

# Mathematische Statistik

Version vom: 20. Februar 2009

Vorläufig

Prof. Dr. Claudia Czado<sup>1</sup> und Prof. Dr. Thorsten Schmidt<sup>2</sup>

---

<sup>1</sup>Technische Universität München, Lehrstuhl für Mathematische Statistik, Boltzmannstr. 3, 85747 Garching, Deutschland

<sup>2</sup>Technische Universität Chemnitz, Mathematisches Institut, Reichenhainer Str. 41, 09126 Chemnitz, Deutschland

# Inhaltsverzeichnis

<b>1 Grundlagen der Wahrscheinlichkeitstheorie und Statistik</b>	<b>1</b>
1.1 Grundbegriffe der Wahrscheinlichkeitstheorie . . . . .	1
1.2 Klassische Verteilungen der Statistik . . . . .	9
1.3 Bedingte Verteilungen . . . . .	16
1.4 Gesetz der großen Zahl . . . . .	20
1.5 Aufgaben . . . . .	20
<b>2 Statistische Modelle</b>	<b>22</b>
2.1 Formulierung von statistischen Modellen . . . . .	22
2.2 Suffizienz . . . . .	27
2.3 Exponentielle Familien . . . . .	33
2.4 Bayesianische Modelle . . . . .	40
2.5 Aufgaben . . . . .	46
<b>3 Schätzmethoden</b>	<b>47</b>
3.1 Substitutionsprinzip . . . . .	48
3.1.1 Häufigkeitssubstitution . . . . .	48
3.1.2 Momentenmethode . . . . .	50
3.2 Methode der Kleinsten Quadrate . . . . .	53
3.2.1 Allgemeine und lineare Regressionsmodelle . . . . .	53
3.2.2 Methode der kleinsten Quadrate . . . . .	54
3.2.3 Gewichtete Kleinste-Quadrate-Schätzer . . . . .	57
3.3 Maximum-Likelihood-Schätzung . . . . .	58
3.3.1 Maximum-Likelihood in eindimensionalen Modellen . . . . .	60
3.3.2 Maximum Likelihood in mehrdimensionalen Modellen . . . . .	66
3.3.3 Numerische Bestimmung des MLS . . . . .	67
3.4 Vergleich der Maximum Likelihood Methode mit anderen Schätzverfahren .	69
3.5 Aufgaben . . . . .	70
<b>4 Vergleich von Schätzern: Optimalitätstheorie</b>	<b>71</b>

4.1	Schätzkriterien . . . . .	71
4.2	UMVUE Schätzer . . . . .	75
4.3	Die Informationsungleichung . . . . .	81
4.4	Asymptotische Theorie . . . . .	85
4.4.1	Konsistenz . . . . .	85
4.4.2	Asymptotische Normalität und verwandte Eigenschaften . . . . .	86
4.5	Aufgaben . . . . .	87
<b>5</b>	<b>Konfidenzintervalle und Hypothesentest</b>	<b>88</b>
<b>6</b>	<b>Optimale Tests und Konfidenzintervalle, Likelihood Ratio Tests und verwandte Methoden</b>	<b>89</b>
<b>7</b>	<b>Lineare Modelle - Regression und Varianzanalyse (ANOVA)</b>	<b>90</b>
<b>8</b>	<b>Verzeichnisse</b>	<b>91</b>
	Tabellenverzeichnis . . . . .	91
	Abbildungsverzeichnis . . . . .	91
	Verzeichnis der Beispiele . . . . .	92

Vorläufig

# 1 Grundlagen der Wahrscheinlichkeitstheorie und Statistik

Statistik ist die Wissenschaft, die Regeln und Verfahren für die *Erhebung, Beschreibung, Analyse und Interpretation* von numerischen Daten entwickelt.

Dieser Text behandelt *statistische Modelle*, dafür konstruierte Verfahren zur *Schätzung* von *Parametern* und das *Testen von Hypothesen*. Darüber hinaus wird ein Vergleich der vorgestellten Verfahren anhand bestimmter Kriterien angestrebt, weswegen verschiedene Eigenschaften der vorgestellten Verfahren studiert werden.

Die verwendeten Grundlagen entstammen der Wahrscheinlichkeitstheorie und zu Beginn wird eine kurze Einführung gegeben. Für eine ausgiebige Darstellung sei auf Georgii (2004) und Resnick (2003) verwiesen.

## 1.1 Grundbegriffe der Wahrscheinlichkeitstheorie

Dieser Abschnitt beschreibt kurz den Kolmogorovschen Zugang zur Wahrscheinlichkeitstheorie. Jedem zufälligen Ereignis wird hierbei eine Wahrscheinlichkeit zugeordnet. Ein Ereignis ist beschrieben durch eine Menge. Das gleichzeitige Eintreten zweier Ereignisse ist der Schnitt zweier Mengen, welches wieder ein Ereignis sein sollte. Dies erfordert eine Axiomatik, welche im Folgenden kurz vorgestellt wird. Grundlage bildet ein *Wahrscheinlichkeits-Raum*  $(\Omega, \mathcal{A}, \mathbb{P})$ , wobei  $\Omega$  den *Zustandsraum*,  $\mathcal{A}$  die zugehörige  $\sigma$ -Algebra und  $\mathbb{P}$  ein *Wahrscheinlichkeitsmaß* bezeichnet. Die Elemente von  $\mathcal{A}$  beschreiben die Ereignisse welche in einem Zufallsexperiment auftreten können. Mit zwei Ereignissen  $A$  und  $B$  aus  $\mathcal{A}$  möchte man auch das Ereignis "A und B" beobachten, weswegen man von  $\mathcal{A}$  gewisse Eigenschaften fordert. Eine Menge  $\mathcal{A}$ , dessen Elemente Teilmengen von  $\Omega$  sind, heißt  $\sigma$ -Algebra, falls

(i)  $\Omega \in \mathcal{A}$

(ii) Für jedes  $A \in \mathcal{A}$  gilt  $A^C := \Omega \setminus A \in \mathcal{A}$

(iii) Für Elemente  $A_1, A_2, \dots$  von  $\mathcal{A}$  gilt  $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$ .

Weiterhin wird verlangt, dass das Wahrscheinlichkeitsmaß  $\mathbb{P}$  die klassischen Kolmogorovschen Axiome erfüllt. Demnach ist die Abbildung  $\mathbb{P} : \mathcal{A} \rightarrow [0, 1]$  ein *Wahrscheinlichkeitsmaß*, falls die folgenden drei Eigenschaften erfüllt sind:

(i)  $\mathbb{P}(\Omega) = 1$ ,

(ii)  $0 \leq \mathbb{P}(A) \leq 1$  für alle  $A \in \mathcal{A}$ ,

(iii) Für Elemente  $A_1, A_2, \dots$  von  $\mathcal{A}$  mit  $A_i \cap A_j = \emptyset$  für jedes  $i \neq j$  gilt:

$$\mathbb{P}\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Ein Wahrscheinlichkeitsraum heißt *diskret*, falls der Grundraum  $\Omega$  in abzählbar viele disjunkte Ereignisse zerfällt, das heißt  $\Omega = \{\omega_1, \omega_2, \dots\}$ . In diesem Fall heißt  $\omega_i \in \Omega$  *Elementarereignis*.

**Bedingte Wahrscheinlichkeiten und Unabhängigkeit.** Oft hat man bereits zusätzliche Information über die zu erhebenden Daten, etwa dass untersuchte Patienten ein gewisses Merkmal aufweisen. Diese zusätzliche Information nutzt man durch die Verwendung von bedingten Wahrscheinlichkeiten.

Seine  $A, B \in \mathcal{A}$  zwei Ereignisse mit  $\mathbb{P}(B) > 0$ . Die *bedingte Wahrscheinlichkeit* von  $A$  gegeben  $B$  ist definiert durch

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Darüber hinaus definiert  $\mathbb{P}(\cdot|B) : \mathcal{A} \rightarrow [0, 1]$  das bedingte Wahrscheinlichkeitsmaß gegeben  $B$ . Dieses Maß ist in der Tat ein Wahrscheinlichkeitsmaß, siehe Aufgabe 1.1.

Ist  $\Omega = \bigcup_{i=1}^n B_i$  und sind die  $B_i$  paarweise disjunkt, so schreiben wir  $\Omega = \sum_{i=1}^n B_i$ .

**Satz 1.1** (Satz von Bayes). Sei  $\Omega = \sum_{i=1}^n B_i$  mit  $\mathbb{P}(B_i) > 0$  für  $i = 1, \dots, n$ . Dann gilt für  $A \in \mathcal{A}$  mit  $\mathbb{P}(A) > 0$ , dass

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{j=1}^n \mathbb{P}(A|B_j)\mathbb{P}(B_j)}.$$

Zwei Ereignisse  $A$  und  $B$  heißen *unabhängig*, falls

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Dann gilt auch  $\mathbb{P}(A|B) = \mathbb{P}(A)$ . Für  $n$  Ereignisse muss man die (schwächere) paarweise Unabhängigkeit unterscheiden von der folgenden Eigenschaft:  $A_1, \dots, A_n$  heißen *unabhängig*, falls

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \prod_{j=1}^k \mathbb{P}(A_{i_j}) \quad \forall \{i_1, \dots, i_k\} \subset \{1, \dots, n\}$$

**Zufallsvariablen.** Ein Zufallsexperiment wird durch eine Zufallsvariable modelliert. Eine Zufallsvariable  $X$  ist intuitiv gesprochen eine Abbildung, welche die Grundereignisse  $\omega \in \Omega$  auf reelle Zahlen (oder Vektoren) abbildet. Um die Wahrscheinlichkeit etwa für das Ereignis  $A := X \leq 0$  berechnen zu können, ist  $A \in \mathcal{A}$  zu fordern. Das führt zu folgendem Begriff der Meßbarkeit.

Sei  $\mathcal{B}^k$  die Borel- $\sigma$ -Algebra (dies ist die kleinste  $\sigma$ -Algebra, die alle offenen Rechtecke  $(a_1, b_1) \times \dots \times (a_k, b_k)$  enthält). Eine  $k$ -dimensionale *Zufallsvariable* ist eine  $\mathcal{A}$ - $\mathcal{B}^k$  meßbare Abbildung  $X : \Omega \rightarrow \mathbb{R}^k$ , d.h. für jedes  $B \in \mathcal{B}^k$  ist

$$X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}.$$

Die Meßbarkeit wird für alle in diesem Text auftauchenden Funktionen gegeben sein.

Eine Zufallsvariable  $X$  heißt *diskret*, falls sie nur abzählbar viele Werte  $x_1, x_2, \dots$  annimmt. Dann heißt die Funktion  $p_X : \{x_1, x_2, \dots\} \rightarrow [0, 1]$  gegeben durch

$$p_X(x_i) = \mathbb{P}(X = x_i), \quad i = 1, 2, \dots$$

die *Wahrscheinlichkeitsfunktion* von  $X$ . Durch sie ist  $X$  vollständig beschrieben, denn für jedes Ereignis  $A$  ist  $\mathbb{P}(A) = \sum_{x_i \in A} p_X(x_i)$ . Um im folgenden eine einheitliche Schreibweise mit stetigen Zufallsvariablen nutzen zu können setzen wir stets  $p_X(x) = 0$  für  $x \notin \{x_1, x_2, \dots\}$ .

Ist eine Zufallsvariable nicht diskret, so kann man sie oft durch ihre Dichte beschreiben. Eine *Dichte* ist eine nichtnegative Funktion  $p$  auf  $\mathbb{R}^k$ , die Lebesgue-integrierbar ist mit

$$\int_{\mathbb{R}^k} p(\mathbf{x}) d\mathbf{x} = 1.$$

Gilt für eine Zufallsvariable  $X$  für alle  $A \in \mathcal{A}$

$$\mathbb{P}(X \in A) = \int_A p(x) dx,$$

und ist  $p$  eine Dichte, so heißt  $p$  die *Dichte von  $X$* . In diesem Fall heißt  $X$  *stetige Zufallsvariable*.

