|T), X \in B\Big] \\ &= \mathbb{E}_\theta[\mathbb{P}_{\nu^*}(X \in A|T), X \in B]\end{aligned}$$

und die Behauptung ist gezeigt. □

Theorem 2.5 (Fisher-Neyman'scher Faktorisierungssatz). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein reguläres statistisches Modell und $T = t(X)$ mit $t : \mathbb{R}^n \to \mathbb{R}^m$ messbar. Dann sind äquivalent:*

1. *T ist suffizient,*
2. *Es gibt $g_\theta : \mathbb{R}^m \to \mathbb{R}$ und $h : \mathbb{R}^n \to \mathbb{R}$, so dass*

$$p_\theta(x) = g_\theta(t(x))h(x).$$

Bemerkung 2.6 (Diskreter Fall). Im diskreten Fall ist der Beweis einfach. '2. ⇒ 1.': Nach Bemerkung 2.2 gilt $\mathbb{P}_\theta(X = x|t(X) = t) = 0$ für $t \neq t(x)$, was unabhängig von θ ist. Für $t = t(x)$ hingegen ist unter 2.

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

was ebenfalls unabhängig von θ ist. Für '1. ⇒ 2.' setzen wir

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

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

$$p_\theta(x) = \mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(X = x, t(X) = t(x)) = h(x)g_\theta(t(x))$$

und die Behauptung ist gezeigt.

Beweis im stetigen Fall. '1. $\Rightarrow$ 2.': Sei ν^* wie in Lemma 2.4.1. Nach Lemma 2.4.2 gilt $\mathbb{P}_\theta(X \in A|T) = \nu^*(X \in A|T)$ (ν^*-f.s.) für alle $\theta \in \mathcal{P}$ und für $A \in \mathcal{B}(E)$. Wir schreiben nun mittels (1.1)

$$\begin{aligned}\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}, X \in A\Big] &= \mathbb{P}_\theta(X \in A) = \mathbb{E}_\theta[\mathbb{P}_\theta(X \in A|T)] \\ &= \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}\nu^*(X \in A|T)\Big] \\ &= \mathbb{E}_{\nu^*}\Big[\nu^*(X \in A|T)\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}\Big|T\Big]\Big] \\ &= \mathbb{E}_{\nu^*}\Big[\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}\Big|T\Big], X \in A\Big]\end{aligned}$$

wegen (1.1). Da A beliebig war und $T = t(X)$, gilt

$$\frac{d\mathbb{P}_\theta}{d\nu^*} = \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}\Big|t(X)\Big] =: g_\theta(t(X))$$

und mit $h := \frac{d\nu^*}{d\lambda^n}$ gilt

$$p_\theta(x) = \frac{d\mathbb{P}_\theta}{d\nu^*}\frac{d\nu^*}{d\lambda^n} = g_\theta(t(x))h(x).$$

'2. $\Rightarrow$ 1.' Dies ist eine Folgerung aus Lemma 2.4.3 und Lemma 2.4.4. □

Beispiel 2.7 (Beispiel *Norm*). Bei normalverteilten Daten wie in Beispiel 1.7 und 1.11 schreiben wir für die Dichte

$$\begin{aligned}p_\theta(x) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{(x_i-\mu)^2}{2\sigma^2}\Big) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\Big(-\sum_{i=1}^n\frac{(x_i-\mu)^2}{2\sigma^2}\Big) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\Big(-\frac{n\mu^2}{2\sigma^2}\Big)\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i\Big).\end{aligned}$$

Im Fall von unbekanntem μ und σ^2, d.h. im statistischen Modell (Norm 1a) folgt mit dem Fisher-Neyman Kriterium, dass $T = \Big(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\Big)$ suffizient ist. Weiter können wir hier ablesen, dass im statistischen Modell (Norm 1b) (bei bekanntem σ^2) bereits $\sum_{i=1}^n X_i$ suffizient ist.

Beispiel 2.8 (Beispiel *Unif*). Für die Situation aus Beispiel 1.8 sehen wir mit dem Fisher-Neyman'schen Kriterium, dass $T := t(X) := \max_{i=1,\dots,n} X_i$ suffizient ist. Beweis siehe Übung.

Definition 2.9 (Minimalsuffizienz). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell und $T = t(X)$ für $t : E \to E'$ eine suffiziente Statistik. Dann heißt T* minimalsuffizient*, falls es für jede weitere suffiziente Statistik $U = u(X)$ für $u : E \to E''$ eine messbare Abbildung $g : E'' \to E'$ gibt mit $T \overset{\mathbb{P}_\theta - fs}{=} g(U)$ für alle $\theta \in \mathcal{P}$.*

Für obige Beispiele ist es nicht einfach nachzuprüfen, ob die angegebenen suffizienten Statistiken auch minimalsuffizient sind. Folgender Satz erleichtert aber das Auffinden minimalsuffizienter Statistiken.

Theorem 2.10 (Kriterium für Minimalsuffizienz). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein reguläres statistisches Modell mit $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$ und $t : E \to E'$ eine messbare Abbildung, so dass*

$$t(y) = t(x) \iff \text{Es gibt ein } \ell(x,y) > 0, \text{ so dass für alle } \theta \in \mathcal{P}\text{: } p_\theta(y) = p_\theta(x)\ell(x,y).$$

Dann ist $T = t(X)$ minimalsuffizient.

Beweis. Zunächst ist klar, dass die für $t(y) = t(x)$ angegebene Bedingung symmetrisch in x und y, also wohldefiniert ist. (Mit $\ell(x,y) > 0$ ist nämlich auch $\ell(y,x) := 1/\ell(x,y) > 0$.)
Wir zeigen zunächst, dass T suffizient ist. Sei ν^* wie in Lemma 2.4. Es gilt für $t(x) = t(y)$

$$\begin{aligned}\frac{d\mathbb{P}_\theta}{d\nu^*}(x) &= \frac{d\mathbb{P}_\theta}{d\lambda^n}(x) \Big/ \frac{d\nu^*}{d\lambda^n}(x) = \frac{p_\theta(x)}{\sum_{i=1}^\infty c_i p_{\theta_i}(x)} \\ &= \frac{p_\theta(x)\ell(x,y)}{\sum_{i=1}^\infty c_i p_{\theta_i}(x)\ell(x,y)} = \frac{p_\theta(y)}{\sum_{i=1}^\infty c_i p_{\theta_i}(y)} = \frac{d\mathbb{P}_\theta}{d\nu^*}(y).\end{aligned}$$

Damit hängt $\frac{d\mathbb{P}_\theta}{d\nu^*}$ nur von $t(x)$ ab und die Behauptung folgt nach Lemma 2.4.4.
Es folgt nun der Beweis, dass T minimalsuffizient ist. Sei hierzu $U = u(X)$ eine weitere suffiziente Statistik. Nach dem Fisher-Neyman'schen Faktorisierungssatz gibt es g_θ und h, so dass $p_\theta(x) = g_\theta(u(x))h(x)$. (Man kann oBdA annehmen, dass $h > 0$ ist.) Wir müssen zeigen, dass aus $u(x) = u(y)$ folgt, dass $t(x) = t(y)$. Dann nämlich gibt es eine Funktion g mit $t(x) = g(u(x))$. Sei also $u(x) = u(y)$ und damit

$$\frac{p_\theta(y)}{p_\theta(x)} = \frac{g_\theta(u(y))h(y)}{g_\theta(u(x))h(x)} = \frac{h(y)}{h(x)}.$$

Daraus folgt $t(x) = t(y)$ mit der Wahl $\ell(x,y) = h(y)/h(x)$. □

Beispiel 2.11 (Beispiel *Bern*). Wir haben bereits gesehen, dass $T = \sum_{i=1}^n X_i$ suffizient ist. Außerdem gilt $p_\theta(x) = p_\theta(y)$ genau dann, wenn $t(x) = t(y)$. Wählt man also $\ell(x,y) = 1$ in Theorem 2.10, so erhält man die Minimalsuffizienz von T.

Beispiel 2.12 (Beispiel *Unif*). In Beispiel *Unif* ist $t(x) := \max_{i=1,\dots,n} x_i$ suffizient; siehe Beispiel 2.8. Für festes θ ist

$$\frac{p_\theta(y)}{p_\theta(x)} = \frac{\theta^n 1(t(y) \le \theta)}{\theta^n 1(t(x) \le \theta)} = \frac{1(t(y) \le \theta)}{1(t(x) \le \theta)}.$$

Nun $t(x) = t(y)$ genau dann, wenn $\frac{p_\theta(y)}{p_\theta(x)} = 1$ für alle θ. Damit ist $T = t(X)$ minimalsuffizient.

2.2 Vollständigkeit und Verteilungsfreiheit

Manchmal benötigt man suffiziente Statistiken, die weitere Eigenschaften erfüllen. Eine solche Eigenschaft wird nun beschrieben. Sie hilft insbesondere, minimalsuffiziente Statistiken zu finden; siehe Theorem 2.16.

Definition 2.13 (Vollständige Statistik). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell. Eine Statistik $T = t(X)$ für ein $t : E \to E'$ heißt* (beschränkt) vollständig, *falls für alle (beschränkten) messbaren Funktionen g gilt, dass*

$$\mathbb{E}_\theta[g(T)] = 0 \text{ für alle } \theta \in \mathcal{P} \implies g(T) \overset{\mathbb{P}_\theta - fs}{=} 0 \text{ für alle } \theta \in \mathcal{P}.$$

Beispiel 2.14 (X nicht beschränkt vollständig). Wir zeigen nun, dass im Allgemeinen für ein statistisches Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ die Zufallsvariable X nicht vollständig ist. Sei hierzu etwa für ein $n \geq 2$ die Verteilung $\mathbb{P}_\theta = \text{Poi}(\theta)^n$ für $\theta \in \mathcal{P} := \mathbb{R}_+$ die n-dimensionale Poisson-Verteilung mit Parameter θ, d.h. unter $\mathbb{P}_\theta$ hat $X = (X_1, ..., X_n)$ Werte in $\mathbb{R}^n$ und $X_1, ..., X_n$ sind unabhängig und identisch Poisson verteilt zum Parameter $\theta \geq 0$. Für ein beschränktes (aber auf $\mathbb{Z}_+$ nicht konstantes) $f : \mathbb{Z}_+ \to \mathbb{R}$ sei $g : \mathbb{Z}_+^n \to \mathbb{R}$ gegeben durch $g(x_1, ..., x_n) := f(x_1) - f(x_2)$. Dann ist offensichtlich (da $X_1 \sim X_2$ für alle $\theta \in \mathcal{P}$)

$$\mathbb{E}_\theta[g(X)] = \mathbb{E}_\theta[f(X_1)] - \mathbb{E}_\theta[f(X_2)] = 0,$$

aber $\mathbb{P}_\theta(f(X_1) \neq f(X_2)) > 0$, d.h. $g(T) \overset{\mathbb{P}_\theta-\text{fs}}{=} 0$ gilt nicht. Deshalb ist X nicht beschränkt vollständig (und damit nicht vollständig).

Beispiel 2.15 (Poisson-verteilte Statistik T). Für ein statistisches Modell $\{\mathbb{P}_\theta : \theta \in \mathbb{R}_+\}$ sei $T_*\mathbb{P}_\theta = \text{Poi}(\theta)$. Dann ist T beschränkt vollständig.
Denn: Sei g messbar und beschränkt, so dass

$$\mathbb{E}_\theta[g(T)] = e^{-\theta} \sum_{t=0}^{\infty} g(t) \frac{\theta^t}{t!} = 0$$

für alle $\theta \geq 0$. Dies ist die Potenzreihenentwicklung einer Funktion $\theta \mapsto 0$. Da diese Funktion analytisch ist, ist die Darstellung eindeutig und damit gilt $g(t) = 0$ für $t = 0, 1, 2, ...$, d.h. $g(T) \overset{\mathbb{P}_\theta-\text{fs}}{=} 0$.

Theorem 2.16 (Satz von Bahadur). *Sei $\{\mathbb{P}_\theta : \theta \in \mathbb{R}_+\}$ ein statistisches Modell und $T = t(X)$ für $t : E \to \mathbb{R}^k$ für ein $k \geq 0$ beschränkt vollständig und suffizient. Dann ist T minimalsuffizient.*

Beweis. Wir setzen $t(x) = (t_1(x), ..., t_k(x))$. Sei $U = u(X)$ für $u : E \to E''$ eine weitere suffiziente Statistik. Wir müssen zeigen, dass T eine Funktion von U ist. Wir setzen $S = s(T)$ mit $s = (s_1, ..., s_k)$ und $s_i(t) = (1+e^{t_i})^{-1}$. Dann ist $s : \mathbb{R}^k \to \mathbb{R}^k$ injektiv und S ist beschränkt. Wir setzen für $i = 1, ..., k$

$$H_i(U) = \mathbb{E}_\theta[S_i(T)|U],$$
$$L_i(T) = \mathbb{E}_\theta[H_i(U)|T].$$

Da T und U suffizient sind, hängen diese Funktionen nicht von θ ab und mit S sind auch H_i und L_i beschränkt, $i = 1, ..., k$. Weiter gilt für $\theta \in \mathcal{P}$

$$\mathbb{E}_\theta[S_i(T) - L_i(T)] = \mathbb{E}_\theta[H_i(U) - \mathbb{E}_\theta[H_i(U)|T]] = 0$$

und da T beschränkt vollständig ist folgt $S_i(T) \overset{\mathbb{P}_\theta-\text{fs}}{=} L_i(T)$ für alle $\theta \in \mathcal{P}$. Wir zerlegen nun die Varianz von $L_i(T)$ und von $H_i(U)$ mittels

$$\mathbb{V}_\theta[L_i(T)] = \mathbb{E}_\theta[\mathbb{V}_\theta[L_i(T)|U]] + \mathbb{V}_\theta[\mathbb{E}_\theta[L_i(T)|U]] = \mathbb{E}_\theta[\mathbb{V}_\theta[L_i(T)|U]] + \mathbb{V}_\theta[H_i(U)],$$
$$\mathbb{V}_\theta[H_i(U)] = \mathbb{E}_\theta[\mathbb{V}_\theta[H_i(U)|T]] + \mathbb{V}_\theta[L_i(T)]$$

und damit $\mathbb{V}_\theta[L_i(T)|U] = \mathbb{V}_\theta[H_i(U)|T] \overset{\mathbb{P}_\theta-\text{fs}}{=} 0$, also auch $\mathbb{V}_\theta[S_i(T)|U] \overset{\mathbb{P}_\theta-\text{fs}}{=} 0$. Demnach gilt

$$S_i(T) \overset{\mathbb{P}_\theta-\text{fs}}{=} \mathbb{E}_\theta[S_i(T)|U] \overset{\mathbb{P}_\theta-\text{fs}}{=} H_i(U).$$

Also gilt $T_i = s_i^{-1}(H_i(U))$ für $i = 1, ..., k$ und T ist eine Funktion von U. Damit folgt die Behauptung. □

In Theorem 2.32 werden wir sehen, dass die suffizienten Statistiken aus Beispiel *Bern* und *Norm* vollständig sind. Beispiel *Unif* behandeln wir zunächst getrennt.

Beispiel 2.17 (Beispiel *Unif*). Wir haben bereits (siehe Beispiele 1.8, 2.8 und 2.12) gesehen, dass $T := t(X) := \max_{i=1,\dots,n} X_i$ minimalsuffizient ist. Wir zeigen nun, dass T auch vollständig ist. Also würde hier zusammen mit der Suffizienz aus Beispiel 2.8 und dem Satz von Bahadur nochmals folgen, dass T minimalsuffizient ist.
Unter $\mathbb{P}_\theta$ gilt

$$\mathbb{P}_\theta(T \leq t) = 1_{t\leq\theta}\Big(\frac{t}{\theta}\Big)^n,$$

also ist $t \mapsto 1_{t\leq\theta} n t^{n-1}/\theta^n$ die Dichte von T. Gilt nun $\mathbb{E}_\theta[g(T)] = 0$ für alle $\theta > 0$, so folgt

$$\int_0^\theta t^{n-1} g(t)dt = 0.$$

Da dies für alle $\theta \geq 0$ gilt, ist

$$\int_a^b t^{n-1} g(t)dt = 0.$$

Dies ist nur möglich, wenn $g \overset{\lambda-\text{fü}}{=} 0$.

Wir kommen nun zum Begriff der verteilungsfreien Statistik. Eine solche beinhaltet keine (statistisch verwendbaren) Informationen über die Daten. Das Hauptresultat über verteilungsfreie Statistiken ist Theorem 2.19, der den Zusammenhang zu suffizienten Statistiken herstellt; siehe auch die Abbildung am Anfang des Kapitels.

Definition 2.18 (Verteilungsfreie Statistik). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell. Eine Statistik $T = t(X)$ für ein $t : E \to E'$ heißt* verteilungsfrei*, falls $T_*\mathbb{P}_\theta$ unabhängig von θ ist. Eine verteilungsfreie Statistik T heißt* maximal*, wenn es für jede andere verteilungsfreie Statistik U eine Funktion g gibt mit $U = g(T)$.*

Theorem 2.19 (Satz von Basu). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell. Weiter sei $T = t(X)$ für $t : E \to E'$ und $U = u(X)$ für $u : E \to E''$.*[1]

1. *Ist T beschränkt vollständig und suffizient sowie U verteilungsfrei, dann gilt $T \perp_{\mathbb{P}_\theta} U$ für alle $\theta \in \mathcal{P}$.*
2. *Angenommen, für alle $\theta, \theta' \in \mathcal{P}$ gibt es eine Menge $A \in \mathcal{B}(E)$ mit $\mathbb{P}_\theta(X \in A), \mathbb{P}_{\theta'}(X \in A) > 0$. Ist $T \perp_{\mathbb{P}_\theta} U$ für alle $\theta \in \mathcal{P}$ und T suffizient, dann ist U verteilungsfrei.*
3. *Sei $T \perp_{\mathbb{P}_\theta} U$ für alle $\theta \in \mathcal{P}$ und U verteilungsfrei. Wenn $\sigma(T, U) = \sigma(X)$, dann ist T suffizient.*

Beweis. 1. Sei $A \in \mathcal{B}(E'')$. Offenbar gilt

$$\mathbb{P}_\theta(U \in A) = \mathbb{E}_\theta[\mathbb{P}_\theta(U \in A|T)].$$

Weiter hängt (wegen der Verteilungsfreiheit von U) weder $\mathbb{P}_\theta(U \in A)$ noch (wegen der Suffizienz von T) die Größe $\mathbb{P}_\theta(U \in A|T)$ von θ ab. Damit ist $g : t \mapsto \mathbb{P}_\theta(U \in A) - \mathbb{P}_\theta(U \in A|T)$ eine beschränkte messbare Funktion (unabhängig von θ) mit $\mathbb{E}_\theta[g(T)] = 0$ für alle $\theta \in \mathcal{P}$.

[1] Wir erinnern an die Schreibweise $X \perp_{\mathbb{P}_\theta} Y$, falls X und Y unter $\mathbb{P}_\theta$ stochastisch unabhängig sind.

Wegen der beschränkten Vollständigkeit von T ist damit $\mathbb{P}_\theta(U \in A) \overset{\mathbb{P}_\theta-\text{fs}}{=} \mathbb{P}_\theta(U \in A|T)$. Dies aber bedeutet die Unabhängigkeit von U und T.
2. Da T suffizient ist, hängt $\mathbb{P}_\theta(U \in A|T)$ für alle $A \in \mathcal{B}(E'')$ nicht von θ ab. Da $\mathbb{P}_\theta(U \in A|T) = \mathbb{P}_\theta(U \in A)$ wegen der Unabhängigkeit von U und T, hängt also $\mathbb{P}_\theta(U \in A)$ nicht von θ ab. Dies ist aber gerade die Verteilungsfreiheit von U, da A beliebig war.
3. Es ist zu zeigen, dass für alle $A \in \mathcal{B}(E)$ gilt, dass $\mathbb{P}_\theta(X^{-1}(A)|T)$ unabhängig von θ ist. In der Tat genügt es mittels einer monotonen Klasse, für einen schnittstabilen Erzeuger $\mathcal{E}$ von $\sigma(X)$ zu zeigen, dass für alle $E \in \mathcal{E}$ $\mathbb{P}_\theta(E|T)$ unabhängig von θ ist.

Sei nun $B \in \mathcal{B}(E')$ und $C \in \mathcal{B}(E'')$. Es gilt

$$\mathbb{P}_\theta(T \in B, U \in C|T) = 1_{T\in B}\mathbb{P}_\theta(U \in C)$$

und dies ist (wegen der Verteilungsfreiheit von U) unabhängig von θ. Da $\{T^{-1}(B) \cap U^{-1}(C) : B \in \mathcal{B}(E'), C \in \mathcal{B}(E'')\}$ schnittstabil ist und nach Voraussetzung $\sigma(X)$ erzeugt, folgt die Aussage. □

Beispiel 2.20 (Beispiel *Norm*). Für festes σ^2 betrachten wir das statistische Modell aus (Norm 1*b*), also das Normalverteilungsmodell mit bekanntem σ^2. Weiter setzen wir

$$T := \overline{X} := \frac{1}{n}\sum_{i=1}^n X_i, \qquad U = \frac{1}{n-1}\sum_{i=1}^n (X_i - T)^2.$$

Aus der Rechnung in Beispiel 2.7 sieht man zusammen mit dem Fisher-Neyman'schen Faktorisierungssatzes (Theorem 2.5), dass T suffizient ist. (Weiter werden wir in Theorem 2.32 sehen, dass T auch vollständig, also sogar minimalsuffizient ist.) Außerdem ist U verteilungsfrei, denn: Der Vektor $(X_1 - T, ..., X_n - T)$ ist unabhängig von θ normalverteilt mit

$$\begin{aligned}\mathbb{E}_\theta[X_i - T] &= 0,\\ \mathbb{COV}_\theta[X_i - T, X_j - T] &= \sigma^2(\delta_{ij} - \tfrac{2}{n} + \tfrac{1}{n}) = \sigma^2(\delta_{ij} - \tfrac{1}{n}).\end{aligned}$$

Nun ist U eine Funktion dieses Vektors, also verteilungsfrei.) Nach dem Theorem von Basu, Theorem 2.19.2, sind diese beiden Vektoren also unabhängig. (Dieses Ergebnis ist auch Teil des bekannten Satzes von Fisher.)

2.3 Exponentialfamilien

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

Definition 2.21 (Exponentialfamilie). *Sei $\mathcal{P} \subseteq \mathbb{R}^k$. Eine Familie $\{\mathbb{P}_\theta : \theta \in \mathcal{P}\}$ von Wahrscheinlichkeitsmaßen auf $\mathbb{R}^n$ (auf $\mathbb{Z}^n$) heißt* k-parametrige Exponentialfamilie *(mit c, t, d, h) für*

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

falls $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$ und

$$p_\theta(x) = h(x) \cdot \exp\Big(\sum_{j=1}^k c_j(\theta)t_j(x) - d(\theta)\Big) = h(x) \cdot \exp\big(c(\theta)^\top t(x) - d(\theta)\big), \qquad x \in \mathbb{R}^n.$$

Gilt insbesondere $c_j(\theta) = \theta_j$, also

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

so sagt man, die Exponentialfamilie sei in kanonischer Form *und wir definieren den* kanonischen Parameterraum

$$\Gamma := \Big\{\theta \in \mathbb{R}^k : \int h(x) e^{\theta^\top t(x)} d\lambda^n(x) < \infty\Big\}.$$

Bemerkung 2.22 (1-parametrige Exponentialfamilie). Für den Spezialfall einer 1-parametrigen Exponentialfamilie gibt es Funktionen c, d, t, h mit

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

Beispiel 2.23 (Beispiel *Norm*). Wir betrachten das statistische Modell (Norm $1a$) für $n = 1$, also $\theta := (\mu, \sigma^2)$ mit $\mathcal{P} = \mathbb{R} \times \mathbb{R}_+$ und $\mathbb{P}_\theta := \mathcal{N}(\mu, \sigma^2)$, und damit

$$p_{(\mu,\sigma^2)}(x) = \exp\Big(\frac{\mu}{\sigma^2}x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\Big(\frac{\mu}{\sigma^2} + \log(2\pi\sigma^2)\Big)\Big).$$

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

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

Diese ist nun allerdings nicht in kanonischer Form.

