KAPITEL 7

Suffizienz und Vollständigkeit

7.1. Definition der Suffizienz im diskreten Fall

Beispiel 7.1.1. Betrachten wir eine unfaire Münze, wobei die Wahrscheinlichkeit θ, dass die Münze Kopf zeigt, geschätzt werden soll. Dafür werde die Münze n mal geworfen. Falls die Münze beim i-ten Wurf Kopf zeigt, definieren wir $x_i = 1$, sonst sei $x_i = 0$. Die komplette Information über unser Zufallsexperiment ist somit in der Stichprobe $(x_1, \ldots, x_n)$ enthalten. Es erscheint aber intuitiv klar, dass für die Beantwortung der statistischen Fragen über θ nur die Information darüber, *wie oft* die Münze Kopf gezeigt hat (also die Zahl $x_1 + \ldots + x_n$) relevant ist. Hingegen ist die Information, *bei welchen Würfen* die Münze Kopf gezeigt hat, nicht nützlich. Deshalb nennt man in diesem Beispiel die Stichprobenfunktion $T(x_1, \ldots, x_n) = x_1 + \ldots + x_n$ eine suffiziente (d.h. ausreichende) Statistik. Anstatt das Experiment durch die ganze Stichprobe $(x_1, \ldots, x_n)$ zu beschreiben, können wir es lediglich durch den Wert von $x_1 + \ldots + x_n$ beschreiben, ohne dass dabei nützliche statistische Information verloren geht.

Es sei im Weiteren $\{h_\theta : \theta \in \Theta\}$ eine Familie von Zähldichten. Den Fall der Dichten werden wir später betrachten. Ist S ein Zufallsvektor mit Werten in $\mathbb{R}^d$, so bezeichnen wir mit $\operatorname{Im} S = \{z \in \mathbb{R}^d : \mathbb{P}[S = z] \neq 0\}$ die Menge aller Werte z, die der Zufallsvektor S mit einer strikt positiven Wahrscheinlichkeit annehmen kann. Wirr nennen $\operatorname{Im} S$ den *Träger* von S. Im folgenden Abschnitt werden wir annehmen, dass der Träger von $X_1, \ldots, X_n$ nicht von θ abhängt.

Definition 7.1.2. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte diskrete Zufallsvariablen mit Zähldichte h_θ. Eine Funktion $T : \mathbb{R}^n \to \mathbb{R}^m$ heißt eine *suffiziente Statistik*, wenn für alle $x_1, \ldots, x_n \in \mathbb{R}$ und für alle $t \in \operatorname{Im} T(X_1, \ldots, X_n)$ der Ausdruck

$$\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n | T(X_1, \ldots, X_n) = t]$$

eine von θ unabhängige Funktion ist. D.h., für alle $\theta_1, \theta_2 \in \Theta$ soll gelten:

$$\mathbb{P}_{\theta_1}[X_1 = x_1, \ldots, X_n = x_n | T(X_1, \ldots, X_n) = t] = \mathbb{P}_{\theta_2}[X_1 = x_1, \ldots, X_n = x_n | T(X_1, \ldots, X_n) = t].$$

Beispiel 7.1.3 (Fortsetzung von Beispiel 7.1.1). Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $X_i \sim \operatorname{Bern}(\theta)$, wobei $\theta \in (0, 1)$ zu schätzen sei. Die Zähldichte von X_i ist somit gegeben durch

$$h_\theta(x) = \theta^x (1 - \theta)^{1-x}, \quad x \in \{0, 1\}.$$

Wir zeigen nun, dass $T(X_1, \ldots, X_n) = X_1 + \ldots + X_n$ eine suffiziente Statistik ist. Sei $t \in \{0, \ldots, n\}$. Betrachte den Ausdruck

$$\begin{aligned} P(\theta) &:= \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n | X_1 + \ldots + X_n = t] \\ &= \frac{\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n, X_1 + \ldots + X_n = t]}{\mathbb{P}_\theta[X_1 + \ldots + X_n = t]}. \end{aligned}$$

Fall 1. Ist $x_1 + \ldots + x_n \neq t$ oder $x_i \notin \{0, 1\}$ für mindestens ein i, dann gilt $P(\theta) = 0$. In diesem Fall hängt $P(\theta)$ von θ nicht ab.

Fall 2. Sei nun $x_1 + \ldots + x_n = t$ mit $x_1, \ldots, x_n \in \{0, 1\}$. Dann gilt

$$P(\theta) = \frac{\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n, X_1 + \ldots + X_n = t]}{\mathbb{P}_\theta[X_1 + \ldots + X_n = t]} = \frac{\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n]}{\mathbb{P}_\theta[X_1 + \ldots + X_n = t]}.$$

Indem wir nun benutzen, dass $X_1, \ldots, X_n$ unabhängig sind und $X_1 + \ldots + X_n \sim \text{Bin}(n, \theta)$ ist, erhalten wir, dass

$$P(\theta) = \frac{\theta^{x_1}(1-\theta)^{1-x_1} \cdot \ldots \cdot \theta^{x_n}(1-\theta)^{1-x_n}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}.$$

Dieser Ausdruck ist ebenfalls von θ unabhängig.
Somit ist $T(X_1, \ldots, X_n) = X_1 + \ldots + X_n$ eine suffiziente Statistik. Ein guter Schätzer für θ muss eine Funktion von $X_1 + \ldots + X_n$ sein. Das garantiert nämlich, dass der Schätzer nur nützliche statistische Information verwendet und nicht durch die Verwendung von unnützlichem Zufallsrauschen die Varianz des Schätzers gesteigert wird. In der Tat, wir haben bereits gezeigt, dass $\bar{X}_n$ die Cramér–Rao–Schranke erreicht und somit der beste erwartungstreue Schätzer ist.

Bemerkung 7.1.4. Im obigen Beispiel haben wir gezeigt, dass für jedes $t \in \{0, 1 \ldots, n\}$ die bedingte Verteilung von $(X_1, \ldots, X_n)$ gegeben, dass $X_1 + \ldots + X_n = t$, eine Gleichverteilung auf der Menge

$$\{(x_1, \ldots, x_n) \in \{0, 1\}^n : x_1 + \ldots + x_n = t\}$$

ist. Diese Menge besteht aus $\binom{n}{t}$ Elementen.

Aufgabe 7.1.5. Seien $X_1, \ldots, X_n$ unabhängige und mit Parameter $\theta > 0$ Poisson–verteilte Zufallsvariablen. Zeigen Sie, dass $T(X_1, \ldots, X_n) = X_1 + \ldots + X_n$ eine suffiziente Statistik ist. Bestimmen Sie für $t \in \mathbb{N}_0$ die bedingte Verteilung von $(X_1, \ldots, X_n)$ gegeben, dass $X_1 + \ldots + X_n = t$.

Im obigen Beispiel haben wir gezeigt, dass $X_1 + \ldots + X_n$ eine suffiziente Statistik ist. Ist dann z.B. auch $\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$ eine suffiziente Statistik? Im folgenden Lemma zeigen wir, dass die Antwort positiv ist.

Lemma 7.1.6. *Sei T eine suffiziente Statistik und sei $g : \operatorname{Im} T(X_1, \ldots, X_n) \to \mathbb{R}^k$ eine injektive Funktion. Dann ist auch*

$$g \circ T : \mathbb{R}^n \to \mathbb{R}^k, \quad (X_1, \ldots, X_n) \mapsto g(T(X_1, \ldots, X_n))$$

eine suffiziente Statistik.

Beweis. Sei $t \in \operatorname{Im} g(T(X_1, \dots, X_n))$. Da T eine suffiziente Statistik ist, hängt

$$\begin{aligned} P(\theta) &:= \mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n | g(T(X_1, \dots, X_n)) = t] \\ &= \mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n | T(X_1, \dots, X_n) = g^{-1}(t)] \end{aligned}$$

nicht von θ ab. Dabei ist $g^{-1}(t)$ wohldefiniert, da g injektiv ist. □

7.2. Faktorisierungssatz von Neyman–Fisher

In diesem Abschnitt beweisen wir den Faktorisierungssatz von Neyman–Fisher. Dieser Satz bietet eine einfache Methode zur Überprüfung der Suffizienz. Sei $\{h_\theta : \theta \in \Theta\}$ eine Familie von Zähldichten und $X_1, \dots, X_n$ unabhängige Zufallsvariablen mit Zähldichte h_θ. Sei $T : \mathbb{R}^n \to \mathbb{R}^m$ eine Funktion. Im nächsten Lemma benutzen wir folgende Notation:

(1) $L(x_1, \dots, x_n; \theta) = \mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n]$ ist die Likelihood-Funktion.
(2) $q(t; \theta) = \mathbb{P}_\theta[T(X_1, \dots, X_n) = t]$, wobei $t \in \mathbb{R}^m$, ist die Zähldichte von $T(X_1, \dots, X_n)$ unter $\mathbb{P}_\theta$.

