KAPITEL 4

Erwartungstreue Schätzer

Ein statistisches Modell ist ein Tripel $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$, wobei $\mathfrak{X}$ (genannt der Stichprobenraum) die Menge aller möglichen Stichproben, $\mathcal{A}$ eine σ–Algebra auf $\mathfrak{X}$, und $(\mathbb{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf $(\mathfrak{X}, \mathcal{A})$ ist. In diesem Kapitel sei der Parameterraum Θ eine Teilmenge von $\mathbb{R}^r$. Eine Stichprobe X wird gemäß einem Wahrscheinlichkeitsmaß $\mathbb{P}_\theta$ zufällig aus $\mathfrak{X}$ gezogen, wobei $\theta \in \Theta$ unbekannt ist. Unsere Aufgabe besteht darin, θ anhand von X zu schätzen.

Definition 4.0.1. Ein *Schätzer* ist eine beliebige (Borel–messbare) Funktion

$$\hat{\theta} : \mathfrak{X} \to \Theta, \quad x \mapsto \hat{\theta}(x).$$

Bemerkung 4.0.2. Manchmal werden wir erlauben, dass ein Schätzer auch Werte außerhalb von Θ annimmt.

Man möchte nun Schätzer konstruieren, für die $\hat{\theta}(X)$ möglichst "nah" an θ liegt. Dabei ist es ganz natürlich zu fordern, dass der Erwartungswert von $\hat{\theta}(X)$ mit dem zu schätzenden Wert θ übereinstimmen soll. Solche Schätzer heißen erwartungstreu. In diesem Kapitel werden wir versuchen, unter allen erwartungstreuen Schätzern in einem gewissen Sinne "den besten" zu finden.

4.1. Erwartungstreue, Bias, mittlerer quadratischer Fehler

Definition 4.1.1. Ein Schätzer $\hat{\theta}$ heißt *erwartungstreu* (oder *unverzerrt*), falls

$$\mathbb{E}_\theta[\hat{\theta}(X)] = \theta \text{ für alle } \theta \in \Theta.$$

Der *Bias* (die *Verzerrung*) eines Schätzers $\hat{\theta}$ ist

$$\mathrm{Bias}_\theta(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}(X)] - \theta.$$

Wir betrachten $\mathrm{Bias}_\theta(\hat{\theta})$ als eine Funktion von $\theta \in \Theta$.

Bemerkung 4.1.2. Ein Schätzer $\hat{\theta}$ ist genau dann erwartungstreu, wenn $\mathrm{Bias}_\theta(\hat{\theta}) = 0$ für alle $\theta \in \Theta$.

Aufgabe 4.1.3. Zeigen Sie, dass die Menge aller erwartungstreuen Schätzer ein affiner Unterraum des Vektorraumes aller Schätzer ist. D.h. sind $\hat{\theta}_1$ und $\hat{\theta}_2$ erwartungstreu, so ist auch $t\hat{\theta}_1 + (1-t)\hat{\theta}_2$ für alle $t \in \mathbb{R}$ erwartungstreu.

Manchmal möchte man nicht den Parameter θ, sondern eine Funktion $g(\theta)$ schätzen.

Definition 4.1.4. Ein Schätzer φ heißt *erwartungstreu für* $g(\theta)$, falls
$$\mathbb{E}_\theta[\varphi(X)] = g(\theta) \text{ für alle } \theta \in \Theta.$$

Aufgabe 4.1.5. Seien $X_1, \ldots, X_n$ unabhängig und mit Parameter $\theta \in [0,1]$ Bernoulliverteilt. Zeigen Sie, dass es keinen erwartungstreuen Schätzer für $\frac{1}{\theta}$ gibt. Es gibt also statistische Modelle ohne erwartungstreue Schätzer.

Beispiel 4.1.6. In diesem Beispiel werden wir verschiedene Schätzer für den Endpunkt der Gleichverteilung konstruieren. Es seien $X_1, \ldots, X_n \sim \mathrm{U}[0,\theta]$ unabhängige und auf dem Intervall $[0,\theta]$ gleichverteilte Zufallsvariablen, wobei $\theta > 0$ der zu schätzende Parameter sei. Es seien $X_{(1)} < \ldots < X_{(n)}$ die Ordnungsstatistiken von $X_1, \ldots, X_n$.

Erster Schätzer. Zuerst betrachten wir den Maximum–Likelihood–Schätzer
$$\hat{\theta}_1(X_1, \ldots, X_n) = X_{(n)} = \max\{X_1, \ldots, X_n\}.$$
Es ist offensichtlich, dass $\hat{\theta}_1 < \theta$. Somit hat $\hat{\theta}_1$ einen negativen Bias.

Zweiter Schätzer. Wir versuchen nun den Schätzer $\hat{\theta}_1$ zu verbessern, indem wir ihn etwas vergrößern. Wir würden ihn gerne um $\theta - X_{(n)}$ vergrößern, allerdings ist θ unbekannt. Deshalb machen wir den folgenden Ansatz. Wir gehen davon aus, dass die beiden Intervalle $(0, X_{(1)})$ und $(X_{(n)}, \theta)$ ungefähr gleich lang sind, d.h.
$$X_{(1)} \stackrel{!}{=} \theta - X_{(n)}.$$
Lösen wir diese Gleichung bzgl. θ, so erhalten wir den Schätzer
$$\hat{\theta}_2(X_1, \ldots, X_n) = X_{(n)} + X_{(1)}.$$

Dritter Schätzer. Es gibt aber auch einen anderen natürlichen Ansatz. Wir können davon ausgehen, dass die Intervalle
$$(0, X_{(1)}), (X_{(1)}, X_{(2)}), \ldots, (X_{(n)}, \theta)$$
ungefähr gleich lang sind. Dann kann man die Länge des letzten Intervalls durch das arithmetische Mittel der Längen aller vorherigen Intervalle schätzen, was zu folgender Gleichung führt:
$$\theta - X_{(n)} \stackrel{!}{=} \frac{1}{n}(X_{(1)} + (X_{(2)} - X_{(1)}) + (X_{(3)} - X_{(2)}) + \ldots + (X_{(n)} - X_{(n-1)})).$$
Da auf der rechten Seite eine Teleskop–Summe steht, erhalten wir die Gleichung
$$\theta - X_{(n)} \stackrel{!}{=} \frac{1}{n} X_{(n)}.$$
Auf diese Weise ergibt sich der Schätzer
$$\hat{\theta}_3(X_1, \ldots, X_n) = \frac{n+1}{n} X_{(n)}.$$

Vierter Schätzer. Wir können auch den Momentenschätzer betrachten. Setzen wir den Erwartungswert von X_i dem empirischen Mittelwert gleich, so erhalten wir

$$\mathbb{E}_\theta[X_i] = \frac{\theta}{2} \stackrel{!}{=} \bar{X}_n.$$

Dies führt zum Schätzer

$$\hat{\theta}_4(X_1, \ldots, X_n) = 2\bar{X}_n.$$

Aufgabe 4.1.7. Zeigen Sie, dass $\hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4$ erwartungstreu sind, $\hat{\theta}_1$ jedoch nicht.

Man sieht an diesem Beispiel, dass es für ein parametrisches Problem mehrere natürliche (und sogar mehrere erwartungstreue) Schätzer geben kann. Die Frage ist nun, welcher Schätzer der beste ist.

Definition 4.1.8. Sei $\Theta = (a, b) \subset \mathbb{R}$ ein Intervall. Der *mittlere quadratische Fehler* (*mean square error*, MSE) eines Schätzers $\hat{\theta} : \mathfrak{X} \to \mathbb{R}$ ist definiert durch

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

In dieser Definition wird stillschweigend vorausgesetzt, dass $\hat{\theta}$ *quadratisch integrierbar* ist, d.h.

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

Wir bezeichnen mit L^2 die Menge aller quadratisch integrierbaren Schätzer.

Wir fassen $\mathrm{MSE}_\theta(\hat{\theta})$ als eine Funktion von $\theta \in (a, b)$ auf.

Lemma 4.1.9. Es gilt der folgende Zusammenhang zwischen dem mittleren quadratischen Fehler und dem Bias:

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Var}_\theta\, \hat{\theta} + (\mathrm{Bias}_\theta(\hat{\theta}))^2.$$

Beweis. Um die Notation zu vereinfachen, benutzen wir $\hat{\theta}$ als eine Abkürzung für die Zufallsvariable $\hat{\theta}(X)$. Wir benutzen die Definition des mittleren quadratischen Fehlers, erweitern mit $\mathbb{E}_\theta[\hat{\theta}]$ und quadrieren:

$$\begin{aligned}
\mathrm{MSE}_\theta(\hat{\theta}) &= \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \\
&= \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}] + \mathbb{E}_\theta[\hat{\theta}] - \theta)^2] \\
&= \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2] + 2\mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]) \cdot (\mathbb{E}_\theta[\hat{\theta}] - \theta)] + \mathbb{E}_\theta[(\mathbb{E}_\theta[\hat{\theta}] - \theta)^2] \\
&= \mathrm{Var}_\theta(\hat{\theta}) + 2(\mathbb{E}_\theta[\hat{\theta}] - \theta) \cdot \mathbb{E}_\theta[\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]] + (\mathrm{Bias}_\theta(\hat{\theta}))^2.
\end{aligned}$$

Dabei haben wir benutzt, dass $\mathbb{E}_\theta[\hat{\theta}] - \theta$ nicht zufällig ist. Der mittlere Term auf der rechten Seite verschwindet, denn $\mathbb{E}_\theta[\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]] = \mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_\theta[\hat{\theta}] = 0$. Daraus ergibt sich die gewünschte Identität. □

Bemerkung 4.1.10. Ist $\hat{\theta}$ erwartungstreu, so gilt $\mathrm{Bias}_\theta(\hat{\theta}) = 0$ für alle $\theta \in \Theta$ und somit vereinfacht sich Lemma 4.1.9 zu

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}).$$

Bemerkung 4.1.11. Der Bias ist der "systematische Fehler" eines Schätzers, die Standardabweichung $\sqrt{\mathrm{Var}_\theta\,\hat{\theta}}$ kann als der "zufällige Fehler" eines Schätzers angesehen werden. Der mittlere quadratische Fehler MSE berücksichtigt beide Arten von Fehlern.

Mit dem Begriff des mittleren quadratischen Fehlers können wir nun Schätzer vergleichen: je kleiner der Fehler, umso besser der Schätzer.

Definition 4.1.12. Seien $\hat{\theta}_1$ und $\hat{\theta}_2$ zwei Schätzer. Wir sagen, dass $\hat{\theta}_1$ *gleichmäßig besser* als $\hat{\theta}_2$ ist, falls

$$\mathrm{MSE}_\theta(\hat{\theta}_1) \leq \mathrm{MSE}_\theta(\hat{\theta}_2) \text{ für alle } \theta \in \Theta.$$

Bemerkung 4.1.13. Falls $\hat{\theta}_1$ und $\hat{\theta}_2$ erwartungstreu sind, dann ist $\hat{\theta}_1$ gleichmäßig besser als $\hat{\theta}_2$, wenn

$$\mathrm{Var}_\theta(\hat{\theta}_1) \leq \mathrm{Var}_\theta(\hat{\theta}_2) \text{ für alle } \theta \in \Theta.$$

Aufgabe 4.1.14. Es sei X eine Zufallsvariable mit

$$\mathbb{P}_\theta[X = n] = \frac{e^{-\theta}}{1 - e^{-\theta}} \frac{\theta^n}{n!}, \quad n \in \mathbb{N}, \quad \theta > 0.$$

Bestimmen Sie alle erwartungstreuen Schätzer für $g(\theta) = e^{-\theta}$ und den mittleren quadratischen Fehler für jeden solchen Schätzer.