Beispiel 2.24 (Beispiel *Bern*). Sei $(X, \{\mathbb{P}_\theta : \theta \in (0,1)\})$ wie in Beispiel 1.6. Dann ist also mit $(Bern\ 0)$

$$\begin{aligned} p_\theta(x) &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ &= \exp\Big(\sum_{i=1}^n x_i \log\theta + \Big(n - \sum_{i=1}^n x_i\Big)\log(1-\theta)\Big) \\ &= \exp\Big(\log\frac{\theta}{1-\theta}\sum_{i=1}^n x_i + n\log(1-\theta)\Big) \end{aligned}$$

und damit ist $\{\mathbb{P}_\theta : \theta \in (0,1)\}$ eine 1-parametrige Exponentialfamilie.

Beispiel 2.25 (Beispiel *Unif*). Die Familie $\{\mathbb{P}_\theta : \theta \in \mathbb{R}_+\}$ aus Beispiel 1.8 ist keine Exponentialfamilie.

Beispiel 2.26 (Beispiel *Exp*). Sei $\mathcal{P} = \mathbb{R}_+$ und $\mathbb{P}_\theta = p_\theta \cdot \lambda$ die Exponentialverteilung mit Parameter θ. Dann ist

$$p_\theta(x) = 1_{x \geq 0} e^{-\theta x + \log\theta}$$

und damit ist die Familie der Exponentialverteilungen eine 1-parametrige Exponentialfamilie und obige Darstellung ist die kanonische Form.

Beispiel 2.27 (Beispiel *Pois*). Sei $\Lambda = \mathbb{R}_+$. Für die Poisson-Verteilung mit Parameter λ schreiben wir $\mathbb{P}_\lambda = p_\lambda \cdot \mu$ mit

$$p_\lambda(x) = \frac{1_{x\geq 0}}{x!} \exp\big((\log\lambda)x - \lambda\big)$$

Diese ist damit eine 1-parametrige Exponentialfamilie. Um obige Darstellung in kanonische Form zu bringen, setzen wir $\theta := \log\lambda$ und schreiben für $\theta \in \mathbb{R}$

$$p_\theta(x) = \frac{1_{x\geq 0}}{x!} \exp\big(\theta x - e^\theta\big).$$

Bemerkung 2.28 (Sample aus einer Exponentialfamilie). Ist $\{\mathbb{P}_\theta : \theta \in \mathcal{P}\}$ eine k-parametrige Exponentialfamilie auf $\mathbb{R}^n$ mit Funktionen $c_1, ..., c_k, d, t_1, ..., t_k, h$, und sind $X_1, ..., X_N$ unabhängig und identisch nach $\mathbb{P}_\theta$ verteilt. Dann ist die gemeinsame Verteilung von $X_1, ..., X_N$ eine Exponentialfamilie mit

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

mit der Projektion $\pi_i(x) = x_i$.
Denn: Als gemeinsame Dichte schreiben wir

$$p_\theta(x_1, ..., x_N) = h(x_1)\cdots h(x_N) \cdot \exp\Big(\sum_{j=1}^k c_j(\theta) \sum_{i=1}^N t_j(x_i) - Nd(\theta)\Big).$$

Proposition 2.29 (Suffiziente Statistik bei Exponentialfamilien).
Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ eine Exponentialfamilie (mit c, t, d, h). Dann ist die Statistik $T := t(X) = (t_1(X), ..., t_k(X))$ suffizient.

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

$$p_\theta(x) = h(x) g_\theta(t(x))$$

für

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

Damit folgt die Aussage aus dem Fisher-Neyman'schen Faktorisierungssatz, Theorem 2.5. □

Beispiel 2.30 (Lineare Regression). Bei der linearen Regression beobachten wir $(x_1, Y_1), ..., (x_n, Y_n)$ (mit $x_i = (x_{i0} = 1, x_{i1}, ..., x_{im}) \in \mathbb{R}^{m+1}$) und gehen davon aus, dass

$$Y_i = x_i\beta + \varepsilon_i \qquad (\text{also } Y = x^\top\beta + \varepsilon)$$

für $\beta_0, ..., \beta_m \in \mathbb{R}$ mit $\varepsilon_1, ..., \varepsilon_n$ unabhängig und identisch nach $\mathcal{N}(0, \sigma^2)$ verteilt. Hierbei sind $x_1, ..., x_n$ bekannt (und fest, also keine Parameter des Modells) und $Y_1, ..., Y_n$ werden als

Zufallsvariablen betrachtet. Wir haben also ein statistisches Modell $(Y, \{\mathbb{P}_\theta = p_\theta \cdot \lambda^n : \theta = (\beta, \sigma^2) \in \mathbb{R}^{m+2}\})$ und wir schreiben für die gemeinsame Dichte von $Y = (Y_1, ..., Y_n)$

$$\begin{aligned} p_\theta(y) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y_i - x_i\beta)^2}{2\sigma^2}\Big) \\ &= \exp\Big(- \frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 + \frac{1}{\sigma^2}\beta^\top \sum_{i=1}^n y_i x_i - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i\beta)^2 - \tfrac{n}{2}\log(2\pi\sigma^2)\Big) \end{aligned}$$

Wählen wir

$$\begin{aligned} c_j(\beta) &= \frac{\beta_j}{\sigma^2}, & t_j(y) &= \sum_{i=1}^n y_i x_{ij}, & j &= 0, ..., m, \\ c_{m+1}(\beta) &= -\frac{1}{2\sigma^2}, & t_{m+1}(y) &= \sum_{i=1}^n y_i^2, \end{aligned}$$

$$d(\beta) = \frac{1}{2\sigma^2}\sum_{i=1}^n (\beta^\top x_i)^2 + \tfrac{n}{2}\log(2\pi\sigma^2),$$

so sehen wir, dass $\{\mathbb{P}_\theta : \theta \in \mathbb{R}^{m+2}\}$ eine $m+2$-parametrige Exponentialfamilie ist. Weiter ist

$$\Big(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i x_{i1}, ..., \sum_{i=1}^n Y_i x_{im}, \sum_{i=1}^n Y_i^2\Big)$$

suffizient.

Lemma 2.31 (Kanonischer Parameterraum konvex). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ eine Exponentialfamilie in kanonischer Form (mit c, t, d, h). Dann ist der kanonische Parameterraum konvex und $\theta \mapsto e^{d(\theta)}$ ist konvex.*

Beweis. Sei $\theta, \eta \in \Gamma$ und $0 < \alpha < 1$. Dann gilt, da die Exponentialfunktion konvex ist,

$$\begin{aligned} e^{d(\alpha\theta + (1-\alpha)\eta)} &:= \int h(x) \exp\big((\alpha\theta + (1-\alpha)\eta)^\top t(x)\big) d\lambda^n(x) \\ &\le \int h(x)\big(\alpha e^{\theta^\top t(x)} + (1-\alpha) e^{\eta^\top t(x)}\big) d\lambda^n(x) \\ &= \alpha e^{d(\theta)} + (1-\alpha) e^{d(\eta)}. \end{aligned}$$

Dies zeigt alle Behauptungen. □

Theorem 2.32 (Vollständigkeit der suffizienten Statistik). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ eine k-parametrige Exponentialfamilie in kanonischer Form (mit (θ, t, d, h)). Ist[2] $\mathbb{R}^k \supseteq \mathcal{P}^\circ \neq \emptyset$, so ist $T = t(X)$ vollständig (und wegen der Suffizienz aus Proposition 2.29 mit Theorem 2.16 auch minimalsuffizient).*

[2] Mit A° bezeichnen wir das Innere der Menge A.

Beweis. Wir beweisen die Behauptung nur im Fall $k = 1$, da die anderen Fälle mit Induktion folgen. Angenommen, T sei nicht vollständig. Dann gibt es $g : \mathbb{R} \to \mathbb{R}$ messbar und $\mathbb{E}_\theta[g(T)] = 0$ für alle $\theta \in \mathcal{P}$, aber $\mathbb{P}_{\theta_0}(g(T) > 0) > 0$ für ein $\theta_0 \in \mathcal{P}^\circ$. (Wir können oBdA θ_0 im Inneren von $\mathcal{P}$ wegen der Stetigkeit der Dichte von T und monotoner Konvergenz wählen.) Anders ausgedrückt heißt das, dass für alle $\theta \in \mathcal{P}$

$$\mathbb{E}_\theta[g^+(T)] = \mathbb{E}_\theta[g^-(T)] \tag{$*$}$$

gilt, aber

$$\mathbb{E}_{\theta_0}[g^+(T)] = \mathbb{E}_{\theta_0}[g^-(T)] =: w \in (0, \infty)$$

für ein $\theta_0 \in \mathcal{P}^\circ$. Wir definieren die beiden Wahrscheinlichkeitsmaße

$$P := \tfrac{1}{w} g^+ \circ T \cdot \mathbb{P}_{\theta_0}, \qquad Q := \tfrac{1}{w} g^- \circ T \cdot \mathbb{P}_{\theta_0}.$$

Nun schreiben wir $(*)$ um in

$$\begin{aligned}
\mathbb{E}_P[e^{(\theta-\theta_0)T}] &= \tfrac{1}{w}\mathbb{E}_{\theta_0}[e^{(\theta-\theta_0)T} g^+(T)] \\
&= \frac{1}{w} \int e^{(\theta-\theta_0)t(x)} g^+(t(x)) h(x) e^{\theta_0 t(x) - d(\theta_0)} d\lambda(x) \\
&= \frac{e^{d(\theta)-d(\theta_0)}}{w} \int g^+(t(x)) h(x) e^{\theta t(x) - d(\theta)} d\lambda(x) \\
&= \frac{e^{d(\theta)-d(\theta_0)}}{w} \mathbb{E}_\theta[g^+(T)] = \frac{e^{d(\theta)-d(\theta_0)}}{w} \mathbb{E}_\theta[g^-(T)] \\
&= \tfrac{1}{w}\mathbb{E}_{\theta_0}[e^{(\theta-\theta_0)T} g^-(T)] \\
&= \mathbb{E}_Q[e^{(\theta-\theta_0)T}].
\end{aligned}$$

Wegen der Eindeutigkeit der Laplace-Transformierten bedeutet dies $P = Q$ und damit auch $g^+ \overset{\mathbb{P}_{\theta_0}-\text{fs}}{=} g^-$, also $g \overset{\mathbb{P}_{\theta_0}-\text{fs}}{=} 0$ und damit (wegen der Form der Exponentialfamilie) auch $g \overset{\mathbb{P}_\theta-\text{fs}}{=} 0$ für alle $\theta \in \mathcal{P}$. □

Bemerkung 2.33 (Exponentialfamilien in nicht-kanonischer Form). Das Ergebnis aus Theorem 2.32 lässt sich durch Umparametrisierung auf Exponentialfamilien übertragen, die nicht in kanonischer Form vorliegen. Man sieht etwa, dass die suffizienten Statistiken aus Beispiel 2.23, 2.24, 2.26 und 2.27 minimalsuffizient sind.

Exponentialfamilien erlauben oft explizite Berechnungen. Dies illustrieren wir an zwei Ergebnissen, von denen wir das erste ohne Beweis angeben.

Lemma 2.34 (Differenzierbarkeit nach θ). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ eine Exponentialfamilie in kanonischer Form (mit θ, t, d, h) und der kanonische Parameterraum Γ habe nicht-leeres Inneres, $\Gamma^\circ \neq \emptyset$. Weiter sei $\phi : \mathbb{R}^k \to \mathbb{R}$ messbar und $\theta \in \Gamma^\circ$ so, dass $\eta \mapsto \mathbb{E}_\eta[|\phi(t(X)|)e^{\eta^\top t(X)}] < \infty$ für η in einer Umgebung von θ. Dann ist die Abbildung*

$$f : \eta \mapsto \mathbb{E}_\eta[\phi(t(X))e^{\eta^\top t(X)}]$$

in einer Umgebung von θ analytisch mit

$$\frac{\partial f(\eta)}{\partial \eta_i} = \mathbb{E}_\eta[t_i(X)\phi(t(X))e^{\eta^\top t(X)}].$$

Proposition 2.35 (Laplace-Transformierte von Exponentialfamilien). *Sei* $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ *eine* k*-parametrige Exponentialfamilie in kanonischer Form. Ist* $\theta \in \mathcal{P}^\circ$*, dann existiert die Laplace-Transformierte von* $t(X)$ *mit*

$$\mathbb{E}_\theta[e^{\eta^\top t(X)}] = \exp(d(\eta+\theta) - d(\theta)),$$

falls $|\eta|$ *klein genug ist und für* $l_1, ..., l_k \geq 0$ *mit* $l_1 + \cdots + l_k = l$

$$\mathbb{E}_\theta\Big[\prod_{i=1}^k t_i(X)^{\ell_i}\Big] = e^{-d(\theta)} \frac{\partial^\ell}{\partial \eta_1^{\ell_1} \cdots \partial \eta_k^{\ell_k}} e^{d(\eta)}\Big|_{\eta=\theta}.$$

Insbesondere gilt also

$$\mathbb{E}_\theta[t_i(X)] = \frac{\partial d(\eta)}{\partial \eta_i}\Big|_{\eta=\theta},$$
$$\mathbb{COV}_\theta[t_i(X), t_j(X)] = \frac{\partial^2 d(\eta)}{\partial \eta_i \partial \eta_j}\Big|_{\eta=\theta}.$$

Beweis. Wir berechnen

$$\begin{aligned}\mathbb{E}[e^{\eta^\top t(X)}] &= \int \exp(\eta^\top t(x)) \cdot h(x) \cdot \exp(\theta^\top t(x) - d(\theta))\lambda^n(dx) \\ &= e^{d(\eta+\theta)-d(\theta)} \int h(x) \cdot \exp((\theta+\eta)^\top t(x) - d(\theta+\eta))\lambda^n(dx) \\ &= e^{d(\eta+\theta)-d(\theta)},\end{aligned}$$

da das Integral über eine Dichte eins ist. Die zweite Behauptung folgt aus Lemma 2.34. □

2.4 Bayes'sche Modelle

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

Definition 2.36 (A-priori-Verteilung, a-posteriori-Verteilung). *Sei* $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ *ein statistisches Modell und* $\mathcal{P}$ *ein vollständiger, separabler metrischen Raum. Eine* a-priori-Verteilung *ist eine Wahrscheinlichkeitsverteilung* π *auf* $\mathcal{P}$*. Ist für alle* $A \in \mathcal{B}(E)$ *die Abbildung* $\theta \mapsto \mathbb{P}_\theta(X \in A)$ *messbar, so wird durch* $(\theta, A) \mapsto \mathbb{P}_\theta(X \in A)$ *ein Markov-Kern von* $\mathcal{P}$ *nach* $\mathcal{B}(E)$ *definiert und für* $\Theta \sim \pi$ *wird durch*

$$P(\Theta \in A, X \in B) := \int_A \pi(d\theta)\mathbb{P}_\theta(X \in B)$$

die gemeinsame Verteilung von Θ *und* X *auf* $\mathcal{B}(\mathcal{P})\otimes\mathcal{B}(E)$ *definiert. Die* a-posteriori-*Verteilung ist dann die reguläre Version der bedingten Verteilung* $P(\Theta \in .|X)$*, also*

$$\pi_x = P(\Theta \in .|X = x).$$

Bemerkung 2.37 (A-posteriori-Verteilung bei regulären Modellen). Ist $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell und $\mathcal{P} \subseteq \mathbb{R}^k$, und ist $\pi = g \cdot \lambda^k$, so hat die gemeinsame Verteilung von Θ und X die Dichte $g(d\theta) \cdot p_\theta(dx)$. Die a-posteriori Verteilung π_x hat dann die Dichte

$$p_x(\theta) = \frac{p_\theta(x)g(\theta)}{\int p_\eta(x)g(\eta)d\eta}.$$

Beispiel 2.38 (Beispiel *Bern*). Wieder ist $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$ durch (*Bern* 0) gegeben. Weiter sei die a-priori-Verteilung die Beta-Verteilung $\pi = \beta(k, l)$, d.h. $\pi = p_{kl} \cdot \lambda$ mit[3]

$$p_{kl}(x) = \frac{\Gamma(k)\Gamma(l)}{\Gamma(k+l)} x^{k-1}(1-x)^{l-1} \sim x^{k-1}(1-x)^{l-1},$$

wobei wir mit $\sim$ ausdrücken, dass der restliche Faktor unabhängig von x ist. Mit Bemerkung 2.37 folgt für die Dichte der a-posteriori-Verteilung π_x

$$p_x(\theta) \sim \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\theta^{k-1}(1-\theta)^{l-1} = \theta^{k-1+\sum x_i}(1-\theta)^{l+n-1-\sum x_i}.$$

Dies ist also eine $\beta(k + \sum x_i, l + n - \sum x_i)$-Verteilung.

Das letzte Beispiel lässt sich auf allgemeine Exponentialfamilien verallgemeinern. (Aus Beispiel 2.24 wissen wir, dass das letzte Beispiel eine solche Verteilungsklasse behandelt.)

Proposition 2.39 (Konjugierte Familie bei Exponentialfamilie). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ eine k-parametrige Exponentialfamilie (mit c, t, d, h). Weiter sei $\pi_s = p_s \cdot \lambda^k$ eine a-priori-Verteilung mit*

$$p_s(\theta) = \frac{\exp\Big(\sum_{j=1}^k c_j(\theta)s_j - s_{k+1}d(\theta)\Big)}{\int \exp\Big(\sum_{j=1}^k c_j(\eta)s_j - s_{k+1}d(\eta)\Big)\lambda^k(d\eta)}$$

gegeben. Dann ist die a-posteriori-Verteilung gerade $P(\Theta \in .|X = x) = p_x \cdot \lambda^k$ mit

$$p_x(\theta) = p_{(s_1+t_1,\ldots,s_k+t_k,s_{k+1}+1)}(\theta).$$

Beweis. Wir schreiben $a \sim_x b$, falls a/b nur von x abhängt (d.h. a und b sind proportional). Es gilt

$$\begin{aligned} p_x(\theta) \sim_x p_\theta(x)\pi_s(x) &\sim_x \exp\Big(\sum_{j=1}^k c_j(\theta)(t_j(x) + s_j) - (s_{k+1}+1)d(\theta)\Big) \\ &\sim_x p_{(s_1+t_1,\ldots,s_k+t_k,s_{k+1}+1)}(\theta). \end{aligned}$$

□

Beispiel 2.40 (A-posteriori-Verteilung bei der Normalverteilung). Wir betrachten das Normalverteilungsmodell bei bekanntem σ^2, gegeben durch (Norm 1b). Angenommen, die a-priori-Verteilung π für θ ist selbst eine Normalverteilung, nämlich $\mathcal{N}(a, b^2)$ für $a, b \in \mathbb{R}$. Wie

[3] Für die Gamma-Funktion gilt etwa $\Gamma(k) = (k-1)!$ für $k = 1, 2, \ldots$ sowie $\Gamma(x+1) = x\Gamma(x)$ für $x \in \mathbb{R}$.

sieht dann die a-posteriori-Verteilung aus? Und ist diese um den wahren Wert θ konzentriert? In der Übung wird gezeigt werden, dass dies wieder eine Normalverteilung $\mathcal{N}(\alpha, \beta)$ ist mit

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

Insbesondere ist für große n die a-posteriori-Verteilung um $\overline{x}$ konzentriert.

3 Entscheidungstheorie

Statistische Fragestellungen formuliert man oft als Entscheidungsprobleme. Diese zeichnen sich dadurch aus, dass aufgrund von Daten in einem (statistischen) Modell immer eine Entscheidung über die verwendeten Parameter getroffen werden muss. Bei einem statistischen Test ist diese Entscheidung etwa, ob der Parameter größer oder kleiner als ein vorgegebener Wert ist, bei einem Schätzproblem fällt eine Entscheidung über die (vermutete) Größe eines Parameters.

3.1 Einführung

Zentral sind bei Entscheidungen die Begriffe des Entscheidungsraumes und der Entscheidungsfunktion. Um über die Qualität der Entscheidung(sfunktion) zu urteilen, gibt es außerdem eine Verlustfunktion.

Definition 3.1 (Entscheidungsraum, -funktion, Verlustfunktion). *Sei* $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ *ein statistisches Modell.*

1. *Für einen* Entscheidungsraum $\aleph$ *heißt jede messbare Abbildung* $d : E \to \aleph$ nicht-randomisierte Entscheidungsfunktion. *Eine* (randomisierte) Entscheidungsfunktion *ist ein Markov-Kern* $\delta(.,.)$ *von* E *nach* $\mathcal{B}(\aleph)$. *Die Menge der nicht-randomisierten Entscheidungsfunktionen bezeichnen wir mit* $\mathcal{D}_{nr}$, *die Menge der randomisierten Entscheidungsfunktionen mit* $\mathcal{D}$. *Für* $d \in \mathcal{D}_{nr}$ *sei* $\delta_d(x, A) := 1_{d(x)\in A}$ *die zugehörige randomisierte Entscheidungsfunktion.*

2. *Eine* Verlustfunktion *ist eine messbare Abbildung* $\ell : \mathcal{P} \times \aleph \to \mathbb{R}_+$.

3. *Ein* statistisches Entscheidungsproblem *ist ein Tripel* $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ *aus einem statistischen Modell, einem Entscheidungsraum und einer Verlustfunktion.*

4. *Für ein statistisches Entscheidungsproblem* $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ *heißt die Abbildung*

$$R_\delta(\theta) := R(\theta, \delta) := \mathbb{E}_\theta\Big[\int \ell(\theta, y)\delta(X, dy)\Big] \tag{3.1}$$

Risikofunktion. *Für eine nicht-randomisierte Entscheidung ist dies gleich*

$$R_d(\theta) := R(\theta, d) := \mathbb{E}_\theta\big[\ell(\theta, d(X))\big].$$

Die Menge $\mathcal{R} := \{R_\delta : \delta \in \mathcal{D}\}$ *heißt* Risikomenge.

Bemerkung 3.2 (Interpretation). In einem statistischen Modell steht X für die (zufällig entstandenen) Daten. Bei Vorliegen des (unbekannten Parameters) $\theta \in \mathcal{P}$ sind diese nach $\mathbb{P}_\theta$ verteilt. Für eine Entscheidungsfunktion δ ist nun $\delta(x, A)$ die Wahrscheinlichkeit, sich für ein $a \in A$ zu entscheiden, wenn $X = x$ vorliegt. (Bei einer nicht-randomisierten Entscheidungsfunktion d ist $d(x) = a$ die Wahl der Entscheidung a im Entscheidungsraum $\aleph$.) Bei einer Entscheidung für a hat man, falls θ vorliegt, einen Verlust zu verzeichnen. Dieser wird mit $\ell(\theta, a)$ bezeichnet. Da die Daten als zufällig angesehen werden, kann man sich fragen, welchen Verlust man denn (bei Vorliegen von θ und für die Entscheidung δ) erwartet. Dies ist gerade die Risikofunktion $R_\delta(\theta)$.

Bemerkung 3.3 (Punkt-Schätzproblem). 1. Aus der Vorlesung *Stochastik* bekannt ist folgendes Problem:
Ein Versuch, der entweder einen Erfolg oder einen Misserfolg bringt, wird n-mal wiederholt, wobei $S = k$-mal ein Erfolg eintritt. Nun soll die Wahrscheinlichkeit für einen Erfolg (in einem der n Versuchen) geschätzt werden.
Dies erinnert an das statistische Modell aus Beispiel *Bern*. Wir setzen hier $\mathcal{P} = [0, 1]$, $\mathbb{P}_\theta$ wie in (*Bern* 0), $S = X_1 + \cdots + X_n$ und $\aleph = \mathcal{P}$. Eine offensichtlich Wahl für eine nicht-randomisierte Entscheidungsfunktion ist $d(x_1, ..., x_n) = (x_1 + \cdots + x_n)/n =: \overline{x}$. Als Verlustfunktion könnte etwa $\ell(\theta, a) = (\theta - a)^2$ dienen, also $\ell(\theta, d(x_1, ..., x_n)) = (\theta - \overline{x})^2$. Die Risikofunktion berechnet sich dann zu

$$R_d(\theta) = \mathbb{E}_\theta[(\theta - \overline{X})^2] = \mathbb{V}_\theta[\overline{X}] = \frac{\theta(1-\theta)}{n}.$$

2. Allgemeiner ist ein Punkt-Schätzproblem ein statistisches Entscheidungsproblem $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ mit dem statistischen Raum $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$. Im Allgemeinen wollen wir nicht θ schätzen, sondern $g(\theta)$ für eine Funktion $g : \mathcal{P} \to \aleph$, wobei $\aleph$ auch der Entscheidungsraum ist. (Wie hierbei die Entscheidungsfunktion δ (oder d im nicht-randomisierten Fall) aussieht, hängt vom jeweiligen Beispiel ab.) Verlustfunktionen, die der Bemessung einer falschen Entscheidung dienen sollen, sind (im Falle eines normierten Raumes $\aleph$) etwa der

$$\begin{aligned} \text{Laplace-Verlust } \ell(\theta, a) &:= |a - g(\theta)|, \\ \text{Gauß-Verlust } \ell(\theta, a) &:= |a - g(\theta)|^2, \\ \text{0-1-Verlust } \ell(\theta, a) &:= 1_{|a-g(\theta)|>\varepsilon}. \end{aligned}$$

Die Risikofunktion $R_\delta(\theta)$ ist dann der (unter $\mathbb{P}_\theta$) erwartete Verlust unter der (etwa nicht-randomisierten) Entscheidungsfunktion d, also beim Laplace-Verlust etwa

$$R_\delta(\theta) = \mathbb{E}_\theta[|d(X) - g(\theta)|].$$

Man nennt Entscheidungsfunktionen bei solchen Schätzproblemen *(Punkt-)Schätzer* und $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ ein Schätzproblem.