Unabhängig davon, ob eine Zufallsvariable diskret ist oder etwa eine Dichte besitzt, lässt sie sich stets durch ihre Verteilungsfunktion beschreiben. Die *Verteilungsfunktion* einer Zufallsvariable  $\mathbf{X}$  ist definiert durch

$$F_{\mathbf{X}}(\mathbf{x}) = F_{\mathbf{X}}(x_1, \dots, x_k) := \mathbb{P}(X_1 \leq x_1, \dots, X_k \leq x_k).$$

Die Verteilungsfunktion hat, wie man leicht sieht, folgende Eigenschaften. Zur Einfachheit betrachten wir nur den eindimensionalen Fall. Dann gilt:  $0 \leq F \leq 1$ .  $F$  ist monoton wachsend, rechtsseitig stetig,  $\lim_{x \rightarrow \infty} F(x) = 1$  und  $\lim_{x \rightarrow -\infty} F(x) = 0$ . Neben der Verteilungsfunktion spricht man allgemeiner von der Verteilung einer Zufallsvariable. Die *Verteilung* einer Zufallsvariable  $\mathbf{X}$  ist ein Wahrscheinlichkeitsmaß  $P_{\mathbf{X}}$ , gegeben durch

$$P_{\mathbf{X}}(B) := \mathbb{P}(\{\omega \in \Omega : \mathbf{X}(\omega) \in B\}) = \mathbb{P}(\mathbf{X} \in B) \quad \forall B \in \mathcal{B}^k$$

Die Verteilung einer Zufallsvariable ist je nach Typ der Zufallsvariable unterschiedlich darstellbar. Ist  $\mathbf{X}$  eine diskrete Zufallsvariable mit Werten  $\mathbf{x}_1, \mathbf{x}_2, \dots$  und mit Wahrscheinlichkeitsfunktion  $p$ , so ist

$$\mathbb{P}(\mathbf{X} \in A) = \sum_{\mathbf{x}_i \in A} p(\mathbf{x}_i)$$

Hat  $\mathbf{X}$  hingegen die Dichte  $p$ , so ist

$$\mathbb{P}(\mathbf{X} \in A) = \int_A p(\mathbf{x}) d\mathbf{x}$$

**Transformationsatz** Eine Transformation einer  $k$ -dimensionalen Zufallsvariablen  $\mathbf{X}$  ist eine meßbare Abbildung  $\mathbf{g} : \mathbb{R}^k \rightarrow \mathbb{R}^m$ , d.h.  $\mathbf{g}^{-1}(B) \in \mathcal{B}^k$  für alle Mengen  $B$  aus der Borel- $\sigma$ -Algebra  $\mathcal{B}^m$ . Die Verteilung der transformierten Zufallsvariable  $\mathbf{g}(\mathbf{X})$  ist bestimmt durch  $\mathbb{P}(\mathbf{g}(\mathbf{X}) \in B) = \mathbb{P}(\mathbf{X} \in \mathbf{g}^{-1}(B))$  für alle  $B \in \mathcal{B}^m$ .

**Beispiel 1.1.** (*Mittelwert und Stichprobenvarianz*) Betrachtet man eine Stichprobe gegeben durch reellwertige Zufallsvariablen  $\mathbf{X} = (X_1, \dots, X_n)^\top$ , so ist der Vektor gegeben durch den arithmetischen Mittelwert und die Stichprobenvarianz eine Transformation: In diesem Fall ist  $\mathbf{g} = (g_1, g_2)$ ; der *arithmetische Mittelwert* ist  $g_1$  und die *Stichprobenvarianz* ist  $g_2$  mit

$$g_1(\mathbf{X}) := \frac{1}{k} \sum_{i=1}^k X_i = \bar{X}$$

$$g_2(\mathbf{X}) := \frac{1}{k-1} \sum_{i=1}^k (X_i - \bar{X})^2 =: s^2(\mathbf{X}).$$

Die besondere Normierung mit  $(k-1)$  sorgt dafür, dass die Stichprobenvarianz erwartungstreu ist, eine Eigenschaft welche man verliert, wenn statt dessen man mit  $k$  normiert. Dies werden wir in Aufgabe 1.3 diskutieren.

Für stetige Zufallsvariablen hat man folgenden wichtigen Satz:

**Satz 1.2** (Transformationsatz). *Sei  $X$  eine stetig Zufallsvariable mit Dichte  $p_X$ . Die Transformation  $g : \mathbb{R} \rightarrow \mathbb{R}$  sei bijektiv auf einer offenen Menge  $S$  mit  $\mathbb{P}(X \in S) = 1$ . Ferner sei  $g$  differenzierbar und  $g'(X) \neq 0 \forall X \in S$ . Dann ist  $Y = g(X)$  eine stetige Zufallsvariable und die Dichte von  $Y$  ist gegeben durch*

$$p_{g(X)}(y) = \frac{\mathbb{P}_X(g^{-1}(y))}{|g'(g^{-1}(y))|}.$$

Diese Behauptung lässt sich leicht durch Differenzieren der Verteilungsfunktion von  $Y$  und Anwenden der Kettenregel zeigen.

Im mehrdimensionalen Fall gilt ein analoges Resultat: Sei  $\mathbf{h} : \mathbb{R}^k \rightarrow \mathbb{R}^k$ ,  $\mathbf{h} = (h_1, \dots, h_k)$ ,  $h_i : \mathbb{R}^k \rightarrow \mathbb{R}$  und die Jacobi Determinante gegeben durch

$$J_{\mathbf{h}}(\mathbf{x}) := \begin{vmatrix} \frac{\partial}{\partial t_1} h_1(\mathbf{x}) & \dots & \frac{\partial}{\partial t_1} h_k(\mathbf{x}) \\ \vdots & & \vdots \\ \frac{\partial}{\partial t_k} h_1(\mathbf{x}) & \dots & \frac{\partial}{\partial t_k} h_k(\mathbf{x}) \end{vmatrix}$$

**Satz 1.3** (Transformationsatz für Zufallsvektoren). Sei  $\mathbf{h}$  auf einer offenen Menge  $B \subset \mathbb{R}^k$  definiert, so dass

- (i)  $\mathbf{h}$  hat stetige erste partielle Ableitungen auf  $B$
- (ii)  $\mathbf{h}$  ist bijektiv auf  $B$
- (iii)  $J_{\mathbf{h}}(\mathbf{x}) \neq 0 \quad \forall \mathbf{x} \in B$

Sei  $\mathbf{X}$  eine stetige Zufallsvariable mit  $\mathbb{P}(\mathbf{X} \in B) = 1$ . Dann ist die Dichte von  $\mathbf{Y} = \mathbf{h}(\mathbf{X})$  gegeben durch

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{h}^{-1}(\mathbf{y})) |J_{\mathbf{h}^{-1}}(\mathbf{y})|$$

**Unabhängigkeit.** Die Unabhängigkeit von Zufallsvariablen geht maßgeblich auf die Unabhängigkeit von Ereignissen zurück. Zwei Zufallsvariablen  $\mathbf{X}_1 \in \mathbb{R}^k$  und  $\mathbf{X}_2 \in \mathbb{R}^m$  heißen *unabhängig*, falls die Ereignisse  $\{\mathbf{X}_1 \in A\}$  und  $\{\mathbf{X}_2 \in B\}$  unabhängig sind für *alle*  $A \in \mathcal{B}^k$  und  $B \in \mathcal{B}^m$ .

Unabhängigkeit kann man dadurch charakterisieren, dass die Dichte, die Wahrscheinlichkeitsfunktion oder die Verteilungsfunktion in Produktgestalt zerfällt:

**Satz 1.4.** Ist die Zufallsvariable  $\mathbf{X} = (X_1, \dots, X_n)^\top$  stetig mit Dichte  $p_{\mathbf{X}}$  oder diskret mit Wahrscheinlichkeitsfunktion  $p_{\mathbf{X}}$ , so sind die folgenden drei Aussagen äquivalent:

- (i)  $X_1, \dots, X_n$  sind unabhängig
- (ii)  $F_{\mathbf{X}}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$  für alle  $x_1, \dots, x_n$
- (iii)  $p_{\mathbf{X}}(x_1, \dots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n)$  für alle  $x_1, \dots, x_n$

Wir bezeichnen Zufallsvariablen  $X_1, \dots, X_n$  oder auch etwa eine ganze Folge  $X_1, X_2, \dots$  als unabhängig, falls für jede beliebige Kombination  $i, j$  welche sich nicht überschneiden, die Vektoren  $(X_{i(1)}, \dots, X_{i(n_1)})^\top$  und  $(X_{j(1)}, \dots, X_{j(n_2)})^\top$  unabhängig sind. Im Allgemeinen ist dies schärfer als die Annahme der paarweisen Unabhängigkeit, unter welcher jedes  $X_i$  und  $X_j$  mit  $i \neq j$  unabhängig sind.

Zufallsvariablen, welche unabhängig und identische verteilt sind, bezeichnen wir kurz als i.i.d. (independent and identically distributed). Dies ist eine in der Statistik häufig gemachte Annahme.

**Momente.** Oft kann man wichtige Charakteristika von Zufallsvariablen bereits durch einfachere Funktionale als die Verteilungsfunktion beschreiben. Die Normalverteilung beispielsweise ist vollständig durch ihr erstes und zweites Moment beschrieben. Dieser Abschnitt führt zentrale Größen wie Erwartungswert und Varianz und darüber hinausgehend die Momente einer Zufallsvariable ein.

Der *Erwartungswert* einer Zufallsvariable  $X$  ist wie folgt definiert: Ist  $X$  diskret mit Werten  $\{x_1, x_2, \dots\}$  so ist der Erwartungswert definiert durch

$$\mathbb{E}(X) := \sum_{i=1}^n x_i \mathbb{P}(X = x_i);$$

ist  $X$  eine stetige Zufallsvariable mit Dichte  $p_X$ , so ist

$$\mathbb{E}(X) := \int xp_X(x)dx.$$

Der Erwartungswert einer Zufallsvariable gibt den Wert an, welchen die Zufallsvariable im Mittel annimmt. Man verifiziert leicht, dass der Erwartungswert ein linearer Operator ist, d.h. für  $a_1, \dots, a_n \in \mathbb{R}$  ist

$$\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathbb{E}(X_i).$$

Darüberhinaus ist der Erwartungswert monoton, d.h. aus  $\mathbb{P}(X \geq Y) = 1$  folgt

$$\mathbb{E}(X) \geq \mathbb{E}(Y). \quad (1.1)$$

Folgende Ungleichung wird sich als äußerst nützlicher Begleiter erweisen. Eine Funktion  $g : \mathbb{R} \rightarrow \mathbb{R}$  heißt konvex, falls  $g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y)$  für alle  $\lambda \in (0, 1)$  und alle  $x, y \in \mathbb{R}$ .

**Satz 1.5** (Jensensche Ungleichung). *Sei  $g$  eine konvexe Funktion und  $X$  eine Zufallsvariable. Dann gilt*

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X)).$$

*Gleichheit gilt genau dann, wenn jede Gerade  $a + bx$  welche tangential zu  $g$  and  $x = \mathbb{E}(X)$  ist gilt, dass  $\mathbb{P}(g(X) = a + bX) = 1$ .*

Ein typisches Beispiel ist  $g(x) = x^2$ : Für eine Zufallsvariable  $X$  mit verschwindenden Erwartungswert folgt bereits aus  $x^2 \geq 0$ , dass  $\mathbb{E}(X^2) \geq (\mathbb{E}(X))^2 = 0$ .

Das  $k$ -te Moment von  $X$  ist  $\mathbb{E}(X^k)$  und das  $k$ -te zentrierte (zentrale) Moment von  $X$  ist definiert durch

$$\mu_k := \mathbb{E}\left((X - \mathbb{E}(X))^k\right).$$

Das zweite zentrierte Moment spielt eine besondere Rolle: Die Varianz von  $X$  ist definiert durch

$$\sigma^2 := \text{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

Die letzte Gleichheit lässt sich wieder leicht zeigen. Die Varianz ist ein Maß für die Streuung einer Zufallsvariable. Um die Abweichung einer Zufallsvariable von einer Normalverteilung zu messen, nutzt man typischerweise noch ein geeignetes drittes und viertes Moment, die Schiefe (skewness):  $\gamma_1 = \frac{\mu_3}{\sigma^3}$  und die Kurtosis:  $\gamma_2 := \frac{\mu_4}{\sigma^4} - 3$ .