4.2. Bester erwartungstreuer Schätzer

In einem statistischen Modell kann es mehrere erwartungstreue Schätzer geben. Wir versuchen nun, unter diesen Schätzern denjenigen mit der kleinsten Varianz zu finden.

Definition 4.2.1. Ein Schätzer $\hat{\theta}$ heißt *bester erwartungstreuer Schätzer* (für θ), falls er erwartungstreu ist und für jeden anderen erwartungstreuen Schätzer $\tilde{\theta}$ gilt, dass

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

Bemerkung 4.2.2. Der entsprechende englische Begriff lautet "UMVU estimator" (uniformly minimal variance unbiased).

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

Satz 4.2.3. Seien $\hat{\theta}_1, \hat{\theta}_2 : \mathfrak{X} \to \Theta$ zwei beste erwartungstreue Schätzer, dann gilt

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

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

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

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 die Cauchy–Schwarz–Ungleichung in Wirklichkeit eine Gleichheit gewesen sein muss, also

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

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 $a(\theta) = 1$, denn $\operatorname{Var}_\theta \hat{\theta}_1 \neq 0$ gemäß Schritt 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 $\theta \in \Theta$. □

Bemerkung 4.2.4. Der beste erwartungstreue Schätzer muss nicht in jedem parametrischen Modell existieren. Z.B. kann es passieren, dass es überhaupt keine erwartungstreuen Schätzer gibt (Aufgabe 4.1.5).

4.3. Bester erwartungstreuer Schätzer im Bernoulli–Modell

Im Folgenden werden wir für mehrere statistische Modelle den besten erwartungstreuen Schätzer konstruieren. Um unsere Vorgehensweise zu erklären, betrachten wir ein Beispiel, das trotz seiner Einfachheit die beiden wichtigsten Ideen, *Suffizienz* und *Vollständigkeit*, beinhaltet, die man für die allgemeine Konstruktion braucht.

Wir werden den besten erwartungstreuen Schätzer für die Erfolgswahrscheinlichkeit im n-fachen Bernoulli–Experiment bestimmen. Es seien $X_1, \dots, X_n \sim \mathrm{Bern}(\theta)$ unabhängige, mit Parameter $\theta \in [0,1]$ Bernoulli–verteilte Zufallsvariablen. Wir beobachten eine Realisierung $(x_1, \dots, x_n)$ und sollen θ schätzen.

Statistisches Modell. Der Stichprobenraum ist $\mathfrak{X} = \{0,1\}^n$. Als σ–Algebra der messbaren Ereignisse nehmen wir die Potenzmenge $\mathcal{A} = 2^{\mathfrak{X}}$. Die möglichen Verteilungen von $(X_1, \dots, X_n)$ sehen wie folgt aus. Für $\theta \in [0,1]$ ist $\mathbb{P}_\theta$ das Wahrscheinlichkeitsmaß auf $\mathfrak{X}$ mit

$$\mathbb{P}_\theta[A] = \sum_{(x_1,\dots,x_n)\in A} \theta^{x_1+\dots+x_n}(1-\theta)^{n-(x_1+\dots+x_n)}, \quad A \subset \mathfrak{X}.$$

Der folgende Satz sollte nicht überraschend sein.

Satz 4.3.1. Der Schätzer $\hat{\theta}(x_1, \dots, x_n) = \bar{x}_n$ ist der beste erwartungstreue Schätzer von θ im n-fachen Bernoulli–Experiment.

Beweis. Sei $\varphi : \mathfrak{X} \to [0,1]$ ein erwartungstreuer Schätzer von θ. Wir wollen zeigen, dass

$$\mathrm{Var}_\theta\, \varphi \geq \mathrm{Var}_\theta\, \bar{X}_n \text{ für alle } \theta \in [0,1].$$

Erste Idee: Suffizienz. Intuitiv erscheint es plausibel, dass ein "guter" Schätzer nur die Information verwenden sollte, *wieviele* Erfolge in den Bernoulli–Experimenten beobachtet wurden. Es sollte egal sein, *wann* die Erfolge eintgereten sind. So sollte z.B. ein Schätzer φ mit $\varphi(0,0,1,1,1) \neq \varphi(1,0,1,0,1)$ kein guter Schätzer sein.

Wie können wir das beweisen? Für $k = 0, 1, \dots, n$ definieren wir die Mengen

$$A_k := \{x = (x_1, \dots, x_n) \in \{0,1\}^n \colon x_1 + \dots + x_n = k\}.$$

Dann ist $A_0 \cup \dots \cup A_n = \mathfrak{X}$ eine disjunkte Zerlegung von $\mathfrak{X}$. Die Anzahl der Elemente in A_k ist $\binom{n}{k}$. Für jeden Schätzer $\varphi : \mathfrak{X} \to [0,1]$ betrachten wir nun seine *Rao–Blackwell–Verbesserung* $\varphi^* : \mathfrak{X} \to [0,1]$ mit

$$\varphi^*(x) = \frac{1}{\binom{n}{k}} \sum_{y\in A_k} \varphi(y), \text{ falls } x = (x_1, \dots, x_n) \in A_k.$$

Der Schätzer φ^* ist konstant auf jeder der Mengen $A_0, \dots, A_n$, und der Wert von φ^* auf A_k ist einfach der Mittelwert von φ über A_k; siehe Abbildung 1.

Wir behaupten nun, dass φ^* ebenfalls erwartungstreu ist. In der Tat, wegen der Definition von φ^* gilt

$$\mathbb{E}\varphi^* = \frac{1}{2^n}\sum_{x\in\mathfrak{X}} \varphi^*(x) = \frac{1}{2^n}\sum_{k=0}^{n}\sum_{x\in A_k} \varphi^*(x) = \frac{1}{2^n}\sum_{k=0}^{n}\sum_{y\in A_k} \varphi(y) = \mathbb{E}\varphi = \theta.$$

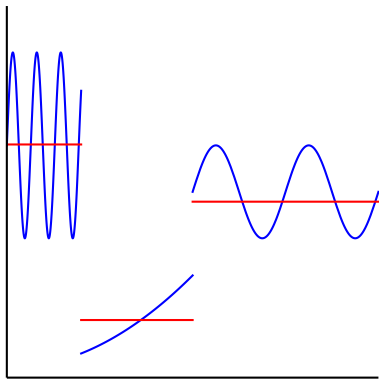

ABBILDUNG 1. Die Idee der Rao–Blackwell–Verbesserung. Eine Funktion (blau) wird durch Mittelwerte (rot) über Mengen aus einer disjunkten Zerlegung ersetzt. Der Erwartungswert bleibt unverändert, die Varianz wird kleiner.

Und nun zeigen wir, dass $\operatorname{Var}_\theta \varphi^* \le \operatorname{Var}_\theta \varphi$. Mit der Ungleichung $\frac{a_1^2+\ldots+a_N^2}{N} \ge \left(\frac{a_1+\ldots+a_N}{N}\right)^2$ (vom arithmetischen und quadratischen Mittel) erhalten wir

$$\sum_{y\in A_k} (\varphi(y)-\theta)^2 = \binom{n}{k} \frac{1}{\binom{n}{k}} \sum_{y\in A_k} (\varphi(y)-\theta)^2 \ge \binom{n}{k} \left(\frac{1}{\binom{n}{k}} \sum_{y\in A_k} (\varphi(y)-\theta) \right)^2 = \sum_{x\in A_k} (\varphi^*(x)-\theta)^2.$$

Da die Wahrscheinlichkeit (unter $\mathbb{P}_\theta$) von jedem Ausgang $y \in A_k$ gleich $\theta^k(1-\theta)^{n-k}$ ist, können wir schreiben:

$$\begin{aligned}\operatorname{Var}_\theta \varphi = \mathbb{E}_\theta[(\varphi-\theta)^2] &= \sum_{k=0}^n \theta^k(1-\theta)^{n-k} \sum_{y\in A_k} (\varphi(y)-\theta)^2 \\ &\ge \sum_{k=0}^n \theta^k(1-\theta)^{n-k} \sum_{x\in A_k} (\varphi^*(x)-\theta)^2 = \mathbb{E}_\theta[(\varphi^*-\theta)^2] = \operatorname{Var}_\theta \varphi^*.\end{aligned}$$

Zweite Idee: Vollständigkeit. Im Rest des Beweises können wir also annehmen, dass φ konstant auf jeder der Mengen $A_0, \ldots, A_n$ bleibt (denn andernfalls können wir φ durch φ^* ersetzen, was den Schätzer verbessert). Außerdem muss φ erwartungstreu sein. Gibt es viele solche Schätzer? Zum Glück gibt es nur einen, nämlich $\bar{X}_n$! Um das zu zeigen, bezeichnen wir den Wert von φ auf A_k mit a_k. Dann lautet die Bedingung der Erwartungstreue wie folgt:

$$\mathbb{E}_\theta \varphi = \sum_{k=0}^n a_k \binom{n}{k} \theta^k(1-\theta)^{n-k} \stackrel{!}{=} \theta \text{ für alle } \theta \in [0,1].$$

Es ist eine (nicht ganz triviale) Übung zu zeigen, dass das nur für $a_k = k/n$ möglich ist. Für die Lösung verweisen wir auf Abschnitt 4.7. □

Im Rest dieses Kapitels werden wir den obigen Beweis auf eine viel größere Klasse von statistischen Modellen erweitern.

4.4. Definition der Suffizienz im diskreten Fall

Beispiel 4.4.1. 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.
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.

Nun werden wir eine allgemeine Definition der Suffizienz geben. Sei $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell. In diesem Abschnitt betrachten wir nur den Fall eines endlichen oder abzählbar unendlichen Stichprobenraumes $\mathfrak{X}$.

Definition 4.4.2. Eine Funktion $T : \mathfrak{X} \to \mathbb{R}^r$ heißt eine *suffiziente Statistik*, wenn für alle $x \in \mathfrak{X}$ und für alle $t \in \mathbb{R}^r$ die Funktion

$$\theta \mapsto \mathbb{P}_\theta[X = x | T(X) = t]$$

konstant ist. D.h. es soll gelten, dass

$$\mathbb{P}_{\theta_1}[X = x | T(X) = t] = \mathbb{P}_{\theta_2}[X = x | T(X) = t]$$

für alle $t \in \mathbb{R}^r$ und alle $\theta_1, \theta_2 \in \Theta$ mit $\mathbb{P}_{\theta_1}[T(X) = t] \neq 0$, $\mathbb{P}_{\theta_2}[T(X) = t] \neq 0$.

Man kann die obige Definition auch wie folgt formulieren: T ist suffizient, wenn die bedingte Verteilung von X gegeben, dass $T(X) = t$ nicht von θ abhängt.

Beispiel 4.4.3 (Fortsetzung von Beispiel 4.4.1)**.** Seien $X_1, \ldots, X_n \sim \text{Bern}(\theta)$ unabhängige Zufallsvariablen, wobei $\theta \in (0, 1)$ zu schätzen sei. Wir behaupten, dass $T(x_1, \ldots, x_n) = x_1 + \ldots + x_n$ eine suffiziente Statistik ist.