Bemerkung 3.4 (Bereichs-Schätzproblem). 1. Aus der Vorlesung *Stochastik* bekannt ist folgendes Problem:
Für eine Stichprobe $x_1, ..., x_n$ normalverteilter Daten (d.h. $X_1, ..., X_n$ sind unabhängig und identisch nach $\mathcal{N}(\theta, \sigma^2)$ verteilt) bei bekannter Varianz σ^2. Es wird ein von den Daten abhängiger Bereich gesucht, so dass θ mit Wahrscheinlichkeit $1 - \alpha$ (etwa für

$\alpha = 5\%$) in diesem liegt. Es ist naheliegend, dass dieses Intervall symmetrisch um $\overline{x} := (x_1 + \cdots + x_n)/n$ liegt.
Dies erinnert an das statistische Modell *Norm* mit bekannter Varianz σ^2, d.h. an das statistische Modell $(X, \{\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\})$, so dass (*Norm* 0) gilt. Um das Intervall aufzustellen, verwenden wir, dass $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, also $(\overline{X} - \mu)/\sqrt{\sigma^2/n} \sim \mathcal{N}(0,1)$. Sei q_α das α-Quantil der Standardnormalverteilung[4]. Dann gilt

$$\begin{aligned} 1 - \alpha &= \mathbb{P}_\theta(q_{\alpha/2} \leq (\overline{X} - \theta)/\sqrt{\sigma^2/n} \leq q_{1-\alpha/2}) \\ &= \mathbb{P}_\theta(|\overline{X} - \theta| \leq q_{1-\alpha/2}\sqrt{\sigma^2/n}) \\ &= \mathbb{P}_\theta(\overline{X} - q_{1-\alpha/2}\sqrt{\sigma^2/n} \leq \theta \leq \overline{X} + q_{1-\alpha/2}\sqrt{\sigma^2/n}) \end{aligned}$$

Der gesuchte Datenbereich ist also $\{y : \overline{x} - q_{1-\alpha/2}\sqrt{\sigma^2/n} \leq \overline{y} \leq \overline{x} + q_{1-\alpha/2}\sqrt{\sigma^2/n}\}$.

2. Allgemeiner ist ein Bereichs-Schätzproblem ein statistisches Entscheidungsproblem $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ mit dem statistischen Raum $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$. Wieder wollen wir nicht θ schätzen, sondern $g(\theta)$ für eine Funktion $g : \mathcal{P} \to \Upsilon$, wobei hier nun $\aleph = \mathcal{B}(\Upsilon)$ ist. (Übrigens ist damit $\aleph$ kein metrischer Raum, und man muss die Messbarkeit der folgenden Abbildungen überprüfen.) Eine nicht-randomisierte Entscheidungsfunktion ist hier wieder eine Funktion $d : E \to \aleph$. (Nun muss etwa für alle $g \in \Upsilon$ die Bedingung $\{x : g \in d(x)\} \in \mathcal{B}(E)$ gelten.) Als Verlustfunktion bietet sich die Wahl

$$\ell(\theta, B) := \begin{cases} 1, & g(\theta) \notin B, \\ 0, & g(\theta) \in B \end{cases}$$

an. Das Risiko ist somit

$$R_d(\theta) = \mathbb{E}_\theta[\ell(\theta, d(X))] = \mathbb{P}_\theta[g(\theta) \notin d(X)].$$

Entscheidungsfunktionen in einer solchen Situation heißen auch *Bereichs-Schätzer*.

Bemerkung 3.5 (Test-Problem). 1. Aus der Vorlesung *Stochastik* bekannt ist folgendes Problem (siehe *einfacher t-Test*):
Für eine Stichprobe $x_1, ..., x_n$ normalverteilter Daten (d.h. $X_1, ..., X_n$ sind unabhängig und identisch nach $\mathcal{N}(\theta = (\mu, \sigma^2))$ verteilt) bei unbekannter Varianz σ^2. Es soll getestet werden, ob die (Null-)Hypothese $\mu = \mu_0$ aufgrund der Daten verworfen werden kann oder nicht. Dabei soll die Wahrscheinlichkeit, die Nullhypothese irrtümlicherweise zu verwerfen, höchstens ein vorgegebenes α sein.
Dies erinnert an das statistische Modell *Norm*, d.h. an das statistische Modell $(X, \{\mathbb{P}_\theta = \mathcal{N}(\theta) : \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\})$, so dass (*Norm* 0) gilt. Um die Hypothese $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$ zu testen, verwendet man, dass für $s^2(X) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ die Statistik

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

unter $\mathbb{P}_{(\mu,\sigma^2)}$ (für jedes $\sigma^2 \in \mathbb{R}_+$) nach t_{n-1}-verteilt ist.[5] Man stellt hier den kritischen (oder Ablehnungs-)Bereich $C \subseteq \mathbb{R}$ (der nur von α abhängt) so auf, dass H_0 gerade dann

[4] Es ist also für $Z \sim \mathcal{N}(0,1)$ gerade $\mathbb{P}(Z \leq q_\alpha) = \alpha$. Insbesondere ist auch $q_\alpha = -q_{1-\alpha}$.
[5] Dies ist eine studentische t-Verteilung mit $n-1$ Freiheitsgraden. Ihr α-Quantil bezeichnen wir mit $t_{n-1,\alpha}$.

abgelehnt wird, wenn $t(X) \in C$ ist. Es soll außerdem $\mathbb{P}_{(\mu_0,\sigma^2)}(t(X) \in C) \leq \alpha$ gelten. In diesem Beispiel ergibt sich mit $C = (-\infty, t_{n-1,\alpha/2}) \cup (t_{n-1,1-\alpha/2}, \infty)$ gerade, dass

$$\mathbb{P}_{(\mu_0,\sigma^2)}(t(X) \in C) = \mathbb{P}_{(\mu_0,\sigma^2)}(t(X) \leq t_{n-1,\alpha/2}) + \mathbb{P}_{(\mu_0,\sigma^2)}(t(X) \geq t_{n-1,1-\alpha/2}) = \alpha,$$

was ja gefordert war.

2. Allgemeiner ist ein Test-Problem ein statistisches Entscheidungsproblem $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ mit dem statistischen Raum $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ mit $\aleph = \{H_0, H_1\}$. Die (randomisierte) Entscheidungsfunktion δ ist eindeutig durch eine Funktion $\varphi : E \to [0,1]$ mittels $\delta(x, \{H_1\}) := \varphi(x)$ und $\delta(x, \{H_0\}) := 1 - \varphi(x)$ bestimmt. (Hier ist $\varphi(x)$ die Wahrscheinlichkeit, sich bei Daten x für die Hypothese H_1 zu entscheiden.) Die Entscheidungsfunktion δ ist genau dann nicht-randomisiert, falls $\varphi = 1_C$ für ein geeignetes $C \in \mathcal{B}(E)$. Als Verlustfunktion bietet sich die Neyman-Pearson-Verlustfunktion an. Hierbei ist für $\mathcal{P}_0, \mathcal{P}_1$ mit $\mathcal{P}_0 \uplus \mathcal{P}_1 = \mathcal{P}$ gerade

$$\ell(\theta, H_1) = \begin{cases} 0, & \theta \in \mathcal{P}_1 \quad \text{richtige Entscheidung,} \\ 1, & \theta \in \mathcal{P}_0 \quad \text{Fehler 1. Art: Entscheidung für } H_1, \text{ aber } H_0 \text{ trifft zu.} \end{cases}$$

$$\ell(\theta, H_0) = \begin{cases} 0, & \theta \in \mathcal{P}_0 \quad \text{richtige Entscheidung,} \\ 1, & \theta \in \mathcal{P}_1 \quad \text{Fehler 2. Art: Entscheidung für } H_0, \text{ aber } H_1 \text{ trifft zu.} \end{cases}$$

Die Hypothese H_i besteht dabei gerade daraus, dass eine der Verteilungen aus $\mathcal{P}_i$ (auf die Daten) zutrifft, $i = 0, 1$. Die Risikofunktion ist damit

$$\begin{aligned} R_\delta(\theta) &= \mathbb{E}_\theta[\ell(\theta, H_0)\delta(X, \{H_0\}) + \ell(\theta, H_1)\delta(X, \{H_1\})] \\ &= 1_{\theta \in \mathcal{P}_1}\mathbb{E}_\theta[1 - \varphi(X)] + 1_{\theta \in \mathcal{P}_0}\mathbb{E}_\theta[\varphi(X)] = \begin{cases} 1 - \mathbb{E}_\theta[\varphi(X)], & \theta \in \mathcal{P}_1, \\ \mathbb{E}_\theta[\varphi(X)], & \theta \in \mathcal{P}_0. \end{cases} \end{aligned}$$

Wie wir sehen, wird das Risiko eindeutig durch die *Gütefunktion*

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

bestimmt. Im Fall einer nicht-randomisierten Entscheidungsfunktion $\varphi = 1_C$ ist also

$$R_\delta(\theta) = \begin{cases} \mathbb{P}_\theta(X \notin C), & \theta \in \mathcal{P}_1, \\ \mathbb{P}_\theta(X \in C), & \theta \in \mathcal{P}_0. \end{cases}$$

Bei diesem allgemeinen Vorgehen fällt auf, dass nirgends das Signifikanzniveau α aus obigem Beispiel eingeht. Das liegt daran, dass wir zunächst alle Entscheidungsfunktionen zugelassen haben. Sinnvoller ist es jedoch, sich auf eine Klasse von Entscheidungsfunktionen einzuschränken, etwa auf solche δ, für die $R_\delta(\theta) \leq \alpha$ für $\theta \in \mathcal{P}_0$ gilt.
Man nennt Entscheidungsfunktionen δ (oder φ) bei solchen Testproblemen auch *Tests*.

Bemerkung 3.6 (Randomisierter Test). Wir geben noch das Beispiel eines randomisierten Tests an. Sei hierzu $(X, \{\mathbb{P}_\theta : \theta \in [0,1]\})$ das statistische Modell aus Beispiel 1.6 sowie $H_0 = \{\theta \leq 0.5\}$ und $H_1 = \{\theta > 0.5\}$. Für ein $\alpha \in (0,1)$ wollen wir einen Test φ mit $\mathbb{E}_\theta[\varphi(X)] \leq \alpha$ für $\theta \in H_0$ angeben, d.h. die Wahrscheinlichkeit, sich für die Alternative zu entscheiden, soll bei Vorliegen von $\theta \in H_0$ höchstens α betragen. Man sagt auch, der Fehler 1.

Art soll höchstens α sein. Sei hierzu $q_{\theta,1-\alpha}$ ein $1-\alpha$-Quantil einer $B(n,\theta)$-Verteilung[6]. Dann könnten wir etwa

$$\varphi(x) = \begin{cases} 0, & \sum_{i=1}^n x_i \le q_{0.5,1-\alpha}, \\ 1, & \sum_{i=1}^n x_i > q_{0.5,1-\alpha} \end{cases}$$

setzen. Für $\theta \in H_0$ gilt dann

$$\mathbb{E}_\theta[\varphi(X)] = \mathbb{P}_\theta\Big[\sum_{i=1}^n X_i > q_{0.5,1-\alpha}\Big] \le \mathbb{P}_{0.5}\Big[\sum_{i=1}^n X_i > q_{0.5,1-\alpha}\Big] \le \alpha.$$

Schön wäre es, wenn wir sogar ein $\psi \ge \varphi$ angeben können, für das ebenfalls $\mathbb{E}_\theta[\psi(X)] \le \alpha$ für $\theta \in H_0$ gilt. Dies können wir erreichen, indem wir

$$\psi(x) = \begin{cases} 0, & \sum_{i=1}^n x_i < q_{0.5,1-\alpha}, \\ p := \frac{\mathbb{P}_{0.5}\left[\sum_{i=1}^n x_i \le q_{0.5,1-\alpha}\right] - (1-\alpha)}{\mathbb{P}_{0.5}\left[\sum_{i=1}^n x_i = q_{0.5,1-\alpha}\right]}, & \sum_{i=1}^n x_i = q_{0.5,1-\alpha}, \\ 1, & \sum_{i=1}^n x_i > q_{0.5,1-\alpha}. \end{cases}$$

Das bedeutet: falls $\sum_{i=1}^n x_i = q_{0.5,1-\alpha}$, so entscheiden wir uns mit Wahrscheinlichkeit p für die Alternative, und mit $1-p$ für die Nullhypothese. Die Entscheidung ist also zufällig oder randomisiert. Dann ist für $\theta \in H_0$

$$\begin{aligned} \mathbb{E}_\theta[\psi(X)] \le \mathbb{E}_{0.5}[\psi(X)] &= 1 - \mathbb{P}_{0.5}\Big[\sum_{i=1}^n X_i \le q_{0.5,1-\alpha}\Big] + p\mathbb{P}_{0.5}\Big[\sum_{i=1}^n X_i = q_{0.5,1-\alpha}\Big] \\ &= 1 - \mathbb{P}_{0.5}\Big[\sum_{i=1}^n X_i \le q_{0.5,1-\alpha}\Big] + \mathbb{P}_{0.5}\Big[\sum_{i=1}^n X_i \le q_{0.5,1-\alpha}\Big] - (1-\alpha) = \alpha, \end{aligned}$$

also ist der Fehler 1. Art ebenfalls höchstens α.

Es sollte klar sein, dass *gute* Entscheidungskriterien gerade solche sind, die eine kleine Risikofunktion aufweisen. Allerdings gibt es verschiedene Möglichkeiten dafür, dass eine Entscheidungsfunktion *klein* ist. Mit diesen werden wir uns in Abschnitt 3.3 beschäftigen. Wir geben nur noch ein Beispiel, in dem klar wird, dass naiv angegebene Entscheidungsfunktionen auch *schlecht* sein können.

Beispiel 3.7 (Schätzung des Lageparameters einer Uniformverteilung).
Sei $U(\theta-1/2,\theta+1/2)$ die Uniformverteilung auf $(\theta-1/2,\theta+1/2)$ und $\mathbb{P}_\theta = U(\theta-1/2,\theta+1/2)^n$ das n-fache Produkt, $\theta \in \mathbb{R}$. Wir wollen für das statistische Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathbb{R}\})$ den Lageparameter θ schätzen. Hierzu geben wir zwei verschiedene, nicht-randomisierte Entscheidungsfunktionen an. Wir wählen (mit $g(\theta) := \theta$)[7]

$$d_1(X) = \overline{X}, \qquad d_2(X) = \tfrac{1}{2}(X_{(1)} + X_{(n)})$$

[6] Wir erinnern daran, dass für eine Verteilung $\mathbb{P}$ mit Verteilungsfunktion F das α-Quantil durch jede Zahl x mit $F(x-) \le \alpha \le F(x)$ gegeben ist.

[7] Für einen Vektor $x = (x_1,...,x_n) \in \mathbb{R}^n$ ist $x_{(i)}$ das it-kleinste Element. Insbesondere ist also $x_{(1)}$ das kleinste und $x_{(n)}$ das größte Element. Die Zahlen $x_{(1)},...,x_{(n)}$ heißen auch *Ordnungsstatistiken*.

und verwenden den Gauß-Verlust. Das Risiko berechnet sich für d_1 zu

$$R_{d_1}(\theta) = \mathbb{E}_\theta[(\overline{X} - \theta)^2] = \mathbb{V}_\theta[\overline{X}] = \frac{1}{n}\mathbb{V}_\theta[X_1] = \frac{1}{12n}.$$

Für d_2 benötigen wir ein paar Vorüberlegungen. Um die gemeinsame Verteilung von $X_{(1)}$ und $X_{(n)}$ zu berechnen, schreiben wir (für $\theta = 1/2$)

$$\mathbb{P}_{1/2}(X_{(1)} > x, X_{(n)} < y) = (y - x)^n.$$

Also hat $(X_{(1)}, X_{(n)})$ die gemeinsame Dichte $(x, y) \mapsto n(n-1)(y-x)^{n-2}$. Daraus berechnen wir

$$\begin{aligned}
\mathbb{E}_{1/2}[X_{(1)}] &= 1 - \mathbb{E}_{1/2}[1 - X_{(1)}] = 1 - \int_0^1 n(1-x)^n dx = \frac{1}{n+1},\\
\mathbb{E}_{1/2}[X_{(n)}] &= \int_0^1 nyy^{n-1} dy = \frac{n}{n+1},\\
\mathbb{V}_{1/2}[X_{(1)}] &= \mathbb{V}_{1/2}[1 - X_{(1)}] = \int_0^1 n(1-x)^{n+1} dx - \frac{n^2}{(n+1)^2} = \frac{n}{n+2} - \frac{n^2}{(n+1)^2}\\
&= \frac{n}{(n+2)(n+1)^2},\\
\mathbb{V}_{1/2}[X_{(n)}] &= \frac{n}{(n+2)(n+1)^2},\\
\mathbb{COV}_{1/2}[X_{(1)}, X_{(n)}] &= \int_0^1 \int_0^y n(n-1)xy(y-x)^{n-2} dxdy - \frac{n}{(n+1)^2}\\
&= \int_0^1 \int_0^y ny(y-x)^{n-1} dxdy - \frac{n}{(n+1)^2}\\
&= \int_0^1 y^{n+1} dy - \frac{n}{(n+1)^2} = \frac{1}{n+2} - \frac{1}{(n+1)^2} = \frac{1}{(n+1)^2(n+2)},
\end{aligned}$$

wobei die Varianzen und Covarianz unabhängig von θ sind. Wir erhalten

$$\begin{aligned}
R_{d_1}(\theta) &= \mathbb{E}_\theta[(\tfrac{1}{2}(X_{(1)} + X_{(n)}) - \theta)^2] = \mathbb{V}_\theta[\tfrac{1}{2}(X_{(1)} + X_{(n)})]\\
&= \tfrac{1}{4}(\mathbb{V}[X_{(1)}] + \mathbb{V}[X_{(n)}] + 2\mathbb{COV}[X_{(1)}, X_{(n)}])\\
&= \frac{1}{2}\Big(\frac{n}{(n+2)(n+1)^2} + \frac{1}{(n+1)^2(n+2)}\Big) = \frac{1}{2(n+1)(n+2)}.
\end{aligned}$$

Wir sehen also, dass das Risiko von d_2 gerade für große n deutlich kleiner ist als das von d_1.

3.2 Die Rolle suffizienter Statistiken

Suffiziente Statistiken enthalten alle wichtigen Informationen über die Daten. Deswegen ist es sinnvoll, dass Entscheidungsfunktionen nur von solchen Statistiken abhängen.

Proposition 3.8. *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem, δ eine (randomisierte) Entscheidungsfunktion (also ein Markov-Kern von E nach $\mathcal{B}(\aleph)$) und R_δ ihre Risikofunktion. Ist $T = t(X)$ (für $t : E \to E'$) eine suffiziente Statistik, so gibt es einen Markov-Kern ε von E' nach $\mathcal{B}(\aleph)$, so dass $\varepsilon \circ t$ (mit $\varepsilon \circ t(x, A) := \varepsilon(t(x), A)$) eine Entscheidungsfunktion ist mit $R_{\varepsilon \circ t} = R_\delta$.*

Beweis. Für $A \in \mathcal{B}(\aleph)$ setzen wir

$$\varepsilon(t, A) := \mathbb{E}_\theta[\delta(X, A)|T = t].$$

(Da T suffizient ist, hängt die rechte Seite nicht von θ ab. Wir unterdrücken den Subskript θ im Folgenden.) Deshalb gilt für integrierbare Funktionen $h : \aleph \to \mathbb{R}$, dass

$$\mathbb{E}\Big[\int h(a)\delta(X, da)\Big|T = t\Big] = \int h(a)\varepsilon(t, da).$$

(Zunächst überprüft man die Aussage mit Indikatorfunktionen, dann mit einfachen Funktionen und anschließend mit allgemeinen messbaren Funktionen h.) Nach Definition gilt dann

$$\begin{aligned} R_{\varepsilon\circ t}(\theta) &= \mathbb{E}_\theta\Big[\int \ell(\theta, a)\varepsilon(t(X), da)\Big] = \mathbb{E}_\theta\Big[\mathbb{E}\Big[\int \ell(\theta, a)\delta(X, da)\Big|T\Big]\Big] \\ &= \mathbb{E}_\theta\Big[\int \ell(\theta, a)\delta(X, da)\Big] = R_\delta(\theta). \end{aligned}$$

□

Man beachte, dass auch bei nicht-randomisierten δ der Markov-Kern ε (und damit $\varepsilon \circ t$) randomisiert sein kann.

Proposition 3.9. *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem. Sei $\aleph \subseteq \mathbb{R}^m$ konvex und $\ell(\theta, .)$ für alle $\theta \in \mathcal{P}$ eine konvexe Verlustfunktion. Für eine (randomisierte) Entscheidungsfunktion δ definieren wir die nicht-randomisierte Entscheidungsfunktion*

$$d(x) := \int a\delta(x, da),$$

zumindest für $x \in E$, für die dieses Integral existiert. Für solche $x \in E$ und alle $\theta \in \mathcal{P}$ ist dann

$$\ell(\theta, d(x)) \leq \int \ell(\theta, a)\delta(x, da).$$

Beweis. Zunächst ist $d(x) \in \aleph$, da $\aleph$ konvex ist. Weiter gilt mit der Jensen'schen Ungleichung

$$\ell(\theta, d(x)) = \ell\Big(\theta, \int a\delta(x, da)\Big) \leq \int \ell(\theta, a)\delta(x, da).$$

□

Man kann mit Hilfe suffizienter Statistiken Entscheidungsfunktionen besser machen. Vor allem für Schätzprobleme wird folgendes Resultat schöne Anwendungen haben.