Lemma 7.2.1. *Eine Funktion $T : \mathbb{R}^n \to \mathbb{R}^m$ ist genau dann eine suffiziente Statistik, wenn für alle $x_1, \dots, x_n \in \operatorname{Im} X_1$ die Funktion*

$$\frac{L(x_1, \dots, x_n; \theta)}{q(T(x_1, \dots, x_n); \theta)} \tag{7.2.1}$$

nicht von θ abhängt.

Beweis. Betrachte den Ausdruck

$$P(\theta) := \mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n | T(X_1, \dots, X_n) = t].$$

Im Falle $t \neq T(x_1, \dots, x_n)$ ist die bedingte Wahrscheinlichkeit gleich 0, was unabhängig von θ ist. Sei deshalb $t = T(x_1, \dots, x_n)$. Dann gilt

$$\begin{aligned} P(\theta) &= \frac{\mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n, T(X_1, \dots, X_n) = t]}{\mathbb{P}_\theta[T(X_1, \dots, X_n) = T(x_1, \dots, x_n)]} \\ &= \frac{\mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n]}{\mathbb{P}_\theta[T(X_1, \dots, X_n) = T(x_1, \dots, x_n)]} \\ &= \frac{L(x_1, \dots, x_n; \theta)}{q(T(x_1, \dots, x_n); \theta)}. \end{aligned}$$

Somit ist T eine suffiziente Statistik genau dann, wenn (7.2.1) nicht von θ abhängt.

Satz 7.2.2 (Faktorisierungssatz von Neyman–Fisher). *Eine Funktion $T : \mathbb{R}^n \to \mathbb{R}^m$ ist eine suffiziente Statistik genau dann, wenn es Funktionen $g : \mathbb{R}^m \times \Theta \to \mathbb{R}$ und $h : \mathbb{R}^n \to \mathbb{R}$ gibt, so dass für alle $x_1, \dots, x_n \in \mathbb{R}$ und alle $\theta \in \Theta$ die folgende Faktorisierung gilt:*

$$L(x_1, \dots, x_n; \theta) = g(T(x_1, \dots, x_n); \theta) \cdot h(x_1, \dots, x_n). \tag{7.2.2}$$

Beweis von "$\Longrightarrow$". Sei T eine suffiziente Statistik. Nach Lemma 7.2.1 ist die Funktion

$$h(x_1, \dots, x_n) := \begin{cases} \frac{L(x_1, \dots, x_n; \theta)}{q(T(x_1, \dots, x_n); \theta)}, & \text{falls } x_1, \dots, x_n \in \operatorname{Im} X_1, \\ 0, & \text{sonst} \end{cases}$$

unabhängig von θ. Mit diesem h und $g(t;\theta) = q(t;\theta)$ gilt die Faktorisierung (7.2.2). $\square$

Beweis von "$\Longleftarrow$". Es gelte die Faktorisierung (7.2.2). Wir bezeichnen mit $\sum^*$ die Summe über alle $(y_1, \ldots, y_n)$ mit $T(y_1, \ldots, y_n) = T(x_1, \ldots, x_n)$. Dann gilt

$$
\begin{aligned}
\frac{L(x_1, \ldots, x_n; \theta)}{q(T(x_1, \ldots, x_n); \theta)} &= \frac{g(T(x_1, \ldots, x_n); \theta) h(x_1, \ldots, x_n)}{\sum^* L(y_1, \ldots, y_n; \theta)} \\
&= \frac{g(T(x_1, \ldots, x_n); \theta) h(x_1, \ldots, x_n)}{\sum^* g(T(y_1, \ldots, y_n); \theta) h(y_1, \ldots, y_n)} \\
&= \frac{h(x_1, \ldots, x_n)}{\sum^* h(y_1, \ldots, y_n)}.
\end{aligned}
$$

Dieser Ausdruck hängt nicht von θ ab. Nach Lemma 7.2.1 ist T suffizient. $\square$

Beispiel 7.2.3. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $X_i \sim \text{Bern}(\theta)$, wobei $\theta \in (0, 1)$. Für die Likelihood–Funktion gilt

$$
\begin{aligned}
L(x_1, \ldots, x_n; \theta) &= h_\theta(x_1) \ldots h_\theta(x_n) \\
&= \theta^{x_1}(1-\theta)^{1-x_1} \mathbb{1}_{x_1 \in \{0,1\}} \cdot \ldots \cdot \theta^{x_n}(1-\theta)^{1-x_n} \mathbb{1}_{x_n \in \{0,1\}} \\
&= \theta^{x_1 + \ldots + x_n}(1-\theta)^{n-(x_1+\ldots+x_n)} \mathbb{1}_{x_1, \ldots, x_n \in \{0,1\}}.
\end{aligned}
$$

Daraus ist ersichtlich, dass die Neyman–Fisher–Faktorisierung (7.2.2) mit

$$
T(x_1, \ldots, x_n) = x_1 + \ldots + x_n, \quad g(t;\theta) = \theta^t (1-\theta)^{n-t}, \quad h(x_1, \ldots, x_n) = \mathbb{1}_{x_1, \ldots, x_n \in \{0,1\}}
$$

gilt. Nach dem Faktorisierungssatz von Neyman–Fisher ist T suffizient.

7.3. Definition der Suffizienz im absolut stetigen Fall

Bisher haben wir nur diskrete Zufallsvariablen betrachtet. Seien nun $X_1, \ldots, X_n$ absolut stetige Zufallsvariablen mit Dichte h_θ. Die Funktion $T : \mathbb{R}^n \to \mathbb{R}^m$ sei Borel–messbar. Außerdem nehmen wir an, dass $T(X_1, \ldots, X_n)$ ebenfalls eine absolut stetige Zufallsvariable mit Dichte $q(t;\theta)$ ist. Zum Beispiel darf T nicht konstant sein. Die vorherige Definition der Suffizienz macht im absolut stetigen Fall keinen Sinn, denn das Ereignis $T(X_1, \ldots, X_n) = t$ hat Wahrscheinlichkeit 0. Wir benötigen also eine andere Definition. Ein möglicher Zugang besteht darin, den Satz von Neyman–Fisher im absolut stetigen Fall als Definition zu benutzen. Die Likelihood–Funktion sei

$$
L(x_1, \ldots, x_n; \theta) = h_\theta(x_1) \ldots h_\theta(x_n).
$$

Definition 7.3.1. Im absolut stetigen Fall heißt eine Statistik T suffizient, wenn es eine Faktorisierung der Form

$$
L(x_1, \ldots, x_n; \theta) = g(T(x_1, \ldots, x_n); \theta) \cdot h(x_1, \ldots, x_n)
$$

gibt.

Beispiel 7.3.2. Seien $X_1, \dots, X_n$ unabhängige und auf dem Intervall $[0, \theta]$ gleichverteilte Zufallsvariablen, wobei $\theta > 0$ der unbekannte Parameter sei. Somit ist die Dichte von X_i gegeben durch

$$h_\theta(x) = \frac{1}{\theta} \mathbb{1}_{x \in [0,\theta]}.$$

Wir werden nun zeigen, dass $T(X_1, \dots, X_n) = \max\{X_1, \dots, X_n\}$ eine suffiziente Statistik ist. Für die Likelihood–Funktion gilt

$$\begin{aligned} L(x_1, \dots, x_n; \theta) &= h_\theta(x_1) \dots h_\theta(x_n) \\ &= \frac{1}{\theta^n} \mathbb{1}_{x_1 \in [0,\theta]} \cdot \ldots \cdot \mathbb{1}_{x_n \in [0,\theta]} \\ &= \frac{1}{\theta^n} \mathbb{1}_{\max(x_1, \dots, x_n) \le \theta} \cdot \mathbb{1}_{x_1, \dots, x_n \ge 0} \\ &= g(T(x_1, \dots, x_n); \theta) \cdot h(x_1, \dots, x_n), \end{aligned}$$

wobei

$$g(t; \theta) = \frac{1}{\theta^n} \mathbb{1}_{t \le \theta}, \quad h(x_1, \dots, x_n) = \mathbb{1}_{x_1, \dots, x_n \ge 0}.$$

Somit ist $T(X_1, \dots, X_n) = \max\{X_1, \dots, X_n\}$ eine suffiziente Statistik. Ein guter Schätzer für θ muss also eine Funktion von $\max\{X_1, \dots, X_n\}$ sein. Insbesondere ist der Schätzer $2\bar{X}_n$ in diesem Sinne nicht gut, dennn er benutzt überflüssige Information. Diese überflüssige Information steigert die Varianz des Schätzers. Das ist der Grund dafür, dass der Schätzer $\frac{n+1}{n} \max\{X_1, \dots, X_n\}$ (der suffizient und erwartungstreu ist) eine kleinere Varianz als der Schätzer $2\bar{X}_n$ (der nur erwartungstreu ist) hat.