Beweis. Sei $t \in \{0, \ldots, n\}$, denn für alle anderen Werte von t ist das Ereignis $T(X) = t$ unmöglich. Betrachte für $(x_1, \ldots, x_n) \in \{0, 1\}^n$ 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 $x_1 + \ldots + x_n \neq t$, so gilt $P(\theta) = 0$. In diesem Fall hängt $P(\theta)$ von θ nicht ab.

FALL 2. Sei nun $x_1 + \ldots + x_n = t$. 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 $X_1 + \ldots + X_n \sim \mathrm{Bin}(n, \theta)$ 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. □

Bemerkung 4.4.4. Wir haben gezeigt, dass für alle $t \in \{0, \ldots, n\}$

$$\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n | X_1 + \ldots + X_n = t] = \frac{\mathbb{1}_{x_1+\ldots+x_n=t}}{\binom{n}{t}}, \quad (x_1, \ldots, x_n) \in \{0,1\}^n.$$

Somit ist die bedingte Verteilung von $(X_1, \ldots, X_n)$ gegeben, dass $X_1 + \ldots + X_n = t$, eine Gleichverteilung auf der Menge

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

Diese Menge besteht aus $\binom{n}{t}$ Elementen. Die bedingte Verteilung hängt nicht von θ ab (Suffizienz).

Aufgabe 4.4.5. Seien $X_1, \ldots, X_n$ unabhängige und

(a) mit Parameter $\theta > 0$ Poisson–verteilte Zufallsvariablen;
(b) mit Parameter $\theta \in (0,1)$ geometrisch–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 $X_1 + \ldots + X_n = t$.

Aufgabe 4.4.6. Seien $X_1, \ldots, X_n$ unabhängige, auf der endlichen Menge $\{1, \ldots, \theta\}$ gleichverteilte Zufallsvariablen, wobei $\theta \in \mathbb{N}$ ein Parameter sei. Zeigen Sie, dass $T(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\}$ eine suffiziente Statistik ist und bestimmen Sie für $t \in \mathbb{N}$ die bedingte Verteilung von $(X_1, \ldots, X_n)$ gegeben, dass $\max\{X_1, \ldots, X_n\} = t$.

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 4.4.7. Sei $T : \mathfrak{X} \to \mathbb{R}^r$ eine suffiziente Statistik und sei $g : \mathrm{Im}\, T \to \mathbb{R}^k$ eine *injektive* Funktion. Dann ist auch

$$g \circ T : \mathfrak{X} \to \mathbb{R}^k, \quad x \mapsto g(T(x))$$

eine suffiziente Statistik.

Beweis. Seien $t \in \mathbb{R}^k$ und $\theta_1, \theta_2 \in \Theta$ mit $\mathbb{P}_{\theta_i}[g(T(X)) = t] \neq 0$, $i = 1, 2$. Wegen der Suffizienz von T ist

$$P(\theta_i) := \mathbb{P}_{\theta_i}[X = x | g(T(X)) = t] = \mathbb{P}_{\theta_i}[X = x | T(X) = g^{-1}(t)]$$

unabhängig von der Wahl von $i = 1, 2$. Dabei ist das Urbild $g^{-1}(t)$ wohldefiniert, da g injektiv ist. □

4.5. 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 $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ ein statistisches Modell mit einer höchstens abzählbaren Stichprobenmenge $\mathfrak{X}$. Sei $T : \mathfrak{X} \to \mathbb{R}^r$ eine Statistik. Im nächsten Lemma benutzen wir die folgende Notation:

- $L(x; \theta) = \mathbb{P}_\theta[X = x]$ ist die Likelihood-Funktion.
- $q(t; \theta) = \mathbb{P}_\theta[T(X) = t]$, wobei $t \in \mathbb{R}^r$, ist die Zähldichte von $T(X)$ unter $\mathbb{P}_\theta$.

Lemma 4.5.1. Eine Funktion $T : \mathfrak{X} \to \mathbb{R}^r$ ist genau dann eine suffiziente Statistik, wenn für alle $x \in \mathfrak{X}$ die Funktion

$$\theta \mapsto \frac{L(x; \theta)}{q(T(x); \theta)} \tag{4.5.1}$$

konstant ist. (Der Definitionsbereich besteht aus allen $\theta \in \Theta$ mit $q(T(x); \theta) \neq 0$).

Beweis. Betrachte den Ausdruck

$$P(\theta) := \mathbb{P}_\theta[X = x | T(X) = t].$$

Im Falle $t \neq T(x)$ ist $P(\theta) = 0$, was unabhängig von θ ist. Sei deshalb $t = T(x)$. Dann gilt

$$P(\theta) = \frac{\mathbb{P}_\theta[X = x, T(X) = t]}{\mathbb{P}_\theta[T(X) = T(x)]} = \frac{\mathbb{P}_\theta[X = x]}{\mathbb{P}_\theta[T(X) = T(x)]} = \frac{L(x; \theta)}{q(T(x); \theta)}.$$

Somit ist T eine suffiziente Statistik genau dann, wenn (4.5.1) nicht von θ abhängt.

Satz 4.5.2 (Faktorisierungssatz von Neyman–Fisher, diskreter Fall)**.** Eine Funktion $T : \mathfrak{X} \to \mathbb{R}^r$ ist eine suffiziente Statistik genau dann, wenn es Funktionen $g : \mathbb{R}^r \times \Theta \to \mathbb{R}$ und $h : \mathfrak{X} \to \mathbb{R}$ gibt, so dass die folgende Faktorisierung gilt:

$$L(x; \theta) = g(T(x); \theta) \cdot h(x) \text{ für alle } x \in \mathfrak{X}, \theta \in \Theta. \tag{4.5.2}$$

Beweis von "$\Longrightarrow$". Sei T eine suffiziente Statistik. Definiere die Funktion

$$h(x) := \frac{L(x; \theta)}{q(T(x); \theta)}, \quad x \in \mathfrak{X}.$$

Dabei können wir auf der rechten Seite ein beliebiges θ mit $q(T(x); \theta) \neq 0$ einsetzen, denn nach Lemma 4.5.1 ist der Term unabhängig von θ. Gibt es kein θ mit $q(T(x); \theta) \neq 0$, so definieren wir einfach $h(x) = 0$.
Mit diesem h und $g(t; \theta) = q(t; \theta)$ gilt die Faktorisierung (4.5.2). □

Beweis von "$\Longleftarrow$". Es gelte die Faktorisierung (4.5.2). Sei $x \in \mathfrak{X}$ fest. Es bezeichne

$$A := \{y \in \mathfrak{X} \colon T(y) = T(x)\}$$

die Niveaumenge von T, die x enthält. Dann gilt für alle $\theta \in \Theta$ mit $q(T(x);\theta) \neq 0$, dass

$$\frac{L(x;\theta)}{q(T(x);\theta)} = \frac{g(T(x);\theta)h(x)}{\sum_{y\in A} L(y;\theta)} = \frac{g(T(x);\theta)h(x)}{\sum_{y\in A} g(T(y);\theta)h(y)} = \frac{h(x)}{\sum_{y\in A} h(y)}.$$

Dieser Ausdruck hängt nicht von θ ab. Nach Lemma 4.5.1 ist T suffizient. □

Beispiel 4.5.3. Seien $X_1, \ldots, X_n \sim \mathrm{Bern}(\theta)$ unabhängig, wobei $\theta \in (0,1)$. Für die Likelihood–Funktion gilt

$$\begin{aligned} L(x_1,\ldots,x_n;\theta) &= \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n] \\ &= \theta^{x_1}(1-\theta)^{1-x_1}\mathbb{1}_{x_1\in\{0,1\}} \cdot \ldots \cdot \theta^{x_n}(1-\theta)^{1-x_n}\mathbb{1}_{x_n\in\{0,1\}} \\ &= \theta^{x_1+\ldots+x_n}(1-\theta)^{n-(x_1+\ldots+x_n)}\mathbb{1}_{x_1,\ldots,x_n\in\{0,1\}}. \end{aligned}$$

Daraus ist ersichtlich, dass die Neyman–Fisher–Faktorisierung (4.5.2) 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.

4.6. Definition der Suffizienz im absolut stetigen Fall

Bisher haben wir nur den Fall eines höchstens abzählbaren Stichprobenraums $\mathfrak{X}$ betrachtet. Sei nun $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell mit einem beliebigen Stichprobenraum $\mathfrak{X}$. Die Funktion $T : \mathfrak{X} \to \mathbb{R}^r$ sei Borel–messbar. Im diskreten Fall haben wir T suffizient genannt, wenn die bedingte Verteilung von X gegeben, dass $T(X) = t$ nicht von θ abhängt. Im Allgemeinen kann die Wahrscheinlichkeit des Ereignisses $T(X) = t$ gleich 0 sein und es ist nicht klar, wie man die bedingte Verteilung definiert. Dieses Problem hat eine Lösung, die im Abschnitt 4.11 besprochen wird.

Definition 4.6.1. Sei $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell. Eine Statistik $T : \mathfrak{X} \to \mathbb{R}^r$ heißt *suffizient*, falls es eine (von θ unabhängige!) Familie von Wahrscheinlichkeitsmaßen $\{\pi_t \colon t \in \mathbb{R}^r\}$ gibt mit

(1) Für jedes $t \in \mathbb{R}^r$ ist π_t ist ein Wahrscheinlichkeitsmaß auf der Niveaumenge

$$T^{-1}(t) = \{x \in \mathfrak{X} \colon T(x) = t\}.$$

(2) Für alle $\theta \in \Theta$ gilt "die Formel der totalen Wahrscheinlichkeit"

$$\mathbb{P}_\theta[A] = \int_{\mathbb{R}^r} \pi_t(A \cap T^{-1}(t))\, \mu_\theta(dt), \quad A \in \mathcal{A},$$

wobei $\mu_\theta(B) = \mathbb{P}_\theta[T \in B]$, $B \subset \mathbb{R}^r$ Borel, die Verteilung von T unter $\mathbb{P}_\theta$ ist.

Bemerkung 4.6.2. Man kann sich π_t als die "bedingte Verteilung" von X gegeben, dass $T(X) = t$, vorstellen. Entscheidend für die Suffizienz ist die Forderung, dass π_t keine Funktion von θ sein darf. Stellen wir uns vor, es wurde eine Stichprobe X gemäß $\mathbb{P}_\theta$ gezogen, uns wurde allerdings lediglich der Wert $T(X)$ mitgeteilt. Wir wissen also, dass X irgendwo in

der Niveaumenge $T^{-1}(t)$ liegt. Die bedingte Verteilung von X ist π_t. Da aber diese Verteilung nicht mehr von θ abhängt, würde uns die Information über die genaue Position von X auf der Niveaumenge nichts nützen. Wir könnten aus dieser Information keine Rückschlüsse auf θ ziehen. Deshalb heißt T suffizient.

Beispiel 4.6.3. Im n-fachen Bernoulli–Model mit der suffizienten Statistik $T(x_1, \ldots, x_n) = x_1 + \ldots + x_n$ ist π_t die Gleichverteilung auf der Menge $\{(x_1, \ldots, x_n) \in \{0,1\}^n \colon x_1 + \ldots + x_n = t\}$, für alle $t \in \{0, \ldots, n\}$. Es ist egal, wie man π_t für andere Werte von t definiert, da diese eine Nullmenge bzgl. μ_θ bilden.