Theorem 3.10 (Rao-Blackwell). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem. Sei $\aleph \subseteq \mathbb{R}^m$ konvex und $\ell(\theta, .)$ für alle $\theta \in \mathcal{P}$ eine konvexe Verlustfunktion, $T = t(X)$ eine suffiziente Statistik und d eine nicht-randomisierte Entscheidungsfunktion mit $\mathbb{E}_\theta[|d(X)|] < \infty$ für alle $\theta \in \mathcal{P}$ sowie R_d ihre Risikofunktion. Für*

$$e(t) := \mathbb{E}_\theta[d(X)|T = t]$$

(wobei die rechte Seite unabhängig von θ ist) gilt dann für alle $\theta \in \mathcal{P}$

$$R_{e\circ t}(\theta) \leq R_d(\theta).$$

Beweis. Wir setzen

$$\varepsilon(t, A) := \mathbb{P}(d(X) \in A | T = t),$$

also

$$\int h(a)\varepsilon(t, da) = \mathbb{E}[h(d(X))|T = t],$$

falls das Integral existiert. (Wieder zeigt man das, indem man zuerst einfache Funktionen einsetzt und danach durch ein Approximationsargument messbare Funktionen.) Daraus folgt insbesondere

$$\int a\varepsilon(t, da) = \mathbb{E}[d(X)|T = t] = e(t).$$

Dann gilt wegen Proposition 3.9

$$\ell(\theta, e \circ t(x)) \leq \int \ell(\theta, a)\varepsilon(t(x), da),$$

also

$$R_{eot}(\theta) \leq R_\varepsilon(\theta) = \mathbb{E}_\theta\Big[\int \ell(\theta, a)\varepsilon(T, da)\Big] = \mathbb{E}_\theta[\mathbb{E}_\theta[\ell(\theta, d(X))|T]] = R_d(\theta).$$

□

Beispiel 3.11 (Beispiel *Unif*). Sei $(X, \{\mathcal{P}_\theta : \theta \in (0, 1)\})$ wie in Beispiel 1.8, $\aleph = (0, 1)$, $g(\theta) = \theta$ und ℓ der Gauß-Verlust $\ell(\theta, a) = (\theta - a)^2$. Wir betrachten den Schätzer $d(x) = 2\overline{x}$. Für diesen ist immerhin $\mathbb{E}_\theta[d(X)] = \theta$. Wir wissen aus Beispiel 2.8, dass $T = t(X) = \max_{i=1,\ldots,n} X_i$ suffizient ist. Wir verwenden Theorem 3.10 und schreiben aus Symmetriegründen

$$\mathbb{E}_\theta[2\overline{X}|t(X)] = 2\mathbb{E}_\theta[X_1|t(X)] = 2\Big(\frac{1}{n}t(X) + \frac{n-1}{n}\frac{t(X)}{2}\Big) = \frac{n+1}{n}t(X).$$

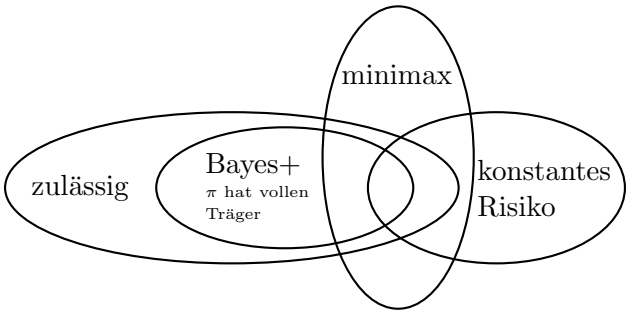
Nun lesen wir ab, dass der Schätzer

$$d'(x) := \frac{n+1}{n}\max_{i=1,\ldots,n} x_i$$

eine kleinere Risikofunktion hat.

3.3 Zulässige, Bayes, Minimax-Entscheidungsfunktionen

In diesem Abschnitt geht es um den Vergleich von Entscheidungsfunktionen mittels deren Risikofunktionen, sowie um bestimmte Optimalitätskriterien. Wie werden hier speziell zulässige, minimax und Bayes-Entscheidungsfunktionen betrachten. Oftmals macht es dabei keinen Sinn, optimale Entscheidungskriterien unter allen möglichen Entscheidungsfunktionen zu suchen, sondern nur aus einer Teilmenge. (Wir erinnern daran, dass wir die Menge aller Entscheidungsfunktionen mit $\mathcal{D}$ bezeichnet hatten.) Wir illustrieren einige Ergebnisse dieses Abschnitts in einer Grafik.



In Lemma 3.14 zeigen wir etwa, dass zulässige Entscheidungsfunktionen mit konstantem Risiko minimax sind. Lemma 3.19 besagt, dass Bayes-Entscheidungsfunktionen (unter einer Voraussetzung an die a-priori-Verteilung) zulässig sind. Die minimax-Eigenschaft von (allgemeinen) Bayes-Entscheidungsfunktionen wird dann mit Theorem 3.20 geklärt.

Beispiel 3.12. 1. Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathbb{R}\}), \aleph, \ell)$ mit $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$, $\aleph = \mathbb{R}$ und ℓ der Gauß-Verlust. Wir wollen also den Mittelwert einer Normalverteilung (bei bekannter Varianz) schätzen, wenn wir nur einmal aus ihr ziehen. Die offensichtliche Wahl $\delta_1(x, a) = x$ (d.h. wir schätzen den Parameter θ durch die Beobachtung x) führt zur Risikofunktion

$$\theta \mapsto R_{\delta_1}(\theta) = \mathbb{E}_\theta[(X - \theta)^2] = 1.$$

Hingegen führt die Wahl $\delta_2(x, a) = b$ für ein festes $b \in \mathbb{R}$ (d.h. wir schätzen den Parameter θ immer durch dieselbe Zahl b, unabhängig von unserer Beobachtung) zur Risikofunktion

$$\theta \mapsto R_{\delta_2}(\theta) = \mathbb{E}_\theta[(b - \theta)^2] = (b - \theta)^2.$$

Insbesondere sehen wir, dass weder $R_{\delta_1} \leq R_{\delta_2}$ noch $R_{\delta_2} \leq R_{\delta_1}$ gilt. Man kann also nicht sagen, dass δ_1 besser wäre als δ_2. Ein Ausweg hier ist es, sich nur auf erwartungstreue Schätzer einzuschränken, d.h. auf Entscheidungsfunktionen δ mit $\mathbb{E}_\theta[\int \delta(X, da)] = \theta$, und unter diesen die kleinste Risikofunktion zu suchen.

2. Ganz ähnlich verhält es sich bei einem Test. Verwenden wir hierzu die Situation aus Bemerkung 3.5.2. Es ist verlockend, einfach $C = E$ zu wählen (d.h. die Hypothese wird immer verworfen), weil dadurch zumindest für $\theta \in \mathcal{P}_0$ die Risikofunktion verschwindet. Allerdings wird durch diese Wahl die Risikofunktion für $\theta \in \mathcal{P}_1$ maximal. Wählt man andererseits $C = \emptyset$, so ist die Situation genau andersherum. Deshalb fragt man hier eher nach dem Minimum aller Risikofunktionen, die $R_\delta(\theta) \leq \alpha$ für alle $\theta \in \mathcal{P}_0$ erfüllen. Das bedeutet, dass man das Signifikanzniveau α festlegt und damit den maximalen Fehler erster Art.

Definition 3.13 (Zulässige, Minimax-Entscheidungsfunktionen). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem, $\delta, \delta' \in \mathcal{D}$ Entscheidungsfunktionen und $R_\delta, R_{\delta'}$ ihre Risikofunktionen. Weiter sei $\mathcal{D}_0 \subseteq \mathcal{D}$.*

1. *Gilt für δ, δ', dass $R_\delta \leq R_{\delta'}$ und $R_\delta(\theta) < R_{\delta'}(\theta)$ für mindestens ein $\theta \in \mathcal{P}$, so sagen wir, δ* dominiert *δ'.*

2. *Wir sagen, die Entscheidungsfunktion δ hat* konstantes Risiko*, falls $\theta \mapsto R_\delta(\theta)$ konstant ist.*

3. *Gibt es für eine Entscheidungsfunktion $\delta \in \mathcal{D}_0$ ein $\delta' \in \mathcal{D}_0$, das δ dominiert, so heißt δ auch* unzulässig für $\mathcal{D}_0$, *andersfalls* zulässig für $\mathcal{D}_0$.

4. *Die Entscheidungsfunktion $\delta \in \mathcal{D}_0$ heißt* minimax für $\mathcal{D}_0$, *falls*

$$\sup_{\theta\in\mathcal{P}} R_\delta(\theta) = \inf_{\delta'\in\mathcal{D}_0} \sup_{\theta\in\mathcal{P}} R_{\delta'}(\theta).$$

Für $\mathcal{D}$ zulässige Entscheidungsfunktionen heißen auch zulässig, für $\mathcal{D}$ unzulässige auch unzulässig, und für $\mathcal{D}$ minimax-Entscheidungsfunktionen heißen auch minimax-Entscheidungsfunktionen.

Lemma 3.14 (Zusammenhänge zulässige, minimax-Entscheidungsfunktionen). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem, $\delta \in \mathcal{D}_0 \subseteq \mathcal{D}$ eine Entscheidungsfunktion und R_δ ihre Risikofunktion.*

1. *Ist δ die einzige minimax-Entscheidungsfunktion für $\mathcal{D}_0$, so ist δ zulässig für $\mathcal{D}_0$.*

2. *Ist δ zulässig für $\mathcal{D}_0$ mit konstantem Risiko, so ist δ minimax für $\mathcal{D}_0$.*

Beweis. 1. Angenommen, δ ist nicht zulässig für $\mathcal{D}_0$. Dann gibt es ein $\delta' \in \mathcal{D}_0$ mit ($\delta' \neq \delta$ und) $R_{\delta'} \leq R_\delta$ und ein $\theta \in \mathcal{P}$ mit $R_{\delta'}(\theta) < R_\delta(\theta)$. Deshalb ist

$$\inf_{\varepsilon\in\mathcal{D}_0} \sup_{\theta\in\mathcal{P}} R_\varepsilon(\theta) = \sup_{\theta\in\mathcal{P}} R_\delta(\theta) \geq \sup_{\theta\in\mathcal{P}} R_{\delta'}(\theta) \geq \inf_{\varepsilon\in\mathcal{D}_0} \sup_{\theta\in\mathcal{P}} R_\varepsilon(\theta).$$

Damit ist δ' ebenfalls minimax für $\mathcal{D}_0$ im Widerspruch zur Voraussetzung.
2. Angenommen, es ist $R_\delta(\theta) =: c$ unabhängig von θ und δ ist nicht minimax für $\mathcal{D}_0$. Dann gibt es $\delta' \in \mathcal{D}_0$ mit

$$\sup_{\theta\in\mathcal{P}} R_{\delta'}(\theta) < \sup_{\theta\in\mathcal{P}} R_\delta(\theta) = c.$$

Daraus folgt $R_{\delta'} < R_\delta$ im Widerspruch zur Zulässigkeit für $\mathcal{D}_0$ von δ. □

Definition 3.15 (Bayes'sches Risiko). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem, δ eine Entscheidungsfunktion und R_δ ihre Risikofunktion. Sei weiter π ein Wahrscheinlichkeitsmaß auf $\mathcal{P}$ und $\Theta \sim \pi$. Dann ist das* Bayes-Risiko *(bezüglich π) gegeben als*

$$r_\delta(\pi) := \mathbb{E}_\pi[R_\delta(\Theta)] := \int R_\delta(\theta)\pi(d\theta).$$

Für $\mathcal{D}_0 \subseteq \mathcal{D}$ heißt $\delta \in \mathcal{D}_0$ eine $\mathcal{D}_0$-Bayes-Entscheidungsfunktion *bezüglich π, falls*

$$r_\delta(\pi) \leq r_{\delta'}(\pi)$$

für alle $\delta' \in \mathcal{D}_0$. Eine $\mathcal{D}$-Bayes-Entscheidungsfunktion *bezüglich π heißt auch* Bayes-Entscheidungsfunktion *bezüglich π.*

Oftmals sind Bayes-Entscheidungsfunktionen gar nicht so schwer zu finden. Hierbei besonders hilfreich ist das nächste Resultat.

Theorem 3.16 (Konstruktion von Bayes-Entscheidungsfunktionen). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem. Sei π eine a-priori-Verteilung, π_x die zugehörige a-posteriori-Verteilung und $\Theta_x \sim \pi_x$.*

1. *Sei δ eine Entscheidungsfunktion, so dass für alle x*

$$\mathbb{E}\Big[\int \ell(\Theta_x, a)\delta(x, da)\Big] = \inf_{m\in\mathcal{M}_1(\aleph)} \mathbb{E}\Big[\int \ell(\Theta, a)m(da)\Big].$$

Dann ist δ eine Bayes-Entscheidungsfunktion bezüglich π.

2. *Sei $d \in \mathcal{D}_{nr}$ eine nicht-randomisierte Entscheidungsfunktion, so dass für alle x*

$$\mathbb{E}[\ell(\Theta_x, d(x)] = \inf_{a\in\aleph} \mathbb{E}[\ell(\Theta, a)].$$

Dann ist δ eine Bayes-Entscheidungsfunktion bezüglich π für $\mathcal{D}_{nr}$.

Beweis. Wir zeigen nur 1. da der Beweis für 2. analog funktioniert. Wir schreiben das Bayes-Risiko für $\Theta \sim \pi$ als

$$r_\delta(\pi) = \mathbb{E}[R_\delta(\Theta)] = \mathbb{E}\Big[\mathbb{E}\Big[\int \ell(\Theta, a)\delta(X, da)\Big|X\Big]\Big].$$

Da die Verteilung von Θ gegeben X gerade π_X ist, die Erwartung sicher dann minimiert, wenn

$$\mathbb{E}\Big[\int \ell(\Theta, a)\delta(X, da)\Big|X = x\Big] = \mathbb{E}\Big[\int \ell(\Theta_x, a)\delta(x, da)\Big]$$

minimal ist. Daraus folgt die Behauptung. □

Korollar 3.17 (Bayes-Schätzer). *Wir betrachten das Schätzproblem $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph = \mathbb{R}, g, \ell)$ für den Gauß-Verlust ℓ, ein reguläres statistisches Modell $\{\mathbb{P}_\theta : \theta \in \mathcal{P}\}$ mit $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$ und $\mathcal{P} \subseteq \mathbb{R}^k$. Weiter sei $\pi = p \cdot \lambda^k \in \mathcal{M}_1(\mathcal{P})$ die a-priori-Verteilung und π_x die a-posteriori-Verteilung. (Nach Bemerkung 2.37 hat diese die Dichte*

$$p_x(\theta) = \frac{p_\theta(x)p(\theta)}{\int p_\eta(x)p(\eta)d\eta}.\Big)$$

Dann ist der nicht-randomisierte Schätzer (falls das Integral existiert)

$$d(x) := \int g(\theta)\pi_x(d\theta).$$

ein Bayes-Schätzer.

Beweis. Nach Proposition 3.9 müssen wir nur zeigen, dass d ein Bayes-Schätzer bezüglich π für $\mathcal{D}_{nr}$ ist. Nach Theorem 3.16 ist ein Bayes-Schätzer gegeben, wenn

$$\mathbb{E}[\ell(\Theta_x, d(x))] = \min_a \mathbb{E}[\ell(\Theta_x, a)].$$

Nun ist

$$\mathbb{E}[\ell(\Theta_x, a)] = \mathbb{E}[(g(\Theta_x) - a)^2] \geq \mathbb{E}[(g(\Theta_x) - \mathbb{E}[g(\Theta_x)])^2]$$

mit '=' genau dann, wenn

$$a = \mathbb{E}[g(\Theta_x)] = \int g(\theta)\pi_x(d\theta) = d(x).$$

Daraus folgt die Behauptung. □

Beispiel 3.18 (Beispiel *Norm* mit σ^2 bekannt). Wir betrachten für (bekanntes) $\sigma^2 > 0$ wie in (Norm 1*b*) das Schätzproblem $((X, \{\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)^n : \theta \in \mathbb{R}\}), \aleph = \mathbb{R}, g = \mathrm{id}, \ell)$ für den Gauß-Verlust ℓ. Wir zeigen, dass für die a-priori-Verteilung $\pi_b = \mathcal{N}(0, b^2)$ der Bayes-Schätzer durch

$$d_b(x) := \frac{nb^2}{nb^2 + \sigma^2}\overline{x}$$

gegeben ist. (Diesen liest man am besten als Konvexkombintation aus dem a-priori-Schätzer 0 – gegeben durch die a-priori-Verteilung, die Erwartungswert 0 hat – und dem Mittelwert.) Aus Beispiel 2.40 können wir die a-posteriori-Verteilung ablesen. Diese ist gerade

$$\pi_x := \mathcal{N}\Big(\frac{nb^2}{nb^2 + \sigma^2}\overline{x}, \frac{\sigma^2 b^2}{nb^2 + \sigma^2}\Big).$$

Nach Proposition 3.17 ist nun der Bayes-Schätzer bezüglich π gegeben durch

$$d(x) = \int \theta \pi_x(d\theta) = \frac{nb^2}{nb^2 + \sigma^2}\overline{x}.$$

Lemma 3.19 (Zusammenhänge zulässige, Bayes-Entscheidungsfunktionen). *Betrachte dieselbe Situation wie in Lemma 3.14. Weiter sei $\pi \in \mathcal{M}_1(\mathcal{P})$.*

1. *Sei $\theta \mapsto R_{\delta'}(\theta)$ für alle $\delta \in \mathcal{D}_0$ stetig und π habe vollen Träger. Ist δ Bayes-Entscheidungsfunktion bezüglich π für $\mathcal{D}_0$, so ist δ zulässig bezüglich $\mathcal{D}_0$.*
2. *Ist δ die einzige Bayes-Entscheidungsfunktion bezüglich π und $\mathcal{D}_0$, so ist δ zulässig bezüglich $\mathcal{D}_0$.*

Beweis. 1. Angenommen, δ ist nicht zulässig $\mathcal{D}_0$. Dann gibt es ein $\delta' \in \mathcal{D}_0$ mit ($\delta' \neq \delta$ und) $R_{\delta'} \leq R_\delta$ und ein $\theta \in \mathcal{P}$ mit $R_{\delta'}(\theta) < R_\delta(\theta)$. Wegen der Stetigkeit von $R_{\delta'}$ gibt es ein $r > 0$, so dass für $\theta' \in B_r(\theta)$ gerade $R_{\delta'}(\theta') < R_\delta(\theta) - r$. Damit gilt

$$\begin{aligned} r_{\delta'}(\pi) &= \mathbb{E}_\pi[R_{\delta'}(\Theta), \Theta \in B_r(\theta)] + \mathbb{E}_\pi[R_{\delta'}(\Theta), \Theta \notin B_r(\theta)] \\ &\leq \mathbb{E}_\pi[R_\delta(\Theta)] - r\mathbb{P}(\Theta \in B_r(\theta)) < \mathbb{E}_\pi[R_\delta(\Theta)] = r_\delta(\pi) \end{aligned}$$

im Widerspruch dazu, dass δ Bayes-Entscheidungsfunktion bezüglich π für $\mathcal{D}_0$ ist.
2. Sei δ' so, dass $R_{\delta'} \leq R_\delta$. Dann gilt

$$r_{\delta'}(\pi) = \mathbb{E}[R_{\delta'}(\Theta)] \leq \mathbb{E}[R_\delta(\Theta)] = r_\delta(\pi) = \inf_{\varepsilon \in \mathcal{D}_0} r_\varepsilon(\pi).$$

Damit ist auch δ' eine Bayes-Entscheidungsfunktion für $\mathcal{D}_0$. Diese ist aber nach Voraussetzung eindeutig und damit $\delta' = \delta$. Damit ist δ zulässig für $\mathcal{D}_0$. □

Theorem 3.20 (Hodges-Lehmann). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem, $\delta, \delta_1, \delta_2, ... \in \mathcal{D}_0 \subseteq \mathcal{D}$ Entscheidungsfunktionen, $R_\delta, R_{\delta_1}, R_{\delta_2}, ...$ ihre Risikofunktionen und $\pi, \pi_1, \pi_2, ...$ Wahrscheinlichkeitsmaße auf $\mathcal{P}$ sowie $\Theta \sim \pi, \Theta_1 \sim \pi_1, ...$*

1. *Ist δ_n eine Bayes-Entscheidungsfunktion bezüglich π_n für $\mathcal{D}_0$, $n = 1, 2, ...$ und*

$$\sup_{\theta \in \mathcal{P}} R_\delta(\theta) \leq \limsup_{n \to \infty} r_{\delta_n}(\pi_n),$$

dann ist δ auch eine minimax-Entscheidungsfunktion für $\mathcal{D}_0$.

2. *Ist δ eine Bayes-Entscheidungsfunktion bezüglich π für $\mathcal{D}_0$ mit konstantem Risiko, so ist δ auch minimax für $\mathcal{D}_0$.*

Beweis. 1. Für $\delta \in \mathcal{D}_0$ und $k = 1, 2, ...$ gilt

$$\sup_{\theta\in\mathcal{P}} R_\delta(\theta) \geq \mathbb{E}_{\pi_k}[R_\delta(\Theta_k)] = r_\delta(\pi_k) \geq r_{\delta_k}(\pi_k).$$

Daraus folgt

$$\inf_{\delta'\in\mathcal{D}_0} \sup_{\theta\in\mathcal{P}} R_{\delta'}(\theta) \limsup_{k\to\infty} r_{\delta_k}(\pi_k) \geq \sup_{\theta\in\mathcal{P}} R_\delta(\theta) \geq \inf_{\delta'\in\mathcal{D}_0} \sup_{\theta\in\mathcal{P}} R_{\delta'}(\theta),$$

woraus insbesondere $\sup_{\theta\in\mathcal{P}} R_\delta(\theta) = \inf_{\delta'\in\mathcal{D}_0} \sup_{\theta\in\mathcal{P}} R_{\delta'}(\theta)$ folgt. Deshalb ist δ eine minimax-Entscheidungsfunktion für $\mathcal{D}_0$.
2. ist ein Spezialfall von 1. Für δ mit konstantem Risiko ist nämlich $\sup_{\theta\in\mathcal{P}} R_\delta(\theta) = \mathbb{E}_\pi[R_\delta(\Theta)]$. Damit ist δ auch eine Bayes-Entscheidungsfunktion für $\mathcal{D}_0$. □

Als Beispiel zeigen wir nun die Optimalität (im Sinne der Zulässigkeit) des arithmetischen Mittels zur Schätzung des Erwartungswertes bei normalverteilten Daten.

Proposition 3.21 (Zulässigkeit des arithmetischen Mittels). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathbb{R}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem mit $\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)^n$ für $\sigma^2 > 0$, $n \in \mathbb{N}$, sowie $\aleph = \mathbb{R}$ und ℓ der Gauß-Verlust für die Funktion $g(\theta) = \theta$. Dann ist die Entscheidungsfunktion $d(x) := \overline{x}$ zulässig und minimax.*

Beweis. Die Risikofunktion von d ist

$$R_d(\theta) = \mathbb{E}_\theta[(\overline{X} - \theta)^2] = \frac{\sigma^2}{n}.$$

Angenommen, d ist nicht zulässig. Dann gibt es einen – nach Proposition 3.9 nichtrandomisierte – Schätzer d', der d dominiert. Es gilt also $R_{d'} \leq \sigma^2/n$ und es gibt ein $\theta \in \mathbb{R}$ mit $R_{d'}(\theta) < \sigma^2/n$. Wegen der Stetigkeit der Risikofunktion (in θ für alle Entscheidungsregeln) gibt es also ein $r > 0$, so dass $R_{d'}(\eta) < \sigma^2/n - r$ für $|\eta - \theta| < r$.