Für einen Zufallsvektor  $\mathbf{X} = (X_1, \dots, X_k)$  sei der Erwartungswert definiert durch

$$\mathbb{E}(\mathbf{X}) := (\mathbb{E}(X_1), \dots, \mathbb{E}(X_k))^\top.$$

Betrachtet man zwei Zufallsvariablen  $X_1$  und  $X_2$ , so kann man deren lineare Abhängigkeit durch die Kovarianz erfassen. Dieses Maß zeigt allerdings außerhalb der Normalverteilungsfamilien prekäre Eigenheiten und sollte dort nur mit Vorsicht angewendet werden, siehe Aufgabe 1.8 und Schmidt (2007). Für zwei Zufallsvariablen  $X_1$  und  $X_2$  definiert man die Kovarianz von  $X_1$  und  $X_2$  durch

$$\text{Cov}(X_1, X_2) := \mathbb{E}\left((X_1 - \mathbb{E}(X_1)) \cdot (X_2 - \mathbb{E}(X_2))\right) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2).$$

Die Kovarianz ist dabei abhängig von den Varianzen der einzelnen Zufallsvariablen. Ein skalenunabhängiges Maß für die lineare Abhängigkeit ist die Korrelation zwischen  $X_1$  und  $X_2$ . Sie ist definiert durch

$$\text{Corr}(X_1, X_2) := \frac{\text{Cov}(X_1, X_2)}{(\text{Var}(X_1) \text{Var}(X_2))^{1/2}};$$

es gilt  $\text{Corr}(X_1, X_2) \in [-1, 1]$ . Zwei Zufallsvariablen  $X_1, X_2$  mit  $\text{Cov}(X_1, X_2) = 0$  nennt man unkorreliert. Sind  $X_1$  und  $X_2$  unabhängig und gilt  $\mathbb{E}|X_i| < \infty$ , so folgt aus  $\mathbb{E}(X_1 X_2) = \mathbb{E}(X_1)\mathbb{E}(X_2)$ , dass

$$\text{Cov}(X_1, X_2) = \text{Corr}(X_1, X_2) = 0.$$

Die Umkehrung gilt typischerweise nicht. Weiterhin gilt die so genannte Cauchy-Schwarz Ungleichung

$$\text{Cov}(X, Y) \leq \text{Var}(X) \cdot \text{Var}(Y). \quad (1.2)$$

Falls  $X_1, \dots, X_n$  endliche Varianzen haben, dann gilt

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i,j=1, i < j}^n \text{Cov}(X_i, X_j).$$

Sind also  $X_1, \dots, X_n$  darüber hinaus paarweise unkorreliert (das folgt natürlich aus deren Unabhängigkeit) so gilt also die wichtige Regel von *Bienaymé*

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i).$$

**Momentenerzeugende Funktion.** Mitunter ist es günstig, zur Beschreibung der Verteilung einer Zufallsvariable ein weiteres Hilfsmittel zur Verfügung zu haben. Ein solches ist die so genannte momentenerzeugende Funktion  $\Psi_X$ . Ist  $X$  eine reellwertige Zufallsvariable, so ist  $\Psi_X$  definiert durch

$$\Psi_X(s) := \mathbb{E}(e^{sX}), \quad s \in \mathbb{R}.$$

Ist  $\Psi_X$  wohldefiniert, so ist  $\Psi_X(s)$  eindeutig und bestimmt eindeutig die Verteilung von  $X$ . Darüber hinaus ist

$$\left. \frac{d^k}{ds^k} \Psi_X(s) \right|_{s=0} = \mathbb{E}(X^k).$$

$\Psi_X$  wird sich für die Beschreibung der Verteilung von Summen von unabhängigen Zufallsvariablen als extrem nützlich erweisen. Denn, sind  $X_1, \dots, X_n$  unabhängig, so folgt

$$\Psi_{X_1 + \dots + X_n}(s) = \prod_{i=1}^n \Psi_{X_i}(s).$$

Anders als die momentenerzeugende Funktion existiert die *charakteristische Funktion*  $\varphi_X(s) := \mathbb{E}(\exp(isX))$  stets für alle  $s \in \mathbb{R}$ . Sie charakterisiert ebenso die Verteilung eindeutig.

## 1.2 Klassische Verteilungen der Statistik

In diesem Abschnitt werden die klassischen Verteilungen kurz und knapp definiert.

**Diskrete Verteilungen.** Wir betrachten eine diskrete Zufallsvariable  $X$  mit Wahrscheinlichkeitsfunktion  $p$ .

- *Binomialverteilung:* Wir schreiben  $X \sim \text{Bin}(n, p)$  falls  $k \in \mathbb{N}$ ,  $p \in (0, 1)$  und für  $k \in \{0, \dots, n\}$

$$p(k) = \binom{n}{k} \theta^k (1 - p)^{n-k}.$$

Als Spezialfall erhält man die *Bernoulli-Verteilung*  $\text{Bin}(1, p)$ . Dies ist eine Zufallsvariable welche nur die Werte 0 oder 1 annimmt. Jede Binomialverteilung läßt sich als Summe von Bernoulli-Zufallsvariablen schreiben, siehe Beispiel 1.3 und Aufgabe 1.4.

- *Poissonverteilung:* Wir schreiben  $X \sim \text{Poiss}(\lambda)$  falls  $\lambda > 0$  und für  $k \in \{0, \dots, n\}$

$$p(k) = \frac{e^{-\lambda} \lambda^k}{k!}. \quad (1.3)$$

- *Multinomialverteilung:* Wir schreiben  $\mathbf{X} \sim M(n, p_1, \dots, p_d)$ , falls  $\mathbf{X} \in \mathbb{N}^d$  und die Wahrscheinlichkeit, dass für Zahlen  $k_1, \dots, k_d \in \{0, \dots, n\}$  mit  $\sum_{i=1}^d k_i = n$  gilt, dass

$$\mathbb{P}(\mathbf{X} = (k_1, \dots, k_d)^\top) = \frac{n!}{k_1! \dots k_d!} p_1^{k_1} \dots p_d^{k_d}.$$

Diese Verteilung entsteht durch die Klassifizierung von  $n$  Objekten in  $d$  Klassen.  $k_i$  repräsentiert die Anzahl der Objekte in Klasse  $i$ ,  $i = 1, \dots, d$ .

**Laplacesche Modelle.** Betrachtet man  $\Omega = \{\omega_1, \dots, \omega_n\}$  so erhält man die wichtige Teilklasse der *Laplaceschen Modelle* falls  $\mathbb{P}(\{\omega_i\}) = \mathbb{P}(\{\omega_j\}) = n^{-1}$ . Alle Elementarereignisse haben also die gleiche Wahrscheinlichkeit. Damit ergibt sich

$$\mathbb{P}(A) = \sum_{w_i \in A} \mathbb{P}(\{w_i\}) = \sum_{w_i \in A} \frac{1}{|\Omega|} = \frac{|A|}{|\Omega|},$$

also die wohlbekannte Regel, wonach die Wahrscheinlichkeit eines Ereignisses durch die Formel „Günstige durch Mögliche“ berechnet werden kann. Dies gilt also nur unter der Annahme, dass alle Elementarereignisse die gleiche Wahrscheinlichkeit haben. Das folgende Beispiel werden wir in Kapitel 2 auf Seite 22 wieder aufgreifen.

**Beispiel 1.2.** (*Hypergeometrische Verteilung.*) Man betrachtet eine Menge mit  $N$  Elementen, wobei jedes Element den Wert 0 oder 1 annehmen kann. Alle Möglichkeiten seien gleich wahrscheinlich, es handelt sich also um ein Laplacesches Modell mit  $\Omega = \{0, 1\}^N$ . Es werde eine Teilmenge aus  $n$  Elementen ausgewählt und man untersucht die Anzahl  $X$  der Elemente in der Teilmenge, welche den Wert 1 haben. Man erhält die *Hypergeometrische Verteilung*

$$P(X = k) = \frac{\binom{N\theta}{k} \binom{N-N\theta}{n-k}}{\binom{N}{n}},$$

oder kurz  $X \sim \text{Hypergeo}(N, n, \theta)$  durch Abzählen der möglichen Kombinationen: Insgesamt gibt es  $\binom{N}{n}$  Möglichkeiten, aus  $N$  Teilen eine Stichprobe des Umfangs  $n$  zu ziehen (Mögliche). Sollen davon  $k \in \{0, \dots, n\}$  Teile defekt sein, so verbleiben zum einen  $\binom{N\theta}{k}$  Möglichkeiten,  $k$  defekte Teile in der Stichprobe aus  $N\theta$  defekten Teilen der Ladung zu ziehen. Zum anderen gibt es  $\binom{N-N\theta}{n-k}$  Möglichkeiten  $n-k$  nicht defekte Teile aus insgesamt  $N - N\theta$  nicht defekten Teilen auszuwählen. Diese Verteilung ist in Abbildung 1.1 dargestellt.

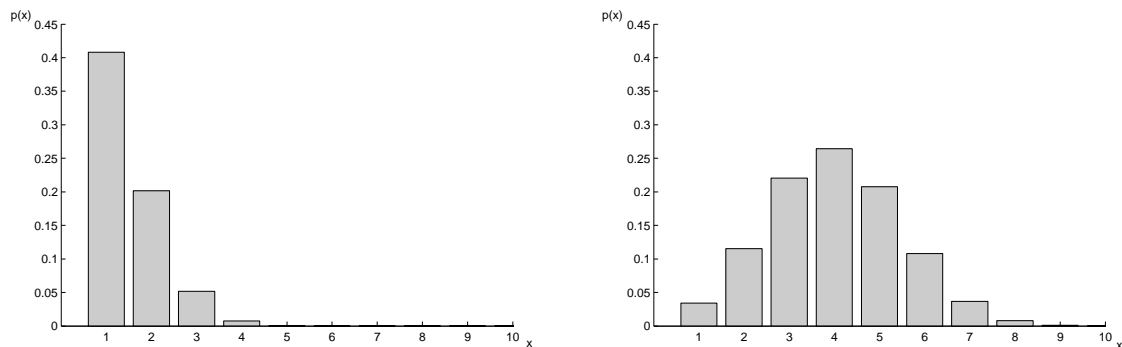


Abbildung 1.1: Wahrscheinlichkeitsfunktion der hypergeometrischen Verteilung aus Beispiel 1.2 mit  $N = 100$ ,  $n = 10$  und  $\theta = 0.1$  (links) bzw.  $\theta = 0.4$  (rechts).

**Stetige Verteilungen.** Wenn die beobachteten Daten nicht einer diskreten Wertemenge unterliegen, so wird man mit stetigen Verteilungen arbeiten. Zu Beginn seien einige wichtige Beispiele von Zufallsvariablen mit Dichte  $p$  vorgestellt.

- *Exponentialverteilung:* Wir schreiben  $X \sim \text{Exp}(\lambda)$  falls  $\lambda > 0$  und

$$p(x) = 1_{\{x>0\}} \lambda e^{-\lambda x}.$$

- *Gleichverteilung:* Wir schreiben  $X \sim U(a, b)$  falls  $a < b$  und

$$p(x) = \frac{1}{b-a} 1_{\{x \in [a,b]\}}.$$

- *Normalverteilung:* Wir schreiben  $X \sim \mathcal{N}(\mu, \sigma^2)$  falls  $\mu \in \mathbb{R}$  und  $\sigma > 0$  und

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} \quad (1.4)$$

Dann gilt, dass  $\mathbb{E}(X) = \mu$  und  $\text{Var}(X) = \sigma^2$ . Die Dichte ist in Abbildung 1.2 dargestellt. Die Normalverteilung ist mit Abstand die wichtigste Verteilung in der Statistik, da sie durch den Zentralen Grenzwertsatz, siehe Georgii (2004) Seite ..., zur Approximation der Verteilung von einer hinreichend großen Zahl unabhängiger und identisch verteilter Zufallsvariablen benutzt werden kann. Die Normalverteilung ist stabil unter Summenbildung, siehe Aufgabe 1.12.

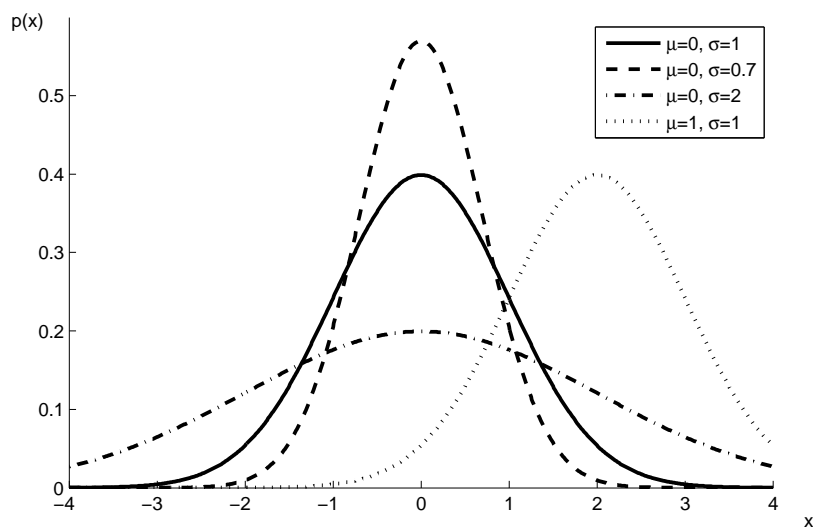


Abbildung 1.2: Dichte der Normalverteilung für verschiedene Parameterkonstellationen.