Bemerkung 4.6.4. Es lässt sich zeigen, dass auch die folgende "Formel der totalen Erwartung" gilt:

$$\mathbb{E}_\theta[f(X)] = \int_{\mathbb{R}^r} \left(\int_{T^{-1}(t)} f d\pi_t \right) \mu_\theta(dt),$$

für alle Funktionen $f : \mathfrak{X} \to \mathbb{R}$ mit $\mathbb{E}_\theta |f(X)| < \infty$.

Leider lassen die Bedingungen der obigen Definition nicht so einfach überprüfen. Zum Glück gibt es eine allgemeine Version des Satzes von Neyman–Fisher, die als eine alternative Definition der Suffizienz benutzt werden kann. Wir nehmen an, dass das statistische Modell $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ dominiert ist, d.h. es gibt ein σ–endliches Maß λ auf $\mathfrak{X}$ (typischerweise das Lebesgue– oder das Zählmaß), so dass jedes $\mathbb{P}_\theta$ eine Dichte bezüglich λ besitzt. Diese Dichte wird mit $L(x; \theta)$ bezeichnet und heißt die Likelihood–Funktion.

Satz 4.6.5 (Faktorisierungssatz von Neyman–Fisher, allgemeiner Fall)**.** Eine Statistik $T : \mathfrak{X} \to \mathbb{R}^r$ ist suffizient genau dann, wenn es messbare Funktionen $g : \mathbb{R}^r \times \Theta \to \mathbb{R}$ und $h : \mathfrak{X} \to \mathbb{R}$ gibt mit

$$L(x; \theta) = g(T(x); \theta) \cdot h(x), \text{ für alle } x \in \mathfrak{X}, \theta \in \Theta.$$

Beispiel 4.6.6. Seien $X_1, \ldots, X_n$ unabhängige und auf dem Intervall $[0, \theta]$ gleichverteilte Zufallsvariablen, wobei $\theta > 0$ der unbekannte Parameter sei. Wir zeigen, dass $T(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\}$ eine suffiziente Statistik ist.

Die Dichte von X_i gegeben durch

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

Für die Likelihood–Funktion (also die Dichte von $(X_1, \ldots, X_n)$ bzgl. des Lebesgue–Maßes auf $\mathbb{R}^n$) gilt

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

wobei

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

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

Beispiel 4.6.7. Seien $X_1,\ldots,X_n$ unabhängige und mit Parameter $\theta > 0$ exponentialverteilte Zufallsvariablen. Somit ist die Dichte von X_i gegeben durch

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

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

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

wobei

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

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

Beispiel 4.6.8. Seien $X_1,\ldots,X_n$ 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

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

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 4.6.9. 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 4.6.10. Es seien $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit Dichte h_θ, wobei θ unbekannt sei. 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 die Likelihood–Funktion gilt

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

Diese Funktion ändert sich bei Permutationen von $x_1, \ldots, x_n$ nicht und kann somit als eine Funktion von $(x_{(1)}, \ldots, x_{(n)})$ und θ dargestellt werden. Somit haben wir eine Neyman–Fisher–Faktorisierung angegeben.

4.7. Vollständigkeit

Sei $(\mathbb{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum $(\mathfrak{X}, \mathcal{A})$.

Definition 4.7.1. Eine Stichprobenfunktion $\varphi : \mathfrak{X} \to \mathbb{R}$ heißt *erwartungstreuer Schätzer von* 0, falls

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

Beispiel 4.7.2. Sine $\hat{\theta}_1$ und $\hat{\theta}_2$ beide erwartungstreue Schätzer von θ, so ist ihre Differenz $\hat{\theta}_1 - \hat{\theta}_2$ erwartungstreuer Schätzer von 0.

Definition 4.7.3. Eine Statistik $T : \mathfrak{X} \to \mathbb{R}^r$ heißt *vollständig*, falls für alle Borel–Funktionen $g : \mathbb{R}^r \to \mathbb{R}$ aus der Gültigkeit von

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

folgt, dass $g(T) = 0$ 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 4.7.4. 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 4.7.5. 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 4.7.6. 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

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

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

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

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

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

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 $[0, \theta]$. □

4.8. Eine Charakterisierung des besten erwartungstreuen Schätzers

Sei $(\mathbb{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum $(\mathfrak{X}, \mathcal{A})$, wobei $\Theta \subset \mathbb{R}$. [1]

Satz 4.8.1. Sei $\hat{\theta} : \mathfrak{X} \to \mathbb{R}$ ein erwartungstreuer Schätzer für θ. Dann sind die folgenden Bedingungen äquivalent:

(1) $\hat{\theta}$ ist der beste erwartungstreue Schätzer für θ.
(2) Für jeden (quadratisch integrierbaren) erwartungstreuen Schätzer φ für 0 gilt, dass $\mathrm{Cov}_\theta(\hat{\theta}, \varphi) = 0$ 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

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

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

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

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

Beweis von "$\Longleftarrow$". Sei $\hat{\theta}$ ein erwartungstreuer Schätzer für θ. Sei außerdem $\mathrm{Cov}_\theta(\varphi, \hat{\theta}) = 0$ für alle erwartungstreuen Schätzer φ für 0. Jetzt werden wir zeigen, dass $\hat{\theta}$ der beste

[1] Die Ergebnisse dieses Abschnitts werden im Folgenden nicht verwendet.

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. □

Aufgabe 4.8.2. Seien $\hat{\nu}_1, \ldots, \hat{\nu}_k$ beste erwartungstreue Schätzer für die Funktionen $\nu_1(\theta), \ldots, \nu_k(\theta)$. Zeigen Sie, dass für beliebige Konstanten $c_1, \ldots, c_k \in \mathbb{R}$ der beste erwartungstreue Schätzer für $c_1\nu_1(\theta) + \ldots + c_k\nu_k(\theta)$ durch $c_1\hat{\nu}_1 + \ldots + c_k\hat{\nu}_k$ gegeben ist.

4.9. Exponentialfamilien

In diesem Abschnitt führen wir den Begriff der Exponentialfamilie ein. Dieser Begriff ist aus mindestens zwei Gründen sehr nützlich. 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 4.9.1. 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 4.9.2. Betrachten wir die Familie der Binomialverteilungen mit Parametern n (bekannt) und $\theta \in (0, 1)$ (unbekannt). Für $x \in \{0, \ldots, n\}$ 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 4.9.3. 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

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

Aufgabe 4.9.4. Zeigen Sie, dass folgende Familien von Verteilungen Exponentialfamilien sind:

(1) $\{\mathrm{Exp}(\theta)\colon \theta > 0\}$.
(2) $\{\mathrm{Poi}(\theta)\colon \theta > 0\}$.

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 4.9.5. 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 4.9.6. 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, die später eingeführt werden.

4.10. Vollständige und suffiziente Statistik für Exponentialfamilien

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

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

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

$$\begin{aligned} L(x_1, \dots, x_n; \theta) &= h_\theta(x_1) \dots h_\theta(x_n) \\ &= (a(\theta))^n b(x_1) \dots b(x_n) \exp\left(\sum_{i=1}^{m} c_i(\theta) d_i(x_1)\right) \dots \exp\left(\sum_{i=1}^{m} c_i(\theta) d_i(x_n)\right) \\ &= (a(\theta))^n b(x_1) \dots b(x_n) \exp\left(T_1 c_1(\theta) + \dots + T_m c_m(\theta)\right). \end{aligned}$$

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

$$h(x_1, \dots, x_n) = b(x_1) \dots b(x_n), \quad g(T_1, \dots, T_m; \theta) = (a(\theta))^n \exp\left(T_1 c_1(\theta) + \dots + T_m c_m(\theta)\right).$$

Man kann zeigen, dass diese Statistik auch vollständig ist, wenn die Menge

$$\{(c_1(\theta), \dots, c_m(\theta)) \colon \theta \in \Theta\} \subset \mathbb{R}^m$$

mindestens einen m-dimensionalen Ball enthält (ohne Beweis).

Beispiel 4.10.1. 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.

4.11. Bedingter Erwartungswert und bedingte Wahrscheinlichkeiten

In diesem Abschnitt werden wir eine allgemeine Definition der bedingten Wahrscheinlichkeiten und Erwartungswerte vorstellen.

Elementare bedingte Wahrscheinlichkeiten und Erwartungswerte. Zuerst erinnern wir uns an die Definition, die bereits mehrmals benutzt wurde.

Definition 4.11.1. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Seien $A \in \mathcal{A}$ und $B \in \mathcal{A}$ zwei Ereignisse mit $\mathbb{P}[B] \neq 0$. 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 ergibt nur dann Sinn, wenn $\mathbb{P}[B] \neq 0$. Analog kann man den bedingten Erwartungswert definieren.

Definition 4.11.2. 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 mit $\mathbb{P}[B] \neq 0$. 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 ergibt nur dann Sinn, wenn $\mathbb{P}[B] \neq 0$. Bedingte Wahrscheinlichkeiten können als Spezialfall des bedingten Erwartungswerts angesehen werden, denn

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

Wir haben gesehen (z.B. bei der Definition der Suffizienz im absolut stetigen Fall), dass man bedingte Wahrscheinlichkeiten oder Erwartungswerte oft 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.

Bedingter Erwartungswert gegeben eine σ-Algebra. Sei $X : \Omega \to \mathbb{R}$ eine auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ definierte Zufallsvariable. Wir nehmen an, dass X integrierbar ist, d.h. $\mathbb{E}|X| < \infty$. 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 $\mathbb{E}|X| < \infty$. 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 4.11.3. 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 4.11.4. 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.

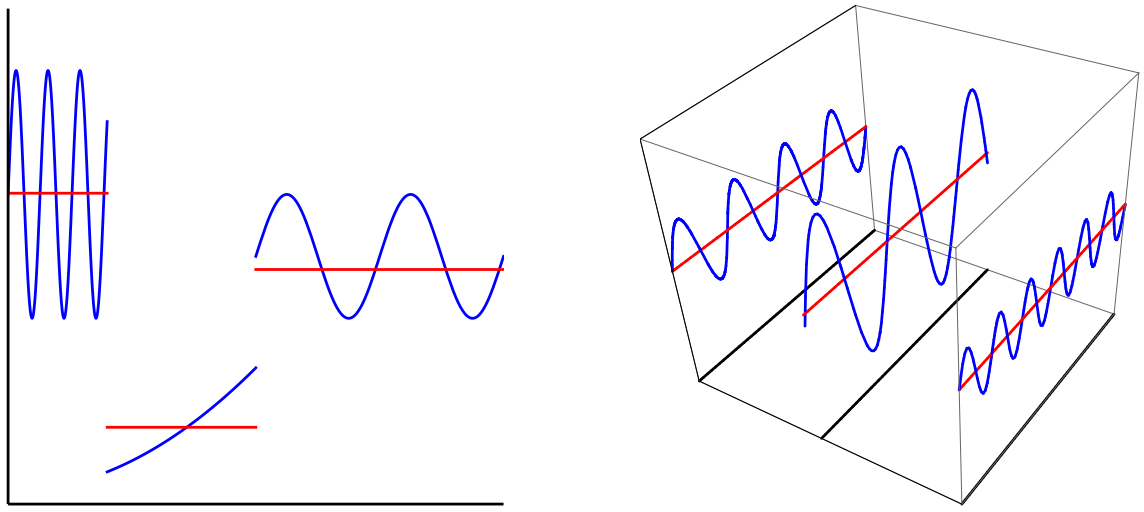

ABBILDUNG 2. Bedingter Erwartungswert in Beispiel 4.11.5 (links) und Beispiel 4.11.6 (rechts). Blau: Der Graph von X. Rot: Der bedingte Erwartungswert.