Beispiel 7.3.3. Seien $X_1, \dots, X_n$ unabhängige und mit Parameter $\theta > 0$ exponentialverteilte Zufallsvariablen. Somit ist die Dichte von X_i gegeben durch

$$h_\theta(x) = \theta \exp(-\theta x) \mathbb{1}_{x \ge 0}.$$

Wir zeigen, dass $T(x_1, \dots, x_n) = x_1 + \ldots + x_n$ eine suffiziente Statistik ist. Für die Likelihood–Funktion gilt

$$\begin{aligned} L(x_1, \dots, x_n; \theta) &= h_\theta(x_1) \dots h_\theta(x_n) \\ &= \theta^n \exp(-\theta(x_1 + \ldots + x_n)) \mathbb{1}_{x_1, \dots, x_n \ge 0} \\ &= g(T(x_1, \dots, x_n); \theta) \cdot h(x_1, \dots, x_n), \end{aligned}$$

wobei

$$g(t; \theta) = \theta^n \exp(-\theta t), \quad h(x_1, \dots, x_n) = \mathbb{1}_{x_1, \dots, x_n \ge 0}.$$

Ein guter Schätzer für θ muss also eine Funktion von $X_1 + \ldots + X_n$ sein.

Beispiel 7.3.4. Seien $X_1, \dots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $X_i \sim N(\mu, \sigma^2)$. Der unbekannte Parameter ist $\theta = (\mu, \sigma^2)$, wobei $\mu \in \mathbb{R}$ und $\sigma^2 > 0$. Die Aufgabe besteht nun darin, eine suffiziente Statistik zu finden. Da wir normalverteilte Zufallsvariablen betrachten, gilt für die Dichte

$$h_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

Somit ist die Likelihood–Funktion gegeben durch

$$\begin{aligned} L(x_1, \ldots, x_n; \mu, \sigma^2) &= h_{\mu,\sigma^2}(x_1) \ldots h_{\mu,\sigma^2}(x_n) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right]\right). \end{aligned}$$

Nun betrachten wir die Statistik $T : \mathbb{R}^n \to \mathbb{R}^2$ mit

$$(x_1, \ldots, x_n) \mapsto \left(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i\right) = (T_1, T_2).$$

Diese Statistik T ist suffizient, denn wir haben die Neyman–Fisher–Faktorisierung

$$L(x_1, \ldots, x_n; \mu, \sigma^2) = g(T_1, T_2; \mu, \sigma^2)h(x_1, \ldots, x_n)$$

mit $h(x_1, \ldots, x_n) = 1$ und

$$g(T_1, T_2; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{T_1 - 2\mu T_2 + n\mu^2}{2\sigma^2}\right).$$

Allerdings ist T_1 oder T_2 allein betrachtet nicht suffizient.

Bemerkung 7.3.5. Im obigen Beispiel ist die Statistik $(\bar{x}_n, s_n^2)$ mit

$$\bar{x}_n = \frac{x_1 + \ldots + x_n}{n} \text{ und } s_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^n x_i^2 - n\bar{x}_n^2\right)$$

ebenfalls suffizient, denn

$$T_1 = n\bar{x}_n \text{ und } T_2 = (n-1)s_n^2 + n\bar{x}_n^2.$$

Wir können also $g(T_1, T_2; \mu, \sigma^2)$ auch als eine Funktion von $\bar{x}_n, s_n^2$ und μ, σ^2 schreiben.

Beispiel 7.3.6. Sei $\{h_\theta : \theta \in \Theta\}$ eine beliebige Familie von Dichten bzw. Zähldichten, dann ist die identische Abbildung $T : \mathbb{R}^n \to \mathbb{R}^n$ mit $(x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_n)$ suffizient. Die Suffizienz folgt aus dem Faktorisierungssatz von Neyman–Fisher, denn für die Likelihood–Funktion gilt

$$L(x_1, \ldots, x_n; \theta) = h_\theta(x_1) \cdot \ldots \cdot h_\theta(x_n) =: g(x_1, \ldots, x_n; \theta).$$

Beispiel 7.3.7. Sei $\{h_\theta : \theta \in \Theta\}$ eine beliebige Familie von Dichten oder Zähldichten, dann ist die Statistik

$$T : (x_1, \ldots, x_n) \mapsto (x_{(1)}, \ldots, x_{(n)})$$

suffizient. Das heißt, die Angabe der Werte der Stichprobe ohne die Angabe der Reihenfolge, in der diese Werte beobachtet wurden, ist suffizient. In der Tat, für Likelihood–Funktion gilt

$$L(x_1, \ldots, x_n; \theta) = h_\theta(x_1) \cdot \ldots \cdot h_\theta(x_n).$$

Diese Funktion ändert sich bei Permutationen von $x_1, \ldots, x_n$ nicht und kann somit als eine Funktion von T und θ dargestellt werden. Somit haben wir eine Neyman–Fisher–Faktorisierung angegeben.

7.4. Vollständigkeit

Sei $\{h_\theta(x) : \theta \in \Theta\}$ eine Familie von Dichten bzw. Zähldichten und seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte bzw. Zähldichte h_θ.

Definition 7.4.1. Eine Statistik $T : \mathbb{R}^n \to \mathbb{R}^m$ heißt *vollständig*, falls für alle Borel–Funktionen $g : \mathbb{R}^m \to \mathbb{R}$ aus der Gültigkeit von

$$\mathbb{E}_\theta g(T(X_1, \ldots, X_n)) = 0 \text{ für alle } \theta \in \Theta$$

folgt, dass $g(T(X_1, \ldots, X_n)) = 0$ fast sicher bezüglich $\mathbb{P}_\theta$ für alle $\theta \in \Theta$. Mit anderen Worten: Es gibt keinen nichttrivialen erwartungstreuen Schätzer von 0, der nur auf dem Wert der Statistik T basiert.

Beispiel 7.4.2. Seien $X_1, \ldots, X_n$ unabhängige und mit Parameter $\theta \in (0, 1)$ Bernoulli–verteilte Zufallsvariablen. In diesem Fall ist die Statistik

$$T : (X_1, \ldots, X_n) \to (X_1, \ldots, X_n)$$

nicht vollständig für $n \geq 2$. Um die Unvollständigkeit zu zeigen, betrachten wir die Funktion $g(X_1, \ldots, X_n) = X_2 - X_1$. Dann gilt für den Erwartungswert

$$\mathbb{E}_\theta g(T(X_1, \ldots, X_n)) = \mathbb{E}_\theta g(X_1, \ldots, X_n) = \mathbb{E}_\theta[X_2 - X_1] = 0,$$

denn X_2 hat die gleiche Verteilung wie X_1. Dabei ist $X_2 - X_1 \neq 0$ fast sicher, also ist die Bedingung aus der Definition der Vollständigkeit nicht erfüllt.

Beispiel 7.4.3. Seien $X_1, \ldots, X_n$ unabhängige und mit Parameter $\theta \in (0, 1)$ Bernoulli–verteilte Zufallsvariablen. Dann ist die Statistik

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

vollständig.

Beweis. Sei $g : \mathbb{R} \to \mathbb{R}$ eine Funktion mit $\mathbb{E}_\theta g(X_1 + \ldots + X_n) = 0$ für alle $\theta \in (0, 1)$. Somit gilt

$$0 = \sum_{i=0}^{n} g(i) \binom{n}{i} \theta^i (1-\theta)^{n-i} = (1-\theta)^n \sum_{i=0}^{n} g(i) \binom{n}{i} \left(\frac{\theta}{1-\theta}\right)^i.$$

Betrachte die Variable $z := \frac{\theta}{1-\theta}$. Nimmt θ alle möglichen Werte im Intervall $(0, 1)$ an, so nimmt z alle möglichen Werte im Intervall $(0, \infty)$ an. Es folgt, dass

$$\sum_{i=0}^{n} g(i) \binom{n}{i} z^i = 0 \text{ für alle } z > 0.$$

Also gilt für alle $i = 0, \ldots, n$, dass $g(i)\binom{n}{i} = 0$ und somit auch $g(i) = 0$. Hieraus folgt, dass $g = 0$ und die Vollständigkeit ist bewiesen. □

Beispiel 7.4.4. Seien $X_1, \ldots, X_n$ unabhängige und auf $[0, \theta]$ gleichverteilte Zufallsvariablen, wobei $\theta > 0$. Dann ist die Statistik $T(X_1, \ldots, X_n) = \max \{X_1, \ldots, X_n\}$ vollständig.