Die Exponentialverteilung ist ein Spezialfall der Gamma-Verteilung und die Gleichverteilung ein Spezialfall der Beta-Verteilung welche auf Seite 14 eingeführt werden.

Rund um die Normalverteilung und die Schätzung von  $\mu$  und  $\sigma^2$  gibt es eine Familie von unerlässlichen Verteilungen, welche nun kurz vorgestellt werden.

### Die $\chi^2$ , F und t-Verteilung:

Die  $\chi^2$ -Verteilung entsteht als Summe von quadrierten normalverteilten Zufallsvariablen.

**Lemma 1.6.** (und Definition) Für  $X_1, \dots, X_n$  unabhängig und  $\mathcal{N}(0, 1)$ -verteilt definieren wir die Verteilung von  $V := \sum_{i=1}^n X_i^2$  als  $\chi^2$ -Verteilung mit  $n$  Freiheitsgraden, kurz  $\chi_n^2$ . Die Dichte von  $V$  ist gegeben durch

$$f_{\chi_n^2}(x) = 1_{\{x>0\}} \frac{1}{2^{n/2} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}. \quad (1.5)$$

Hierbei ist  $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$ ,  $a > 0$  die Gamma-Funktion mit  $\Gamma(n) = (n-1)!$ ,  $n \in \mathbb{N}$  und  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ . Weiterhin gilt  $\mathbb{E}(V) = n$  und  $\text{Var}(V) = 2n$ .

**Bemerkung 1.7.** Wir erhalten, dass  $V$  für  $n = 2$  gerade exponentialverteilt mit Parameter  $\frac{1}{2}$  ist. Außerdem ist  $\mathbb{E}(\chi_n^2) = n$  und  $\text{Var}(\chi_n^2) = 3n - n = 2n$ . Aus dem ZGWS folgt damit, dass

$$\frac{\chi_n^2 - n}{\sqrt{2n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Möchte man ein Konfidenzintervall für den Mittelwert einer Normalverteilung mit unbekannter Varianz bilden, so muss man die Varianz schätzen. Dabei taucht die Wurzel einer Summe von Normalverteilungsquadraten (mit Faktor  $\frac{1}{n}$ ) im Nenner auf. Hierüber gelangt man zur  $t$ -Verteilung.

**Definition 1.8.** Ist  $X$  standardnormalverteilt und  $V$  gerade  $\chi_n^2$ -verteilt, so heißt die Verteilung von

$$T := \frac{X}{\sqrt{\frac{1}{n}V}} \tag{1.6}$$

$t$ -Verteilung mit  $n$  Freiheitsgraden, kurz  $t_n$ -Verteilung.

**Lemma 1.9.** Die Dichte der  $t_n$ -Verteilung ist gegeben durch

$$t_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(n/2)\Gamma(1/2)\sqrt{n}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

für alle  $x \in \mathbb{R}$ .

Für Vergleiche von Varianzen werden wir Quotienten der Schätzer betrachten und gelangen so zur  $F$ -Verteilung.

**Definition 1.10.** Sind  $V_1$  und  $V_2$  unabhängig und  $\chi_n^2$  bzw.  $\chi_m^2$ -verteilt, so heißt die Verteilung von

$$F := \frac{V_1/n}{V_2/m}$$

$F$ -Verteilung mit  $(n, m)$ -Freiheitsgraden, kurz  $F_{n,m}$ -Verteilung.

Für die Dichte sei an die Formel für die *Beta-Funktion*  $B(a, b)$  erinnert: Für  $a, b > 0$  ist

$$B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt. \tag{1.7}$$

Dann ist  $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ . Damit erhalten wir folgende Darstellung.

**Lemma 1.11.** Die Dichte der  $F_{n,m}$ -Verteilung ist

$$f_{F_{n,m}}(x) = 1_{\{x>0\}} \frac{n^{n/2} m^{m/2}}{B(n/2, m/2)} \frac{x^{\frac{n}{2}-1}}{(m+nx)^{n+m/2}}.$$

*Beweis.* Für die Verteilungsfunktion erhalten wir aufgrund der Unabhängigkeit von  $V_1$  und  $V_2$

$$\begin{aligned} \mathbb{P}\left(\frac{V_1/n}{V_2/m} \leq t\right) &= \int_{x \leq tym/n} f_{\chi_n^2}(x) f_{\chi_m^2}(y) dx dy \\ &= \int_0^\infty f_{\chi_m^2}(y) \left[ \int_0^{tym/n} f_{\chi_n^2}(x) dx \right] dy. \end{aligned}$$

Da wir die Dichte bestimmen wollen, transformieren wir das zweite Integral mittels  $w = mx/(ny)$  und erhalten, dass

$$\begin{aligned} \mathbb{P}\left(\frac{V_1/n}{V_2/m} \leq t\right) &= \int_0^\infty f_{\chi_m^2}(y) \int_0^t f_{\chi_n^2}(w^{ny/m}) \frac{ny}{m} dx dy \\ &= \int_0^t \left[ \int_0^\infty f_{\chi_m^2}(y) f_{\chi_n^2}(w^{ny/m}) \frac{ny}{m} dy \right] dx. \end{aligned}$$

Der Ausdruck in der Klammer gibt also die Dichte an. Unter Verwendung von (1.5) ergibt sich die Behauptung.  $\square$

**Bemerkung 1.12.** Die *Rayleigh-Verteilung*. Ein Rayleigh-verteilte Zufallsvariable  $X$  ist nicht negativ und hat die Dichte

$$p(x, \sigma) = \frac{x}{\sigma^2} \exp\left(-\frac{x}{2\sigma^2}\right) 1_{\{x \geq 0\}}.$$

Die Rayleigh-Verteilung entsteht als Betrag einer zweidimensionalen, zentrierten Normalverteilung: Seien  $Y, Z$  unabhängig und jeweils  $\mathcal{N}(0, \sigma^2)$ -verteilt. Dann ist  $\sqrt{Y^2 + Z^2}$  Rayleigh-verteilt, siehe Aufgabe 1.13. Deswegen ist  $X^2$  gerade  $\chi_2^2$ -verteilt.

## Die Beta- und die Gamma-Verteilung

**Definition 1.13.** Eine Zufallsvariable  $X$  heißt *Gamma-verteilt* zu den Parametern  $a, \lambda > 0$ , falls sie die Dichte

$$p_{a,\lambda}(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}.$$

besitzt.

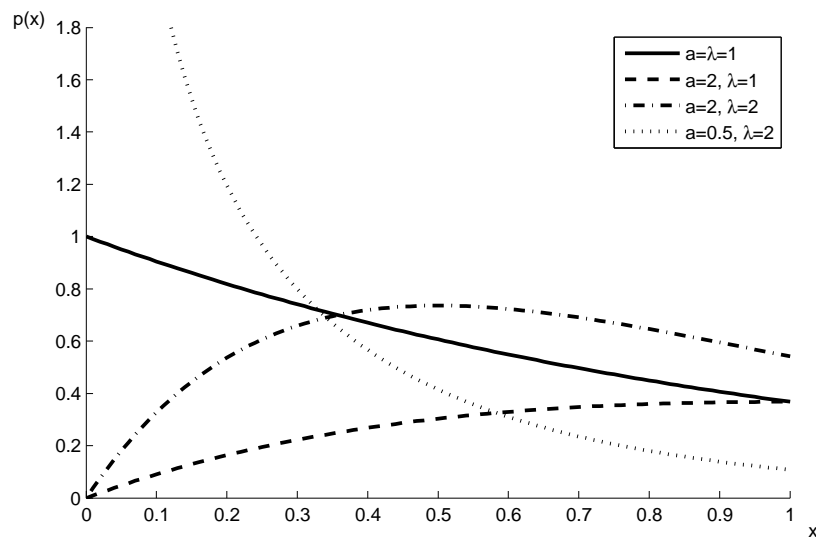


Abbildung 1.3: Dichte der  $\text{Gamma}(a, \lambda)$ -Verteilung für verschiedene Parameterkonstellationen. Für  $a = 1$  erhält man eine Exponentialverteilung.

Ist  $X$  Gamma-verteilt, so schreiben wir kurz  $X \sim \text{Gamma}(a, \lambda)$ . Die Summe von unabhängigen  $\text{Gamma}(\cdot, \lambda)$ -verteilten Variablen ist wieder Gamma-verteilt: Seien  $X_1, \dots, X_n$  unabhängig mit  $X_i \sim \text{Gamma}(a_i, \lambda)$ , so ist

$$\sum_{i=1}^n X_i \sim \text{Gamma}\left(\sum_{i=1}^n a_i, \lambda\right).$$

Der Beweis kann über die momentenerzeugende Funktion erfolgen, siehe Aufgabe 1.9. Weiterhin ist eine  $\chi_n^2$ -verteilte Zufallsvariable gerade  $\text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$ -verteilt. Als weiteren Spezialfall erhält man die Exponentialverteilung zum Parameter  $\lambda$  für  $a = 1$ .

**Definition 1.14.** Eine Zufallsvariable heißt *Beta*-verteilt zu den Parametern  $a, b > 0$ , falls sie die Dichte

$$p_{a,b}(x) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} 1_{\{x \in [0,1]\}}$$

hat.

Hierbei ist  $B(a, b)$  die Beta-Verteilung, siehe Gleichung (1.7). Für  $a = b = 1$  erhält man die Gleichverteilung auf  $[0, 1]$  als Spezialfall. Erwartungswert einer  $\text{Beta}(a, b)$ -Verteilung

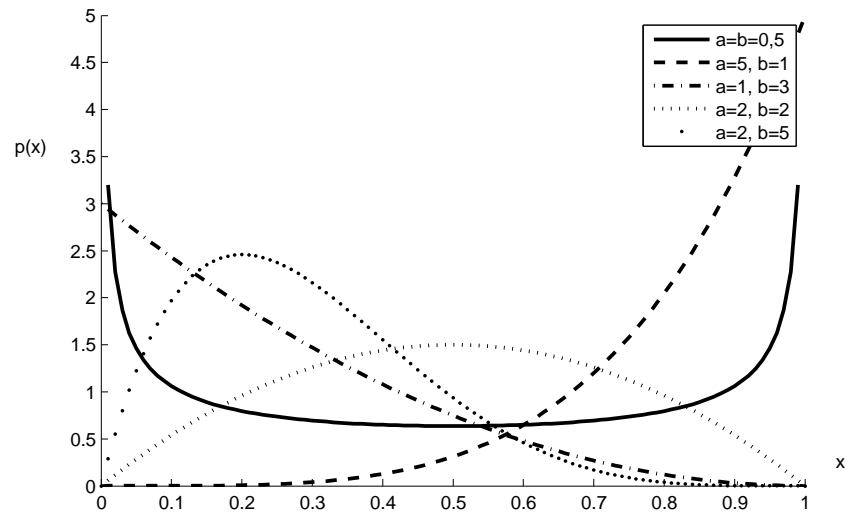


Abbildung 1.4: Dichte der Beta-Verteilung für verschiedene Parameterkonstellationen.

ist  $a/a+b$  und die Varianz ist

$$\frac{ab}{(1+a+b)(a+b)^2}.$$

**Bemerkung 1.15.** Sind  $X, Y$  unabhängig und  $Gamma(a, b)$  bzw.  $Gamma(a, c)$ -verteilt, so ist  $X/(X+Y)$   $Beta(b, c)$ -verteilt, siehe Aufgabe 1.7.

### 1.3 Bedingte Verteilungen

Wir setzen die Einführung in die notwendigen Hilfsmittel mit bedingten Verteilungen und dem wichtigen bedingten Erwartungswert fort.

**Bedingte Verteilungen.** Bedingten Verteilungen verallgemeinern den Begriff der bedingten Wahrscheinlichkeit wesentlich und bilden ein wichtiges Hilfsmittel.

Im diskreten Fall geht man eigentlich analog zu dem schon eingeführten Begriff der bedingten Wahrscheinlichkeit vor. Seien  $X, Y$  diskrete Zufallsvariablen. Die *bedingte Verteilung* von  $X$  gegeben  $Y = y$  mit  $\mathbb{P}(Y = y) > 0$  ist definiert durch die Wahrscheinlichkeitsfunktion

$$p(x|y) := \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p(x, y)}{p_Y(y)}. \quad (1.8)$$

Für stetige Zufallsvariablen  $X, Y$  mit gemeinsamer Dichte  $p(x, y)$  definiert man analog für  $p_y(y) > 0$

$$p(x|y) := \frac{p(x, y)}{p_Y(y)}. \quad (1.9)$$

**Beispiel 1.3.** (*Bernoulli-Verteilung*) Die Summe von unabhängigen Bernoulli-Zufallsvariablen ist Binomialverteilt: Eine Zufallsvariable  $X$  heißt *Bernoulli-verteilt* falls  $X \in \{0, 1\}$ . Seien  $X_1, \dots, X_n$  unabhängig und Bernoulli-verteilt mit  $\mathbb{P}(X_1 = 1) = p$ . Dann ist  $Y := \sum_{i=1}^n X_i$  gerade  $\text{Bin}(n, p)$ -verteilt, siehe Aufgabe 1.4.