Beispiel 4.11.5. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Betrachte eine disjunkte abzählbare Zerlegung $\Omega = \cup_{n\in\mathbb{N}}\Omega_n$, wobei $\Omega_n \in \mathcal{A}$ und $\mathbb{P}[\Omega_n] \neq 0$. (Wir betrachten hier eine unendliche Zerlegung, der Fall einer endlichen Zerlegung ist völlig analog). Sei $\mathcal{B}$ die σ–Algebra, die von

der Familie $\{\Omega_1, \Omega_2, \ldots\}$ erzeugt wird. Somit ist

$$\mathcal{B} = \left\{ \bigcup_{n\in\mathbb{N}} \Omega_n^{\varepsilon_n} : \varepsilon_1, \varepsilon_2, \ldots \in \{0,1\} \right\},$$

wobei $\Omega_n^1 = \Omega_n$ und $\Omega_n^0 = \varnothing$. 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:

$$\mathbb{E}[X|\mathcal{B}](\omega) = \frac{\mathbb{E}[X\mathbb{1}_{\Omega_n}]}{\mathbb{P}[\Omega_n]}, \quad \text{falls } \omega \in \Omega_n.$$

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

$$\mathbb{E}[X\mathbb{1}_{\Omega_n}] = \mathbb{E}[Z\mathbb{1}_{\Omega_n}] = c_n\mathbb{P}[\Omega_n].$$

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

Beispiel 4.11.6. 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:

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

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 4.11.7. 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) $\mathbb{E}[X|\{\Omega, \varnothing\}] = \mathbb{E}X$.
(2) $\mathbb{E}[X|\mathcal{A}] = X$.

Beweis. Übung. □

Satz 4.11.8. Seien $X, Y : \Omega \to \mathbb{R}$ 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. □

Bedingter Erwartungswert gegeben eine Zufallsvariable. Besonders oft wird die Definition der bedingten Erwartung im folgenden Spezialfall benutzt.

Definition 4.11.9. Sei Y eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Die *von Y erzeugte σ–Algebra* ist definiert durch

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

Man kann sich die σ–Algebra $\sigma(Y)$ folgendermaßen vorstellen. Zunächst einmal liegen alle Niveaumengen der Form $Y^{-1}(t) := \{\omega \in \Omega \colon Y(\omega) = t\}$ in $\sigma(Y)$, für alle $t \in \mathbb{R}$. Dabei ist Ω eine disjunkte Vereinigung dieser Niveaumengen: $\Omega = \cup_{t\in\mathbb{R}} Y^{-1}(t)$. Außerdem beinhaltet $\sigma(Y)$ alle Vereinigungen der Niveaumengen der Form $\cup_{t\in C} Y^{-1}(t)$, wobei $C \subset \mathbb{R}$ eine beliebige Borel–Menge ist.

Definition 4.11.10. Seien X und Y zwei $\mathcal{A}$–messbare Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Sei X integrierbar. Der bedingte Erwartungswert von X gegeben Y ist definiert durch

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

Bemerkung 4.11.11. $\mathbb{E}[X|Y]$ ist eine Zufallsvariable! Aus der Messbarkeit von $\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)]$ bzgl. $\sigma(Y)$ (s. Definition 4.11.3) kann man herleiten, dass $\mathbb{E}[X|Y]$ eine Borel–Funktion von Y sein muss (s. das Faktorisierungslemma 4.11.16 im nächsten Abschnitt). Es gibt also eine Borel–Funktion $\varphi : \mathbb{R} \to \mathbb{R}$ mit

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

D.h. der bedingte Erwartungswert $\mathbb{E}[X|Y]$ bleibt konstant auf jeder Niveaumenge $\Omega_t = \{\omega \in \Omega \colon Y(\omega) = t\}$, für alle $t \in \mathbb{R}$:

$$\mathbb{E}[X|Y](\omega) = \varphi(t) \text{ falls } Y(\omega) = t.$$

Dabei ist Ω eine disjunkte Vereinigung dieser Niveaumengen: $\Omega = \cup_{t\in\mathbb{R}} \Omega_t$. Den Wert von $\mathbb{E}[X|Y]$ auf einer Niveaumenge Ω_t kann man sich als den "Mittelwert" von X über Ω_t vorstellen, vgl. Beispiele 4.11.5 und 4.11.6.

Definition 4.11.12. Wir definieren den bedingten Erwartungswert von X gegeben, dass $Y = t$ ist, durch

$$\mathbb{E}[X|Y = t] = \mathbb{E}[X|Y](\omega) = \varphi(t), \quad t \in \mathbb{R},$$

wobei $\omega \in \Omega$ ein beliebiges Element mit $Y(\omega) = t$ sei.

Dabei darf $\mathbb{P}[Y = t]$ auch 0 sein! Wir müssen hier allerdings die Frage der Eindeutigkeit klären. Die Zufallsvariable $\mathbb{E}[X|Y]$ ist nur bis auf $\mathbb{P}$–Nullmengen eindeutig definiert. Inwieweit ist die Funktion φ eindeutig? Sei μ_Y die Verteilung von Y, d.h. μ_Y ist ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ mit $\mu_Y(A) = \mathbb{P}[Y \in A]$. Verändern wir nun φ auf einer μ_Y–Nullmenge, so verändert sich $\varphi(Y)$ auf einer $\mathbb{P}$–Nullmenge. Somit ist die Funktion φ nur bis auf μ_Y–Nullmengen eindeutig definiert. Es folgt, dass auch die Borel–Funktion $t \mapsto \mathbb{E}[X|Y = t]$ bis auf Nullmengen von μ_Y eindeutig definiert ist. Für einen vorgegebenen Wert $t \in \mathbb{R}$ kann man also in der Regel leider nicht sagen, was $\mathbb{E}[X|Y = t]$ ist! Man muss die Funktion $t \mapsto \mathbb{E}[X|Y = t]$ immer als Ganzes betrachten.

Nun können wir auch bedingte Wahrscheinlichkeuten als Spezialfall des bedingten Erwartungswerts definieren.

Definition 4.11.13. Sei $Y : \Omega \to \mathbb{R}$ eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$. Für ein Ereignis $A \in \mathcal{A}$ definieren wir die die bedingte Wahrscheinlichkeit von A gegeben, dass $Y = t$, durch

$$\mathbb{P}[A|Y = t] := \mathbb{E}[\mathbb{1}_A|Y](\omega), \quad t \in \mathbb{R},$$

wobei $\omega \in \Omega$ beliebig mit $Y(\omega) = t$ ist.

Beispiel 4.11.14. Sei Y eine diskrete Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$. Das Bild von Y, also die Menge

$$\operatorname{Im} Y = \{t \in \mathbb{R} \colon \mathbb{P}[Y = t] > 0\}$$

ist somit höchstens abzählbar. Die von Y erzeugte σ–Algebra $\sigma(Y)$ wird von den Mengen $\Omega_t = \{Y = t\}$, $t \in \operatorname{Im} Y$, erzeugt und hat somit die gleiche Gestalt wie in Beispiel 4.11.5. Für der bedingten Erwartungswert einer integrierbaren Zufallsvariable $X : \Omega \to \mathbb{R}$ gilt somit

$$\mathbb{E}[X|Y = t] = \mathbb{E}[X|Y](\omega) = \frac{\mathbb{E}[X\,\mathbb{1}_{\Omega_t}]}{\mathbb{P}[\Omega_t]},$$

wobei $\omega \in \Omega_t$ beliebig ist.
Seien nun X und Y beide diskret mit gemeinsamer Zähldichte $f_{X,Y}(s,t) = \mathbb{P}[X = s, Y = t]$ und die Zähldichte von Y sei $f_Y(t) = \mathbb{P}[Y = t]$. Dann gilt für den bedingten Erwartungswert

$$\mathbb{E}[X|Y = t] = \frac{\sum_{s \in \operatorname{Im} X} \mathbb{P}[X = s \cap Y = t]s}{\mathbb{P}[Y = t]} = \frac{\sum_{s \in \operatorname{Im} X} f_{X,Y}(s,t)s}{f_Y(t)}.$$

Beispiel 4.11.15. Seien X, Y 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:

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

wobei $\omega \in \Omega$ beliebig mit $Y(\omega) = t$ ist. Diese Formel ergibt Sinn, wenn $f_Y(t) \neq 0$.

Faktorisierungslemma. Hier beweisen wir die im vorherigen Abschnitt angekündigte Aussage. Es ist klar, dass eine Zufallsvariable der Form $\varphi(Y)$, wobei $\varphi : \mathbb{R} \to \mathbb{R}$ eine Borel–Funktion ist, $\sigma(Y)$–messbar ist. Wir beweisen nun, dass jede $\sigma(Y)$–messbare Zufallsvariable von dieser Form ist.

Lemma 4.11.16 (Faktorisierungslemma)**.** Sei $Y : \Omega \to \mathbb{R}$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Ist $Z : \Omega \to \mathbb{R}$ eine $\sigma(Y)$–messbare Zufallsvariable, so gibt es eine Borel–Funktion $\varphi : \mathbb{R} \to \mathbb{R}$ mit $Z = \varphi(Y)$.

Beweis. Schritt 1. Sei zuerst $Z \geq 0$. Wir zeigen, dass es Mengen $A_1, A_2, \ldots \in \sigma(Y)$ und Konstanten $\alpha_1, \alpha_2, \ldots \geq 0$ gibt mit

$$Z = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{A_n}.$$

Definiere $Z_n := 2^{-n} \lfloor 2^n Z \rfloor \wedge n$, wobei $\wedge$ für das Minimum steht. Dann gilt $Z_n \uparrow Z$ punktweise und Z_n nimmt Werte in der Menge $\{0, \frac{1}{2^n}, \frac{2}{2^n}, \ldots, \frac{n2^n}{2^n}\}$ an. Definiere $\sigma(Y)$–messbare Mengen

$$B_{n,i} = \{\omega \in \Omega \colon Z_n(\omega) - Z_{n-1}(\omega) = i2^{-n}\}, \quad n \in \mathbb{N}, \;\; i = 1, 2, \ldots, 2^n.$$

Es gilt $\cup_{i=1}^{2^n} B_{n,i} = \Omega$ und somit $Z_n - Z_{n-1} = \sum_{i=1}^{2^n} \frac{i}{2^n} \mathbb{1}_{B_{n,i}}$. Wir können nun Z als eine Teleskop–Summe schreiben:

$$Z = \sum_{n=1}^{\infty} (Z_n - Z_{n-1}) = \sum_{n=1}^{\infty} \sum_{i=1}^{2^n} \frac{i}{2^n} \mathbb{1}_{B_{n,i}},$$

denn $Z_0 = 0$. Nach einer Umnummerierung der Mengen $B_{n,i}$ und der Konstanten $\frac{i}{2^n}$ erhalten wir die gesuchte Darstellung.

Schritt 2. Sei nach wie vor $Z \geq 0$ mit der Darstellung $Z = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{A_n}$ wie in Schritt 1. Nach Definition der σ–Algebra $\sigma(Y)$ gibt es zu jeder Menge $A_n \in \sigma(Y)$ eine Borel–Menge $B_n \subset \mathbb{R}$ mit $A_n = Y^{-1}(B_n)$ und somit

$$\mathbb{1}_{A_n}(\omega) = \mathbb{1}_{B_n}(Y(\omega)).$$

Definiere nun die Funktion $\varphi(t) = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{B_n}$. Dann gilt

$$Z(\omega) = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{A_n}(\omega) = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{B_n}(Y(\omega)) = \varphi(Y(\omega)),$$

was die gesuchte Darstellung von Z ist.