Sei nun $\pi_b = \mathcal{N}(0, b^2)$ für $b > 0$. Einerseits ist damit $r_{d'}(\pi_b) < \sigma^2/n$. Andererseits ist nach Beispiel 3.18 der Bayes-Schätzer bezüglich π_b gerade

$$d_b(x) := \frac{nb^2}{nb^2 + \sigma^2}\overline{x}$$

mit Bayes-Risiko

$$\begin{aligned} r_{d_b}(\pi_b) &= \int \mathbb{E}_\theta[(d_b(X) - \theta)^2]\pi_b(d\theta) = \int \mathbb{V}_\theta[d_b(X)] + (\mathbb{E}_\theta[d_b(X)] - \theta)^2\pi_b(d\theta) \\ &= \frac{n^2b^4}{(nb^2+\sigma^2)^2}\frac{\sigma^2}{n} + \frac{\sigma^4b^2}{(nb^2+\sigma^2)^2} = \frac{\sigma^2}{n}\frac{n^2b^4 + n\sigma^2b^2}{(nb^2+\sigma^2)^2} \\ &= \frac{\sigma^2}{n}\Big(\frac{n^2b^4 + 2nb^2\sigma^2 + \sigma^4 - nb^2\sigma^2 - \sigma^4}{(nb^2+\sigma^2)^2}\Big) = \frac{\sigma^2}{n}\Big(1 - \frac{nb^2\sigma^2 + \sigma^4}{(nb^2+\sigma^2)^2}\Big). \end{aligned}$$

Da d_b ein Bayes-Schätzer bezüglich π_b ist, gilt

$$\frac{2nb^2\sigma^2 + \sigma^4}{(nb^2+\sigma^2)^2} \geq \tfrac{\sigma^2}{n} - r_{d_b}(\pi_b) \geq \tfrac{\sigma^2}{n} - r_{d'}(\pi_b) \geq \frac{1}{\sqrt{2\pi b^2}} \int_{\theta-r}^{\theta+r} re^{-\eta^2/(2b^2)}d\eta.$$

Die linke Seite ist offenbar $O(1/b^2)$ für $b \to \infty$, die rechte jedoch $O(1/b)$. Dies ist offenbar ein Widerspruch und damit ist die Zulässigkeit von d gezeigt. Nun ist d auch minimax nach Lemma 3.14. □

Korollar 3.22 (Unbekannte Varianz). *Betrachte dieselbe Situation wie in Proposition 3.21, jedoch mit unbekannter Varianz σ^2, d.h. das statistische Modell $(X, \{\mathbb{P}_\theta : \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\})$ und $g(\mu, \sigma^2) = \mu$. Dann ist $d(x) := \overline{x}$ zulässig und minimax-Schätzer.*

Beweis. Wie im Beweis von Proposition 3.21 ist nur die Zulässigkeit von d zu zeigen. Angenommen, d wäre unzulässig. Dann gibt es einen (oBdA nicht-randomisierten) Schätzer d' mit $R_{d'} \leq R_d$ und ein $\theta = (\mu, \sigma^2)$ mit $R_{d'}(\theta) < R_d(\theta)$. Da $R_{d'}$ und R_d nicht davon abhängen, ob σ^2 bekannt oder unbekannt ist, würde d' den Schätzer nun bei bekannter Varianz σ^2 dominieren. Dies wiederspricht aber der Aussage von Proposition 3.21 und die Behauptung ist gezeigt. □

Bemerkung 3.23 (Höher-dimensionale Normalverteilungen). Interessanterweise lässt sich Proposition 3.21 (und damit Korollar 3.22) nur bedingt auf höhere Dimensionen verallgemeinern. Ist nämlich in der gleichen Situation $X = (X_1, ..., X_n)$ mit $X_i = (X_{i1}, ..., X_{ik}) \in \mathbb{R}^k$ nach $\mathbb{E}_\theta[X_{ij}] = \theta_j$ und $\mathbb{COV}[X_{ij}, X_{il}] = \delta_{jl}$, so kann man zeigen, dass für die Verlustfunktion $\ell(\theta, a) = \sum_{i=1}^k (\theta_i - a_i)^2 = (\theta - a)^\top (\theta - a)$ der Schätzer $d(x) = \overline{x}$ für $k \geq 3$ nicht mehr zulässig ist. Ein dominierender Schätzer ist durch den *James-Stein-Schätzer*

$$d'(x) := \Big(1 - \frac{k-2}{x^\top x}\Big)\overline{x}$$

gegeben. (Übrigens ist d' ebenfalls nicht zulässig und ist von

$$d''(x) := \Big(1 - \frac{k-2}{x^\top x}\Big)^+ \overline{x}$$

dominiert. Aber auch d'' ist nicht zulässig.)
Nun zum Beweis der Unzulässigkeit von d im Fall $k > 2$. OBdA sei $n = 1$, also $d(x) = x$, andersfalls ändert sich nur die Varianz von $d(X)$. Wir berechnen die Risikofunktion von d

$$R_d(\theta)\mathbb{E}_\theta[(\theta - d(X))^\top(\theta - d(X))] = k,$$

da $(\theta - X)^\top(\theta - X) \sim \chi^2_k$. Wir werden nun $R_{d'}(\theta) < k$ zeigen. Es ist nämlich

$$\begin{aligned} R_{d'}(\theta) &= \mathbb{E}_\theta\Big[\Big(X - \theta - \frac{(k-2)X}{X^\top X}\Big)^\top \Big(X - \theta - \frac{(k-2)X}{X^\top X}\Big)\Big] \\ &= \mathbb{E}_\theta[(X-\theta)^\top(X-\theta)] - 2(k-2)\mathbb{E}_\theta\Big[\frac{(X-\theta)^\top X}{X^\top X}\Big] + (k-2)^2\mathbb{E}_\theta\Big[\frac{1}{X^\top X}\Big] \\ &= k - 2(k-2)\sum_{i=1}^k \mathbb{E}_\theta\Big[\frac{(X_i - \theta_i)X_i}{X^\top X}\Big] + (k-2)^2\mathbb{E}_\theta\Big[\frac{1}{X^\top X}\Big]. \end{aligned}$$

Nun ist mit partieller Integration

$$\begin{aligned}\sum_{i=1}^k \mathbb{E}_\theta\Big[\frac{(X_i-\theta_i)X_i}{X^\top X}\Big] &= \sum_{i=1}^k \frac{1}{\sqrt{2\pi k}}\int \frac{x_1}{x^\top x}(x_i-\theta_i)\exp\big(-(x-\theta)^\top(x-\theta)/2\big)dx\\ &= \sum_{i=1}^k \frac{1}{\sqrt{2\pi k}}\int \frac{x^\top x-2x_1^2}{(x^\top x)^2}\exp\big(-(x-\theta)^\top(x-\theta)/2\big)dx\\ &= \frac{k-2}{\sqrt{2\pi k}}\int \frac{1}{x^\top x}\exp\big(-(x-\theta)^\top(x-\theta)/2\big)dx\\ &= (k-2)\mathbb{E}_\theta\Big[\frac{1}{X^\top X}\Big]\end{aligned}$$

und damit

$$R_\theta(d') = k-(k-2)\mathbb{E}_\theta\Big[\frac{1}{X^\top X}\Big] < k = R_\theta(d).$$

Wir wollen nun die entwickelte Theorie auf die Beispiele *Bern* und *Unif* anwenden.

Beispiel 3.24 (Beispiel *Bern*). Wir betrachten das Schätzproblem $((X,\{\mathbb{P}_\theta : \theta\in(0,1)\},\aleph=(0,1),g=\mathrm{id},\ell)$ für den Gauß-Verlust ℓ. Speziell werden wir zeigen, dass $d(x)=(\sum x_i)/n$ eine zulässige, aber keine minimax-Entscheidungsfunktion ist.
Wir wollen zunächst die Zulässigkeit von d zeigen. Offenbar minimiert d das Risiko genau dann für $\ell(\theta,a)=(\theta-a)^2$, wenn d das Risiko für $\ell'(\theta,a)=(\theta-a)^2/(\theta(1-\theta))$ minimiert. Sei nun $\pi=U(0,1)$ eine a-priori-Verteilung. Da dies genau die $\beta(1,1)$-Verteilung ist, sehen wir aus Beispiel 2.38, dass $\pi_x=\beta(1+\sum x_i, n+1-\sum x_i)$ die a-posteriori-Verteilung ist. Um den Bayes-Schätzer zu bestimmen, ist hierfür nach Theorem 3.16.2

$$\mathbb{E}[\ell(\Theta_x,a)] = \frac{\Gamma(n+2)}{\Gamma(1+\sum x_i)\Gamma(n+1-\sum x_i)}\int_0^1 (\theta-a)^2\theta^{\sum x_i-1}(1-\theta)^{n-1-\sum x_i}d\theta$$

zu minimieren. Offenbar liegt dieses Minimum genau beim Erwartungswert der $\beta(\sum x_i, n-\sum x_i)$-Verteilung, also bei[8] $a=(\sum x_i)/n$. Damit ist d Bayes-Schätzer und nach Lemma 3.19 auch zulässig.
Um zu zeigen, dass d nicht minimax ist sei $d_{a,b}(x) := \frac{a+\sum x_i}{a+b+n}$. Für die a-priori-Verteilung $\pi=\beta(a,b)$ ist die a-posteriori-Verteilung $\beta(a+\sum x_i, b+n-\sum x_i)$ und der Bayes-Schätzer ist gerade $d_{a,b}(x)$. Das Risiko dieses Schätzers ist

$$\begin{aligned}R_{d_{a,b}}(\theta) &= \mathbb{E}_\theta[(d_{a,b}(X)-\theta)^2] = \mathbb{V}_\theta[d_{a,b}(X)] + (\mathbb{E}_\theta[d_{a,b}(X)]-\theta)^2\\ &= \frac{1}{(a+b+n)^2}(n\theta(1-\theta)+(a+n\theta-(a+b+n)\theta)^2)\\ &= \frac{1}{(a+b+n)^2}(\theta^2((a+b)^2-n)+\theta(n-2a(a+b))+a^2).\end{aligned}$$

Für $a=b=\sqrt{n}/2$ ist dieser konstant $\frac{n/4}{(n+\sqrt{n})^2}$, und damit ist $d_{\sqrt{n}/2,\sqrt{n}/2}$ nach Theorem 3.20 minimax. Allerdings ist

$$\sup_{\theta\in(0,1)} R_d(\theta) = \sup_{\theta\in(0,1)} R_{d_{0,0}}(\theta) = \sup_{\theta\in(0,1)}\frac{n\theta(1-\theta)}{n^2} = \frac{1}{4n} > \frac{n/4}{(n+\sqrt{n})^2}$$

und damit ist d nicht minimax.

[8] Wir verwenden, dass für $\beta(p,q)$ der Erwartungswert durch $p/(p+q)$ und die Varianz durch $pq/((p+q+1)(p+q)^2)$ gegeben ist.

4 Testtheorie

In diesem Kapitel sei immer $((X, \{\mathbb{P}_\theta : \theta \in \Theta\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem mit $\aleph = \{H_0, H_1\}$ (d.h. die Entscheidung besteht immer zwischen der Null-Hypothese H_0 und der Alternativ-Hypothese H_1). Wir erinnern an die Neyman-Pearson'schen Verlustfunktion

$$\ell(\theta, H_1) = \begin{cases} 0, & \theta \in \mathcal{P}_1, \\ \ell_0, & \theta \in \mathcal{P}_0 \end{cases},$$
$$\ell(\theta, H_0) = \begin{cases} 0, & \theta \in \mathcal{P}_0, \\ \ell_1, & \theta \in \mathcal{P}_1 \end{cases}.$$

für $\mathcal{P}_0, \mathcal{P}_1$ mit $\mathcal{P} = \mathcal{P}_0 \uplus \mathcal{P}_1$. Wir werden immer $\ell_0 = \ell_1 = 1$ betrachten. (Der allgemeinere Fall ergibt sich dann meist ebenfalls.) Die Entscheidung $a = H_0$ bedeutet, dass man sich für die Nullhypothese entscheidet (bzw. dass man sie nicht verwirft) und $a = H_1$ bedeutet, dass man die Nullhypothese verwirft und sich für die Alternative entscheidet. Eine Entscheidungsfunktion (bzw. der Test) δ wir eindeutig durch die Funktion

$$\varphi(x) := \delta(x, \{H_1\})$$

definiert, d.h. durch die Wahrscheinlichkeit, sich für die Alternativ-Hypothese zu entscheiden bei Vorliegen der Daten x. (Damit ist automatisch $\delta(x, \{H_0\}) = 1 - \varphi(x)$.) Das Entscheidungsproblem (oder auch Test-Problem) heißt *einfach*, wenn sowohl $\mathcal{P}_0$ als auch $\mathcal{P}_1$ ein-elementig sind. Andernfalls heißt es *zusammengesetzt*. Wir erinnern an die Gütefunktion $\beta_\varphi : \theta \mapsto \mathbb{E}_\theta[\varphi(X)]$. Das Risiko von δ (oder φ) ist gegeben durch

$$R_\varphi(\theta) := R_\delta(\theta) = \mathbb{E}_\theta\Big[\int \ell(\theta, a)\delta(X, da)\Big] = \begin{cases} \mathbb{E}_\theta[\varphi(X)] = \beta_\varphi(\theta), & \theta \in \mathcal{P}_0, \\ 1 - \mathbb{E}_\theta[\varphi(X)] = 1 - \beta_\varphi(\theta), & \theta \in \mathcal{P}_1. \end{cases}$$

Weiter sei Φ die Menge aller möglichen Tests (gegeben entweder durch δ oder φ) und $\{\theta \mapsto R_\delta(\theta) : \delta \in \Phi\}$ die Risikomenge, d.h. die Menge der Risikofunktionen aller Tests. Für einfache Tests mit $\mathcal{P}_0 = \{\mathbb{P}_0\}, \mathcal{P}_1 = \{\mathbb{P}_1\}$ identifizieren wir

$$\mathcal{R} = \{(\mathbb{E}_0[\varphi(X)], 1 - \mathbb{E}_1[\varphi(X)]) : \varphi \in \Phi\}. \tag{R}$$

4.1 Bayes-Tests

Wie wir in Theorem 3.16 gesehen haben, ist es oftmals gar nicht schwierig, Bayes-Entscheidungsfunktionen aufzustellen. Dies wollen wir nun für Tests durchführen, die entsprechenden Entscheidungsfunktionen heißen dann Bayes-Tests.

Proposition 4.1 (Bayes-Tests). *Sei π ein Wahrscheinlichkeitsmaß auf $\mathcal{P}$. Dann ist $\varphi \in \Phi$ genau dann ein Bayes-Test (d.h. eine Bayes-Entscheidungsfunktion für das Test-Problem) bezüglich π, wenn (für die a-posteriori-Verteilung π_x)*

$$\varphi(x) = \begin{cases} 1, & \pi_x(\mathcal{P}_0) < \pi_x(\mathcal{P}_1), \\ 0, & \pi_x(\mathcal{P}_0) > \pi_x(\mathcal{P}_1). \end{cases}$$

(Auf dem Bereich $\{x : \pi_x(\mathcal{P}_0) = \pi_x(\mathcal{P}_1)\}$ ist φ nicht eindeutig bestimmt.)

Beweis. Wir berechnen das Bayes-Risiko für $\Theta \sim \pi$ für $\Theta_x \sim \pi_x$ und den Test $\psi \in \Phi$ als

$$\begin{aligned} r_\delta(\pi) &= \mathbb{E}\Big[\mathbb{E}\Big[\int \ell(\Theta, a)\delta(X, da)\Big|X\Big]\Big] = \mathbb{E}\Big[\mathbb{E}\Big[\ell(\Theta, H_1)\psi(X) + \ell(\Theta, H_0)(1-\psi(X))\Big|X\Big]\Big] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{1}_{\Theta_X \in \mathcal{P}_0}\psi(X) + \mathbb{1}_{\Theta_X \in \mathcal{P}_1}(1-\psi(X))|X]] \\ &= \mathbb{E}[\pi_X(\mathcal{P}_1) + (\pi_X(\mathcal{P}_0) - \pi_X(\mathcal{P}_1))\psi(X)]. \end{aligned}$$

Die Funktion ψ, die die rechte Seite minimiert, muss genau die angegebene Form haben. Daraus folgt die Behauptung. □

Korollar 4.2 (Bayes-Tests für einfache Tests). *Sei $\mathcal{P}_0 = \{\mathbb{P}_0 = p_0 \cdot \lambda\}, \mathcal{P}_1 = \{\mathbb{P}_1 = p_1 \cdot \lambda^n\}$ (d.h. wir betrachten einen einfachen Test für ein reguläres statistisches Modell) und $\pi_k(\mathcal{P}_0) = \frac{k}{k+1}$ (und damit $(\pi(\mathcal{P}_1) = \frac{1}{k+1})$ für ein $k \in [0, \infty]$[9]. Dann ist φ_k genau dann ein Bayes-Test bezüglich π_k, wenn φ_k gegeben ist als*

$$\begin{aligned} \varphi_k(x) &:= \begin{cases} 1, & \frac{p_1(x)}{p_0(x)} > k, \\ 0, & \frac{p_1(x)}{p_0(x)} < k \end{cases}, \\ \varphi_0(x) &:= 1_{p_1(x)>0}, \\ \varphi_\infty(x) &:= 1_{p_0(x)>0}. \end{aligned}$$

Beweis. Wir zeigen nur den Fall $0 < k < \infty$, die anderen beiden Fälle zeigt man direkt. Es gilt

$$\pi_x(\{H_0\}) = \frac{p_0(x)k}{p_0(x)k + p_1(x)}, \qquad \pi_x(\{H_1\}) = \frac{p_1(x)}{p_0(x)k + p_1(x)}.$$

Aus Proposition 4.1 folgt, dass φ_k Bayes-Tests sind mit

$$\pi_x(\{\mathbb{P}_0\}) < \pi_x(\{\mathbb{P}_1\}) \text{ genau dann, wenn } \frac{p_1(x)}{p_0(x)} > k.$$

□

4.2 Likelihood-Quotienten-Tests

Die soeben konstruierten Bayes-Tests werden gerade durch die Quotienten der Dichten der Alternative und der Nullhypothese bestimmt. Diese Dichten sind – bei gegebenen Daten – außerdem als Likelihood (des Parameters) bekannt. Deshalb nennt man diese Tests auch Likelihood-Quotienten-Tests.

Definition 4.3 (Likelihood-Quotienten-Test). *Sei $\mathcal{P}_0 = \{\mathbb{P}_0 = p_0 \cdot \lambda^n\}, \mathcal{P}_1 = \{\mathbb{P}_1 = p_1 \cdot \lambda^n\}$. Für $k \in [0, \infty]$ heißt φ_k aus Korollar 4.2* Likelihood-Quotienten-Test *(oder LQ-Test) mit kritischem Wert k. Genauer setzen wir*

$$\varphi_k(x) := \varphi_{k,\gamma}(x) := \begin{cases} 1, & \frac{p_1(x)}{p_0(x)} > k, \\ \gamma(x), & \frac{p_1(x)}{p_0(x)} = k, \\ 0, & \frac{p_1(x)}{p_0(x)} < k. \end{cases}$$

Die Menge $\left\{x : \frac{p_1(x)}{p_0(x)} = k\right\}$ heißt Randomisierungsbereich *von φ_k.*

[9] Wir setzen $\infty/\infty = 1$.

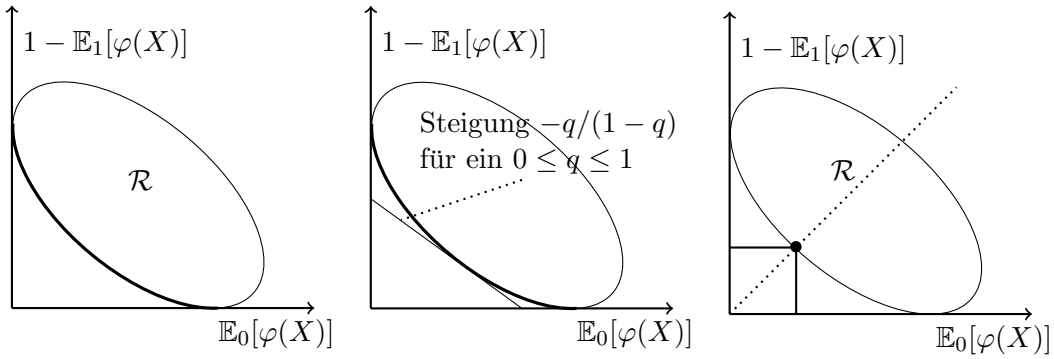
Beispiel 4.4 (Beispiel *Bern*). Im Beispiel *Bern* seien $\theta_0, \theta_1 \in (0,1)$ verschieden. Ein Likelihood-Quotiententest ist von der Form

$$\varphi(x) = \begin{cases} 1, & \frac{\theta_1^{\sum x_i}(1-\theta_1)^{n-\sum x_i}}{\theta_0^{\sum x_i}(1-\theta_0)^{n-\sum x_i}} > k, \\ 0, & \frac{\theta_1^{\sum x_i}(1-\theta_1)^{n-\sum x_i}}{\theta_0^{\sum x_i}(1-\theta_0)^{n-\sum x_i}} < k, \end{cases}$$

also für ein geeignetes k'

$$\varphi(x) = \begin{cases} 1, & \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum x_i} > k', \\ 0, & \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum x_i} < k', \end{cases}$$

Bemerkung 4.5 (Grafische Darstellung der Risikomenge für einfache Tests). Die Menge aller Tests Φ ist konvex, da die Konvexkombination zweier $[0,1]$-wertiger Funktionen wieder eine solche Funktion ergibt. Demnach ist auch die in (R) definierte Risikomenge für einfache Tests eine konvexe Teilmenge des $\mathbb{R}^2_+$. Weiter ist sowohl $\varphi = 0$ als auch $\varphi = 1$ erlaubt, so dass $\mathcal{R}$ sowohl die x- als auch die y-Achse berührt. Hieraus lassen sich die zulässigen, Bayes-Tests (also LQ-Tests) und minimax-Tests ablesen.



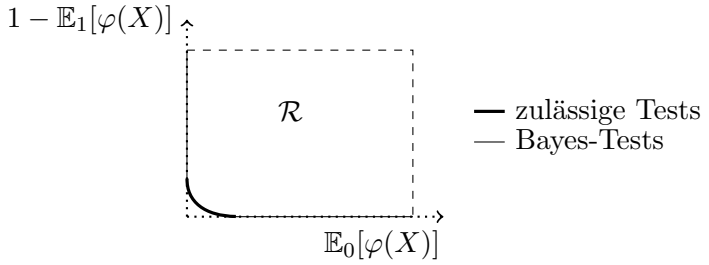
Zulässige Tests Bayes-Tests Minimax-Test

Zulässige Tests ergeben sich für solche φ, für die zwar $R_\varphi \in \mathcal{R}$, aber kein $(x,y) \le R_\varphi$ (in der üblichen Halbordnung in $\mathbb{R}^2_+$) in $\mathcal{R}$ ist.
Ein *Bayes-Test* φ bezüglich $\pi = (q, 1-q)$ minimiert gerade das Skalarprodukt

$$r_\varphi(\pi) = (q,1-q)(\mathbb{E}_0[\varphi(X)], 1 - \mathbb{E}_1[\varphi(X)]) = \min_{(x,y)\in\mathcal{R}} (q,1-q)(x,y).$$

Schreibt man $(x,y) = a(0,1) + b(1-q,-q)$ (d.h. eine parallelverschobene Gerade mit Steigung $-q/(1-q)$, so ist $(q,1-q)(x,y) = a(1-q)$. Dieses Minimum ergibt sich also gerade als die am wenigsten verschobene Gerade mit Steigung $-q/(1-q)$, die $\mathcal{R}$ berührt.
Wir wissen aus Theorem 3.20, dass ein zulässiger Test mit konstantem Risiko ein *minimax-Test* ist. Dies ist gerade für $\mathbb{E}_0[\varphi(X)] = 1 - \mathbb{E}_1[\varphi(X)]$ gegeben, also für Schnittpunkte von $\mathcal{R}$ mit $\{(x,x) : x \ge 0\}$.

Es gibt in diesem Bild die Möglichkeit, dass die Menge der Bayes-Tests größer ist als die der zulässigen Tests, was wir nun veranschaulichen wollen.



Hier bestehen die minimalen Geraden, die $\mathcal{R}$ berühren, ebenfalls aus den horizontalen und vertikalen Begrenzungen von $\mathcal{R}$. Diese sind jedoch keine zulässigen Tests, weil es Punkte $(x, y) \in \mathcal{R}$ gibt, die diese Tests dominieren.