Beweis. Die Verteilungsfunktion von T unter $\mathbb{P}_\theta$ ist gegeben durch

$$\mathbb{P}_\theta[T \le x] = \begin{cases} 0, & x \le 0, \\ (\frac{x}{\theta})^n, & 0 \le x \le \theta, \\ 1, & x \ge \theta. \end{cases}$$

Die Dichte von T unter $\mathbb{P}_\theta$ erhält man indem man die Verteilungsfunktion ableitet:

$$q(x;\theta) = nx^{n-1}\theta^{-n}\mathbb{1}_{0\le x\le\theta}.$$

Sei nun $g : \mathbb{R} \to \mathbb{R}$ eine Borel–Funktion mit $\mathbb{E}_\theta g(T(X_1, \dots, X_n)) = 0$ für alle $\theta > 0$. Das heißt, es gilt

$$\theta^{-n}\int_0^\theta nx^{n-1}g(x)dx = 0 \text{ für alle } \theta > 0.$$

Wir können durch θ^{-n} teilen:

$$\int_0^\theta nx^{n-1}g(x)dx = 0 \text{ für alle } \theta > 0.$$

Nun können wir nach θ ableiten: $n\theta^{n-1}g(\theta) = 0$ und somit $g(\theta) = 0$ für Lebesgue–fast alle $\theta > 0$. Somit ist $g(x) = 0$ fast sicher bzgl. der Gleichverteilung auf $[0, \theta]$ für alle $\theta > 0$. Es sei bemerkt, dass g auf der negativen Halbachse durchaus ungleich 0 sein kann, allerdings hat die negative Halbachse Wahrscheinlichkeit 0 bzgl. der Gleichverteilung auf $[0, \theta]$. □

7.5. Exponentialfamilien

In diesem Abschnitt führen wir den Begriff der Exponentialfamilie ein. Auf der einen Seite, lässt sich für eine Exponentialfamilie sehr schnell eine suffiziente und vollständige Statistik (und somit, wie wir später sehen werden, der beste erwartungstreue Schätzer) konstruieren. Auf der anderen Seite, sind praktisch alle Verteilungsfamilien, die wir bisher betrachtet haben, Exponentialfamilien.
Sei $\{h_\theta(x) : \theta \in \Theta\}$ eine Familie von Dichten bzw. Zähldichten.

Definition 7.5.1. Die Familie $\{h_\theta(x) : \theta \in \Theta\}$ heißt *Exponentialfamilie*, falls es Funktionen $a(\theta)$, $b(x)$, $c(\theta)$, $d(x)$ gibt mit

$$h_\theta(x) = a(\theta)b(x)e^{c(\theta)d(x)}.$$

Beispiel 7.5.2. Betrachten wir die Familie der Binomialverteilungen mit Parametern n (bekannt) und $\theta \in (0, 1)$ (unbekannt). Für $x \in \{0, \dots, n\}$ ist die Zähldichte gegeben durch

$$h_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = (1-\theta)^n\binom{n}{x}\left(\frac{\theta}{1-\theta}\right)^x = (1-\theta)^n\binom{n}{x}\exp\left(\log\left(\frac{\theta}{1-\theta}\right)x\right).$$

Somit haben wir die Darstellung $h_\theta(x) = a(\theta)b(x)e^{c(\theta)d(x)}$ mit

$$a(\theta) = (1-\theta)^n, \quad b(x) = \binom{n}{x}, \quad c(\theta) = \log\left(\frac{\theta}{1-\theta}\right), \quad d(x) = x.$$

Beispiel 7.5.3. Für die Normalverteilung mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ ist die Dichte gegeben durch:

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

Unbekanntes μ, bekanntes σ^2. Betrachten wir den Parameter μ als unbekannt und σ^2 als gegeben und konstant, so gilt die Darstellung $h_{\mu,\sigma^2}(x) = a(\mu)b(x)e^{c(\mu)d(x)}$ mit

$$a(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right), \quad b(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad c(\mu) = \frac{\mu}{\sigma^2}, \quad d(x) = x.$$

Bekanntes μ, unbekanntes σ^2. Betrachten wir μ als gegeben und konstant und σ^2 als unbekannt, so gilt die Darstellung $h_{\mu,\sigma^2}(x) = a(\sigma^2)b(x)e^{c(\sigma^2)d(x)}$ mit

$$a(\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right), \quad b(x) = 1, \quad c(\sigma^2) = \frac{1}{2\sigma^2}, \quad d(x) = 2x\mu - x^2.$$

Weitere Beispiele von Exponentialfamilien sind die Familie der Poissonverteilungen und die familie der Exponentialverteilungen. Kein Beispiel hingegen ist die Familie der Gleichverteilungen auf $[0, \theta]$. Das liegt daran, dass der Träger der Gleichverteilung von θ abhängt.

Leider bildet die Familie der Normalverteilungen, wenn man sowohl μ als auch σ^2 als unbekannt betrachtet, keine Exponentialfamilie im Sinne der obigen Definition. Deshalb werden wir die obige Definition etwas erweitern.

Definition 7.5.4. Eine Familie $\{h_\theta : \theta \in \Theta\}$ von Dichten oder Zähldichten heißt eine *m-parametrige Exponentialfamilie*, falls es eine Darstellung der Form

$$h_\theta(x) = a(\theta)b(x)e^{c_1(\theta)d_1(x)+\ldots+c_m(\theta)d_m(x)}$$

gibt.

Beispiel 7.5.5. Die Familie der Normalverteilungen mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ (die beide als unbekannt betrachtet werden) ist eine 2-parametrige Exponentialfamilie, denn

$$h_{\mu,\sigma^2}(x) = a(\mu, \sigma^2)b(x)e^{c_1(\mu,\sigma^2)d_1(x)+c_2(\mu,\sigma^2)d_2(x)}$$

mit

$$a(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu^2}{\sigma^2}\right), \quad b(x) = 1,$$

$$c_1(\mu, \sigma^2) = -\frac{1}{2\sigma^2}, \quad d_1(x) = x^2, \quad c_2(\mu, \sigma^2) = \frac{\mu}{\sigma^2}, \quad d_2(x) = x.$$

Weitere Beispiele zwei-parametriger Exponentialfamilien sind die Familie der Gammaverteilungen und die Familie der Betaverteilungen.

7.6. Vollständige und suffiziente Statistik für Exponentialfamilien

Für eine Exponentialfamilie lässt sich sehr leicht eine suffiziente und vollständige Statistik angeben. Nämlich ist die Statistik $(T_1, \ldots, T_m)$ mit

$$T_1(X_1, \ldots, X_n) = \sum_{j=1}^n d_1(X_j), \quad \ldots, \quad T_m(X_1, \ldots, X_n) = \sum_{j=1}^n d_m(X_j)$$

suffizient. Um dies zu zeigen, schreiben wir die Likelihood–Funktion wie folgt um:

$$\begin{aligned}L(x_1,\dots,x_n;\theta) &= h_\theta(x_1)\dots h_\theta(x_n)\\ &= (a(\theta))^n b(x_1)\dots b(x_n)\exp\left(\sum_{i=1}^m c_i(\theta)d_i(x_1)\right)\dots\exp\left(\sum_{i=1}^m c_i(\theta)d_i(x_n)\right)\\ &= (a(\theta))^n\exp\left(T_1c_1(\theta)+\dots+T_mc_m(\theta)\right).\end{aligned}$$

Die Suffizienz von $(T_1,\dots,T_m)$ folgt aus dem Faktorisierungssatz von Neyman–Fisher mit

$$h(x_1,\dots,x_n)=b(x_1)\dots b(x_n),\quad g(T_1,\dots,T_m;\theta)=(a(\theta))^n\exp\left(T_1c_1(\theta)+\dots+T_mc_m(\theta)\right).$$

Man kann zeigen, dass diese Statistik auch vollständig ist (ohne Beweis).

Beispiel 7.6.1. Betrachten wir die Familie der Normalverteilungen mit Parametern $\mu\in\mathbb{R}$ und $\sigma^2>0$, wobei beide Parameter als unbekannt betrachtet werden. Wir haben bereits gesehen, dass diese Familie eine zweiparametrige Exponentialfamilie mit $d_1(x)=x^2$ und $d_2(x)=x$ ist. Somit st die Statistik (T_1,T_2) mit

$$\begin{aligned}T_1(X_1,\dots,X_n)&=\sum_{j=1}^n d_1(X_j)=\sum_{j=1}^n X_j^2,\\ T_2(X_1,\dots,X_n)&=\sum_{j=1}^n d_2(X_j)=\sum_{j=1}^n X_j\end{aligned}$$

suffizient und vollständig.