**Beispiel 1.4.** (*Fortsetzung.*) Setze  $\mathbf{X} = (X_1, \dots, X_n)^\top$ . Dann ist die Verteilung von  $\mathbf{X}$  gegeben  $Y$  gerade eine Gleichverteilung: Für  $\mathbf{x} \in \{0, 1\}^n$  mit  $\sum x_i = y$  gilt

$$\mathbb{P}(\mathbf{X} = \mathbf{x} | Y = y) = \frac{\mathbb{P}(\mathbf{X} = \mathbf{x}, Y = y)}{\mathbb{P}(Y = y)} = \frac{p^y(1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} = \binom{n}{y}^{-1}.$$

Also hat  $\mathbf{X} | Y = y$  eine (diskrete) Gleichverteilung auf  $\{\mathbf{x} \in \{0, 1\}^n : \sum_{i=1}^n x_i = y\}$ .

Sei  $X$  eine Zufallsvariable mit Werten  $x_1, x_2, \dots$  und  $\mathbb{E}(|X|) < \infty$ . Der *bedingte Erwartungswert* von  $X$  gegeben  $Y = y$  ist definiert durch

$$\mathbb{E}(X | Y = y) := \sum_{i \geq 1} x_i p(y_i | y).$$

Ist  $X$  eine stetige Zufallsvariable mit  $\mathbb{E}(|X|) < \infty$  so ist der *bedingte Erwartungswert* von  $X$  gegeben  $Y = y$  definiert durch

$$\mathbb{E}(X | Y = y) := \int xp(x|y), dx.$$

Sei  $g(y) := \mathbb{E}(X | Y = y)$ , dann heißt die Zufallsvariable

$$g(Y) := \mathbb{E}(X | Y)$$

bedingter Erwartungswert von  $X$  gegeben  $Y$ .

**Beispiel 1.5.** (*Suffiziente Statistik in der Bernoulli-Verteilung*) Wir setzen Beispiel 1.3 fort. Mit der dortigen Notation gilt

$$\begin{aligned} \mathbb{E}(X_1 | Y = y) = \mathbb{P}(X_1 = 1 | Y = y) &= \frac{p \binom{n-1}{p-1} p^{y-1} (1-p)^{(n-1)-(y-1)}}{\binom{n}{y} p^y (1-p)^{n-y}} \\ &= \binom{n-1}{y-1} \cdot \binom{n}{y}^{-1} = \frac{y}{n} \end{aligned}$$

Damit ergibt sich der mittlere Wert von  $X_1$  gegeben  $Y$  durch  $\mathbb{E}(X_1|Y) = Yn^{-1}$ . Man beachte, dass dies eine Zufallsvariable ist.

**Bemerkung 1.16.** Sind  $X$  und  $Y$  unabhängig, so gibt  $Y$  keine neue Information über  $X$  und der bedingte Erwartungswert ist gleich dem unbedingten Erwartungswert, da

$$p(x|y) = \frac{p(x, y)}{p_Y(y)} = \frac{p_X(x)p_Y(y)}{p_Y(y)} = p_X(x)$$

und somit  $\mathbb{E}(X|Y = y) = \mathbb{E}(X)$  und auch  $\mathbb{E}(X|Y) = \mathbb{E}(X)$ .

Bedingte Erwartungswerte lassen sich analog auf Zufallsvektoren verallgemeinern. Betrachtet man zwei Zufallsvektoren  $\mathbf{X} = (X_1, \dots, X_n)^T$  und  $\mathbf{Y} = (Y_1, \dots, Y_m)^T$  und sind entweder beide diskret mit gemeinsamer Wahrscheinlichkeitsfunktion  $\mathbb{P}(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = p(\mathbf{x}, \mathbf{y})$  oder beide stetig mit gemeinsamer Dichte  $p(\mathbf{x}, \mathbf{y})$  so definiert man analog zu (1.8) und (1.9) die bedingte Wahrscheinlichkeitsfunktion bzw. Dichte von  $\mathbf{X}$  gegeben  $\mathbf{Y} = \mathbf{y}$  durch

$$p(\mathbf{x}|\mathbf{y}) := \frac{p(\mathbf{x}, \mathbf{y})}{p_{\mathbf{Y}}(\mathbf{y})} 1_{\{p_{\mathbf{Y}}(\mathbf{y}) > 0\}}.$$

Der *bedingte Erwartungswert* ist nun definiert durch

$$\mathbb{E}(\mathbf{X}|\mathbf{Y} = \mathbf{y}) = (\mathbb{E}(X_1|\mathbf{Y} = \mathbf{y}), \dots, \mathbb{E}(X_n|\mathbf{Y} = \mathbf{y}))^T.$$

**Satz 1.17** (Substitutionssatz). *Sei  $q : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$  eine messbare Abbildung. Gilt für  $\mathbf{y} \in \mathbb{R}^m$ , dass  $p_{\mathbf{Y}}(\mathbf{y}) > 0$  und  $\mathbb{E}|q(\mathbf{X}, \mathbf{y})| < \infty$ , so ist*

$$\mathbb{E}(q(\mathbf{X}, \mathbf{Y})|\mathbf{Y} = \mathbf{y}) = \mathbb{E}(q(\mathbf{X}, \mathbf{y})|\mathbf{Y} = \mathbf{y}).$$

Ein typischer Spezialfall ist  $q(\mathbf{X}, \mathbf{y}) = r(\mathbf{X})h(\mathbf{y})$  mit einer beschränkten Funktion  $h$ . Hat  $r(\mathbf{X})$  eine endliche Erwartung, so ist

$$\mathbb{E}(r(\mathbf{X})h(\mathbf{Y})|\mathbf{Y} = \mathbf{y}) = \mathbb{E}(r(\mathbf{X})h(\mathbf{y})|\mathbf{Y} = \mathbf{y}) = h(\mathbf{y})\mathbb{E}(r(\mathbf{X})|\mathbf{Y} = \mathbf{y}).$$

Daraus folgt nun  $\mathbb{E}(r(\mathbf{X})h(\mathbf{Y})|\mathbf{Y}) = h(\mathbf{Y})\mathbb{E}(r(\mathbf{X})|\mathbf{Y})$ . Oft hat man die zusätzliche Annahme, dass  $\mathbf{X}$  und  $\mathbf{Y}$  unabhängig sind. Dann folgt unter den obigen Annahmen sogar, dass

$$\mathbb{E}(q(\mathbf{X}, \mathbf{Y})|\mathbf{Y} = \mathbf{y}) = \mathbb{E}(q(\mathbf{X}, \mathbf{y})).$$

Der Erwartungswert der bedingten Erwartung ist gleich dem Erwartungswert selbst. Dies ist Inhalt des Satzes vom doppelten Erwartungswert.

**Satz 1.18.** *Unter  $\mathbb{E}(\mathbf{X}) < \infty$  gilt*

$$\mathbb{E}(\mathbf{X}) = \mathbb{E}(\mathbb{E}(\mathbf{X}|\mathbf{Y})).$$

*Beweis.* Wir beweisen den eindimensionalen Fall, der mehrdimensionale Fall folgt analog. Zunächst seien  $X$  und  $Y$  diskrete Zufallsvariablen welche die Werte  $\{x_1, x_2, \dots\}$  bzw.  $\{y_1, y_2, \dots\}$  annehmen. Dann gilt

$$\begin{aligned} \mathbb{E}(\mathbb{E}(X|Y = y)) &= \sum_{i \geq 1} p_Y(y_i) \left( \sum_{j \geq 1} x_j p(x_j|y_i) \right) \\ &= \sum_{i,j \geq 1} \frac{x_j p(x_j, y_i)}{p_Y(y_i)} p_Y(y_i) = \sum_{i,j \geq 1} p(x_i, y_j) = \sum_{j \geq 1} x_j p_X(x_j) = \mathbb{E}(X). \end{aligned}$$

Für den Beweis des stetigen Fall sei auf Aufgabe 1.10 verwiesen.  $\square$

**Beispiel 1.6.** (*Minima und Maxima von gleichverteilten Zufallsvariablen.*) Seien  $X_1, X_2$  unabhängig und jeweils  $U(0, 1)$ -verteilt. Setze  $Y := \min(X_1, X_2)$  und  $Z := \max(X_1, X_2)$ . Im folgenden sei  $x, y, z$  stets in  $(0, 1)$ . Die gemeinsame Verteilungsfunktion von  $Y$  und  $Z$  ist

$$\begin{aligned} F(y, z) &= \mathbb{P}(Y \leq y, Z \leq z) = 2 \mathbb{P}(X_1 < X_2, X_1 \leq y, X_2 \leq z) \\ &= 2 \int_0^z \int_0^{\min(x_2, y)} dx_1 dx_2 = 2 \cdot \begin{cases} \frac{z^2}{2} & z < y \\ zy - \frac{y^2}{2} & z \geq y \end{cases} \end{aligned}$$

Die gemeinsame Dichte erhält man durch partielles Ableiten der Verteilungsfunktion:

$$p(y, z) = \frac{\partial F(y, z)}{\partial y \partial z} = 2 \begin{cases} 0 & z < y \\ 1 & z \geq y \end{cases} = 2 \mathbf{1}_{\{z \geq y\}}.$$

Die Dichte von  $Y$  ist

$$p_Y(y) = \int_0^1 p(y, z) dz = \int_y^1 2 dz = 2(1 - y).$$

Damit zeigt sich, dass das Maximum  $Z$  gegeben  $Y$  auf  $(y, 1)$  gleichverteilt ist:

$$p(z|Y = y) = \frac{p(y, z)}{p_Y(y)} = \frac{2}{2(1 - y)} \mathbf{1}_{\{z \geq y\}}.$$

## 1.4 Gesetz der großen Zahl

(XXX) TODO

**Satz 1.19.** *Schwaches Gesetz der großen Zahl*

**Satz 1.20.** *Starkes Gesetz der großen Zahl*

**Satz 1.21.** *Continuous Mapping Theorem*

## 1.5 Aufgaben

**Aufgabe 1.1.** Sei  $B \in \mathcal{A}$  ein Ereignis mit  $\mathbb{P}(B) > 0$ . Dann ist durch  $\mu(A) := \mathbb{P}(A|B) : \mathcal{A} \rightarrow [0, 1]$  ein Wahrscheinlichkeitsmaß definiert.

**Aufgabe 1.2.** Die Potenzmenge  $\mathcal{P}(\Omega)$  ist eine  $\sigma$ -Algebra.

**Aufgabe 1.3.** Seien  $X_1, \dots, X_n$  i.i.d. mit Varianz  $\sigma^2$ . Die Stichprobenvarianz ist  $s^2(\mathbf{X}) := (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ . Dann gilt  $\mathbb{E}(s^2(\mathbf{X})) = \sigma^2$ , d.h. die Stichprobenvarianz ist erwartungstreu.

**Aufgabe 1.4.** *Darstellung der Binomialverteilung als Summe von unabhängigen Bernoulli-Zufallsvariablen.* Seien  $X_1, \dots, X_n$  i.i.d. mit  $X_i \in \{0, 1\}$  und  $\mathbb{P}(X_i = 1) = p \in (0, 1)$ . Dann ist

$$\sum_{i=1}^n X_i \sim \text{Bin}(n, p).$$

**Aufgabe 1.5.** Zeigen Sie, dass für eine Poisson-verteilte Zufallsvariable  $X$

$$\mathbb{E}(X) = \text{Var}(X) = \lambda.$$

**Aufgabe 1.6.** Die Exponentialverteilung ist gedächtnislos: Sei  $X$  exponentialverteilt mit Intensität  $\lambda$ . Dann gilt für  $x, h > 0$

$$\mathbb{P}(X > x + h \mid X > x) = \mathbb{P}(X > h).$$

**Aufgabe 1.7.** Seien  $X \sim \text{Gamma}(a, \lambda)$  und  $Y \sim \text{Gamma}(b, \lambda)$  zwei unabhängige Zufallsvariablen. Dann ist  $X + Y \sim \text{Gamma}(a + b, \lambda)$  und

$$\frac{X}{X + Y} \sim \text{Beta}(a, b).$$

**Aufgabe 1.8.** *Unkorreliertheit impliziert nicht Unabhängigkeit.* Sei  $X \sim \mathcal{N}(0, 1)$  eine standardnormalverteilte Zufallsvariable und  $Y = X^2$ . Dann ist  $\text{Cov}(X, Y^2) = 0$ , aber  $X$  und  $Y$  sind nicht unabhängig.