Proposition 4.6 (Zulässige und minimax-LQ-Tests). *Sei $\mathcal{P}_0 = \{\mathbb{P}_0 = p_0 \cdot \lambda^n\}, \mathcal{P}_1 = \{\mathbb{P}_1 = p_1 \cdot \lambda^n\}$ und φ_k für $k \in [0, \infty]$ ein Likelihood-Quotienten-Test mit kritischem Parameter k. Dann gilt:*

1. *Ist φ ein zulässiger Test, so gibt es $k \in [0, \infty]$ mit $\varphi = \varphi_k$.*
2. *Gilt $\mathbb{E}_0[\varphi_k(X)] > 0, \mathbb{E}_1[\varphi_k(X)] < 1$, so ist φ_k zulässig.*
3. *Genau dann ist φ ein minimax-Test, wenn es $k \in [0, \infty]$ gibt mit $\varphi = \varphi_k$ und $\mathbb{E}_0[\varphi_k(X)] = \mathbb{E}_1[1 - \varphi_k(X)]$.*

Beweis. 1. Wir verwenden die grafische Darstellung aus Bemerkung 4.5. Ist φ mit $(\mathbb{E}_0[\varphi(X), 1 - \mathbb{E}_1[\varphi(X)]) \in \mathcal{R}$ zulässig (also Element des unteren Randes der Risikomenge), so lesen wir ab, dass es eine minimale berührende Gerade gibt, die $\mathcal{R}$ gerade in $(\mathbb{E}_0[\varphi(X), 1 - \mathbb{E}_1[\varphi(X)])$ berührt. Damit ist φ ein Bayes-Test, also nach Korollar 4.2 auch ein LQ-Test.
2. Wir müssen ausschließen, dass $k = 0$ oder $k = \infty$. Für $k \in (0, \infty)$ ist nämlich φ_k Bayes-Test zu einer a-priori-Verteilung mit vollem Träger und damit nach Lemma 3.19 zulässig. Offenbar ist $\mathbb{P}_i[p_i(X) > 0] = 1$, $i = 0, 1$. Wäre $k = 0$, so gilt aber $\mathbb{E}_1[\varphi_0(X)] = \mathbb{P}_1(p_1(X) > 0) = 1$ im Widerspruch zur Voraussetzung und wäre $k = \infty$, so gilt $\mathbb{E}_0[\varphi_\infty(X)] = \mathbb{P}_0[p_0(X) > 0] = 1$.
3. '$\Leftarrow$': Offenbar ist φ_k Bayes-Test mit konstantem Risiko. Deshalb folgt die Behauptung aus Theorem 3.20.2.
'$\Rightarrow$': Wir zeigen zunächst, dass $\mathbb{E}_0[\varphi(X)] = 1 - \mathbb{E}_1[\varphi(X)]$ gelten muss. Angenommen, es wäre $\lambda := 1 - \mathbb{E}_1[\varphi(X)] - \mathbb{E}_0[\varphi(X)] > 0$. Wir definieren dann die neue Entscheidungsfunktion

$$\psi(X) := \frac{1}{1+\lambda}\varphi(X) + \frac{\lambda}{1+\lambda} \in \Phi.$$

Nun gilt

$$\begin{aligned}\mathbb{E}_0[\psi(X)] &= \frac{1}{1+\lambda}\mathbb{E}_0[\varphi(X)] + \frac{1}{1+\lambda}(1 - \mathbb{E}_1[\varphi(X)] - \mathbb{E}_0[\varphi(X)]) \\ &= \frac{1}{1+\lambda}(1 - \mathbb{E}_1[\varphi(X)]) \\ &= 1 - \frac{1}{1+\lambda}\mathbb{E}_1[\varphi(X)] - \frac{\lambda}{1+\lambda} = 1 - \mathbb{E}_1[\psi(X)].\end{aligned}$$

Weiter gilt

$$\begin{aligned}\max\big(\mathbb{E}_0[\psi(X)], 1-\mathbb{E}_1[\psi(X)]\big) &= \frac{1}{1+\lambda}(1-\mathbb{E}_1[\varphi(X)]) < 1-\mathbb{E}_1[\varphi(X)] \\ &= \max\big(\mathbb{E}_0[\varphi(X)], 1-\mathbb{E}_1[\varphi(X)]\big).\end{aligned}$$

Dies steht im Widerspruch dazu, dass φ minimax ist. Im Fall $\lambda < 0$ führt eine ähnliche Konstruktion zum Ziel. Nun ist also $(\mathbb{E}_0[\varphi(X)], 1-\mathbb{E}_0[\varphi(X)])) \in \mathcal{R} \cap \{(x,x) : x \in \mathbb{R}_+\}$. Aus der grafischen Darstellung aus Bemerkung 4.5 und der Konvexität der Risikomenge $\mathcal{R}$ folgt, dass φ ein Bayes-Test sein muss und damit ein LQ-Test. □

Auch für zusammengesetzte Hypothesen kann man Likelihood-Quotiententests definieren. Dies wollen wir nun tun, auch wenn wir in diesem Fall nur Beispiele angeben und die Optimalität der Tests nicht weiter untersuchen werden.

Definition 4.7 (Likelihood-Quotienten-Tests für zusammengesetzte Alternativen). *Sei $(X, \{\mathcal{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}_0\}, \aleph, \ell\})$ ein reguläres statistisches Modell, $\mathcal{P} = \mathcal{P}_0 \uplus \mathcal{P}_1$ eine Partition von $\mathcal{P}$ und $H_0 : \theta \in \mathcal{P}_0$, $H_1 : \theta \in \mathcal{P}_1$, sowie $\aleph = \{H_0, H_1\}$ und ℓ der Neyman-Pearon-Verlust. Dann heißt ein Test φ_k mit*

$$\varphi_k(x) := \varphi_{k,\gamma}(x) := \begin{cases} 1, & \lambda(x) := \frac{\sup_{\theta\in\mathcal{P}_0} p_\theta(x)}{\sup_{\theta\in\mathcal{P}} p_\theta(x)} < k, \\ \gamma(x), & \lambda(x) = k, \\ 0, & \lambda(x) > k \end{cases}$$

Likelihood-Quotienten-Test *mit kritischem Wert k.*

Bemerkung 4.8 (Mögliche Werte für k). Anders als bei Likelihood-Quotienten-Tests von einfachen Hypothesen ist die Definition bei zusammengesetzten Hypothesen so, dass immer $\lambda \leq 1$ ist. Es machen also nur $k \in [0,1]$ Sinn. Diese Werte legen natürlich auch das Signifikanzniveau fest.

Beispiel 4.9 (Test auf den Parameter einer Exponentialverteilung). Sei $\mathcal{P} = [\theta_0, \infty)$ und $\mathbb{P}_\theta = \exp(\theta)^n$. Es ist unter $\mathbb{P}_\theta$ also $X = (X_1, ..., X_n)$ unabhängig und $X_i \sim \exp(\theta)$. Weiter sei $\mathcal{P}_0 = \{\theta_0\}$ und $\mathcal{P}_1 = (\theta_0, \infty)$. Nun ist φ genau dann ein Likelihood-Quotienten-Test, falls

$$\varphi(x) := 1_{\{\sum_{i=1}^n x_i > c\}}$$

für ein $c > 0$.
Denn: Zunächst berechnen wir $\sup_{\theta\in\mathcal{P}} p_\theta(x)$. Wir schreiben hierfür

$$\log p_\theta(x) = \log(\theta^n e^{-\theta(x_1+\cdots+x_n)}) = n\log\theta - \theta(x_1+\cdots+x_n)$$

und damit $\frac{d}{d\theta}\log p_\theta(x) = \frac{n}{\theta} - (x_1+\cdots+x_n)$, also

$$\log\frac{p_{\theta_0}(x)}{\sup_{\theta\in\mathcal{P}} p_\theta(x)} = \begin{cases} \log p_{\theta_0(x)} - \log p_{1/\overline{x}}(x) = n\log\theta_0 - \theta_0 n\overline{x} + n\log\overline{x} - n, & \overline{x} < 1/\theta_0, \\ 0, & \overline{x} \geq 1/\theta_0. \end{cases}$$

Damit ist ein Likelihood-Quotienten-Test von der Form $\varphi(x) = 1_{\log\overline{x}-\theta_0\overline{x}<k} 1_{\overline{x}>1/\theta_0}$. Nun ist $\overline{x} \mapsto \log\overline{x} - \theta_0\overline{x}$ für $\overline{x} \geq 1/\theta_0$ monoton fallend, d.h. φ ist von der Form $\varphi(x) = 1_{\overline{x}>k'}$.

Bereits bekannte Tests, etwa der t-Test, sind ebenfalls Likelihood-Quotienten-Tests.

Proposition 4.10 (Einfacher t-Test ist ein Likelihood-Quotienten-Test). *Sei $(X, \{\mathbb{P}_\theta = \mathcal{N}(\mu,\sigma^2)^n : \theta = (\mu,\sigma)^2 \in \mathcal{P}_0 := \mathbb{R}\times\mathbb{R}_+\})$ das Normalverteilungsmodell, $\mathcal{P}_0 = \{\mathbb{P}_\theta : \theta = (\mu_0,\sigma^2) \in \{\mu_0\}\times\mathbb{R}_+\}$, $\mathcal{P}_1 = \mathcal{P}\setminus\mathcal{P}_0$, $\aleph = \{H_0, H_1\}$ und ℓ der Neyman-Pearson-Verlust. Dann ist φ genau dann ein Likelihood-Quotienten-Test, wenn $\varphi(x) = 1_{|t(x)|\geq c}$ für ein geeignetes c mit*

$$t(x) = \frac{\overline{x}-\mu_0}{\sqrt{s^2(x)/n}}.$$

Beweis. Sei $\mathbb{P}_\theta = p_\theta\cdot\lambda^n$ mit

$$p_\theta(x) = (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{\sum_{i=1}^n(x_i-\theta)^2}{2\sigma^2}\Big).$$

Bekannt ist, dass der Maximum-Likelihood-Schätzer für $\theta = (\mu,\sigma^2)$ immer durch $(\overline{x}, (n-1)s^2(x)/n)$ gegeben ist. Bei gegebenen μ ist der Maximum-Likelihood-Schätzer von σ^2 durch $\hat{s}^2(x) := \frac{1}{n}\sum_{i=1}^n(x_i-\mu)^2$ gegeben. Um also p_θ zu maximieren, berechnen wir erst

$$\begin{aligned}\sup_{\sigma^2\in\mathbb{R}_+} p_\theta(x) &= (2\pi\hat{s}^2(x))^{-n/2}\exp\Big(-\frac{\sum_{i=1}^n(x_i-\mu)^2}{2\hat{s}^2(x)}\Big)\\ &= \Big(\frac{2\pi}{n}\sum_{i=1}^n(x_i-\mu)^2\Big)^{-n/2}e^{-n/2}\end{aligned}$$

und damit

$$\begin{aligned}\lambda(x) &= \Big(\frac{\sum_{i=1}^n(x_i-\mu_0)^2}{\sum_{i=1}^n(x_i-\overline{x})^2}\Big)^{-n/2} = \Big(\frac{\sum_{i=1}^n(x_i-\overline{x})^2+(\overline{x}-\mu_0)^2}{\sum_{i=1}^n(x_i-\overline{x})^2}\Big)^{-n/2}\\ &= \Big(1+|t(x)|^2\frac{n-1}{n}\Big)^{-n/2}.\end{aligned}$$

Die Behauptung folgt nun daraus, dass $x\mapsto(1+(n-1)x/n)^{-n/2}$ monoton ist. □

Bemerkung 4.11 (Approximative Verteilung des Likelihood-Quotienten). Um das Signifikanzniveau eines Tests φ zu bestimmen, benötigt man $\sup_{\theta\in\mathcal{P}_0}\mathbb{E}_\theta[\varphi(X)]$. Da bei Likelihood-Quotienten-Tests $\varphi = \varphi_{k,\gamma}$ von den Parametern k und γ abhängt, heißt das, dass man diese so bestimmen muss, dass

$$\mathbb{E}_\theta[\varphi_{k,\gamma}(X)] = \mathbb{P}_\theta[\lambda(X)\leq k] + \mathbb{E}_\theta[\gamma(X),\lambda(X)=k] \leq \alpha,\quad \theta\in\mathcal{P}_0.$$

Hierzu benötigt man also die Verteilung des Likelihood-Quotienten λ.

Sei $\mathcal{P}\subseteq\mathbb{R}^p$ und $\mathcal{P}_0 = \{\eta\}$. Wir werden (mit Hilfe der noch zu zeigenden Aussagen aus Abschnitt 5.4) nun zeigen, dass (unter gewissen Regularitätsannahmen)

$$-2\log\lambda(X) =: \Lambda(X) \overset{n\to\infty}{\Longrightarrow} Z\sim\chi^2_p.$$

Sei $\hat{\theta}$ der Maximum-Likelihood-Schätzer für θ. Dann schreiben wir

$$\begin{aligned}\sum_{k=1}^n\log p_\eta^n(x_k) = \sum_{k=1}^n\Big(&\log p_{\hat{\theta}}(x_k) + \sum_{i=1}^p\frac{\partial}{\partial\theta_i}\log p_{\hat{\theta}}(x_k)(\eta_i-\hat{\theta}_i)\\ &+\frac{1}{2}\sum_{i,j=1}^k\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\log p_{\hat{\theta}}(x_k)(\eta_i-\hat{\theta}_i)(\eta_j-\hat{\theta}_j)\Big) + o((\eta-\hat{\theta})^2).\end{aligned}$$

Nun verschwindet der zweite Term, da $\hat{\theta}$ die Maximalstelle von $\log p_\theta$ ist, und damit ist für $X \sim \mathbb{P}_\eta^n$

$$\begin{aligned}\Lambda(X) &= 2\sum_{k=1}^{n} \log p_{\hat{\theta}}(X_k) - 2\log p_\eta(X_k) \\ &= -\sum_{k=1}^{n}\sum_{i,j=1}^{p} \frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j} \log p_{\hat{\theta}}(X_k)(\eta_i - \hat{\theta}_i)(\eta_j - \hat{\theta}_j) + o((\eta-\hat{\theta})^2) \\ &\overset{n\to\infty}{\approx} -n\sum_{i,j=1}^{p} \mathbb{E}_\eta\Big[\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\log p_\eta(X)\Big](\eta_i - \hat{\theta}_i)(\eta_j - \hat{\theta}_j),\end{aligned}$$

da $\hat{\theta}$ konsistent ist (siehe Theorem 5.30). Nun ist nach Theorem 5.31 und Bemerkung 5.18 für

$$I(\theta) := -\Big(\mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\Big]\Big)_{i,j=1,\dots,k}$$

gerade (für die Einheitsmatrix E_k)

$$\sqrt{n}(I(\eta))^{1/2}(\hat{\theta}-\eta) \overset{n\to\infty}{\Longrightarrow} \mathcal{N}(0, E_p),$$

also

$$\Lambda(X) \approx n(\hat{\theta}-\eta)I(\eta)(\hat{\theta}-\eta) \overset{n\to\infty}{\Longrightarrow} Z_1^2 + \cdots + Z_p^2 \sim \chi_p^2$$

4.3 Beste Tests

Optimale Tests minimieren die Risikofunktion. Allerdings ist es nicht sinnvoll, alle Tests zuzulassen, sondern nur solche, deren Fehler erster Art kleiner als ein vorgegebenes α sind. Diese Tests heißen auch beste Tests.

Definition 4.12 (Niveau eines Tests). *1. Für $\alpha \in [0,1]$ ist*

$$\Phi_\alpha := \{\varphi \in \Phi : \mathbb{E}_\theta[\varphi(X)] \le \alpha \text{ für alle } \theta \in \mathcal{P}_0\}$$

die Menge aller Tests zum Niveau α.

2. Ein Test φ heißt (gleichmäßig) bester Tests zum Niveau α *(oder* (Uniformly) Most Powerful Test oder UMP-Test*), falls*

$$\mathbb{E}_\theta[\varphi(X)] = \sup_{\psi\in\Phi_\alpha} \mathbb{E}_\theta[\psi(X)], \qquad \theta \in \mathcal{P}_1.$$

Das folgende Resultat erlaubt die Konstruktion bester Tests im Falle von einfachen Hypothesen.

Theorem 4.13 (Neyman-Pearson-Lemma). *Wir betrachten das reguläre statistische Modell $(X, \{\mathbb{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ mit $\mathcal{P}_i = \{\mathbb{P}_i\}, i = 0,1$ und $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$. Sei $\alpha \in (0,1)$.*

1. Es gibt einen LQ-Test $\varphi_{k,\gamma}$ mit $\gamma \in [0,1]$ konstant und $\mathbb{E}_0[\varphi(X)] = \alpha$.

2. Ist φ ein LQ-Test mit $\mathbb{E}_0[\varphi(X)] = \alpha$, so ist φ bester Test zum Niveau α.

3. *Ist φ ein bester Test zum Niveau α, so ist φ ein LQ-Test. Es gilt dann entweder $\mathbb{E}_0[\varphi(X)] = \alpha$ oder $\mathbb{E}_1[\varphi(X)] = 1$.*

Beweis. 1. Wir setzen

$$k := \inf\{y : \mathbb{P}_0(p_1(X)/p_0(X) > y) \leq \alpha\}$$

und haben damit k bereits bestimmt. Es gilt $\mathbb{P}_0(p_1(X)/p_0(X) \geq k) \geq \alpha \geq \mathbb{P}_0(p_1(X)/p_0(X) > k)$. Für die Wahl von γ gibt es nun zwei Möglichkeiten. Ist $\mathbb{P}_0(p_1(X)/p_0(X) = k) = 0$, so setzen wir $\gamma = 0$. Andernfalls setzen wir

$$\gamma := \frac{\alpha - \mathbb{P}_0(p_1(X)/p_0(X) > k)}{\mathbb{P}_0(p_1(X)/p_0(X) = k)} \in (0, \infty).$$

Nun gilt nach Definition des LQ-Tests $\varphi^*_{k,\gamma}$

$$\mathbb{E}_0[\varphi^*_{k,\gamma}(X)] = \mathbb{P}_0(p_1(X)/p_0(X) > k) + \gamma \mathbb{P}_0(p_1(X)/p_0(X) = k) = \alpha.$$

2. Sei $\varphi = \varphi^*_k$ für ein $k \in [0, \infty]$ sowie $\psi \in \Phi_\alpha$. Wir müssen zeigen, dass

$$\mathbb{E}_1[\varphi^*_k(X)] \geq \mathbb{E}_1[\psi(X)]$$

und schreiben hierfür

$$\begin{aligned}
&\mathbb{E}_1[\varphi^*(X)] - \mathbb{E}_1[\psi(X)] \\
&\qquad = \int (\varphi^*(x) - \psi(x))(p_1(x) - k p_0(x))\lambda^n(dx) + k \int (\varphi^*_k(x) - \psi(x)) p_0(x) \lambda^n(dx) \\
&\qquad = \int (1_{p_1(x) > k p_0(x)} - \psi(x))(p_1(x) - k p_0(x))\lambda^n(dx) + k \mathbb{E}_0[\varphi^*_k(X) - \psi(X)] \\
&\qquad \geq \int 1_{p_1(x) > k p_0(x)} (1 - \psi(x))(p_1(x) - k p_0(x))\lambda^n(dx) \geq 0.
\end{aligned}$$

3. Sei φ bester Test zum Niveau α und φ^*_k der LQ-Test zum Niveau α aus 1. Nach 2. ist φ^*_k ebenfalls bester Test zum Niveau α und wir müssen zeigen, dass $\varphi = \varphi^*_k$. In der Rechnung aus dem Beweis von 2. muss Gleichheit in beiden Abschätzungen gelten, es gilt also $\mathbb{E}_0[\varphi(X)] = \mathbb{E}_0[\varphi^*_k(X)] = \alpha$ sowie $(\varphi^*_k - \varphi)(p_1 - k p_0) \overset{\lambda^n\text{-f.ü.}}{=} 0$. Daraus folgt $\{x : \varphi^*_k(x) \neq \varphi(x)\} \subseteq \{x : p_1(x)/p_0(x) = k\}$, d.h. φ muss ein LQ-Test sein, der höchstens auf dem Randomisierungsbereich nicht mit φ^*_k übereinstimmt. □

Ziel des restlichen Kapitels ist es, die Optimalität von Tests bei zusammengesetzten Hypothesen zu zeigen. Dies ist vor allem dann möglich, wenn das statistische Modell stochastisch geordnet ist.

Definition 4.14 (Monotone Dichtequotienten). *Sei $(\mathcal{P}, \leq)$ total geordnet und $(X, \{\mathcal{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ ein reguläres statistisches Modell und $T = t(X)$ für $t : \mathbb{R}^n \to \mathbb{R}$. Dann hat $\mathcal{P}$ einen monotonen Dichtequotienten in t, wenn für $\theta, \theta' \in \mathcal{P}$ mit $\theta \leq \theta'$ eine monotone Abbildung $p_{\theta,\theta'}$ existiert mit*

$$\frac{p_{\theta'}}{p_\theta} = p_{\theta,\theta'} \circ t$$

($(p_\theta + p_{\theta'}) \cdot \lambda^n$-fast sicher).

Beispiel 4.15 (Exponentialfamilie). Sei $p_\theta(x) = h(x)\exp\big(c(\theta)^\top t(x) - d(\theta)\big)$, d.h. $(X, \{\mathbb{P}_\theta = p_\theta \cdot \lambda^n\})$ ist eine ein-parametrige Exponentialfamilie. Dann ist

$$\frac{p_{\theta'}(x)}{p_\theta(x)} = \exp\big((c(\theta') - c(\theta))^\top t(x) - (d(\theta') - d(\theta))\big),$$

also hat $\mathcal{P}$ genau dann einen monotonen Dichtequotienten in t, wenn $\mathcal{P}$ total geordnet ist und c monoton ist.

Beispiel 4.16 (Beispiel *Norm*). Für das Modell aus Beispiel 2.23 mit bekanntem $\sigma^2 > 0$ ist

$$p_\theta(x) = \exp\Big(-\frac{x^2}{2\sigma^2}\Big) \cdot \exp\Big(\frac{\theta x}{\sigma^2} - \frac{1}{2}\Big(\frac{\theta^2}{\sigma^2} + \log(2\pi\sigma^2)\Big)\Big),$$

es handelt sich also um eine 1-parametrige Exponentialfamilie mit

$$c(\theta) = \frac{\theta}{\sigma^2}, \qquad t(x) = x,$$
$$h(x) = \exp\Big(-\frac{x^2}{2\sigma^2}\Big), \qquad d(\theta) = -\frac{1}{2}\Big(\frac{\theta^2}{\sigma^2} + \log(2\pi\sigma^2)\Big).$$

insbesondere ist c monoton, und damit hat $\mathcal{P}$ einen monotonen Dichtequotienten.

Proposition 4.17 (Stochastische Monotonie). *Sei $(\mathcal{P}, \leq)$ total geordnet und $(X, \{\mathcal{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ ein reguläres statistisches Modell und $T = t(X)$ für $t : \mathbb{R}^n \to \mathbb{R}$. Sei f isoton (und so, dass $\mathbb{E}_\theta[f(t(X))]$ für alle $\theta \in \mathcal{P}$ existiert), und hat $\mathcal{P}$ einen monotonen Dichtequotienten in t, dann ist $\theta \mapsto \mathbb{E}_\theta[f(t(X))]$ ebenfalls isoton.*

Beweis. Wir schreiben für $\theta \leq \theta'$

$$\begin{aligned}
\mathbb{E}_{\theta'}[f(t(X))] - \mathbb{E}_\theta[f(t(X))] &= \int (f(t(y)) - f(t(x)))p_{\theta'}(y)p_\theta(x)\lambda^n(dx)\lambda^n(dy) \\
&= \int 1_{t(y)>t(x)}((f(t(y)) - f(t(x)))p_{\theta'}(y)p_\theta(x) \\
&\qquad + (f(x) - f(y))p_{\theta'}(x)p_\theta(y))\lambda^n(dx)\lambda^n(dy) \\
&= \int 1_{t(y)>t(x)}(f(t(y)) - f(t(x)))(p_{\theta'}(y)p_\theta(x) - p_{\theta'}(x)p_\theta(y))\lambda^n(dx)\lambda^n(dy) \\
&\geq 0,
\end{aligned}$$

da $f \circ t$ isoton ist und $\frac{p_{\theta'}(y)}{p_\theta(y)} = p_{\theta,\theta'}(t(y)) \geq p_{\theta,\theta'}(t(x)) = \frac{p_{\theta'}(x)}{p_\theta(x)}$ auf $\{t(y) > t(x)\}$. □

Theorem 4.18 (Beste Tests bei einseitigen, zusammengesetzten Hypothesen). *Sei $(\mathcal{P}, \leq)$ total geordnet und $(X, \{\mathcal{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ ein reguläres statistisches Modell, $T = t(X)$ für $t : \mathbb{R}^n \to \mathbb{R}$ und habe $\mathcal{P}$ einen streng monotonen Dichtequotienten in t. Weiter sei $\alpha \in [0, 1]$ und $\theta_0 \in \mathcal{P}$ so, dass $\mathcal{P}_0 = \{\theta \leq \theta_0\}, \mathcal{P}_1 = \{\theta > \theta_0\}$. Dann gilt:*

1. *Sei*

$$\psi(x) := \psi_{m,\gamma}(x) := \begin{cases} 1, & t(x) > m, \\ \gamma, & t(x) = m, \\ 0, & t(x) < m \end{cases}$$

für ein $m \in [0, \infty]$ und $\gamma \in [0, 1]$, so dass $\mathbb{E}_{\theta_0}[\psi(X)] = \alpha$. Dann ist ψ bester Test zum Niveau α.