7.7. Der beste erwartungstreue Schätzer

Sei $\Theta=(a,b)\subset\mathbb{R}$. Wir betrachten eine Familie von Dichten bzw. Zähldichten $\{h_\theta:\theta\in\Theta\}$. Seien $X_1,\dots,X_n$ unabhängige und identisch verteilte Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega,\mathcal{A},\mathbb{P}_\theta)$ mit Dichte bzw. Zähldichte h_θ.
Wir bezeichnen mit L^2 die Menge aller Stichprobenfunktionen (Schätzer) $f:\mathbb{R}^n\to\mathbb{R}$ mit

$$\mathbb{E}_\theta[f(X_1,\dots,X_n)^2]<\infty \text{ für alle } \theta\in\Theta.$$

Im Folgenden werden wir nur Schätzer aus L^2 betrachten. Außerdem benutzen wir die folgende Abkürzung: Wir benutzen $\hat\theta$ als eine Bezeichnung für die Zufallsvariable $\hat\theta(X_1,\dots,X_n)$.

Definition 7.7.1. Ein Schätzer $\hat\theta:\mathbb{R}^n\to\mathbb{R}$ heißt *erwartungstreu* (für θ), falls

$$\mathbb{E}_\theta\hat\theta=\theta \text{ für alle } \theta\in\Theta.$$

Wir bezeichnen mit H die Menge der erwartungstreuen Schätzer, d.h.

$$H=\{\hat\theta\in L^2:\hat\theta \text{ ist erwartungstreu}\}.$$

Aufgabe 7.7.2. Zeigen Sie, dass H ein affiner Unterraum ist, d.h. für alle $\hat\theta_1,\hat\theta_2\in H$ und alle $t\in\mathbb{R}$ ist auch $t\hat\theta_1+(1-t)\hat\theta_2\in H$.

Wir haben bereits gesehen, dass es in parametrischen Modellen typischerweise mehrere erwartungstreue Schätzer für den Parameter existieren. Wie wählt man unter diesen Schätzer den besten?

Definition 7.7.3. Ein Schätzer $\hat{\theta}$ heißt *bester erwartungstreuer Schätzer* (für θ), falls er erwartungstreu ist und wenn für jeden anderen erwartungstreuen Schätzer $\tilde{\theta} \in H$ gilt, dass

$$\operatorname{Var}_\theta \hat{\theta} \leq \operatorname{Var}_\theta \tilde{\theta} \text{ für alle } \theta \in \Theta.$$

Im nächsten Satz zeigen wir, dass es höchstens einen besten erwartungstreuen Schätzer geben kann.

Satz 7.7.4. *Seien $\hat{\theta}_1, \hat{\theta}_2 : \mathbb{R}^n \to \Theta$ zwei beste erwartungstreue Schätzer, dann gilt*

$$\hat{\theta}_1 = \hat{\theta}_2 \textit{ fast sicher unter } \mathbb{P}_\theta \textit{ für alle } \theta \in \Theta.$$

Bemerkung 7.7.5. Der beste erwartungstreue Schätzer muss nicht in jedem parametrischen Modell existieren. Es kann nämlich durchaus passieren, dass der erwartungstreue Schätzer, der die kleinste Varianz unter $\mathbb{P}_\theta$ hat, nicht die kleinste Varianz unter einem anderen Wahrscheinlichkeitsmaß $\mathbb{P}_{\theta'}$ hat.

Beweis. Schritt 1. Da beide Schätzer beste erwartungstreue Schätzer sind, stimmen die Varianzen dieser beiden Schätzer überein, d.h.

$$\operatorname{Var}_\theta \hat{\theta}_1 = \operatorname{Var}_\theta \hat{\theta}_2 \text{ für alle } \theta \in \Theta.$$

Ist nun $\operatorname{Var}_\theta \hat{\theta}_1 = \operatorname{Var}_\theta \hat{\theta}_2 = 0$ für ein $\theta \in \Theta$, so sind $\hat{\theta}_1$ und $\hat{\theta}_2$ fast sicher konstant unter $\mathbb{P}_\theta$. Da beide Schätzer erwartungstreu sind, muss diese Konstante gleich θ sein und somit muss $\hat{\theta}_1 = \hat{\theta}_2$ fast sicher unter $\mathbb{P}_\theta$ gelten. Die Behauptung des Satzes wäre somit gezeigt. Wir können also im Folgenden annehmen, dass die beiden Varianzen $\operatorname{Var}_\theta \hat{\theta}_1 = \operatorname{Var}_\theta \hat{\theta}_2$ strikt positiv sind.

Schritt 2. Da beide Schätzer erwartungstreu sind, ist auch $\theta^* = \frac{\hat{\theta}_1 + \hat{\theta}_2}{2}$ erwartungstreu und für die Varianz von θ^* gilt

$$\begin{aligned}
\operatorname{Var}_\theta \theta^* &= \frac{1}{4} \operatorname{Var}_\theta \hat{\theta}_1 + \frac{1}{4} \operatorname{Var}_\theta \hat{\theta}_2 + \frac{1}{2} \operatorname{Cov}_\theta(\hat{\theta}_1, \hat{\theta}_2) \\
&\leq \frac{1}{2} \operatorname{Var}_\theta \hat{\theta}_1 + \frac{1}{2} \sqrt{\operatorname{Var}_\theta \hat{\theta}_1} \sqrt{\operatorname{Var}_\theta \hat{\theta}_2} \\
&= \operatorname{Var}_\theta \hat{\theta}_1.
\end{aligned}$$

Dabei wurde die Cauchy–Schwarzsche Ungleichung angewendet. Somit folgt, dass $\operatorname{Var}_\theta \theta^* \leq \operatorname{Var}_\theta \hat{\theta}_1$. Allerdings ist $\hat{\theta}_1$ der beste erwartungstreue Schätzer, also muss $\operatorname{Var}_\theta \theta^* = \operatorname{Var}_\theta \hat{\theta}_1$ gelten. Daraus folgt, dass

$$\operatorname{Cov}_\theta(\hat{\theta}_1, \hat{\theta}_2) = \operatorname{Var}_\theta \hat{\theta}_1 = \operatorname{Var}_\theta \hat{\theta}_2.$$

Schritt 3. Der Korrelationskoeffizient von $\hat{\theta}_1$ und $\hat{\theta}_2$ ist also gleich 1. Somit besteht ein linearer Zusammenhang zwischen $\hat{\theta}_1$ und $\hat{\theta}_2$, d.h. es gibt $a = a(\theta)$, $b = b(\theta)$ mit

$$\hat{\theta}_2 = a(\theta) \cdot \hat{\theta}_1 + b(\theta) \text{ fast sicher unter } \mathbb{P}_\theta \text{ für alle } \theta \in \Theta.$$

Setzen wir diesen Zusammenhang bei der Betrachtung der Kovarianz ein und berücksichtigen zusätzlich, dass wie oben gezeigt $\operatorname{Var}_\theta \hat{\theta}_1 = \operatorname{Cov}_\theta(\hat{\theta}_1, \hat{\theta}_2)$, so erhalten wir, dass

$$\operatorname{Var}_\theta \hat{\theta}_1 = \operatorname{Cov}_\theta(\hat{\theta}_1, \hat{\theta}_2) = \operatorname{Cov}_\theta(\hat{\theta}_1, a(\theta) \cdot \hat{\theta}_1 + b(\theta)) = a(\theta) \cdot \operatorname{Var}_\theta \hat{\theta}_1.$$

Also ist $a(\theta) = 1$.

Schritt 4. Somit gilt $\hat{\theta}_2 = \hat{\theta}_1 + b(\theta)$. Auf Grund der Erwartungstreue der Schätzer ist $b(\theta) = 0$, denn

$$\theta = \mathbb{E}_\theta \hat{\theta}_2 = \mathbb{E}_\theta \hat{\theta}_1 + b(\theta) = \theta + b(\theta).$$

Somit folgt, dass $\hat{\theta}_1 = \hat{\theta}_2$ fast sicher unter $\mathbb{P}_\theta$ für alle $\theta \in \Theta$. □

Definition 7.7.6. Ein Stichprobenfunktion $\varphi : \mathbb{R}^n \to \Theta$ heißt *erwartungstreuer Schätzer für* 0, falls

$$\mathbb{E}_\theta \varphi = 0 \text{ für alle } \theta \in \Theta.$$

Satz 7.7.7. *Sei $\hat{\theta}$ ein erwartungstreuer Schätzer für θ. Dann sind die folgenden Bedingungen äquivalent:*

(1) *$\hat{\theta}$ ist der beste erwartungstreue Schätzer für θ.*
(2) *Für alle erwartungstreuen Schätzer φ für* 0 *gilt, dass* $\operatorname{Cov}_\theta(\hat{\theta}, \varphi) = 0$ *für alle $\theta \in \Theta$.*

Also ist ein erwartungstreuer Schätzer genau dann der beste erwartungstreue Schätzer, wenn er zu jedem erwartungstreuen Schätzer für 0 orthogonal ist.