**Aufgabe 1.9.** Die Summe von unabhängigen  $\text{Gamma}(\cdot, \lambda)$ -verteilten Variablen ist wieder Gamma-verteilt: Seien  $X_1, \dots, X_n$  unabhängig mit  $X_i \sim \text{Gamma}(a_i, \lambda)$ , so ist

$$\sum_{i=1}^n X_i \sim \text{Gamma}\left(\sum_{i=1}^n a_i, \lambda\right).$$

**Aufgabe 1.10.** Sei  $\mathbf{X}$  eine Zufallsvariable mit Dichte  $p_{\mathbf{X}}$  und  $\mathbb{E}(\mathbf{X}) < \infty$ . Dann gilt  $\mathbb{E}(\mathbf{X}) = \mathbb{E}(\mathbb{E}(\mathbf{X}|\mathbf{Y}))$ .

Um die Verteilung von Summen von unabhängigen Zufallsvariablen zu bestimmen, kann man zum einen mit der momentenerzeugenden Funktion oder der charakteristischen Funktion arbeiten, zum anderen auch mit der so genannten *Faltungsformel*:

**Aufgabe 1.11.** Haben  $X$  und  $Y$  die Dichten  $p_X$  und  $p_Y$  und sind beide unabhängig, so ist die Dichte von  $Z = X + Y$  gegeben durch

$$p_Z(z) = \int_{\mathbb{R}} p_X(x) p_Y(z - x) dx.$$

**Aufgabe 1.12.** *Die Summe von normalverteilten Zufallsvariablen ist wieder normalverteilt.* Seien  $X_1, \dots, X_n$  unabhängig und normalverteilt mit  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ . Dann ist die Summe wieder normalverteilt:

$$\sum_{i=1}^n X_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

**Aufgabe 1.13.** Seien  $X$  und  $Y$  unabhängig und  $\mathcal{N}(0, \sigma^2)$ -verteilt. Dann ist  $Z := \sqrt{X^2 + Y^2}$  Rayleigh-verteilt, d.h.  $Z$  hat Dichte  $x\sigma^{-2} \exp(-x^2/2\sigma^2)$ . Es gilt  $\mathbb{E}(Z) = \sigma\sqrt{\pi/2}$ ,  $\mathbb{E}(Z^2) = 2\sigma^2$  und  $\text{Var}(Z) = \sigma\sqrt{2 - \pi/2}$ .

## 2 Statistische Modelle

### 2.1 Formulierung von statistischen Modellen

Die Formulierung von statistischen Modellen wird anhand der folgenden beiden Beispiele illustriert.

**Beispiel 2.1.** (*Qualitätssicherung*) Eine Ladung von  $N$  Teilen soll auf ihre Qualität untersucht werden. Die Ladung enthält defekte und nicht defekte Teile. Mit  $\theta$  sei der Anteil der defekten Teile bezeichnet, von insgesamt  $N$  Teilen sind  $N\theta$  defekt. Aus Kostengründen wird nur eine Stichprobe von  $n \leq N$  Teilen an Stelle der ganzen Ladung untersucht. Zur Modellierung verwenden wir einen Wahrscheinlichkeitsraum mit Grundmenge  $\Omega = \{0, 1, \dots, n\}$ .  $\mathcal{A}$  sei die Potenzmenge<sup>1</sup> von  $\Omega$ . Die Zufallsvariable  $X$  bezeichne die Anzahl der defekten Teile in der Stichprobe. Erfolgt die Auswahl der Stichprobe zufällig, so kann man ein Laplacesches Modell (vergleiche Seite 10) rechtfertigen und erhält eine hypergeometrische Verteilung für  $X$ , siehe Beispiel 1.2:

$$P(X = k) = \frac{\binom{N\theta}{k} \binom{N-N\theta}{n-k}}{\binom{N}{n}} \quad (2.1)$$

für  $\max\{0, n - N(1 - \theta)\} \leq k \leq \min\{N\theta, n\}$ ; oder kurz  $X \sim \text{Hypergeo}(N, n, \theta)$ . Insgesamt kann man dieses Modell wie folgt zusammenfassen:

$$\{(\Omega, \mathcal{A}, \text{Hypergeo}(N, \Omega, \theta)), \theta \text{ unbekannt}\}.$$

Dies ist der erste Prototyp eines statistischen Modells bestehend aus einer Familie von Wahrscheinlichkeitsräumen. Der wesentliche Unterschied zum Wahrscheinlichkeitsraum besteht darin, dass das Wahrscheinlichkeitsmaß nur bis auf den Parameter  $\theta$  bekannt ist.

In dem zweiten Beispiel werden wir Messfehler untersuchen. Eine typische Annahme hierbei ist, dass der Messfehler symmetrisch um 0 verteilt ist.

**Definition 2.1.** Eine Zufallsvariable  $X$  heißt *symmetrisch um  $c$  verteilt*, falls  $X - c$  und  $-(X - c)$  die gleiche Verteilung besitzen. Dafür schreiben wir

$$X - c \stackrel{\mathcal{L}}{=} -(X - c) \quad (2.2)$$

---

<sup>1</sup>Dies ist stets eine  $\sigma$ -Algebra nach Aufgabe 1.2.

Hat  $X$  die Verteilungsfunktion  $F$  und ist  $F$  stetig, so ist (2.2) äquivalent zu  $F(c+t) = 1 - F(c-t)$  für alle  $t > 0$ . Hieraus folgt, dass für die Dichte  $p(c+t) = p(c-t)$  für alle  $t \geq 0$  gilt. Ist  $X$  diskret, so die Symmetrie von  $X$  um  $c$  sogar äquivalent zu  $p(c+t) = p(c-t)$  für alle  $t \geq 0$ .

Insbesondere gilt, dass eine Normalverteilung  $\mathcal{N}(\mu, \sigma^2)$  symmetrisch um  $\mu$  und eine Binomialverteilung  $\text{Bin}(n, \frac{1}{2})$  symmetrisch um  $\frac{n}{2}$  verteilt ist.

Das zweite Beispiel beschreibt typische Ergebnisse einer Messreihe, in welcher wiederholt eine Messung vorgenommen wird und die Messwerte um den gesuchten Parameter schwanken.

**Beispiel 2.2.** (*Messmodell*)  $n$  Messungen einer physikalischen Konstante  $\mu$  werden vorgenommen. Die Messergebnisse werden mit  $X_1, \dots, X_n$  bezeichnet. Man nimmt an, dass die Messungen einem Messfehler unterworfen sind, der *additiv* um  $\mu$  variiert:

$$X_i = \mu + \epsilon_i, \quad i = 1, \dots, n.$$

$\epsilon_i$  bezeichnet den Messfehler der  $i$ -ten Messung. Wir unterscheiden typische Annahmen, welche geringe, oft erfüllte Annahmen an physikalische Messungen beschreiben und weitere Annahmen, welche darüber hinaus die Berechnungen erleichtern. Diese weiteren Annahmen sollten allerdings einer kritischen Überprüfung unterzogen werden.

#### Typische Annahmen:

- (i) Die Verteilung von  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$  ist unabhängig von  $\mu$  (kein systematischer Fehler).
- (ii) Der Meßfehler der  $i$ -ten Messung beeinflusst den Meßfehler der  $j$ -ten Messung nicht, d.h.  $\epsilon_1, \dots, \epsilon_n$  sind unabhängig.
- (iii) Die Verteilung der einzelnen Meßfehler sind gleich, d.h.  $\epsilon_1, \dots, \epsilon_n$  sind identisch verteilt
- (iv) Die Verteilung von  $\epsilon_i$  ist stetig und symmetrisch um 0.

Aus diesen Annahmen folgt, dass  $X_i = \mu + \epsilon_i$  ist wobei  $\epsilon_i$  nach  $F$  verteilt ist, wobei  $F$  von  $\mu$  unabhängig ist, symmetrisch um 0 und Dichte  $f$  besitzt.

#### Weitere Annahmen:

- (v)  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ .
- (vi)  $\sigma^2$  ist bekannt.

Aus Annahme (v) folgt, dass  $X_i \sim N(\mu, \sigma^2)$  und  $X_1, \dots, X_n$  sind i.i.d. Unter Annahme (vi) ist  $\mu$  der einzige unbekannt Parameter, was die Handhabung des Modells wesentlich erleichtert. Bei einem konkreten Messdatensatz ist zu diskutieren, welche Annahmen realistisch für das Experiment sind.

## Statistische Modelle

Das Ergebnis eines Zufallsexperiments ist ein Zufallsvektor  $\mathbf{X} = (X_1, \dots, X_n)^\top$ . Falls man konkrete Daten  $\mathbf{x} = (x_1, \dots, x_n)^\top$  beobachtet, so ist dies gleichbedeutend mit dem Ereignis  $\{\mathbf{X}(\omega) = \mathbf{x}\}$ . Wir verwenden stets die Bezeichnung  $\mathbf{X}$  für die Zufallsvariable und  $\mathbf{x}$  für konkrete, nicht zufällige Daten.

**Definition 2.2.** Unter einem *statistischen Modell* verstehen wir ganz allgemein eine Familie  $\mathcal{P}$  von Verteilungen. Wir setzen voraus, dass  $\mathcal{P}$  die Darstellung

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\}$$

besitzt.  $\Theta$  heißt Parameterraum.

In dem Beispiel 2.1 von der Qualitätssicherung ist das statistische Modell gerade

$$\mathcal{P} = \{\text{Hypergeo}(N, n, \theta), \theta \in [0, 1]\}.$$

In dem Beispiel 2.2 (Messfehler) führen die unterschiedlichen Annahmen zu jeweils unterschiedlichen statistischen Modellen: Unter den Annahmen (i)-(iv) erhält man als statistisches Modell

$$\mathcal{P} = \{X_1, \dots, X_n \text{ i.i.d. } \sim F, F \text{ ist symmetrisch um } \mu\}.$$

Nimmt man die Normalverteilungssannahme hinzu, erhält man unter (i)-(v)

$$\mathcal{P} = \{X_1, \dots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}.$$

Hierbei sind sowohl  $\mu$  als auch  $\sigma$  unbekannt. Im Gegensatz zu dem *interessierenden* Parameter  $\mu$  ist  $\sigma$  nicht primär von Interesse, muss aber ebenso geschätzt werden. Man nennt einen solchen Parameter *Störparameter* (Nuisance Parameter).. Unter den Annahmen (i)-(vi) ist  $\sigma$  darüber hinaus bekannt und man erhält als statistisches Modell

$$\mathcal{P} = \{X_1, \dots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}\}.$$

Es gibt viele Möglichkeiten ein *Modell zu parametrisieren*. Jede bijektive Funktion  $g(\boldsymbol{\theta})$  kann als Parametrisierung gewählt werden. Es sollten jedoch Parametrisierungen gewählt werden, die eine Interpretation zulassen. Manchmal verlieren solche Parametrisierungen ihre Eindeutigkeit, in diesem Fall spricht man von der *Nichtidentifizierbarkeit* von Parametern.

**Definition 2.3.** Ein statistisches Modell heißt *identifizierbar*, falls für alle  $\theta_1, \theta_2 \in \Theta$  gilt, dass

$$\theta_1 \neq \theta_2 \Rightarrow \mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}.$$

**Beispiel 2.3.** (*Ein nicht identifizierbares Modell*) Seien  $X_1 \sim \mathcal{N}(\mu + \alpha_1, 1)$  und  $X_2 \sim \mathcal{N}(\mu + \alpha_2, 1)$  unabhängig. Setzen wir  $\theta = (\mu, \alpha_1, \alpha_2)^\top$ , so erhalten wir ein statistisches Modell durch<sup>2</sup>

$$\mathcal{P}_\theta = \{\mathcal{N}(\mu + \alpha_1, 1) \otimes \mathcal{N}(\mu + \alpha_2, 1), \mu \in \mathbb{R}, \alpha_i \in \mathbb{R}\}.$$

Hierbei stellt  $\mu$  den Gesamteffekt (Overall Effekt) dar und  $\alpha_i$  den Faktoreffekt. Betrachtet man

$$\theta_1 = (2, 0, 0)^\top \Rightarrow X_1 \sim \mathcal{N}(2, 1), X_2 \sim \mathcal{N}(2, 1)$$

$$\theta_2 = (1, 1, 1)^\top \Rightarrow X_1 \sim \mathcal{N}(2, 1), X_2 \sim \mathcal{N}(2, 1),$$

so folgt, dass  $\mathcal{P}_{\theta_1} = \mathcal{P}_{\theta_2}$ , der Faktoreffekt vermischt sich mit dem Gesamteffekt. Allerdings ist  $\theta_1 \neq \theta_2$ , d.h. dieses statistische Modell ist nicht Identifizierbar. Eine weitere Einschränkung wie etwa  $\alpha_1 + \alpha_2 = 0$  kann zur Identifizierbarkeit genutzt werden.