2. *Die Gütefunktion* $\beta_\psi : \theta \mapsto \mathbb{E}_\theta[\psi(X)]$ *ist isoton und für* $\theta < \theta_0$ *gilt*

$$\mathbb{E}_\theta[\psi(X)] = \inf\{\mathbb{E}_\theta[\varphi(X)] : \varphi \in \Phi, \mathbb{E}_{\theta_0}[\varphi(X)] \geq \alpha\},$$

d.h. ψ *minimiert den Fehler erster Art.*

Beweis. Wir beginnen mit dem Beweis von 1. Sei zunächst $\theta_1 > \theta_0$. Wir betrachten den einfachen Test mit $\mathcal{P}_i = \{\theta_i\}, i = 0, 1$ und zeigen, dass ψ auch in diesem Fall ein LQ-Test ist. Wegen des streng monotonen Dichtequotienten ist für $k := p_{\theta_0,\theta_1}(m)$

$$\begin{aligned}
\{t(x) > m\} &= \{p_{\theta_0,\theta_1}(t(x)) > p_{\theta_0,\theta_1}(m)\} = \{p_{\theta_1}(x)/p_{\theta_0}(x) > k\},\\
\{t(x) = m\} &= \{p_{\theta_0,\theta_1}(t(x)) = p_{\theta_0,\theta_1}(m)\} = \{p_{\theta_1}(x)/p_{\theta_0}(x) = k\},\\
\{t(x) < m\} &= \{p_{\theta_0,\theta_1}(t(x)) < p_{\theta_0,\theta_1}(m)\} = \{p_{\theta_1}(x)/p_{\theta_0}(x) < k\}.
\end{aligned}$$

Damit ist ψ also LQ-Test und $\mathbb{E}_{\theta_0}[\psi(X)] = \alpha$ nach Voraussetzung. Nach dem Neyman-Pearson-Lemma, Theorem 4.13, ist also ψ bester Test zum Niveau α, d.h. es gilt $\mathbb{E}_{\theta_1}[\psi(X)] = \sup_{\varphi\in\Phi_\alpha} \mathbb{E}_{\theta_1}[\varphi(X)]$. Da dies aber für alle $\theta_1 \in \mathcal{P}_1$ gilt, folgt, dass ψ bester Test zum Niveau α für $\mathcal{P}_0 = \{\theta_0\}, \mathcal{P}_1 = \{\theta > \theta_0\}$ ist. Wegen $\psi = 1_{\{.\geq m\}} \circ t$, der Monotonie von $1_{\{.\geq m\}}$ und da $\mathcal{P}$ einen monotonen Dichtequotienten in t hat, folgt aus Proposition 4.17, dass $\theta \mapsto \mathbb{E}_\theta[\psi(X)]$ monoton ist (d.h. die in 2. behauptete Isotonie der Gütefunktion ist gezeigt). Insbesondere gilt für $\theta \leq \theta_0$, dass $\mathbb{E}_\theta[\psi(X)] \leq \mathbb{E}_{\theta_0}[\psi(X)] = \alpha$, also $\psi \in \Phi_\alpha$. Nach Definition ist damit ψ bester Test zum Niveau für $\mathcal{P}_0 = \{\theta \leq \theta_0\}$, $\mathcal{P}_1 = \{\theta > \theta_0\}$. Es ist noch zu zeigen, dass ψ den Fehler erster Art minimiert. Sei hierzu $\theta < \theta_0$. Genau wie in 1. zeigt man, dass $1-\psi$ ein LQ-Test zum Niveau $1-\alpha$ ist für $\mathcal{P}_0 = \{\theta_0\}, \mathcal{P}_1 = \{\theta < \theta_0\}$. Insbesondere ist $1-\psi$ bester Test zum Niveau $1-\alpha$, d.h. für $1-\varphi \in \Phi_{1-\alpha}$ (oder $\mathbb{E}_{\theta_0}[1-\varphi(X)] \leq 1-\alpha$) ist $\mathbb{E}_\theta[1-\varphi] \leq \mathbb{E}_\theta[1-\psi], \theta < \theta_0$, also auch $\mathbb{E}_\theta[\psi] \leq \mathbb{E}_\theta[\varphi]$ für $\mathbb{E}_\theta[\varphi(X)] \geq \alpha$ und damit die Behauptung. □

Korollar 4.19 (Gauß-Test). *Sei* $\mathcal{P} = \mathbb{R}$, $(X, \{\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)^N)$ *für ein* $\sigma^2 > 0$ *wie in* (Norm 1*b*) *und* $\theta_0 \in \mathbb{R}$ *mit* $\mathcal{P}_0 = (-\infty, \theta_0], \mathcal{P}_1 = (\theta_0, \infty)$. *Weiter sei* q_α *das* α*-Quantil von* $\mathcal{N}(0, 1)$. *Dann ist*

$$\psi(x) := 1\Big(\frac{\overline{x} - \theta_0}{\sqrt{\sigma^2/n}} \geq q_\alpha\Big)$$

bester Test zum Niveau α.

Beweis. Nach Beispiel 2.23 und Bemerkung 2.28 ist $\mathcal{N}(\theta, \sigma^2)^n$ eine ein-parametrige Exponentialfamilie mit $t(x) = \overline{x}$. Deshalb hat diese Familie nach Proposition 4.17 einen monotonen Dichtequotienten. Weiter ist $\mathbb{E}_{\theta_0}[\psi(X)] = \mathbb{P}(Z \geq q_\alpha) = \alpha$ für $Z \sim \mathcal{N}(0, 1)$. Damit ist ψ nach Theorem bester Test zum Niveau α. □

Bemerkung 4.20 (Weitere beste Tests). Wir haben nun den einfachsten Fall eines besten Tests behandelt. Ebenfalls beste Tests sind zweiseitige Gauß-Tests (hier genügen relativ einfache Abschätzungen), aber auch einfache t-Tests sowie χ^2-Tests. Bei den t-Tests ist σ^2 unbekannt, und deshalb ist $\mathcal{P}$ nicht mehr total geordnet. Deshalb benötigt man dort den Begriff des bedingten Tests, um die Optimalität des Tests zu zeigen.

5 Schätztheorie

Im Folgenden sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ mit einem statistischen Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ und $g : \mathcal{P} \to \aleph$ eine Abbildung auf den Entscheidungsraum $\aleph$ (typischerweise mit $\aleph \subseteq \mathbb{R}^k$ für ein k). In diesem Kapitel betrachten wir nur nicht-randomisierte Entscheidungsfunktionen oder Schätzer $d : E \to \aleph$. Wir sagen hier, dass d ein Schätzer für g ist.

5.1 Grundlagen

Es gibt verschiedene Prinzipien, mit denen man zu Schätzern kommt. Wir stellen in diesem Abschnitt das Substitutionsprinzip (und darauf aufbauend die Momentenmethode) und das Maximum-Likelihood-Prinzip vor. Aber erst gibt es ein paar Grundbegriffe.

Definition 5.1 (Bias, Mean Squared Error). *Sei d ein Schätzer für g und $\aleph \subseteq \mathbb{R}$. Dann heißt*

$$b_\theta(d) := \mathbb{E}_\theta[d(X)] - g(\theta)$$

der Bias *oder die* Verzerrung *von d. Im Fall $b_\theta(d) = 0$ für alle $\theta \in \mathcal{P}$ heißt d* unverzerrt *oder* erwartungstreu *oder* unbiased. *Weiter heißt*

$$\mathbb{E}_\theta[(d(X) - g(\theta))^2]$$

die mittlere quadratische Abweichung *oder der* Mean Squared Error.

Lemma 5.2 (Zerlegung des Risikos). *Sei $\aleph \subseteq \mathbb{R}$, d ein Schätzer für g und ℓ der Gauß-Verlust, d.h. $\ell(\theta, a) = |a - g(\theta)|^2$. Dann gilt für die Risikofunktion*

$$R_d(\theta) = \mathbb{V}_\theta[d(X)] + b_\theta(d)^2.$$

Beweis. Wir schreiben

$$\begin{aligned} R_d(\theta) &= \mathbb{E}_\theta[(d(X) - g(\theta))^2] = \mathbb{E}_\theta[(d(X)^2] - \mathbb{E}_\theta[(d(X)]^2 + \mathbb{E}_\theta[(d(X)]^2 - 2g(\theta)\mathbb{E}_\theta[d(X)] + g(\theta)^2 \\ &= \mathbb{V}_\theta[d(X)] + b_\theta(d)^2. \end{aligned}$$

□

Definition 5.3 (Plug-in-Schätzer, Maximum-Likelihood-Schätzer). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ ein Schätzproblem.*

1. *Sei $X = (X_1, ..., X_n)$ ein Vektor unabhängiger und identisch verteilter Zufallsvariablen. Dann heißt die (zufällige) Wahrscheinlichkeitsverteilung*

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

   empirische Verteilung von X.

2. *Sei $X = (X_1, ..., X_n)$ ein Vektor identisch verteilter Zufallsvariablen und $g(\theta) = \mathbb{E}_\theta[f(X_1)]$ für eine Funktion f, falls der Erwartungswert existiert. Dann heißt*

$$\widehat{g}(X) := E_X[f] = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

   Plug-in-Schätzer für g. Plug-In-Schätzer für $g(\theta) = \mathbb{E}_\theta[X_1]$ heißen auch momentenbasierte Schätzer.

3. *Sei* $(X, \{\mathbb{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ *ein reguläres statistisches Problem und* $\aleph = \mathcal{P}$. *Existiert eine messbare Abbildung* $\widehat{\theta} : E \to \aleph$ *mit*

$$p_{\widehat{\theta}(x)}(x) = \sup_{\theta \in \mathcal{P}} p_\theta(x),$$

so heißt $\widehat{\theta}$ Maximum-Likelihood-Schätzer *von* $g(\theta) = \theta$. *Weiter heißt für festes* x *die Funktion* $\theta \mapsto p_\theta(x)$ Likelihood *und* $\theta \mapsto \log p_\theta(x)$ Log-Likelihood.

Bemerkung 5.4 (Einfache Eigenschaften). 1. Sei $X = (X_1, ..., X_n)$ ein Vektor identisch verteilter Zufallsvariablen. Dann ist jeder Plug-In-Schätzer $\widehat{g}$ für g (mit $g(\theta) := \mathbb{E}_\theta[f(X_1)]$) unverzerrt.
Denn: Es gilt

$$\mathbb{E}_\theta[\widehat{g}(X)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta[f(X_i)] = \mathbb{E}_\theta[f(X_1)] = g(\theta).$$

2. Sei $d : E \to \aleph$ ein Schätzer für $g : \mathcal{P} \to \aleph$ und $f : \aleph \to \aleph$. Ist d ein Maximum-Likelihood-Schätzer für g, so ist $d \circ f$ ein Maximum-Likelihood-Schätzer für $g \circ f$. Ist d unverzerrt, so ist jedoch der Schätzer $d \circ f$ für $g \circ f$ nicht unbedingt unverzerrt.
Denn: Der Wert, an dem eine Funktion ihr Maxmimum annimmt, verändert sich nicht, wenn man die Funktion mit einer weiteren Funktion verknüpft, was die erste Behauptung zeigt. Für die zweite Behauptung sei etwa d ein Plug-In-Schätzer für $g(\theta) = \mathbb{E}_\theta[X_1]$, sowie $f : x \mapsto x^2$. Wäre $d \circ f$ unverzerrt, so müsste

$$\mathbb{E}_\theta[d(X)^2] = \mathbb{E}_\theta[f(d(X))] = f(g(\theta)) = (g(\theta))^2 = (\mathbb{E}_\theta[d(X)])^2,$$

also $\mathbb{V}_\theta[d(X)] = 0$, was im Allgemeinen sicher falsch ist.

Proposition 5.5 (Maximum-Likelihood-Schätzer in Exponentialfamilien). *Sei* $(X, \{\mathbb{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ *eine* k*-parametrige Exponentialfamilie mit* c, t, d, h *für* $c_1, ..., c_k$ *injektiv, also*

$$p_\theta(x) = h(x) \cdot \exp\big(c(\theta)^\top t(x) - d(\theta)\big).$$

Sei $C = \{c(\theta) : \theta \in \mathcal{P}\}^\circ$. *Falls die Gleichung (in* θ*)*

$$\mathbb{E}_\theta[t_i(X)] = t_i(x), \qquad i = 1, ..., k$$

(in θ*) eine Lösung* $x \mapsto \widehat{\theta}(x)$ *besitzt mit* $c(\widehat{\theta}(x)) \in C$, *dann ist* $\widehat{\theta}$ *der eindeutige Maximum-Likelihood-Schätzer für* θ.

Beweis. Wir zeigen die Behauptung nur für $k = 1$. Sei oBdA $c = \mathrm{id}$. Den allgemeinen Fall zeigt man mit Bemerkung 5.4.2. Wir berechnen

$$\frac{\partial}{\partial \theta_i} \log p_\theta(x) = t(x) - d'(\theta), \qquad \frac{\partial^2}{\partial \theta_i^2} \log p_\theta(x) = -d''(\theta).$$

Daraus folgt mit Proposition 2.35, dass $\mathbb{E}_\theta[t(X)] = d'(\theta)$ und $\mathbb{V}_\theta[t(X)] = -d''(\theta) y 0$. Damit ist gezeigt, dass die Log-Likelihood-Funktion wegen der negativen Ableitung strikt konkav ist und $\widehat{\theta}$ genau dann ein Maximum-Likelihood-Schätzer ist, wenn $t(x) = \mathbb{E}_{\widehat{\theta}(x)}[t(X)]$ gilt. (Die Eindeutigkeit folgt mit der Konkavität.) □

Beispiel 5.6 (Beispiel *Norm*). Für das Beispiel aus Beispiel 2.23 erinnern wir an Bemerkung 2.28 und sehen, dass es sich um eine 2-parametrige Exponentialfamilie handelt mit $t_1(x) = \sum_{i=1}^n x_i, t_2(x) = \sum_{i=1}^n x_i^2$. Wir betrachten also etwa die Gleichung

$$\mathbb{E}_{(\mu(x),\sigma^2(x))}[t_1(X)] = t_1(x),$$

die genau dann erfüllt ist, wenn

$$n\mu(x) = t_1(x), \qquad \text{also für } \mu(x) = \overline{x}.$$

Damit ist bereits $\overline{X}$ der maximum-Likelihood-Schätzer für μ. Weiter betrachten wir die Gleichung

$$\mathbb{E}_{(\mu(x),\sigma^2(x))}[t_2(X)] = t_2(x)$$

die genau dann erfüllt ist, wenn

$$n(\sigma^2(x) + \mu^2(x)) = t_2(x), \qquad \text{also für } \sigma^2(x) = \frac{1}{n}t_2(x) - \overline{x}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2.$$

Damit haben wir den maximum-Likelihood-Schätzer für σ^2 hergeleitet.

Beispiel 5.7 (Lineare Regression). Wir kehren zurück zur linearen Regression aus Beispiel 2.30. Wieder leiten wir Maximum-Likelihood-Schätzer für die Parameter $\beta_0, ..., \beta_m, \sigma^2$ her. Für $t(y)^\top = (t_0(y), ..., t_m(y))$ schreiben wir

$$\mathbb{E}_{(\beta(y),\sigma^2(y))}[t(Y)]^\top = xx^\top\beta(y) = t(y)^\top = xy,$$
$$\mathbb{E}_{(\beta(y),\sigma^2(y))}[t_{m+1}(Y)] = (x^\top\beta)^2 + \sigma^2(y) = t_{m+1}(y) = \sum_{i=1}^n y_i^2 = yy^\top.$$

Aus der ersten Gleichung lesen wir ab, dass

$$\hat{\beta}(y) = (xx^\top)^{-1}xy$$

und aus der zweiten, dass

$$\widehat{\sigma^2}(y) = yy^\top - (x^\top(xx^\top)^{-1}xy)^2$$

die Maximum-Likelihood-Schätzer für die Parameter β und σ^2 sind.

5.2 UMVUE-Schätzer

Um es vorwegzunehmen: UMVUE steht für *Uniformly Minimum Variance Unbiased Estimator*. Solche Schätzer minimieren also unter allen unverzerrten Schätzern die Varianz.

Definition 5.8 (UMVUE-Schätzer). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ ein Schätzproblem. Ein Schätzer d für g heißt* UMVUE *oder* Uniformly Minimum Variance Unbiased Estimator *für g, falls er unverzerrt ist und*

$$\mathbb{V}_\theta[d(X)] = \inf_{e \text{ unverzerrt}} \mathbb{V}_\theta[e(X)], \qquad \theta \in \mathcal{P}.$$

Bemerkung 5.9 (Rao-Blackwell). Wir erinnern kurz an den Satz von Rao-Blackwell, Theorem 3.10, für ein Schätzproblem $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ im Falle des Gauß-Verlustes. Sei $T = t(X)$ suffizient und d ein nicht-randomisierter Schätzer für g mit $\mathbb{E}_\theta[|d(X)|] < \infty$ für alle $\theta \in \mathcal{P}$. Dann hat der Schätzer $e \circ t$ für g mit

$$e(t) := \mathbb{E}_\theta[d(X)|T = t]$$

geringeres Risiko als d.

Theorem 5.10 (Lehmann-Scheffé). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ ein Schätzproblem mit Gauß-Verlust ℓ, $T = t(X)$ eine vollständige, suffiziente Statistik und d ein unverzerrter Schätzer für g. Dann ist $e \circ t$ mit*

$$e(t) := \mathbb{E}_\theta[d(X)|T = t]$$

ein UMVUE für g. Ist außerdem $\mathbb{V}_\theta[e(t((X)))] < \infty$ für alle $\theta \in \mathcal{P}$, so ist $e \circ t$ der einzige UMVUE für g.

Beweis. Da d unverzerrt ist, folgt mit der Turmeigenschaft der bedingten Erwartung, dass auch $e \circ t$ unverzerrt ist. Wir zeigen nun, dass e und damit $e \circ t$ unabhängig von der Wahl von d ist. Dann hat insbesondere $e \circ t$ minimale Varianz unter allen unverzerrten Schätzern, ist also UMVUE. Seien also d_1, d_2 unverzerrte Schätzer für g. Weiter seien $e_i \circ t$ für $e_i(t) = \mathbb{E}_\theta[d_i(X)|T = t]$ zwei unverzerrte Schätzer. Insbesondere gilt

$$\mathbb{E}_\theta[e_1(T) - e_2(T)] = \mathbb{E}_\theta[d_1(X) - d_2(X)] = g(\theta) - g(\theta) = 0$$

für alle $\theta \in \mathcal{P}$. Da T vollständig ist folgt $e_1(T) \overset{\mathbb{P}_\theta-\text{fs}}{=} e_2(T)$ für alle $\theta \in \mathcal{P}$, was zu zeigen war. Für die Eindeutigkeit sei d' ein weiterer UMVUE. Einerseits ist dann nach dem Satz von Rao-Blackwell

$$\mathbb{E}_\theta[(\mathbb{E}_\theta[d'(X)|T] - g(\theta))^2] \leq \mathbb{E}_\theta[(d'(X) - g(\theta))^2],$$

und da d' ein UMVUE ist, gilt hier Gleichheit, also $d'(X) = \mathbb{E}_\theta[d'(X)|T]$. Andererseits haben wir oben gezeigt, dass $e(T) = \mathbb{E}_\theta[d'(X)|T]$, da d' unverzerrt ist. Insgesamt gilt also $d' = e \circ t$. □

Beispiel 5.11 (Beispiel *Unif*). Wir betrachten das Schätzproblem mit $g(\theta) = \theta$ und Gauß-Verlust ℓ. Hierzu verwendet wir die Statistik $T = t(X) = \max_{1 \leq i \leq n} X_i$. Wir haben bereits in Beispiel 2.8 und 2.17 gesehen, dass T vollständig und suffizient ist. Aus Beispiel 3.11 wissen wir außerdem, dass

$$d(X) := \frac{n+1}{n} \max_{1 \leq i \leq n} X_i$$

unverzerrt ist und dieser Schätzer eine kleinere Risikofunktion hat als $2\overline{x}$. Außerdem gilt $\mathbb{E}_\theta[d(X)|T] = d(X)$, da $d(X)$ eine Funktion von T ist. Nach dem Satz von Lehmann und Scheffé ist also d ein UMVUE. Wegen $\mathbb{V}_\theta[d(X)] < \infty$ für alle $\theta \in \mathcal{P}$ ist dieser UMVUE auch eineutig.

Die Argumentation des letzten Beispiels lässt sich verallgemeinern, was wir nun für Exponentialfamilien tun wollen.

Korollar 5.12 (UMVUE bei Exponentialfamilien). *Sei $(X, \{\mathbb{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ eine k-parametrige Exponentialfamilie mit c, t, d, h, also*

$$p_\theta(x) = h(x) \cdot \exp\big(c(\theta)^\top t(x) - d(\theta)\big).$$

Weiter sei $g(\theta) = \mathbb{E}_\theta[f(t(X))]$. Dann ist $d(X) := f(t(X))$ ein UMVUE für g. Ist außerdem $\mathbb{V}_\theta[d(X)] < \infty$ für alle $\theta \in \mathcal{P}$, so gibt es nur diesen einen UMVUE.

Beweis. Zunächst wissen wir aus Proposition 2.29 und Theorem 2.32, dass $T = t(X) = (t_1(X), ..., t_k(X))$ suffizient und vollständig ist. Weiter ist d unverzerrt und $\mathbb{E}_\theta[d(X)|T] = d(X)$, da $d(X)$ messbar bezüglich T ist. Nach dem Satz von Lehmann und Scheffé ist damit d ein UMVUE für g. Die letzte Behauptung wurde schon in Theorem 5.10 bewiesen. □

Beispiel 5.13 (Beispiel *Norm*). Für das Beispiel *Norm* aus Beispiel 2.23 haben wir in Beispiel 5.6 gesehen, dass es sich um eine 2-parametrige Exponentialfamilie handelt mit $t_1(x) = \sum_{i=1}^n x_i, t_2(x) = \sum_{i=1}^n x_i^2$. Wir betrachten nun $g_1(\mu, \sigma^2) = \mu$ und $g_2(\mu, \sigma^2) = \sigma^2$ und schreiben

$$\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2 = \frac{1}{n-1}\Big(\Big(\sum_{i=1}^n x_i^2\Big) - 2n\overline{x}^2 + n\overline{x}^2\Big) = \frac{1}{n-1}t_2(x) - \frac{1}{(n-1)n}t_1(x)^2,$$

also

$$\mathbb{E}_\theta\Big[\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2\Big] = \frac{n}{n-1}\sigma^2 - \frac{1}{n-1}\sigma^2 = \sigma^2$$

und damit sind

$$\begin{aligned} g_1(\theta) &= \mathbb{E}_\theta[\tfrac{1}{n}t_1(X)] = \mathbb{E}_\theta[\overline{X}] = \mu, \\ g_2(\theta) &= \mathbb{E}_\theta\Big[\frac{1}{n-1}t_2(X) - \frac{1}{(n-1)n}t_1^2(X)\Big] = \sigma^2 \end{aligned}$$

eindeutige UMVUEs.

Beispiel 5.14 (Lineare Regression). Wir betrachten noch einmal die lineare Regression aus Beispiel 2.30. Für $t(y)^\top = (t_0(y), ..., t_m(y))$ schreiben wir für $f(t(y)) = (xx^\top)^{-1}t(y)$

$$\mathbb{E}_{(\beta,\sigma^2)}[f(t(Y))] = \mathbb{E}_{(\beta,\sigma^2)}[(xx^\top)^{-1}t(Y)]\mathbb{E}_{(\beta,\sigma^2)}[(xx^\top)^{-1}xY] = (xx^\top)^{-1}xx^\top\beta = \beta.$$

Damit ist $(xx^\top)^{-1}xy$ ein UMVUE für β (und ein Maximum-Likelihood-Schätzer nach Beispiel 5.7).

5.3 Information und die Cramér-Rao-Schranke

Zwar haben wir bereits obere Schranken für das Risiko eines Schätzers ermittelt, jedoch gibt es auch untere Schranken, insbesondere die in Theorem 5.22 vorgestellte Cramér-Rao-Schranke. Für diese benötigen wir den Begriff der (Fisher-)Information. Diese beschreibt die Krümmung (im Sinne der zweiten Ableitung) der log-Likelihood in Bezug auf die Parameter. Ist diese Zahl groß, bedeutet das eine hohe Krümmung, und damit ist die log-Likelihood für nahe Parameter deutlich kleiner. Anders ausgedrückt haben wir über die log-Likelihood viel über den wahren Parameter gelernt, es gibt also viel "Information". Um die zweite Ableitung sinnvoll berechnen zu können, benötigen wir zunächst ein paar Regularitätsannahmen.