Schritt 3. Sei nun Z nicht unbedingt nicht-negativ. Dann gibt es eine Darstellung $Z = Z_+ - Z_-$, wobei $Z_+ = \max\{Z, 0\}$ und $Z_- = \max\{-Z, 0\}$ ebenfalls $\sigma(Y)$–messbar und nicht-negativ sind. Nach Schritt 2 gibt es Darstellungen $Z_+ = \varphi_+(Y)$, $Z_- = \varphi_-(Y)$ für geeignete

Borel–Funktionen φ_+ und φ_-. Es folgt, dass $Z = \varphi_+(Y) - \varphi_-(Y) = \varphi(Y)$ mit $\varphi = \varphi_+ - \varphi_-$. □

Markow–Kerne und reguläre bedingte Wahrscheinlichkeiten. Sei $Y : \Omega \to \mathbb{R}$ eine Zufallsvariable. Für jedes Ereignis $A \in \mathcal{A}$ ist die Borel–Funktion $t \mapsto \mathbb{P}[A|Y = t]$ bis auf eine μ_Y–Nullmenge eindeutig definiert. Stellen wir uns nun vor, dass wir uns für jedes Ereignis A auf irgendeine Version dieser Funktion geeinigt haben. Man kann zeigen, dass folgende Eigenschaften gelten:

(1) $\mathbb{P}[\Omega|Y = t] = 1$ für μ_Y-fast alle $t \in \mathbb{R}$.
(2) Für jede disjunkte Familie von Ereignissen $A_1, A_2, \ldots \in \mathcal{A}$ gilt

$$\mathbb{P}\left[\cup_{n=1}^{\infty} A_n | Y = t\right] = \sum_{n=1}^{\infty} \mathbb{P}[A_n|Y = t] \quad \text{für } \mu_Y\text{–fast alle } t \in \mathbb{R}.$$

Naiv könnte man nun glauben, dass für jedes $t \in \mathbb{R}$ die Zuordnung $A \mapsto \mathbb{P}[A|Y = t]$ ein Wahrscheinlichkeitsmaß auf der Niveaumenge $\{Y = t\}$ definiert (und das man sich sozusagen als "Einschränkung" von $\mathbb{P}$ auf die Niveaumenge vorstellen könnte). Leider gibt es hier ein Problem: alle Relationen gelten lediglich für μ_Y–fast alle $t \in \mathbb{R}$. Noch schlimmer, die Ausnahmemenge derjenigen t, für die die zweite Relation nicht gilt, hängt von $A_1, A_2, \ldots$ ab. Im schlimmsten Fall kann es passieren, dass man für jedes t eine Familie $A_1, A_2, \ldots$ finden kann, für die die σ–Additivität falsch ist. Glücklicherweise kann man dieses Problem durch eine geschickte Wahl von Versionen der Funktionen $t \mapsto \mathbb{P}[A|Y = t]$ vermeiden.

Definition 4.11.17. Ein *Markow–Kern* von $(\Omega, \mathcal{A})$ nach $(\Omega', \mathcal{A}')$ ist eine Familie $\{\pi_\omega : \omega \in \Omega\}$ mit den folgenden zwei Eigenschaften:

- Für jedes $\omega \in \Omega$ ist π_ω ein Wahrscheinlichkeitsmaß auf $(\Omega', \mathcal{A}')$.
- Für alle $A' \in \mathcal{A}'$ ist die Abbildung $\omega \mapsto \pi_\omega(A')$ eine Borel–Funktion auf $(\Omega, \mathcal{A})$.

Beispiel 4.11.18. Sei E eine höchstens abzählbare Menge und $p : E \times E \to [0, 1]$ eine Übergangswahrscheinlichkeit, d.h. $p(i, j) \geq 0$ und $\sum_{j \in E} p(i, j) = 1$ für alle $i \in E$. Dann definiert

$$\pi_i(A) = \sum_{j \in A} p(i, j)$$

einen Markow–Kern von $(E, 2^E)$ nach sich selbst.

Definition 4.11.19. Sei $Y : \Omega \to \mathbb{R}$ eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$. Ein Markow–Kern $\{\pi_t : t \in \mathbb{R}\}$ von $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ nach $(\Omega, \mathcal{A})$ heißt *reguläre Version* von $\mathbb{P}$ gegeben Y, wenn

(a) Das Wahrscheinlichkeitsmaß π_t ist auf der Niveaumenge $\Omega_t = \{Y = t\}$ konzentriert, d.h. $\pi_t(\Omega_t) = 1$ für alle $t \in \mathbb{R}$.

(b) Für jedes Ereignis $A \in \mathcal{A}$ gilt die "Formel der totalen Wahrscheinlichkeit"

$$\mathbb{P}[A] = \int_{\mathbb{R}} \pi_t(A) \mu_Y(dt).$$

Eine reguläre Version von $\mathbb{P}$ existiert unter sehr allgemeinen Voraussetzungen an den Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$.

Definition 4.11.20. Zwei Messräume $(\Omega, \mathcal{A})$ und $(\Omega', \mathcal{A}')$ heißen *Borel–isomorph*, falls es eine bijektive Abbilding $T : \Omega \to \Omega'$ gibt mit

$$A \in \mathcal{A} \Longleftrightarrow T(A) \in \mathcal{A}'.$$

Man kann also die beiden Räume aufeinander bijektiv abbilden, sodass messbare Mengen auf messbare Mengen abgebildet werden.

Definition 4.11.21. Ein Messraum $(\Omega, \mathcal{A})$ heißt *standard Borel*, falls er zu einem der folgenden Messräume isomorph ist:

- $(E, 2^E)$, wobei E eine höchstens abzählbare Menge ist.
- Das Intervall $[0, 1]$ mit der Borel–σ–Algebra.

Satz 4.11.22. Sei (M, ρ) ein vollständiger separabler metrischer Raum (z.B. ein kompakter metrischer Raum). Es sei $\mathcal{B}(M)$ die Borel–σ–Algebra auf M, also die von allen offenen Mengen erzeugte σ–Algebra. Dann ist der Raum $(M, \mathcal{B}(M))$ standard Borel.

Ohne Beweis.

Beispiel 4.11.23. Die folgenden metrischen Räume mit ihren jeweiligen Borel–σ–Algebren sind allesamt isomorph zum Intervall $[0, 1]$ mit der Borel–σ–Algebra:

(a) Der Euklidische Raum $\mathbb{R}^n$.
(b) Der unendlich-dimensionale Hilbert–Raum $L^2[0, 1]$.
(c) Der Raum der stetigen Funktionen $C[0, 1]$ mit der Supremumsmetrik.

Die meisten Meßräume, die man in der Stochastik verwendet, sind standard Borel–Räume. Nun können wir endlich den Satz über die Existenz der regulären Version der bedingten Verteilung formulieren.

Satz 4.11.24. Sei $Y : \Omega \to \mathbb{R}$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ derart, dass $(\Omega, \mathcal{A})$ standard Borel ist. Dann existiert eine reguläre Version von $\mathbb{P}$ gegeben Y, s. Definition 4.11.19.

Ohne Beweis.

4.12. Satz von Rao–Blackwell

Eine suffiziente Statistik beinhaltet alles, was man über das Ergebnis eines statistischen Experiments wissen muss. Es ist deshalb plausibel, dass ein "guter" Schätzer eine Funktion der suffizienten Statistik sein muss. Der nächste Satz bestätigt diese Vermutung.

Satz 4.12.1 (Rao–Blackwell)**.** Sei $(\mathbb{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum $(\mathfrak{X}, \mathcal{A})$, wobei $\Theta \subset \mathbb{R}$ ein Intervall sei. Es seien weiterhin

- $T : \mathfrak{X} \to \mathbb{R}^m$ eine suffiziente Statistik und
- $\hat\theta : \mathfrak{X} \to \mathbb{R}$ ein erwartungstreuer Schätzer von θ mit $\mathbb{E}_\theta\hat\theta^2 < \infty$ für alle $\theta \in \Theta$.

Definiere $\tilde\theta := \mathbb{E}_\theta[\hat\theta|T]$. Dann ist $\tilde\theta$ ebenfalls ein erwartungstreuer Schätzer von θ und es gilt

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

Der Schätzer $\tilde\theta$ ist somit mindestens so gut wie $\hat\theta$ und heißt aus diesem Grunde die *Rao–Blackwell–Verbesserung* von $\hat\theta$. Da $\tilde\theta$ als bedingter Erwartungswert $\sigma(T)$–messbar ist, kann man nach dem Faktorisierungslemma 4.11.16 $\tilde\theta$ als eine Borel–Funktion von T darstellen. Somit basiert $\tilde\theta$ nur auf dem Wert der suffizienten Statistik T.

Beweis. Schritt 1. Zuallererst müssen wir zeigen, dass $\tilde\theta = \mathbb{E}_\theta[\hat\theta|T]$ keine Funktion von θ ist. (Das darf nämlich ein Schätzer auf keinen Fall sein!) Wir schreiben den bedingten Erwartungswert an der Stelle $x \in \mathfrak{X}$ als das Integral bzgl. der bedingten Verteilung $\mathbb{P}_\theta[\cdot|T = T(x)]$:

$$\tilde\theta(x) = \mathbb{E}_\theta[\hat\theta|T](x) = \mathbb{E}_\theta[\hat\theta|T = T(x)] = \int_{\mathfrak{X}} \hat\theta(y)\mathbb{P}_\theta[dy|T = T(x)].$$

Da T suffizient ist, hängt das Wahrscheinlichkeitsmaß $A \mapsto \mathbb{P}_\theta[A|T = T(x)]$ nicht von θ ab! Somit hängt das Integral auf der rechten Seite nicht von θ ab.

Schritt 2. Nun zeigen wir, dass $\tilde\theta$ erwartungstreu ist. Mit der Turmeigenschaft des bedingten Erwartungswerts gilt

$$\mathbb{E}_\theta\tilde\theta = \mathbb{E}_\theta\mathbb{E}_\theta[\hat\theta|T] = \mathbb{E}_\theta\hat\theta = \theta,$$

denn $\hat\theta$ ist erwartungstreu.