Beweis von "$\Longrightarrow$". Sei $\hat{\theta}$ der beste erwartungstreue Schätzer für θ und sei $\varphi : \mathbb{R}^n \to \Theta$ eine Stichprobenfunktion mit $\mathbb{E}_\theta \varphi = 0$ für alle $\theta \in \Theta$. Somit müssen wir zeigen, dass

$$\operatorname{Cov}_\theta(\hat{\theta}, \varphi) = 0 \text{ für alle } \theta \in \Theta.$$

Definieren wir uns hierfür $\tilde{\theta} = \hat{\theta} + a\varphi$, $a \in \mathbb{R}$. Dann ist $\tilde{\theta}$ ebenfalls ein erwartungstreuer Schätzer für θ, denn

$$\mathbb{E}_\theta \tilde{\theta} = \mathbb{E}_\theta \hat{\theta} + a \cdot \mathbb{E}_\theta \varphi = \theta.$$

Es gilt für die Varianz von $\tilde{\theta}$, dass

$$\operatorname{Var}_\theta \tilde{\theta} = \operatorname{Var}_\theta \hat{\theta} + a^2 \operatorname{Var}_\theta \varphi + 2a \operatorname{Cov}_\theta(\hat{\theta}, \varphi) = \operatorname{Var}_\theta \hat{\theta} + g(a),$$

wobei $g(a) = a^2 \operatorname{Var}_\theta \varphi + 2a \operatorname{Cov}_\theta(\hat{\theta}, \varphi)$. Wäre nun $\operatorname{Cov}_\theta(\hat{\theta}, \varphi) \neq 0$, dann hätte die quadratische Funktion g zwei verschiedene Nullstellen bei 0 und $-2\operatorname{Cov}_\theta(\hat{\theta}, \varphi)/\operatorname{Var}_\theta \varphi$. (Wir dürfen hier annehmen, dass $\operatorname{Var}_\theta \varphi \neq 0$, denn andernfalls wäre φ fast sicher konstant unter $\mathbb{P}_\theta$ und dann würde $\operatorname{Cov}_\theta(\hat{\theta}, \varphi) = 0$ trivialerweise gelten). Zwischen diesen Nullstellen gäbe es ein $a \in \mathbb{R}$ mit $g(a) < 0$ und es würde folgen, dass $\operatorname{Var}_\theta \tilde{\theta} < \operatorname{Var}_\theta \hat{\theta}$. Das widerspricht aber der Annahme, dass $\hat{\theta}$ der beste erwartungstreue Schätzer für θ ist. Somit muss $\operatorname{Cov}_\theta(\hat{\theta}, \varphi) = 0$ gelten. □

Beweis von "$\Longleftarrow$". Sei $\hat{\theta}$ ein erwartungstreuer Schätzer für θ. Sei außerdem $\operatorname{Cov}_\theta(\varphi, \hat{\theta}) = 0$ für alle erwartungstreuen Schätzer φ für 0. Jetzt werden wir zeigen, dass $\hat{\theta}$ der beste erwartungstreue Schätzer ist. Mit $\tilde{\theta}$ bezeichnen wir einen anderen erwartungstreuen Schätzer für θ. Somit genügt es zu zeigen, dass

$$\operatorname{Var}_\theta \hat{\theta} \leq \operatorname{Var}_\theta \tilde{\theta}.$$

Um das zu zeigen, schreiben wir $\tilde{\theta} = \hat{\theta} + (\tilde{\theta} - \hat{\theta}) =: \hat{\theta} + \varphi$. Da $\hat{\theta}$ und $\tilde{\theta}$ beide erwartungstreue Schätzer für θ sind, ist $\varphi := (\tilde{\theta} - \hat{\theta})$ ein erwartungstreuer Schätzer für 0. Für die Varianzen von $\tilde{\theta}$ und $\hat{\theta}$ gilt:

$$\operatorname{Var}_\theta \tilde{\theta} = \operatorname{Var}_\theta \hat{\theta} + \operatorname{Var}_\theta \varphi + 2\operatorname{Cov}_\theta(\hat{\theta}, \varphi) = \operatorname{Var}_\theta \hat{\theta} + \operatorname{Var}_\theta \varphi \geq \operatorname{Var}_\theta \hat{\theta}.$$

Die letzte Ungleichung gilt, da $\operatorname{Var}_\theta \varphi \geq 0$. Somit ist $\hat{\theta}$ der beste erwartungstreue Schätzer. $\square$

7.8. Bedingter Erwartungswert

Definition 7.8.1. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Seien $A \in \mathcal{A}$ und $B \in \mathcal{A}$ zwei Ereignisse. Dann ist die *bedingte Wahrscheinlichkeit* von A gegeben B folgendermaßen definiert:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Diese Definition macht nur dann Sinn, wenn $\mathbb{P}[B] \neq 0$.

Analog kann man den bedingten Erwartungswert definieren.

Definition 7.8.2. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine Zufallsvariable mit $\mathbb{E}|X| < \infty$. Sei $B \in \mathcal{A}$ ein Ereignis. Dann ist der *bedingte Erwartungswert* von X gegeben B folgendermaßen definiert:

$$\mathbb{E}[X|B] = \frac{\mathbb{E}[X\mathbb{1}_B]}{\mathbb{P}[B]}.$$

Auch diese Definition macht nur dann Sinn, wenn $\mathbb{P}[B] \neq 0$. Zwischen den beiden Begriffen (bedingte Wahrscheinlichkeit und bedingter Erwartungswert) besteht der folgende Zusammenhang:

$$\mathbb{P}[A|B] = \mathbb{E}[\mathbb{1}_A|B].$$

Wir haben aber gesehen (z.B. bei der Definition der Suffizienz im absolut stetigen Fall), dass man oft bedingte Wahrscheinlichkeiten oder Erwartungswerte auch im Falle $\mathbb{P}[B] = 0$ betrachten muss. In diesem Abschnitt werden wir eine allgemeine Definition des bedingten Erwartungswerts geben, die das (zumindest in einigen Fällen) möglich macht.

Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable, definiert auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit $\mathbb{E}|X| < \infty$. Sei $\mathcal{B} \subset \mathcal{A}$ eine Teil–σ–Algebra von $\mathcal{A}$, d.h. für jede Menge $B \in \mathcal{B}$ gelte auch $B \in \mathcal{A}$. In diesem Abschnitt werden wir den bedingten Erwartungswert von X gegeben die σ–Algebra $\mathcal{B}$ definieren.

Sei zunächst $X \geq 0$ fast sicher.

Schritt 1. Sei Q ein Maß auf dem Messraum $(\Omega, \mathcal{B})$ mit

$$Q(B) = \mathbb{E}[X\mathbb{1}_B] \text{ für alle } B \in \mathcal{B}.$$

Das Maß Q ist endlich, denn $Q(\Omega) = \mathbb{E}X < \infty$ nach Voraussetzung. Es sei bemerkt, dass das Maß Q auf $(\Omega, \mathcal{B})$ und nicht auf $(\Omega, \mathcal{A})$ definiert wurde. Das Wahrscheinlichkeitsmaß

$\mathbb{P}$ hingegen ist auf $(\Omega, \mathcal{A})$ definiert, wir können es aber auch auf die kleinere σ–Algebra $\mathcal{B}$ einschränken und als ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{B})$ betrachten.