Annahme 5.15 (Regularitätsannahmen der Fisher-Information). *Für ein reguläres statistisches Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ auf $\mathbb{R}^n$ mit $\mathcal{P} \subseteq \mathbb{R}^m$ und $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$ geben wir folgende Regularitätsannahmen an:*

(A1) Es gibt ein $N \in \mathcal{B}(\mathbb{R}^n)$, so dass für alle $\theta \in \mathcal{P}$ und $i = 1, ..., k$ die Ableitung $\partial p_\theta(x)/\partial\theta_i$ für $x \notin N$ existiert.

(A2) Für jedes messbare ϕ gilt, falls der Erwartungswert existiert,

$$\frac{\partial}{\partial\theta_i}\mathbb{E}_\theta[\phi(X)] = \int \phi(x)\frac{\partial p_\theta(x)}{\partial\theta_i}d\lambda^n(x).$$

(A3) Die Menge

$$C := \{x : p_\theta(x) > 0\}$$

ist unabhängig von θ.

(A4) Es gilt (A2) und außerdem auch

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathbb{E}_\theta[\phi(X)] = \int \phi(x)\frac{\partial^2 p_\theta(x)}{\partial\theta_i\partial\theta_j}d\lambda^n(x)$$

für jedes messbare ϕ, für das der Erwartungswert existiert.

Beispiel 5.16 (Beispiele *Bern*, *Norm*, *Unif*). Für Exponentialfamilien ist Annahme 5.15 nach Lemma 2.34 immer erfüllt, insbesondere also für Beispiele *Bern* und *Norm*. Für Beispiel 1.8 ist jedoch sowohl (A1) als auch (A3) verletzt.

Definition 5.17 (Fisher-Information). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell, das die Annahmen (A1)–(A3) erfüllt. Dann heißt*

$$I(\theta) := I(\mathbb{P}_\theta) := \Big(\mathbb{COV}\Big[\frac{\partial}{\partial\theta_i}\log p_\theta(X), \frac{\partial}{\partial\theta_j}\log p_\theta(X)\Big]\Big)_{i,j=1,...,k}$$

Fisher-Informations-Matrix.

Bemerkung 5.18 (Berechnung der Fisher-Information unter (A4)). Gilt statt (A2) sogar (A4), so kann man schreiben

$$\begin{aligned}\mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\Big] &= \mathbb{E}_\theta\Big[\frac{1}{p_\theta(X)^2}\Big(p_\theta(X)\frac{\partial^2 p_\theta(X)}{\partial\theta_i\partial\theta_j} - \frac{\partial p_\theta(X)}{\partial\theta_i}\cdot\frac{\partial p_\theta(X)}{\partial\theta_j}\Big)\Big] \\ &= 0 - \mathbb{E}_\theta\Big[\frac{\partial\log p_\theta(X)}{\partial\theta_i}\frac{\partial\log p_\theta(X)}{\partial\theta_i}\Big] = -(I(\theta))_{i,j}.\end{aligned}$$

Dies liefert also eine alternative Berechnung der Fisher-Information.

Beispiel 5.19 (Exponentialfamilie). Im Falle einer Exponentialfamilie in kanonischer Form ist

$$p_\theta(x) = h(x)\exp(\theta^\top t(x) - d(\theta))$$

und damit mit Proposition 2.35

$$\frac{\partial}{\partial\theta_i}\log p_\theta(x) = t_i(x) - \frac{\partial d(\theta)}{\partial\theta_i} = t_i(X) - \mathbb{E}_\theta[t_i(X)].$$

Weiter ist

$$(I(\theta))_{i,j} = \mathbb{COV}_\theta[t_i(X), t_j(X)].$$

Beispiel 5.20 (Fisher-Information einer unabhängigen Stichprobe). Für ein statistisches Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ haben wir die Fisher-Information definiert. Nun betrachten wir den Fall von Daten $X_1, ..., X_n$, bei denen die X_i unter allen $\mathbb{P}_\theta$ unabhängig sind. Anders gesagt betrachten wir das statistische Modell $(X = (X_1, ..., X_n), \{\mathbb{P}_\theta^n : \theta \in \mathcal{P}\})$, wobei $\mathbb{P}_\theta^n$ das n-fache Produktmaß ist. Hier ist die Dichte gegeben durch $x = (x_1, ..., x_n) \mapsto p_\theta(x_1) \cdots p_\theta(x_n)$ und damit gilt

$$\begin{aligned} I(\mathbb{P}_\theta^n) &= \mathbb{COV}_\theta^n\Big[\frac{\partial}{\partial\theta_j}\log p_\theta(X), \frac{\partial}{\partial\theta_k}\log p_\theta(X)\Big]_{jk} \\ &= \mathbb{COV}_\theta^n\Big[\sum_{i=1}^n \frac{\partial}{\partial\theta_j}\log p_\theta(X_i), \sum_{i=1}^n \frac{\partial}{\partial\theta_k}\log p_\theta(X_i)\Big]_{jk} \\ &= \sum_{i=1}^n \mathbb{COV}_\theta\Big[\frac{\partial}{\partial\theta_j}\log p_\theta(X_i), \frac{\partial}{\partial\theta_k}\log p_\theta(X_i)\Big]_{jk} \\ &= nI(\mathbb{P}_\theta). \end{aligned}$$

Proposition 5.21 (Mittlere Ableitung der log-Likelihood verschwindet). *Es gelte (A1)–(A3). Dann ist für $i = 1, ..., m$*

$$\mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta_i}\log p_\theta(X)\Big] = 0.$$

Beweis. Wir schreiben

$$\begin{aligned} \mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta_i}\log p_\theta(X)\Big] &= \int p_\theta(x)\frac{\partial}{\partial\theta_i}\log p_\theta(x)d\lambda^n(dx) = \int p_\theta(x)\frac{1}{p_\theta(x)}\frac{\partial}{\partial\theta_i}p_\theta(x)d\lambda^n(dx) \\ &= \frac{\partial}{\partial\theta_i}\mathbb{E}_\theta[1] = 0, \end{aligned}$$

wobei die Vertauschung von Integral und Ableitung nach (A3) erlaubt ist. □

Theorem 5.22 (Cramér-Rao-Schranke). *Es gelte (A1)–(A3) sowie $\mathcal{P} \subseteq \mathbb{R}$ und $I(\theta) > 0$ für alle $\theta \in \mathcal{P}$. Sei $T = t(X)$ für $t : \mathbb{R}^m \to \mathbb{R}$ eine Statistik und es existiere $\Psi(\theta) := \mathbb{E}_\theta[T]$. Dann gilt*

$$\mathbb{V}_\theta[T] \geq \frac{\Psi'(\theta)^2}{I(\theta)}.$$

Ist insbesondere $\mathbb{E}_\theta[T] = \theta$ (d.h. T ist ein erwartungstreuer Schätzer für θ), so gilt

$$\mathbb{V}_\theta[T] \geq \frac{1}{I(\theta)}.$$

Beweis. Wir schreiben mit Proposition 5.21 und der Cauchy-Schwartz-Ungleichung

$$\begin{aligned} \Psi'(\theta) &= \int t(x)\frac{\partial}{\partial\theta}p_\theta(x)\lambda^n(dx) = \mathbb{E}_\theta\Big[t(x)\frac{\partial}{\partial\theta}\log p_\theta(x)\Big] \\ &= \mathbb{E}_\theta\Big[\big(t(X) - \mathbb{E}_\theta[T]\big)\frac{\partial}{\partial\theta}\log p_\theta(X)\Big] \\ &\leq \big(\mathbb{V}_\theta[T]\cdot I(\theta)\big)^{1/2}. \end{aligned}$$

Daraus folgt die Behauptung. □

Beispiel 5.23 (1-parametrige Exponentialfamilie). Wir betrachten den Fall aus Beispiel 5.19 mit $\theta \in \mathbb{R}$ und der suffizienten Statistik $T = t(X)$. Wir wissen (etwa aus Beispiel 5.19 und Proposition 2.35), dass $\mathbb{E}_\theta[t(X)] = d'(\theta)$ und $I(\theta) = \mathbb{V}_\theta[t(X)] = d''(\theta)$. Damit gilt

$$\mathbb{V}_\theta[t(X)] = d''(\theta) = \frac{d''(\theta)^2}{I(\theta)} = \frac{(\frac{d}{d\theta}\mathbb{E}_\theta[t(X)])^2}{I(\theta)},$$

und damit gilt Gleichheit in der Schranke von Theorem 5.22.

Andresherum ist im Beweis der Cramér-Rao-Schranke klar: Gleichheit gilt genau dann, wenn Gleichheit in der Cauchy-Schwartz-Ungleichung gilt, also wenn $t(X) - \mathbb{E}_\theta[T]$ und $\frac{\partial}{\partial\theta}\log p_\theta(X)$ linear abhängig sind, also wenn für geeignete a und b

$$\frac{\partial}{\partial\theta}\log p_\theta(x) = a(\theta)t(x) + b(\theta)$$

oder

$$\log p_\theta(x) = t(x)\int_{\theta_0}^{\theta} a(\eta)d\eta + \int_{\theta_0}^{\theta} b(\eta)d\eta.$$

Das bedeutet, dass p_θ die Dichte einer Exponentialfamilie ist. Die Cramér-Rao-Schranke ist damit (falls (A1)–(A3) gelten) genau für 1-parametrige Exponentialfamilien scharf.

Beispiel 5.24 (Beispiel *Unif*). Aus Beispiel 2.25 wissen wir, dass es sich nich um eine Exponentialfamilie handelt. Außerdem sind die Regularitätsannahmen (A1) und (A3) nicht erfüllt; siehe Beispiel 5.16. Deshalb ist nicht klar, ob die Cramér-Rao-Schranke in diesem Modell existiert. Um zu sehen, ob die Schranke doch funktioniert, verwenden wir Bemerkung 5.18, so dass

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2}\mathbb{E}_\theta[\log p_\theta(X)] = 1/\theta^2. \tag{5.1}$$

Wir setzen $t(X) = X$, so dass $\mathbb{V}_\theta[t(X)] = \theta^2/12$. Außerdem ist $\frac{d}{d\theta}\mathbb{E}_\theta[X] = 1/2$ und damit

$$\mathbb{V}_\theta[t(X)] = \frac{\theta^2}{12} < \frac{\theta^2}{4} = \frac{(\frac{d}{d\theta}\mathbb{E}_\theta[X])^2}{I(\theta)}.$$

Also gilt die Cramér-Rao-Schranke hier nicht.

Bemerkung 5.25 (Unabhängige Stichprobe). Für das statistische Modell $(X^n = (X_1, ..., X_n), \{\mathbb{P}_\theta^n : \theta \in \mathcal{P}\})$ der n-fachen unabhängigen Versuchswiederholung, $n = 1, 2, ...$ aus Beispiel 5.20 erhalten wir folgendes Resultat (falls (A1)–(A3) gelten):
Ist $\mathcal{P} \subseteq \mathbb{R}$ und ist d^n ein Schätzer für θ im n-ten Modell, so gilt

$$\mathbb{V}_\theta[d^n(X^n)] \geq \frac{(\frac{d}{d\theta}\mathbb{E}_\theta[d^n(X^n)])^2}{nI(\theta)}.$$

Asymptotisch bedeutet das, falls $\sqrt{n}(\mathbb{E}_\theta[d^n(X^n) - g(\theta)] \xrightarrow{n\to\infty} 0$ und $\frac{d}{d\theta}(\mathbb{E}_\theta[d^n(X^n)] - g(\theta)) \xrightarrow{n\to\infty} 0$

$$\begin{aligned}\liminf_{n\to\infty} n\mathbb{E}_\theta[(d^n(X^n) - g(\theta))^2] &= \liminf_{n\to\infty} n\mathbb{V}_\theta[(d^n(X^n)] + n(\mathbb{E}_\theta[d^n(X^n) - g(\theta)])^2 \\ &= \liminf_{n\to\infty} n\mathbb{V}_\theta[(d^n(X^n)] \geq \liminf_{n\to\infty} \frac{(\frac{d}{d\theta}\mathbb{E}_\theta[d^n(X^n)])^2}{I(\theta)} = \frac{(g'(\theta))^2}{I(\theta)}.\end{aligned}$$

5.4 Asymptotik von Maximum-Likelihood-Schätzern

Das letzte Resultat behandelt schon die Asymptotik einer Folge von Schätzern. Dies wollen wir nun noch für Maximum-Likelihood-Schätzer ausbauen. Einfach gesagt konvergieren Maximum-Likelihood-Schätzer (unter einigen Regularitätsannahmen) für unabhängige Stichproben gegen den wahren Parameter (Theorem 5.30). Noch genauer sind sie asymptotisch normalverteilt und die Inverse der Fisher-Information gibt die Covarianzmatrix (Theorem 5.31). Es folgen nun weitere Annahmen, insbesondere die der unabhängigen, identisch verteilten Daten.

Bemerkung 5.26 (Weitere Annahmen). (A5) Es ist $(X^n = (X_1^n, ..., X_n^n), \{\mathbb{P}_\theta^n : \theta \in \mathcal{P}\})$ ein reguläres statistisches Modell, so dass X^n ein unabhängiger, identisch verteilter Vektor ist mit

$$\mathbb{P}^n(X_1^n \in .) = p_\theta \cdot \lambda^m.$$

Insbesondere hat $\mathbb{P}^n$ die Dichte

$$\prod_{i=1}^n p_\theta(x_i).$$

(A6) Für alle x ist

$$\eta \mapsto \frac{\partial^2}{\partial\eta_i \partial\eta_j} \log p_\eta(x)$$

stetig und endlich und es gilt für alle $\theta \in \mathcal{P}$

$$\eta \mapsto \frac{\partial^2}{\partial\eta_i \partial\eta_j} \mathbb{E}_\theta\Big[\log p_\eta(X)\Big] = \mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\eta_i \partial\eta_j} \log p_\eta(X)\Big].$$

Bemerkung 5.27 (Maximum-Likelihood-Schätzer). Gilt (A1)–(A5), so ist der Maximum-Likelihood-Schätzer $d^n(x^n)$ für g, falls er im Inneren von $\mathcal{P}$ liegt, Lösung der Gleichung

$$\sum_{i=1}^n \frac{\partial}{\partial\theta} \log p_\theta(x_i^n)\Big|_{\theta = d^n(x^n)} = 0. \tag{Lik}$$

Definition 5.28 (Konsistenz). *Sei $((X^n, \{\mathbb{P}_\theta^n : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ ein Schätzproblem, und d^n ein Schätzer für g, $n = 1, 2, ...$ Dann heißt die Folge d^n konsistent, falls*

$$\mathbb{P}_\theta(|d^n(X^n) - g(\theta)| \geq \varepsilon) \xrightarrow{n\to\infty} 0$$

für alle $\varepsilon > 0$ und alle $\theta \in \mathcal{P}$. Dies ist insbesondere dann der Fall, wenn $X^1, X^2, ...$ auf demselben Wahrscheinlichkeitsraum definiert sind und $d^n(X^n) \xrightarrow{n\to\infty}_{fs} g(\theta)$.

Bemerkung 5.29 (Konsistenz bei unabhängigen Daten). Unter Annahme (A5) ist unter $\mathbb{P}_\theta^n$ der Vektor X^n unabhängig und identisch nach $\mathbb{P}_\theta$ verteilt. Wir können dann oBdA annehmen, dass alle X^n auf demselben Wahrscheinlichkeitsraum definiert sind. Wir schreiben dann auch X anstelle von X^n (und denken uns, dass d^n ja nur von den ersten n Einträgen von X abhängt). Insbesondere kann dann die fast sichere Konvergenz $d^n(X) \xrightarrow{n\to\infty} g(\theta)$ unter $\mathbb{P}_\theta$ gelten.

Theorem 5.30 (Konsistenz von Maximum-Likelihood-Schätzern). *Sei $\theta \in \mathcal{P}$ und $d^n(X)$ der Maximum-Likelihood-Schätzer für θ (basierend auf den ersten n Beobachtungen). Sei*

$$Z(M,x) := \inf_{\eta\in M} \log\frac{p_\theta(x)}{p_\eta(x)}.$$

Angenommen, für jedes $\eta \neq \theta$ gibt es ein $\varepsilon(\eta) > 0$, so dass $\mathbb{E}_\theta[Z(B_{\varepsilon(\eta)}, X)] > 0$. Weiter existiere ein $\mathcal{C}$ kompakt mit $\theta \in \mathcal{C}$ und $\mathbb{E}_\theta[Z(\mathcal{P}\setminus\mathcal{C}, X)] > 0$. Dann ist

$$\lim_{n\to\infty} d^n(X) \overset{\mathbb{P}_\theta-fs}{=} \theta.$$

Beweis. Für beliebiges $\varepsilon > 0$ müssen wir zeigen, dass

$$\mathbb{P}_\theta(\limsup_{n\to\infty} |d^n(X) - \theta| > \varepsilon) = 0.$$

(Die Aussage folgt dann mit einem Grenzwert $\varepsilon \downarrow 0$.) Da $\mathcal{C}\setminus B_\varepsilon(\theta)$ kompakt ist und $\{B_{\varepsilon(\eta)}(\eta) : \eta \in \mathcal{C}\setminus B_\varepsilon(\theta)\}$ eine offene Überdeckung der kompakten Menge $\mathcal{C}\setminus B_\varepsilon(\theta)$ ist, gibt es eine endliche Teilüberdeckung $A_1 := B_{\varepsilon(\eta_1)}(\eta_1), ..., A_m := B_{\varepsilon(\eta_m)}(\eta_m)$. Mit $A_0 := \mathcal{P}\setminus\mathcal{C}$ ist $\mathcal{P} = B_\varepsilon(\theta)\cup A_0\cup\cdots\cup A_m$ und $\mathbb{E}_\theta[Z(A_j, X_i)] =: c_j > 0$, $j = 0, ..., m$. Nach dem starken Gesetz der großen Zahlen ist

$$\frac{1}{n}\sum_{i=1}^n Z(A_j, X_i) \xrightarrow{n\to\infty,\mathbb{P}_\theta-\text{fs}} c_j.$$

Dann gilt

$$\begin{aligned}
&\mathbb{P}_\theta(\limsup_{n\to\infty}|d^n(X)-\theta| \geq \varepsilon)\\
&\qquad\leq \mathbb{P}_\theta\Big(\text{ Es gibt } \eta_1,\eta_2,... \notin B_\varepsilon(\theta), \text{ so dass}\\
&\qquad\qquad\qquad \frac{1}{n}\sum_{i=1}^n \log p_{\eta_i}(X_i) \geq \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) \text{ für unendlich viele } n\Big)\\
&\qquad\leq \sum_{j=0}^m \mathbb{P}\Big(\inf_{\eta\in A_j}\frac{1}{n}\sum_{i=1}^n \log\frac{p_\theta(X_i)}{p_\eta(X_i)} \leq 0 \text{ unendlich oft}\Big)\\
&\qquad\leq \sum_{j=0}^m \mathbb{P}_\theta\Big(\frac{1}{n}\sum_{i=1}^n Z(A_j,X_i) \leq 0 \text{ unendlich oft}\Big) = 0
\end{aligned}$$

und die Behauptung ist gezeigt. □

Theorem 5.31 (Asymptotische Normalität des Maximum-Likelihood-Schätzers). *Es gelte (A1)–(A6) und es sei für jedes $\theta \in \mathcal{P}^\circ$*[10]

$$\sup_{j,k}\mathbb{E}_\theta\Big[\sup_{|\eta-\theta|<r}\Big|\frac{\partial^2}{\partial\eta_j\partial\eta_k}\log p_\eta(X_1)\Big|_{\eta=\theta} - \frac{\partial^2}{\partial\eta_j\partial\eta_k}\log p_\eta(X_1)\Big|\Big] \xrightarrow{r\to 0} 0. \qquad (*)$$

Weiter sei $d^n(X)$ konsistent und die Fisher-Information $I(\theta)$ sei für alle $\theta \in \mathcal{P}$ nicht-singulär. Dann gilt für $X \sim \mathbb{P}_\theta^\infty$

$$\sqrt{n}(d^n(X) - \theta) \xRightarrow{n\to\infty} \mathcal{N}(0, I^{-1}(\theta)).$$

[10] Für eine Menge A sei A° das Innere.

Beweis. Sei

$$\ell_x^n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i).$$

Dann ist, wegen der stetigen Differenzierbarkeit,

$$\nabla \ell_x^n(\theta) := \Big(\frac{\partial}{\partial \theta_j}\ell_x^n(\theta)\Big)_j = \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_j}\log p_\theta(x_i)\Big)_j.$$

Es gilt

$$\mathbb{E}_\theta[\nabla \ell_X^n(\theta)] = 0$$

wegen Proposition 5.21 sowie

$$n\mathbb{COV}_\theta[\nabla\ell_X^n(\theta), \nabla\ell_X^n(\theta)] = I(\theta) = \Big(-\mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p_\theta(X)\Big]\Big)_{jk}.$$

Damit folgt aus dem mehrdimensionalen Zentralen Grenzwertsatz bereits, dass für $X \sim \mathbb{P}_\theta^\infty$

$$\sqrt{n}\nabla\ell_X^n(\theta) \xRightarrow{n\to\infty} \mathcal{N}(0, I(\theta)). \tag{$**$}$$

Außerdem ist

$$\nabla\ell_x^n(d^n(x)) = 0,$$

da die Funktion $\theta \mapsto \ell_x^n(\theta)$ bei $d^n(x)$ ein Maximum besitzt und mit einer Taylor-Entwicklung von $\nabla\ell_x^n(\eta)$ für $\eta = d^n(x)$ um θ gilt

$$0 = \nabla\ell_x^n(d^n(x)) = \nabla\ell_x^n(\theta) + B_n(d^n(x) - \theta)$$

für

$$B_n = \Big(\frac{\partial^2}{\partial\eta_j\partial\eta_k}\ell_X^n(\eta)\Big|_{\eta=\eta_{nj}^*}\Big)_{jk}$$

für ein η_{nj}^* zwischen θ_j und $d^n(x)_j$ für $j = 1, ..., m$. Da $d^n(X) \xrightarrow{n\to\infty}_p \theta$, gilt auch $\eta_{nj}^* \xrightarrow{n\to\infty}_p \theta_j$. Zusammen mit $(**)$ haben wir damit gezeigt, dass

$$\begin{aligned}
&\sqrt{n}B_n(d^n(X) - \theta)\\
&\quad= \sqrt{n}\Big(\Big(\frac{\partial^2}{\partial\eta_j\partial\eta_k}\ell_X^n(\eta)\Big|_{\eta=\eta_{nj}^*} - \mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\eta_j\partial\eta_k}\log p_\eta(X)\Big|_{\eta=\eta_{nj}^*}\Big]\Big)_{jk}\\
&\qquad+ \mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\eta_j\partial\eta_k}\log p_\eta(X)\Big|_{\eta=\eta_{nj}^*} - \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p_\theta(X)\Big]\Big)_{jk}\\
&\qquad\qquad- I(\theta)\Big)(d^n(X) - \theta) \xRightarrow{n\to\infty} \mathcal{N}(0, I(\theta)).
\end{aligned}$$

Nun gilt für $X \sim \mathbb{P}_\theta^\infty$ wegen dem Gesetz der großen Zahlen und Annahme $(*)$

$$\begin{aligned}
\frac{\partial^2}{\partial\eta_j\partial\eta_k}\ell_X^n(\eta)\Big|_{\eta=\eta_{nj}^*} - \mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\eta_j\partial\eta_k}\log p_\eta(X)\Big|_{\eta=\eta_{nj}^*}\Big] &\xrightarrow{n\to\infty}_p 0,\\
\mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\eta_j\partial\eta_k}\log p_\eta(X)\Big|_{\eta=\eta_{nj}^*} - \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p_\theta(X)\Big] &\xrightarrow{n\to\infty}_p 0.
\end{aligned}$$

Damit folgt zunächst $B_n \xrightarrow{n\to\infty}_p -I(\theta)$, woraus $\sqrt{n}(d^n(X-\theta)) = O(1)$ folgt und damit (aus Slutskys Theorem)

$$-\sqrt{n}I(\theta)(d^n(X)-\theta) \xRightarrow{n\to\infty} \mathcal{N}(0, I(\theta)).$$

Da die Matrix-Multiplikation mit $I(\theta)^{-1}$ eine stetige Abbildung ist, folgt damit das Ergebnis. □