Schritt 3. Für die Varianz von $\tilde\theta$ gilt

$$\operatorname{Var}_\theta \tilde\theta = \mathbb{E}_\theta[(\tilde\theta - \theta)^2] = \mathbb{E}_\theta\left[(\mathbb{E}_\theta[\hat\theta|T] - \theta)^2\right] = \mathbb{E}_\theta\left[(\mathbb{E}_\theta[\hat\theta - \theta|T])^2\right].$$

Die Jensen–Ungleichung für den bedingten Erwartungswert besagt, dass $\varphi(\mathbb{E}[X|\mathcal{F}]) \le \mathbb{E}[\varphi(X)|\mathcal{F}]$ f.s., falls φ konvex ist. Mit dieser Ungleichung für $\varphi(x) = x^2$ ergibt sich

$$\mathbb{E}_\theta\left[(\mathbb{E}_\theta[\hat\theta - \theta|T])^2\right] \le \mathbb{E}_\theta\mathbb{E}_\theta\left[(\hat\theta - \theta)^2|T\right] = \mathbb{E}_\theta[(\hat\theta - \theta)^2] = \operatorname{Var}_\theta \hat\theta,$$

was die behauptete Ungleichung $\operatorname{Var}_\theta \tilde\theta \le \operatorname{Var}_\theta \hat\theta$ beweist. □

Beispiel 4.12.2. Seien $X_1, \ldots, X_n$ unabhängig und Bernoulli–verteilt mit Parameter $\theta \in (0, 1)$. Die Statistik $T(x_1, \ldots, x_n) = x_1 + \ldots + x_n$ ist suffizient. Als einen sehr einfachen

erwartungstreuen Schätzer von θ betrachten wir $\hat{\theta} = X_1$. Für die Varianz dieses Schätzers gilt $\operatorname{Var}_\theta \hat{\theta} = \theta(1-\theta)$. Nun definieren wir die Rao–Blackwell–Verbesserung von $\hat{\theta}$:

$$\tilde{\theta} = \mathbb{E}_\theta[X_1 | X_1 + \ldots + X_n].$$

Diesen bedingten Erwartungswert kann man direkt mir der Definition berechnen, es gibt allerdings eine viel elegantere Methode. Wegen Symmetrie gilt

$$\mathbb{E}_\theta[X_1 | X_1 + \ldots + X_n] = \mathbb{E}_\theta[X_2 | X_1 + \ldots + X_n] = \ldots = \mathbb{E}_\theta[X_n | X_1 + \ldots + X_n].$$

(Man erwartet im ersten Wurf genauso viel, wie im zweiten, auch dann, wenn die Summe $X_1 + \ldots + X_n$ gegeben ist). Es folgt, dass

$$\begin{aligned}\tilde{\theta} = \mathbb{E}_\theta[X_1 | X_1 + \ldots + X_n] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta[X_i | X_1 + \ldots + X_n] \\ &= \frac{1}{n} \mathbb{E}_\theta[X_1 + \ldots + X_n | X_1 + \ldots + X_n] = \frac{1}{n}(X_1 + \ldots + X_n) = \bar{X}_n.\end{aligned}$$

Für die Varianz von $\tilde{\theta} = \bar{X}_n$ gilt $\operatorname{Var}_\theta \bar{X}_n = \frac{\theta(1-\theta)}{n}$, eine klare Verbesserung im Vergleich zu $\hat{\theta} = X_1$ (es sei denn $n = 1$, in welchem Fall beide Schätzer übereinstimmen).

4.13. Satz von Lehmann–Scheffé

Der folgende Satz wird uns in vielen Beispielen erlauben, den besten erwartungstreuen Schätzer zu konstruieren.

Satz 4.13.1 (Lehmann–Scheffé)**.** Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum $(\mathfrak{X}, \mathcal{A})$, wobei $\Theta \subset \mathbb{R}$ ein Intervall sei. Es seien weiterhin

- $T : \mathfrak{X} \to \mathbb{R}^m$ eine suffiziente und vollständige Statistik;
- $\hat{\theta} : \mathfrak{X} \to \mathbb{R}$ ein erwartungstreuer Schätzer von θ mit $\mathbb{E}_\theta \hat{\theta}^2 < \infty$ für alle $\theta \in \Theta$.

Dann ist $\tilde{\theta} := \mathbb{E}_\theta[\hat{\theta} | T]$ der beste erwartungstreue Schätzer von θ.

Beweis. Im Satz von Rao–Blackwell wurde gezeigt, dass $\tilde{\theta}$ wohldefiniert (d.h. keine Funktion von θ) und erwartungstreu ist.

Sei H ein weiterer erwartungstreuer Schätzer von θ mit $\mathbb{E}_\theta H^2 < \infty$ für alle $\theta \in \Theta$. Zu zeigen ist, dass $\tilde{\theta}$ besser ist, als H. Wir betrachten den Schätzer $\tilde{H} := \mathbb{E}_\theta[H | T]$. Nach dem Satz von Rao–Blackwell ist $\tilde{H}$ besser als H. Wir zeigen nun, dass $\tilde{H}$ mit $\tilde{\theta}$ übereinstimmt. Beide Schätzer sind $\sigma(T)$–messbar, da sie als bedingte Erwartungswerte gegeben T definiert sind. Nach dem Faktorisierungslemma 4.11.16 gibt es zwei Borel–Funktionen f und g mit $\tilde{\theta} = f(T)$ und $\tilde{H} = g(T)$. Dann ist $\tilde{\theta} - \tilde{H} = f(T) - g(T)$ ein erwartungstreuer Schätzer von 0, der nur auf dem Wert von T basiert. Wegen der Vollständigkeit von T muss $f(T) - g(T) = 0$ $\mathbb{P}_\theta$–f.s. gelten, woraus sich ergibt, dass $\tilde{\theta} = \tilde{H}$ $\mathbb{P}_\theta$–f.s. für alle $\theta \in \Theta$. □

Korollar 4.13.2. Sei $\hat{\theta}$ ein erwartungstreuer, suffizienter und vollständiger Schätzer von θ mit $\mathbb{E}_\theta \hat{\theta}^2 < \infty$ für alle $\theta \in \Theta$. Dann ist $\hat{\theta}$ der beste erwartungstreue Schätzer von θ.

Beweis. Folgt aus dem Satz von Lehmann–Scheffé mit $T = \hat{\theta}$ und $\tilde{\theta} = \mathbb{E}_\theta[\hat{\theta}|\hat{\theta}] = \hat{\theta}$. □

Beispiel 4.13.3. 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 nach Korollar 4.13.2 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.

Aus dem Satz von Lehmann–Scheffé ergibt sich die folgende Methode zur Konstruktion des besten erwartungstreuen Schätzers:

- Finde eine vollständige, suffiziente Statistik T.
- Finde Funktion g mit der Eigenschaft, dass $g(T)$ ein erwartungstreuer Schätzer von θ ist.
- Dann ist $g(T)$ der beste erwartungstreue Schätzer von θ.

Beweis. Folgt aus dem Satz von Lehmann–Scheffé mit $\hat{\theta} = g(T)$ und $\tilde{\theta} = \mathbb{E}_\theta[g(T)|T] = g(T)$. □

Beispiel 4.13.4. 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

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

der erwartungstreu und eine Funktion von $X_{(n)}$ ist. Nach den obigen Überlegungen ist $\frac{n+1}{n} X_{(n)}$ der beste erwartungstreue Schätzer für θ.

In den folgenden Beispielen werden wir den Satz von Lehmann–Scheffé in einer etwas allgemeineren Form benutzen. Der Parameterraum Θ sei beliebig und sei $\nu : \Theta \to \mathbb{R}$ eine Funktion des Parameters. Ist $\hat{\nu} : \mathfrak{X} \to \mathbb{R}$ ein erwartungstreuer und quadratisch integrierbarer Schätzer von $\nu(\theta)$ und T eine suffiziente und vollständige Statistik, so ist $\tilde{\nu} = \mathbb{E}_\theta[\hat{\nu}|T]$ der beste erwartungstreue Schätzer von $\nu(\theta)$. (Der obige Beweis funktioniert mit minimalen Veränderungen).

Beispiel 4.13.5. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ unabhängig und normalverteilt, wobei beide Parameter unbekannt seien. Wir behaupten, dass

- $\bar{X}_n$ der beste erwartungstreue Schätzer von μ ist;
- $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ der beste erwartungstreue Schätzer von σ^2 ist.

Beweis. Die Statistik $(\bar{X}_n, S_n^2)$ ist vollständig und suffizient. Sowohl $\bar{X}_n$ als auch S_n^2 sind Funktionen dieser Statistik, die erwartungstreu für μ bzw. σ^2 sind. □

Beispiel 4.13.6. Es seien $X_1, \ldots, X_n \sim \mathrm{Poi}(\theta)$ unabhängig. Der beste erwartungstreue Schätzer von θ ist, wie bereits gezeigt wurde, $\bar{X}_n$. Wie sieht nun der beste erwartungstreue Schätzer für

$$\nu(\theta) := e^{-\theta} = \mathbb{P}_\theta[X_i = 0]$$

aus? Ein natürlicher Schätzer von $e^{-\theta}$ ist $e^{-\bar{X}_n}$, dieser Schätzer ist aber nicht erwartungstreu:

Aufgabe 4.13.7. Bestimmen Sie $\mathbb{E}_\theta e^{-\bar{X}_n}$.

Um den besten erwartungstreuen Schätzer für $e^{-\theta}$ zu konstruieren, benutzen wir den Satz von Lehmann–Scheffé. Es ist bekannt, dass die Statistik $T = X_1 + \ldots + X_n$ vollständig und suffizient ist. Nun brauchen wir noch einen erwartungstreuen Schätzer von $e^{-\theta}$. Als solchen nehmen wir z.B.

$$\hat{\nu} = \mathbb{1}_{\{X_1 = 0\}}.$$

Die Erwartungstreue von $\hat{\nu}$ folgt aus $\mathbb{E}_\theta \hat{\nu} = \mathbb{P}_\theta[X_1 = 0] = e^{-\theta}$. Nach dem Satz von Lehmann–Scheffé ist $\tilde{\nu} = \mathbb{E}_\theta[\hat{\nu}|S_n]$ der beste erwartungstreue Schätzer. Wir müssen nur noch diesen bedingten Erwartungswert berechnen. Für $s \in \mathbb{N}_0$ betrachten wir

$$f(s) := \mathbb{E}_\theta[\hat{\nu}|T = s] = \mathbb{P}_\theta[X_1 = 0|T = s] = \frac{\mathbb{P}_\theta[X_1 = 0, T = s]}{\mathbb{P}_\theta[T = s]} = \frac{\mathbb{P}_\theta[X_1 = 0]\mathbb{P}_\theta[X_2 + \ldots + X_n]}{\mathbb{P}_\theta[X_1 + \ldots + X_n = s]}.$$

Nun benutzen wir die Faltungseigenschaft der Poisson–Verteilung: Unter $\mathbb{P}_\theta$ gilt

$$X_1 \sim \mathrm{Poi}(\theta), \quad X_2 + \ldots + X_n \sim \mathrm{Poi}((n-1)\theta), \quad X_1 + \ldots + X_n \sim \mathrm{Poi}(n\theta).$$

Somit erhalten wir, dass

$$f(s) = \frac{e^{-\theta} e^{-(n-1)\theta}((n-1)\theta)^s / s!}{e^{-n\theta}(n\theta)^s / s!} = \left(1 - \frac{1}{n}\right)^s.$$

Aus dem Satz von Lehmann–Scheffé folgt nun, dass

$$\tilde{\nu} = \mathbb{E}_\theta[\hat{\nu}|T] = f(T) = \left(1 - \frac{1}{n}\right)^T$$

der beste erwartungstreue Schätzer von $e^{-\theta}$ ist.

Aufgabe 4.13.8. Es seien $X_1, \ldots, X_n \sim \mathrm{Poi}(\theta)$ unabhängig. Definiere $T := X_1 + \ldots + X_n$. Zeigen Sie, dass der beste erwartungstreue Schätzer von

$$\nu_k = e^{-\theta}\frac{\theta^k}{k!} = \mathbb{P}_\theta[X_1 = k], \quad k \in \mathbb{N}_0,$$

durch

$$\hat{\nu}_{k,n} := \binom{T}{k}\left(\frac{1}{n}\right)^k\left(1 - \frac{1}{n}\right)^{T-k}$$

gegeben ist. Zeigen Sie auch, dass $\hat{\nu}_{k,n}$ ein stark konsistenter Schätzer von ν_k ist.

Beispiel 4.13.9. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ 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

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

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.