Ist  $\Theta \subset \mathbb{R}^k$  so spricht man von einem *parametrischen Modell*, ansonsten von einem *nichtparametrischen Modell*. Zum Beispiel implizieren die Zustandsräume

$$\Theta = \{F : F \text{ ist Verteilungsfunktion symmetrisch um } \mu\} \quad \text{oder}$$

$$\Theta = \{(\mu, f) : \mu \in \mathbb{R}, f \text{ ist Dichte und symmetrisch um } 0\}$$

nichtparametrische Modelle.

In diesem Buch beschränken wir uns im wesentlichen auf parametrische Modelle. Kann die parametrische Annahme verifiziert werden, so ist man dadurch in der Lage, schärfere Aussagen zu treffen. Ist dies nicht der Fall, so müssen nichtparametrische Methoden angewendet werden. Hierfür sei auf ... verwiesen.

**Definition 2.4.** Ein statistisches Modell heißt *regulär*, falls eine der beiden folgenden Bedingungen erfüllt ist:

- (i) Alle  $\mathcal{P}_\theta$ ,  $\theta \in \Theta$ , sind stetig mit Dichte  $p(x, \theta)$
- (ii) Alle  $\mathcal{P}_\theta$ ,  $\theta \in \Theta$ , sind diskret mit Wahrscheinlichkeitsfunktion  $p(x, \theta)$ .

Im folgenden schreiben wir für ein reguläres Modell oft  $\mathcal{P} = \{p(\cdot, \theta) : \theta \in \Theta\}$ , wobei  $p$  die entsprechende Dichte oder Wahrscheinlichkeitsfunktion ist.

<sup>2</sup>Mit  $\otimes$  bezeichnen wir die gemeinsame Verteilung von  $X_1$  und  $X_2$ , die aufgrund der Unabhängigkeit durch das Produkt der Dichten bestimmt ist.

**Beispiel 2.4.** (*Messmodell*) Reguläre Modell erhält man etwa durch das Messmodell aus Beispiel 2.2. Unter den Annahmen (i)-(iv) und der zusätzlichen Annahme, dass das Modell eine Dichte hat ist die gemeinsame Dichte gegeben durch

$$p(\mathbf{x}, \boldsymbol{\theta}) = \prod_{i=1}^n f(x_i - \mu), \quad f \text{ ist symmetrische Dichte um } 0, \quad \mu \in \mathbb{R}.$$

Gilt darüber hinaus die Normalverteilungsannahme (v), so erhält man

$$p(\mathbf{x}, \boldsymbol{\theta}) = \prod_{i=1}^n \frac{1}{\sigma} \varphi\left(\frac{x_i - \mu}{\sigma}\right), \quad \boldsymbol{\theta} = \begin{pmatrix} \mu \\ \sigma \end{pmatrix},$$

wobei  $\varphi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$  die Dichte der Standardnormalverteilung ist.

Das Ziel einer statistischen Analyse ist es, aus den vorliegenden Daten zu schließen, welche Verteilung  $\mathbb{P}_{\boldsymbol{\theta}}$  wirklich vorliegt, oder anders ausgedrückt: welcher Parameter  $\boldsymbol{\theta}$  den beobachteten Daten zugrunde liegt. Im Gegensatz hierzu geht man in der Wahrscheinlichkeitstheorie von einer festen Verteilung  $\mathbb{P}_{\boldsymbol{\theta}}$  aus und berechnet interessierende Wahrscheinlichkeiten eines bestimmten Ereignisses. Um die vorhandenen Daten bestmöglich auszunutzen, muss die statistische Untersuchung für das Problem maßgeschneidert sein, weswegen eine statistische Fragestellung häufig von dem Problem selbst abhängt:

In dem Kontext der Qualitätssicherung, Beispiel 2.1, möchte man wissen, ob die Lieferung zu viele defekte Teile enthält, d.h. gibt es einen kritischen Wert  $\theta_0$ , so dass man die Lieferung akzeptiert, falls  $\theta \leq \theta_0$  und sie ablehnt, falls  $\theta > \theta_0$ . Unter welchen Gesichtspunkten kann man ein  $\theta_0$  bestimmen? Dies führt zu einem *statistischer Hypothesentest*, welche in Kapitel ?? vorgestellt werden.

In dem Messmodell aus Beispiel 2.2 soll der unbekannte Parameter  $\mu$  geschätzt werden. Ein möglicher Punktschätzer ist durch den arithmetischen Mittelwert gegeben:

$$\frac{1}{n} \sum_{i=1}^n X_i =: \bar{X}. \quad (2.3)$$

Wie man einen solchen Schätzer bestimmen kann und welche Optimalitätseigenschaften bestimmte Schätzer haben wird in den Kapiteln 3 und 4 untersucht.

Folgende Problemstellung sind in der Statistik zu untersuchen:

- Wie erhebt man die Daten?
- Welche Fragestellungen möchte man untersuchen?

- Welches statistische Modell nimmt man an?

Diese Fragestellungen kann man nicht getrennt voneinander untersuchen, sondern sie hängen ganz unmittelbar zusammen.

## 2.2 Suffizienz

Nach der Wahl des statistischen Modells möchte man irrelevante Informationen aus der Vielzahl der erhobenen Daten herausfiltern, welches zu einer Datenreduktion führt, etwa wie in Gleichung (2.3) durch den Mittelwert der Daten. Formal gesehen, sind die erhobenen Daten durch den Zufallsvektor  $\mathbf{X} = (X_1, \dots, X_n)^\top$  gegeben, so definiert man eine *Statistik*  $T := T(\mathbf{X})$  als Funktion der Daten.  $T$  wird als eine *Zufallsvariable* auf dem Ereignisraum  $\Omega$  betrachtet. Punktschätzer von Parametern verwenden die erhobenen Daten, um einen Schätzwert für den gesuchten Parameter zu berechnen, sie lassen sich demnach als Statistik auffassen.

Ist die Statistik eine geeignete Reduktion, so reicht es aus,  $T(\mathbf{X})$  und nicht den ganzen Datenvektor  $\mathbf{X}$  zu kennen: Gilt  $T(\mathbf{x}_1) = T(\mathbf{x}_2)$  für alle Realisierungen  $\mathbf{x}_1, \mathbf{x}_2$  mit gleichen Charakteristika des Experimentes, so reicht es aus, nur den Wert der Statistik  $T$  zu kennen. Das heißt, im Vergleich zur Kenntnis von  $\mathbf{X}$  geht für die Statistik  $T$  keine Information verloren. Dies wird in folgendem Beispiel illustriert.

**Beispiel 2.5.** (*Qualitätskontrolle, siehe Beispiel 2.1.*) Wir betrachten eine Stichprobe von  $n$  Objekten einer Population. Die Zufallsvariable  $X_i$  ist 1, falls das  $i$ -te Teil der Stichprobe defekt ist, und Null sonst, die erhobenen Daten sind  $\mathbf{X} = (X_1, \dots, X_n)^\top$ . Wir interessieren uns für die Anzahl der defekten Teile der Stichprobe und betrachten die Statistik

$$T(\mathbf{X}) = \sum_{i=1}^n X_i.$$

Gibt es zwei defekte Teile in der Stichprobe, so ist dies beschrieben durch die drei Fälle

$$\mathbf{x}_1 = (1, 0, 1), \quad \mathbf{x}_2 = (0, 1, 1), \quad \mathbf{x}_3 = (1, 1, 0).$$

Es gilt  $T(\mathbf{x}_1) = T(\mathbf{x}_2) = T(\mathbf{x}_3)$ . Ist man also an der Anzahl der defekten Teile interessiert, so ist diese Information vollständig in der Statistik  $T(\mathbf{X})$  enthalten.

Ein Schätzer  $T(\mathbf{X})$  reduziert die in den Daten  $\mathbf{X}$  enthaltene Information auf eine einzelne Größe. Möchte man einen Parameter schätzen, so ist es wichtig zu wissen, ob durch diese Reduktion wichtige Information verloren geht oder nicht. Ist eine Statistik suffizient für den Parameter  $\theta$ , so ist das nicht der Fall. Betrachtet wird das statistische Modell  $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ .

**Definition 2.5.** Eine Statistik  $T(\mathbf{X})$  heißt *suffizient für  $\theta$* , falls die bedingte Verteilung von  $\mathbf{X}$  gegeben  $T(\mathbf{X}) = t$  nicht von  $\theta$  abhängt.

Die Interpretation dieser Definition ist wie folgt: Falls man den Wert der suffizienten Statistik  $T$  kennt, dann enthält  $\mathbf{X} = (X_1, \dots, X_n)^\top$  keine weiteren Informationen über  $\theta$ . Kurz schreiben wir für die bedingte Zufallsvariable  $\mathbf{X}$  gegeben  $T(\mathbf{X}) = t$

$$X \mid T(\mathbf{X}) = t.$$

**Beispiel 2.6.** (*Qualitätskontrolle*) (siehe Beispiel 2.1). Betrachtet wird die Beobachtung  $\mathbf{X} = (X_1, \dots, X_n)^\top$ , wobei  $X_i \in \{0, 1\}$ .  $X_i$  hat den Wert eins, falls das  $i$ -te Teil defekt ist und Null sonst. Wir nehmen an, dass die  $X_i$  unabhängig sind und  $\mathbb{P}(X_i = 0) = \theta$ , wobei  $\theta$  der unbekannte Parameter ist. Sei  $\mathbf{x} \in \{0, 1\}^n$  der Vektor der beobachteten Werte und setze  $s := s(\mathbf{x}) = \sum_{i=1}^n x_i$ . Das zugrundeliegende statistische Modell ist beschrieben durch  $P_\theta$ ,  $\theta \in [0, 1]$  mit

$$\mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n) = \theta^s (1 - \theta)^{n-s}.$$

Für die bedingte Verteilung von  $\mathbf{X}$  gegeben  $S(\mathbf{X}) = \sum_{i=1}^n X_i$  erhält man nach Beispiel 1.3 von Seite 17

$$P(\mathbf{X} = \mathbf{x} \mid S(\mathbf{X}) = t) = \binom{n}{t}^{-1}.$$

Dieser Ausdruck ist unabhängig von  $\theta$ , also ist  $S(\mathbf{X})$  eine suffiziente Statistik für den Parameter  $\theta$ . Damit ist auch der arithmetische Mittelwert  $\bar{X} = n^{-1}S(\mathbf{X})$  eine suffiziente Statistik für  $\theta$ .

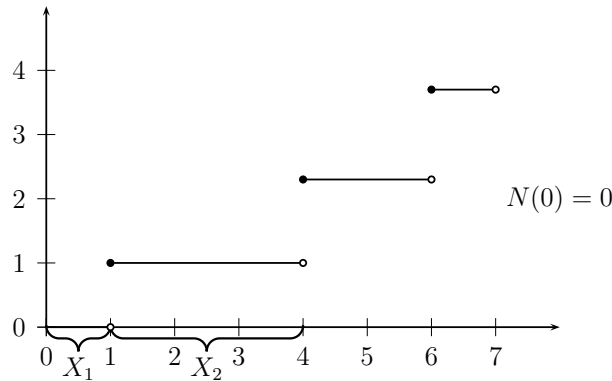
**Bemerkung 2.6.** Falls  $T(\mathbf{X})$  suffizient für  $\theta$  ist, dann kann man wie folgt Daten  $\mathbf{X}'$  mit der gleichen Verteilung wie  $\mathbf{X}$  erzeugen, ohne  $\theta$  zu kennen: Ist  $t = T(\mathbf{X})$ , so erzeuge  $\mathbf{X}'$  nach der Verteilung  $\mathbf{X} \mid T = t$  (welche aufgrund der Suffizienz nicht von  $\theta$  abhängt).

Wir beweisen die Aussage für diskrete Zufallsvariablen. Definiere  $t' := T(\mathbf{x}')$ . Ist  $\mathbb{P}(T = t') > 0$ , so gilt

$$\begin{aligned} P(\mathbf{X}' = \mathbf{x}') &= P(\mathbf{X}' = \mathbf{x}', T = t') = P(\mathbf{X}' = \mathbf{x}' \mid T = t') \cdot P(T = t') \\ &= P(\mathbf{X} = \mathbf{x}' \mid T = t') \cdot P(T = t') && \text{(Def. von } \mathbf{X}') \\ &= P(\mathbf{X} = \mathbf{x}', T = t') = P(\mathbf{X} = \mathbf{x}'), \end{aligned}$$

und somit hat  $\mathbf{X}'$  die gleiche Verteilung wie  $\mathbf{X}$ .