SCHRITT 2. Ist nun $B \in \mathcal{B}$ eine Menge mit $\mathbb{P}[B] = 0$, so folgt, dass $Q(B) = \mathbb{E}[X \mathbb{1}_B] = 0$, denn die Zufallsvariable $X\mathbb{1}_B$ ist $\mathbb{P}$-fast sicher gleich 0. Somit ist Q absolut stetig bezüglich $\mathbb{P}$ auf $(\Omega, \mathcal{B})$. Nach dem Satz von Radon–Nikodym gibt es eine Funktion Z, die messbar bezüglich $\mathcal{B}$ ist, mit

$$\mathbb{E}[Z\mathbb{1}_B] = \mathbb{E}[X\mathbb{1}_B] \text{ für alle } B \in \mathcal{B}.$$

Es sei bemerkt, dass X $\mathcal{A}$–messbar ist, wohingegen Z lediglich $\mathcal{B}$–messbar ist. Wir nennen die Zufallsvariable Z den bedingten Erwartungswert von X gegeben $\mathcal{B}$ und schreiben

$$\mathbb{E}[X|\mathcal{B}] = Z.$$

SCHRITT 3. Sei nun X eine beliebige (nicht unbedingt positive) Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit. Sei $\mathcal{B} \subset \mathcal{A}$ nach wie vor eine Teil–σ–Algebra. Wir haben die Darstellung $X = X^+ - X^-$ mit $X^+ \geq 0$ und $X^- \geq 0$. Die bedingte Erwartung von X gegeben $\mathcal{B}$ ist definiert durch

$$\mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X^+|\mathcal{B}] - \mathbb{E}[X^-|\mathcal{B}].$$

Wir können nun die obigen Überlegungen zu folgender Definition zusammenfassen.

DEFINITION 7.8.3. Sei X eine Zufallsvariable mit $\mathbb{E}|X| < \infty$, definiert auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. (Somit ist X $\mathcal{A}$–messbar). Sei $\mathcal{B} \subset \mathcal{A}$ eine Teil–σ–Algebra. Eine Funktion $Z : \Omega \to \mathbb{R}$ heißt *bedingter Erwartungswert von X gegeben* $\mathcal{B}$, falls

(1) Z ist $\mathcal{B}$–messbar.
(2) $\mathbb{E}[Z\mathbb{1}_B] = \mathbb{E}[X\mathbb{1}_B]$ für alle $B \in \mathcal{B}$.

Wir schreiben dann $\mathbb{E}[X|\mathcal{B}] = Z$.

BEMERKUNG 7.8.4. Die bedingte Erwartung $\mathbb{E}[X|\mathcal{B}]$ ist eine Zufallsvariable, keine Konstante. Die Existenz von $\mathbb{E}[X|\mathcal{B}]$ wurde bereits oben mit dem Satz von Radon–Nikodym bewiesen. Der bedingte Erwartungswert ist bis auf $\mathbb{P}$-Nullmengen eindeutig definiert. Das folgt aus der entsprechenden Eigenschaft der Dichte im Satz von Radon–Nikodym.

BEISPIEL 7.8.5. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Betrachte eine disjunkte Zerlegung $\Omega = \Omega_1 \cup \ldots \cup \Omega_n$, wobei $\Omega_i \in \mathcal{A}$ und $\mathbb{P}[\Omega_i] \neq 0$. Sei $\mathcal{B}$ die σ–Algebra, die von $\{\Omega_1, \ldots, \Omega_n\}$ erzeugt wird. Somit ist

$$\mathcal{B} = \{\Omega_1^{\varepsilon_1} \cup \ldots \cup \Omega_n^{\varepsilon_n} : \varepsilon_1, \ldots, \varepsilon_n \in \{0, 1\}\},$$

wobei $\Omega_i^1 = \Omega_i$ und $\Omega_i^0 = \emptyset$. Sei X eine beliebige ($\mathcal{A}$–messbare) Zufallsvariable auf Ω mit $\mathbb{E}|X| < \infty$. Für den bedingten Erwartungswert von X geegeben $\mathcal{B}$ gilt:

$$Z(\omega) := \mathbb{E}[X|\mathcal{B}](\omega) = \frac{\mathbb{E}[X\mathbb{1}_{\Omega_i}]}{\mathbb{P}[\Omega_i]}.$$

BEWEIS. Beachte, dass $Z := \mathbb{E}[X|\mathcal{B}]$ $\mathcal{B}$–messbar sein muss. Also ist Z konstant auf jeder Menge Ω_i. Sei also $Z(\omega) = c_i$ für $\omega \in \Omega_i$. Es muss außerdem gelten, dass

$$\mathbb{E}[X\mathbb{1}_{\Omega_i}] = \mathbb{E}[Z\mathbb{1}_{\Omega_i}] = c_i\mathbb{P}[\Omega_i].$$

Daraus folgt, dass $c_i = \mathbb{E}[X\mathbb{1}_{\Omega_i}]/\mathbb{P}[\Omega_i]$ sein muss. □

Beispiel 7.8.6. Sei $\Omega = [0,1]^2$. Sei $\mathcal{A}$ die Borel–σ–Algebra auf $[0,1]^2$ und $\mathbb{P}$ das Lebesgue–Maß. Sei $X : [0,1]^2 \to \mathbb{R}$ eine ($\mathcal{A}$–messbare) Zufallsvariable mit $\mathbb{E}|X| < \infty$. Sei $\mathcal{B} \subset \mathcal{A}$ eine Teil–σ–Algebra von $\mathcal{A}$ mit

$$\mathcal{B} = \{C \times [0,1] : C \subset [0,1] \text{ ist Borel}\}.$$

Dann ist der bedingte Erwartungswert von X gegeben $\mathcal{B}$ gegeben durch:

$$Z(s,t) := \mathbb{E}[X|\mathcal{B}](s,t) = \int_0^1 X(s,t)dt, \quad (s,t) \in [0,1]^2.$$

Beweis. Wir zeigen, dass die soeben definierte Funktion Z die beiden Bedingungen aus der Definition der bedingten Erwartung erfüllt. Zunächst ist $Z(s,t)$ eine Funktion, die nur von s abhängt. Somit ist Z messbar bzgl. $\mathcal{B}$. Außerdem gilt für jede $\mathcal{B}$–messbare Menge $B = C \times [0,1]$, dass

$$\mathbb{E}[Z\mathbb{1}_{C\times[0,1]}] = \int_{C\times[0,1]} Z(s,t)dsdt = \int_C \left(\int_0^1 Z(s,t)dt\right) ds = \mathbb{E}[X\mathbb{1}_{C\times[0,1]}].$$

Somit ist auch die zweite Bedingung erfüllt.

Beispiel 7.8.7. Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable mit $\mathbb{E}|X| < \infty$, definiert auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Dann gilt

(1) $\mathbb{E}[X|\{0,\emptyset\}] = \mathbb{E}X$.
(2) $\mathbb{E}[X|\mathcal{A}] = X$.

Beweis. Übung. □

Satz 7.8.8. *Sei $X, Y : \Omega \to \mathbb{R}$ Zufallsvariablen (beide $\mathcal{A}$–messbar) mit $\mathbb{E}|X| < \infty$, $\mathbb{E}|Y| < \infty$, definiert auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Sei $\mathcal{B} \subset \mathcal{A}$ eine Teil–σ–Algebra von $\mathcal{A}$.*

(1) *Es gilt die Formel der totalen Erwartung: $\mathbb{E}[\mathbb{E}[X|\mathcal{B}]] = \mathbb{E}X$.*
(2) *Aus $X \le Y$ fast sicher folgt, dass $\mathbb{E}[X|\mathcal{B}] \le \mathbb{E}[Y|\mathcal{B}]$ fast sicher.*
(3) *Für alle $a, b \in \mathbb{R}$ gilt $\mathbb{E}[aX + bY|\mathcal{B}] = a\,\mathbb{E}[X|\mathcal{B}] + b\,\mathbb{E}[Y|\mathcal{B}]$ fast sicher.*
(4) *Falls Y sogar $\mathcal{B}$–messbar ist und $\mathbb{E}|XY| < \infty$, dann gilt*

$$\mathbb{E}[XY|\mathcal{B}] = Y\mathbb{E}[X|\mathcal{B}] \text{ fast sicher.}$$

Beweis. Übung. □

Besonders oft wird die Definition der bedingten Erwartung im folgenden Spezialfall benutzt.

Definition 7.8.9. Seien X und Y zwei $\mathcal{A}$–messbare Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Sei

$$\sigma(Y) = \{Y^{-1}(C) : C \subset \mathbb{R} \text{ Borel}\}$$

die *von Y erzeugte σ–Algebra*. Der bedingte Erwartungswert von X gegeben Y ist definiert durch

$$\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)].$$

Bemerkung 7.8.10. Aus der Messbarkeit des bedingten Erwartungswertes bzgl. $\sigma(Y)$ kann man herleiten, dass $\mathbb{E}[X|Y]$ eine Borel–Funktion von Y sein muss. Es gibt also eine Borel–Funktion $g : \mathbb{R} \to \mathbb{R}$ mit

$$\mathbb{E}[X|Y] = g(Y) \text{ fast sicher.}$$

Wir schreiben dann $\mathbb{E}[X|Y = t] = g(t)$. Dabei darf $\mathbb{P}[Y = t]$ auch 0 sein.