Aufgabe 4.13.10. Seien $X_1, \ldots, X_n$ unabhängig und Bernoulli-verteilt mit Parameter $p \in (0,1)$. Bestimmen Sie den besten erwartungstreuen Schätzer für p^2.

Aufgabe 4.13.11. Gegeben sei eine Urne mit einer unbekannten Anzahl $N \in \mathbb{N}$ Kugeln, die von 1 bis N durchnummeriert sind. Es werden n Kugeln mit Zurücklegen gezogen und die zugehörigen Nummern $X_1, \ldots, X_n$ notiert. Zeigen Sie, dass $X_{(n)}$ eine suffiziente und vollständige Statistik ist und konstruieren Sie den besten erwartungstreuen Schätzer für N.

Aufgabe 4.13.12. Seien $X_1, \ldots, X_n \sim \mathrm{U}[\theta_1, \theta_2]$ unabhängig, wobei $\theta_1 < \theta_2$ unbekannt seien. Konstruieren Sie die besten erwartungstreuen Schätzer von θ_1 und θ_2.

Aufgabe 4.13.13. Seien $X_1, \ldots, X_n \sim \mathrm{Exp}(\theta)$ unabhängig, $\theta > 0$. Konstruieren Sie den besten erwartungstreuen Schätzer von $\mathbb{P}_\theta[X_1 > a] = e^{-a\theta}$, wobei $a > 0$ gegeben ist.

Aufgabe 4.13.14. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, 1)$ unabhängig, wobei $\mu \in \mathbb{R}$ unbekannt sei. Konstruieren Sie den besten erwartungstreuen Schätzer von $e^{a\mu}$, wobei $a \in \mathbb{R}$ gegeben ist.

4.14. Satz von Basu

Sei $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell.

Definition 4.14.1. Eine Statistik $S : \mathfrak{X} \to \mathbb{R}^p$ heißt *verteilungsfrei*, falls

$$\mathbb{P}_{\theta_1}[S \in A] = \mathbb{P}_{\theta_2}[S \in A] \text{ für alle } \theta_1, \theta_2 \in \Theta, \quad A \subset \mathbb{R}^p(\text{Borel}).$$

D.h., die Verteilung von S unter $\mathbb{P}_\theta$ hängt nicht von θ ab.

Beispiel 4.14.2. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ unabhängig, wobei $\mu \in \mathbb{R}$ unbekannter Parameter ist und σ^2 bekannt sei. Wir behaupten, dass die Spannweite $X_{(n)} - X_{(1)}$ und die empirische Varianz $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ verteilungsfreie Statistiken sind.

Beweis. Die Idee ist, dass μ ein "Verschiebungsparameter" ist, der sich sowohl in der Spannweite, als auch in S_n^2 aufhebt. Seien $\xi_1, \ldots, \xi_n \sim \mathrm{N}(0, \sigma^2)$ unabhängig (mit Erwartungswert 0). Unter $\mathbb{P}_\mu$ hat $(X_1, \ldots, X_n)$ die gleiche Verteilung wie $(\xi_1 + \mu, \ldots, \xi_n + \mu)$. Somit hat S_n^2 unter $\mathbb{P}_\mu$ die gleiche Verteilung wie

$$\frac{1}{n-1}\sum_{i=1}^n \left(\xi_i + \mu - \frac{(\xi_1+\mu) + \ldots + (\xi_n+\mu)}{n}\right)^2 = \frac{1}{n-1}\sum_{i=1}^n \left(\xi_i - \frac{\xi_1 + \ldots + \xi_n}{n}\right)^2.$$

Die Verteilung der rechten Seite hängt aber nicht von μ ab, denn diese Variable taucht dort gar nicht erst auf. Für die Spannweite ist der Beweis analog. □

Der folgende Satz von Basu besagt, dass jede verteilungsfreie Statistik von jeder vollständig suffizienten Statistik unabhängig ist.

Satz 4.14.3 (Basu, 1955)**.** Sei $S : \mathfrak{X} \to \mathbb{R}^p$ eine verteilungsfreie Statistik und $T : \mathfrak{X} \to \mathbb{R}^m$ eine vollständige suffiziente Statistik. Dann sind die Zufallsvektoren S und T unabhängig unter $\mathbb{P}_\theta$ für alle $\theta \in \Theta$.

Beweis. Betrachte das folgende Wahrscheinlichkeitsmaß Q auf $\mathbb{R}^p$:

$$Q(A) = \mathbb{P}_\theta[S \in A], \quad A \subset \mathbb{R}^p \text{ Borel.}$$

Es sei bemerkt, dass Q unabhängig von θ wegen der Verteilungsfreiheit von S ist. Im Folgenden halten wir die Menge A fest. Betrachte die Funktion

$$f_A(t) = \mathbb{P}_\theta[S \in A | T = t] = \mathbb{E}_\theta[\mathbb{1}_{\{S \in A\}} | T = t], \quad t \in \mathbb{R}^m.$$

Diese Funktion ist unabhängig von θ, da die Statistik T suffizient ist! Es gilt dann auch $f_A(T) = \mathbb{E}_\theta[\mathbb{1}_{\{S \in A\}} | T]$ und somit

$$\mathbb{E}_\theta[f_A(T)] = \mathbb{E}_\theta\left[\mathbb{E}_\theta[\mathbb{1}_{\{S \in A\}} | T]\right] = \mathbb{E}_\theta \mathbb{1}_{\{S \in A\}} = \mathbb{P}_\theta[S \in A] = Q(A).$$

Es folgt, dass

$$\mathbb{E}_\theta[f_A(T) - Q(A)] = 0 \text{ für alle } \theta \in \Theta.$$

Somit ist $f_A(T) - Q(A)$ ein erwartungstreuer Schätzer von 0, der auf der Statistik T basiert. Wegen der Vollständigkeit von T folgt daraus, dass

$$f_A(T) = Q(A) \;\; \mathbb{P}_\theta\text{–f.s. für alle } \theta \in \Theta.$$

Wir haben also gezeigt, dass

$$\mathbb{P}_\theta[S \in A | T = t] = \mathbb{P}_\theta[S \in A].$$

Ist nun $B \subset \mathbb{R}^m$ eine Borel–Menge und bezeichnen wir mit μ_T die Verteilung von T, so ergibt sich mit der Formel der totalen Wahrscheinlichkeit, dass

$$\mathbb{P}_\theta[S \in A, T \in B] = \int_B \mathbb{P}_\theta[S \in A | T = t] \mu_T(dt) = \int_B \mathbb{P}_\theta[S \in A] \mu_T(dt) = \mathbb{P}_\theta[S \in A] \mathbb{P}_\theta[T \in B],$$

was die Unabhängigkeit von S und T beweist. □

Hier ist eine typische Anwendung des Satzes von Basu. Dieses Korollar wird sich bei der Konstruktion des Student–t–Tests als sehr wichtig erweisen.

Korollar 4.14.4. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ unabhängig. Dann sind die Zufallsvariablen $\bar{X}_n$ und S_n^2 unabhängig.

Die Behauptung ist sehr überraschend, denn $\bar{X}_n$ taucht in der Definition von S_n^2 explizit auf!

Beweis. Wir fassen $\mu \in \mathbb{R}$ als unbekannten Parameter auf und halten σ^2 konstant. Die Statistik S_n^2 ist verteilungsfrei, während die Statistik $\bar{X}_n$ vollständig suffizient ist. Die Behauptung folgt nun aus dem Satz von Basu. □

Aufgabe 4.14.5. Seien $X_1, X_2, \ldots \sim \mathrm{Exp}(\lambda)$ unabhängig. Betrachten Sie die Ankunftszeiten des Poisson–Punktprozesses $S_k = X_1 + \ldots + X_k$, $k \in \mathbb{N}$. Zeigen Sie, dass

$$\left(\frac{S_1}{S_n}, \frac{S_2}{S_n}, \ldots, \frac{S_{n-1}}{S_n}\right) \text{ und } S_n$$

unabhängig sind.

Aufgabe 4.14.6. Seien $X_1, \ldots, X_n \sim \mathrm{U}[0,1]$ unabhängig und gleichverteilt auf $[0,1]$. Seien $U_{(1)} < \ldots < U_{(n)}$ die Ordnungsstatistiken. Zeigen Sie, dass

$$\left(\frac{U_{(1)}}{U_{(n)}}, \frac{U_{(2)}}{U_{(n)}}, \ldots, \frac{U_{(n-1)}}{U_{(n)}}\right) \text{ und } U_{(n)}$$

unabhängig sind.

4.15. Einige Gegenbeispiele

Kann man vielleicht sogar unter allen quadratisch integrierbaren (nicht unbedingt erwartungstreuen) Schätzern einen finden, der gleichmäßig besser, als alle anderen ist? Die Antwort ist leider "nein". Sei $\theta_0 \in \Theta$ beliebig. Wir können dann den konstanten Schätzer $\varphi = \theta_0$ betrachten. Es gilt

$$\mathrm{MSE}_{\theta_0}(\varphi) = 0.$$

Wäre nun ein Schätzer $\hat{\theta}$ gleichmäßig besser als φ, so müsste $\mathrm{MSE}_{\theta_0}(\hat{\theta}) = 0$ gelten. Das bedeutet aber, dass $\hat{\theta} = \theta_0$ f.s. unter $\mathbb{P}_{\theta_0}$. Wäre $\hat{\theta}$ gleichmäßig besser als alle Schätzer, so müsste $\hat{\theta} = \theta$ f.s. unter $\mathbb{P}_\theta$ für alle $\theta \in \Theta$. Das heißt, der Schätzer $\hat{\theta}$ müsste den richtigen Wert θ mit Wahrscheinlichkeit 1 exakt treffen. Solche Schätzer gibt es aber nur in trivialen Situationen.

Beispiel 4.15.1. Sei X_1 eine Zufallsvariable, die gleichverteilt auf dem Intervall $[\theta, \theta+1]$ ist, wobei $\theta \in \mathbb{Z}$ unbekannt ist. Beobachtet man nun einen Wert x_1 zwischen 2 und 3, so weiß man ganz genau, dass $\theta = 2$ ist. In diesem Fall können wir anhand der Stichprobe den Wert von θ richtig erraten. In allen interessanten statistischen Modellen ist aber so etwas nicht möglich.

Die folgenden Aufgaben zeigen, dass der beste erwartungstreue Schätzer einen gleichmäßig größeren quadratischen Fehler haben kann, als einige nicht-erwartungstreue Schätzer.

Aufgabe 4.15.2. Seien $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ unabhängig. Betrachten Sie den Schätzer $T = \frac{1}{c}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ für σ^2. Zeigen Sie, dass das Minimum des mittleren quadratischen Fehlers für $c = n+1$ erreicht wird. Dabei entspricht der Wert $c = n-1$ dem besten erwartungstreuen Schätzer.

Aufgabe 4.15.3. Seien $X_1, \ldots, X_n$ unabhängig und auf $[0, \theta]$ gleichverteilt, $\theta > 0$. Zeigen Sie, dass der nicht-erwartungstreue Schätzer $\hat{\theta}_1 := X_{(n)}$ für alle $n \geq 3$ einen gleichmäßig kleineren mittleren quadratischen Fehler als der beste erwartungstreue Schätzer $\hat{\theta}_3 := \frac{n+1}{n} X_{(n)}$ hat, nämlich

$$\mathrm{MSE}_\theta(\hat{\theta}_1) = \frac{2\theta^2}{(n+2)(n+1)}, \quad \mathrm{MSE}_\theta(\hat{\theta}_3) = \frac{\theta^2}{3n},$$