Bemerkung 7.8.11. Seien X, Y zwei diskrete Zufallsvariablen mit gemeinsamer Zähldichte $f_{X,Y}(s,t) = \mathbb{P}[X = s, Y = t]$ und die Zähldichte von Y sei $f_Y(t)$. Dann gilt für den bedingten Erwartungswert

$$\mathbb{E}[X|Y](\omega) = \mathbb{E}[X|Y = t] = \sum_s \mathbb{P}[X = s|Y = t]s = \frac{\sum_s f_{X,Y}(s,t)s}{f_Y(t)}, \text{ wenn } Y(\omega) = t.$$

Seien X, Y zwei absolut stetige Zufallsvariablen mit gemeinsamer Dichte $f_{X,Y}(s,t)$ und die Dichte von Y sei $f_Y(t)$. Man kann zeigen, dass dann für den bedingten Erwartungswert eine ähnliche Formel gilt:

$$\mathbb{E}[X|Y](\omega) = \mathbb{E}[X|Y = t] = \frac{\int_{\mathbb{R}} f_{X,Y}(s,t)sds}{f_Y(t)}, \text{ wenn } Y(\omega) = t.$$

Diese Formel hat Sinn, wenn $f_Y(t) \neq 0$.

Beispiel 7.8.12. In diesem Beispiel betrachten wir zwei faire Münzen. Seien X_1 und X_2 zwei Zufallsvariablen, die den Wert 1 annehmen, wenn die erste bzw. die zweite Münze Kopf zeigt, und den Wert 0 sonst. Sei $Y = X_1 + X_2$. Wir bestimmen $\mathbb{E}[X_1|Y]$.

Lösung. Die Grundmenge ist $\Omega = \{0,1\}^2 = \{00, 01, 10, 11\}$. Wir zerlegen die Grundmenge Ω mit Hilfe von Y in

$$\Omega_0 = Y^{-1}(0) = \{00\}, \quad \Omega_1 = Y^{-1}(1) = \{01, 10\}, \quad \Omega_2 = Y^{-1}(2) = \{11\}.$$

Für $\omega \in \Omega_0$ gilt

$$\mathbb{E}[X_1|Y](\omega) = \frac{\mathbb{E}[X_1 \mathbb{1}_{\Omega_0}]}{\mathbb{P}[\Omega_0]} = \frac{0}{1/4} = 0.$$

Für $\omega \in \Omega_1$ gilt

$$\mathbb{E}[X_1|Y](\omega) = \frac{\mathbb{E}[X_1 \mathbb{1}_{\Omega_1}]}{\mathbb{P}[\Omega_1]} = \frac{1/4}{1/2} = \frac{1}{2}.$$

Für $\omega \in \Omega_1$ gilt

$$\mathbb{E}[X_1|Y](\omega) = \frac{\mathbb{E}[X_1 \mathbb{1}_{\Omega_2}]}{\mathbb{P}[\Omega_2]} = \frac{1/4}{1/4} = 1.$$

Zusammenfassend gilt $\mathbb{E}[X_1|Y] = \mathbb{E}[X_1|X_1 + X_2] = \frac{1}{2}(X_1 + X_2)$.

7.9. Satz von Lehmann–Scheffé

Sei $\{h_\theta : \theta \in \Theta\}$ mit $\Theta = (a, b)$ eine Familie von Dichten bzw. Zähldichten und $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte bzw. Zähldichte h_θ.

Satz 7.9.1 (Lehmann–Scheffé). *Sei $\hat{\theta} \in L^2$ ein erwartungstreuer, suffizienter und vollständiger Schätzer für θ. Dann ist $\hat{\theta}$ der beste erwartungstreue Schätzer für θ.*

Bevor wir den Satz beweisen, betrachten wir eine Reihe von Beispielen.

Beispiel 7.9.2. Seien $X_1, \ldots, X_n$ unabhängig und gleichverteilt auf $[0, \theta]$, wobei $\theta > 0$ geschätzt werden soll. Wir haben bereits gezeigt, dass $X_{(n)} = \max\{X_1, \ldots, X_n\}$ eine suffiziente und vollständige Statistik ist. Jedoch ist der Schätzer $X_{(n)}$ nicht erwartungstreu, denn

$$\mathbb{E}_\theta X_{(n)} = \frac{n}{n+1}\theta.$$

Deshalb betrachten wir den Schätzer

$$\hat{\theta}(X_1, \ldots, X_n) := \frac{n+1}{n} X_{(n)} = \frac{n+1}{n} \max\{X_1, \ldots, X_n\}.$$

Dieser Schätzer ist erwartungstreu, suffizient und vollständig. Nach dem Satz von Lehmann–Scheffé ist somit $\frac{n+1}{n} X_{(n)}$ der beste erwartungstreue Schätzer für θ.

Beispiel 7.9.3. Seien $X_1, \ldots, X_n$ unabhängige, mit Parameter $\theta \in [0, 1]$ Bernoulli–verteilte Zufallsvariablen. Der Schätzer $\bar{X}_n$ ist erwartungstreu, suffizient und vollständig und somit bester erwartungstreuer Schätzer für θ. Diese Argumentation greift auch für unabhängige, mit Parameter $\theta > 0$ Poisson–verteilte Zufallsvariablen. Dabei ist der Beweis der Suffizienz und Vollständigkeit eine Übung.

Beispiel 7.9.4. Seien $X_1, \ldots, X_n$ unabhängig und normalverteilt mit bekannter Varianz $\sigma^2 > 0$ und unbekanntem Erwartungswert $\mu \in \mathbb{R}$. Diese Verteilungen bilden eine Exponentialfamilie und die Statistik $\bar{X}_n$ ist vollständig und suffizient. Außerdem ist $\bar{X}_n$ erwartungstreu. Der beste erwartungstreue Schätzer für μ ist somit $\bar{X}_n$.
Versuchen wir nun, μ^2 als Parameter zu betrachten und zu schätzen. Der Schätzer $\bar{X}_n^2$ ist nicht erwartungstreu, denn

$$\mathbb{E}_\mu \bar{X}_n^2 = \operatorname{Var}_\mu \bar{X}_n + (\mathbb{E}_\mu \bar{X}_n)^2 = \frac{1}{n}\sigma^2 + \mu^2.$$

Deshalb betrachten wir den Schätzer $\bar{X}_n^2 - \frac{\sigma^2}{n}$. Dieser Schätzer ist erwartungstreu, suffizient und vollständig (Übung) und somit bester erwartungstreuer Schätzer für μ^2.

Beweis von Satz 7.9.1. Sei $\hat{\theta}$ ein erwartungstreuer, suffizienter und vollständiger Schätzer für θ. Wir wollen zeigen, dass $\hat{\theta}$ bester erwartungstreuer Schätzer für θ ist. Sei $\varphi : \mathbb{R}^n \to \mathbb{R}$ ein erwartungstreuer Schätzer für 0, d.h. $\mathbb{E}_\theta \varphi = 0$ für alle $\theta \in \Theta$. Wir zeigen, dass

$$\mathbb{E}_\theta[\hat{\theta}\varphi] = 0.$$

Dieser Erwartungswert kann mit Hilfe der bedingten Erwartung umgeschrieben werden zu

$$\mathbb{E}_\theta[\hat{\theta}\varphi] = \mathbb{E}_\theta[\mathbb{E}_\theta[\hat{\theta}\varphi|\hat{\theta}]] = \mathbb{E}_\theta[\hat{\theta}\mathbb{E}_\theta[\varphi|\hat{\theta}]] = \mathbb{E}_\theta[\hat{\theta} g(\hat{\theta})],$$

wobei $g(\hat{\theta}) = \mathbb{E}_\theta[\varphi|\hat{\theta}]$. Die Funktion g ist unabhängig von θ, da $\hat{\theta}$ suffizient ist. Es gilt

$$\mathbb{E}_\theta[g(\hat{\theta})] = \mathbb{E}_\theta[\mathbb{E}_\theta[\varphi|\hat{\theta}]] = \mathbb{E}_\theta[\varphi] = 0,$$

da φ ein erwartungstreuer Schätzer für 0 ist. Somit folgt durch die Vollständigkeit von $\hat{\theta}$, dass $g = 0$ fast sicher unter $\mathbb{P}_\theta$ ist. Also gilt

$$\mathbb{E}_\theta[\hat{\theta}\varphi] = \mathbb{E}_\theta[\hat{\theta} \cdot 0] = 0.$$

Wir haben gezeigt, dass $\hat{\theta}$ orthogonal zu jedem erwartungstreuen Schätzer von 0 ist. Laut Satz 7.7.7 ist $\hat{\theta}$ der beste erwartungstreue Schätzer für θ. $\square$