**Beispiel 2.7.** (*Warteschlange.*) Die Ankunft von Kunden an einem Schalter folgt einem *Poisson-Prozess* mit Intensität  $\theta$ , falls folgende Annahmen erfüllt sind. Bezeichne  $N_t$  die Anzahl der Kunden, welche zum Zeitpunkt  $t \geq 0$  angekommen sind. Die Poissonverteilung wurde in Gleichung (1.3) auf Seite 10 definiert.



Abbildungung 2.1: Realisation eines Poisson-Prozesses. Die Sprungzeitpunkte stellen Ankünfte von neuen Kunden an einer Warteschlange dar.  $X_i$  ist die verstrichene Zeite zwischen der Ankunft des  $i$ -ten und des  $i - 1$ -ten Kunden.

- (i)  $N_0 = 0$ ,
- (ii)  $N_{t+h} - N_t$  ist unabhängig von  $N_s$  für alle  $0 \leq s \leq t$  und alle  $h > 0$ ,
- (iii)  $N_{t+h} - N_t \sim \text{Pois}(\theta h)$  für alle  $t \geq 0$  und  $h > 0$ .

Insbesondere folgt aus (iii), dass  $N_t \sim \text{Pois}(\theta t)$ . Ein Illustration des Poisson-Prozesses  $(N_t)_{t \geq 0}$  findet sich in Abbildung 2.1.

Mit  $X_i$  sei die verstrichene Zeit zwischen der Ankunft des  $i$ -ten und des  $i - 1$ -ten Kunden bezeichnet,  $X_1$  sei die Zeit bis zur Ankunft des ersten Kunden. Dann folgt aus (iii), dass  $P(X_1 > t) = P(N(t) = 0) = \exp(-\theta t)$ , also ist  $X_1$  exponentialverteilt mit Parameter  $\theta$ . Aus Aufgabe 2.1 erhält man, dass  $X_i \sim \text{Exp}(\theta)$  und die Unabhängigkeit von  $X_1, X_2, \dots$ .  
Setzte

$$T = X_1 + X_2.$$

Dann ist  $T$  suffizient für  $\theta$ : Wir berechnen die bedingte Dichte durch (1.9). Die gemeinsame Dichte ist

$$p_{\mathbf{X}}(x_1, x_2, \theta) = \exp(-(x_1 + x_2)\theta) \theta^2.$$

Ziel ist es, den Transformationssatz 1.3 in geschickter Weise anzuwenden. Wir wählen folgende Transformation  $g: \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow \mathbb{R}^+ \times [0, 1]$ :

$$g(\mathbf{x}) = \left( x_1 + x_2, \frac{x_1}{x_1 + x_2} \right)^\top.$$

Damit ist  $g^{-1}(\mathbf{y}) = (y_1 y_2, y_1 - y_1 y_2)$  und

$$|J_{g^{-1}}(y_1, y_2)| = \left| \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} \right| = \left| \begin{vmatrix} y_2 & 1 - y_2 \\ y_1 & -y_1 \end{vmatrix} \right| = |-y_1| = y_1.$$

Die Anwendung des Transformationssatzes liefert die Dichte von  $\mathbf{Y} = g(\mathbf{X})$ ,

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}) &= \exp(-\theta(y_1 y_2 + y_1 - y_1 y_2)) \theta^2 y_1 \mathbf{1}_{\{y_1 \geq 0, y_2 \in [0, 1]\}} \\ &= \frac{e^{-y_1 \theta} \theta^2 y_1}{\Gamma(2)} \cdot \mathbf{1}_{[0, 1]}(y_2) = p_{Y_1}(y_1) \cdot p_{Y_2|Y_1}(y_2|y_1), \end{aligned}$$

falls  $y_1 \geq 0$ . Dies ist das Produkt der Dichten einer  $Gamma(2, \theta)$  und einer  $U(0, 1)$ -Verteilung, wonach  $Y_2$  unabhängig von  $Y_1 = X_1 + X_2$  und  $U(0, 1)$ -verteilt ist; wobei die  $U(0, 1)$ -Verteilung unabhängig von  $t$  ist. Man erhält

$$\mathbb{P}(X_1 \leq x | T = t) = \mathbb{P}(TY_2 \leq x | T = t) = \mathbb{P}(tY_2 \leq x) = \frac{x}{t},$$

für  $x \in [0, t]$ . Demnach ist  $X_1$  bedingt auf  $T = t$  gleichverteilt auf  $[0, t]$  ( $U(0, t)$ ). Durch  $X_2 = T - X_1$  erhält man, dass der Vektor  $\mathbf{X}$  bedingt auf  $T = t$  verteilt ist wie

$$(Z, t - Z)$$

wobei  $Z \sim U(0, t)$ . Es folgt, dass  $\mathbf{X}|T = t$  unabhängig von  $\theta$  ist und somit  $T$  suffiziente Statistik für  $\theta$  ist.

Diesem Beispiel liegt allgemeiner die Aussage zugrunde, dass bedingt auf  $N_t = n$  die Zwischenankunftszeiten von  $N$  verteilt sind wie Ordnungssatitistiken von gleichverteilten Zufallsvariablen, siehe Rolski, Schmidli, Schmidt, and Teugels (1999), Seite 502.

Suffizienz kann man mit folgendem wichtigem Satz von Fisher, Neyman, Halmos und Savage nachweisen.

**Satz 2.7** (Faktorisierungssatz). Sei  $\mathcal{P} = \{p(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$  ein reguläres Modell. Das Bild der Statistik  $T(\mathbf{X})$  sei mit  $I$  bezeichnet. Dann sind äquivalent:

(i)  $T(\mathbf{X})$  ist suffizient für  $\boldsymbol{\theta}$

(ii) Es existiert  $g : I \times \Theta \rightarrow \mathbb{R}$  und  $h : \mathbb{R}^n \rightarrow \mathbb{R}$ , so dass

$$p(\mathbf{x}, \boldsymbol{\theta}) = g(T(\mathbf{x}), \boldsymbol{\theta}) \cdot h(\mathbf{x}).$$

*Beweis.* Wieder führen wir den Nachweis nur für den diskreten Fall.  $\mathbf{X}$  nehme also die Werte  $\mathbf{x}_1, \mathbf{x}_2, \dots$  an. Setze  $t_i := T(\mathbf{x}_i)$ . Dann ist auch  $T$  diskret mit Werten  $t_1, t_2, \dots$ . Wir zeigen zunächst, dass  $(ii) \Leftrightarrow (i)$ . Aus (ii) folgt, dass

$$\mathbb{P}_\theta(T = t_i) = \sum_{\{\mathbf{x}: T(\mathbf{x})=t_i\}} p(\mathbf{x}, \theta) = \sum_{\{\mathbf{x}: T(\mathbf{x})=t_i\}} g(t_i, \theta) \cdot h(\mathbf{x}). \quad (2.4)$$

Für  $\theta \in S_i = \{\theta : \mathbb{P}_\theta(T = t_i) > 0\}$  gilt

$$P_\theta(\mathbf{X} = \mathbf{x}_j | T = t_i) = \frac{P_\theta(\mathbf{X} = \mathbf{x}_j, T = t_i)}{P_\theta(T = t_i)}.$$

Dieser Ausdruck verschwindet, falls  $T(\mathbf{x}_j) \neq t_i$ . Gilt  $T(\mathbf{x}_j) = t_i$ , so ist

$$\begin{aligned} P_\theta(\mathbf{X} = \mathbf{x}_j | T = t_i) &= g(t_i, \theta) \cdot \frac{h(\mathbf{x}_j)}{P_\theta(T = t_i)} \\ &\stackrel{(2.4)}{=} \frac{g(t_i, \theta) h(\mathbf{x}_j)}{\sum_{\{\mathbf{x}: T(\mathbf{x})=t_i\}} g(t_i, \theta) \cdot h(\mathbf{x})} = \frac{h(\mathbf{x}_j)}{\sum_{\{\mathbf{x}: T(\mathbf{x})=t_i\}} h(\mathbf{x})}. \end{aligned}$$

Da dieser Ausdruck unabhängig von  $\theta$  ist, ist  $T(\mathbf{X})$  suffizient für  $\theta$ .

Es bleibt zu zeigen, dass  $(i) \Rightarrow (ii)$ . Sei also  $T$  eine suffiziente Statistik für  $\theta$  und setze

$$g(t_i, \theta) := P_\theta(T = t_i), \quad h(\mathbf{x}) := P(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})).$$

Dabei ist  $h$  unabhängig von  $\theta$  da  $T(\mathbf{x})$  suffizient. Es folgt, dass

$$\begin{aligned} p(\mathbf{x}, \theta) &= P_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = T(\mathbf{x})) \\ &= P_\theta(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})) \cdot P_\theta(T(\mathbf{X}) = T(\mathbf{x})) \\ &= h(\mathbf{x}) \cdot g(T(\mathbf{x}), \theta) \end{aligned}$$

und somit die behauptete Faktorisierung (ii).  $\square$

**Beispiel 2.8.** (*Warteschlange, Fortsetzung von Beispiel 2.7.*) Seien  $\mathbf{X} = (X_1, \dots, X_n)^\top$  die ersten  $n$  Zwischenankunftszeiten eines Poisson-Prozesses, d.h.  $X_1, \dots, X_n$  sind unabhängig und  $X_i \sim \text{Exp}(\theta)$ . Die Dichte von  $\mathbf{X}$  ist demnach

$$p(\mathbf{x}, \theta) = \theta^n \exp \left\{ -\theta \sum_{i=1}^n x_i \right\} \cdot \mathbf{1}_{\{x_1, \dots, x_n \geq 0\}}$$

Die Statistik  $T(\mathbf{X}) := \sum_{i=1}^n X_i$  ist suffizient für  $\theta$ : In der Tat, wähle  $g(t, \theta) = \theta^n \exp\{-\theta t\}$  und  $h(\mathbf{x}) = \mathbf{1}_{\{x_1, \dots, x_n \geq 0\}}$ . Dann ist Bedingung (ii) von Satz 2.7 erfüllt und somit  $T$  suffizient für  $\theta$ . Ebenso ist auch das arithmetische Mittel eine suffiziente Statistik für  $\theta$ .

**Beispiel 2.9.** (*Titel fehlt*) Betrachtet werde eine Population mit  $\theta$  Mitgliedern. Dabei seien die Mitglieder geordnet und mit  $1, 2, \dots, \theta$  nummeriert. Man ziehe  $n$ -mal zufällig mit Zurücklegen von der Population.  $X_i$  sei das Ergebnis des  $i$ -ten Zuges. Dies führt zu einem Laplaceschen Modell:  $\mathbb{P}(X_i = k) = \theta^{-1}$  für alle  $k \in \{1, \dots, \theta\}$ . Darüber hinaus sind die  $X_i$  unabhängig. Damit ist die gemeinsame Verteilung ist

$$p(\mathbf{x}, \theta) = \prod_{i=1}^n p(x_i, \theta) = \theta^{-n} \mathbf{1}_{\{x_i \in \{1, \dots, \theta\}, 1 \leq i \leq n\}}.$$

Die Statistik

$$T(\mathbf{X}) := \max_{i=1, \dots, n} X_i$$

ist suffizient für  $\theta$ : durch die Wahl von  $g(t, \theta) := \theta^{-n} \cdot \mathbf{1}_{\{t \leq \theta\}}$  und  $h(\mathbf{x}) := \mathbf{1}_{\{x_i \in \{1, \dots, \theta\}, 1 \leq i \leq n\}}$  erhält man dies aus dem Faktorisierungssatz 2.7.

**Beispiel 2.10.** (*Suffiziente Statistiken für die Normalverteilung.*) Betrachtet man eine Stichprobe von normalverteilten Daten, so sind arithmetisches Mittel *und* Stichprobenvarianz suffizient: Seien  $X_1, \dots, X_n$  i.i.d. mit  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ . Gesucht ist der Parametervektor  $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ , sowohl Erwartungswert  $\mu$  als auch Varianz  $\sigma^2$  sind unbekannt. Das arithmetische Mittel  $\bar{X}$  und die Stichprobenvarianz  $s^2(\mathbf{X})$  wurden in Beispiel 1.1 definiert. Die Dichte von  $\mathbf{X} = (X_1, \dots, X_n)^\top$  ist

$$p(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right).$$

Zunächst betrachten wir den zufälligen Vektor  $T_1(\mathbf{X}) := (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)^\top$ . Mit  $h(\mathbf{x}) := 1$  und

$$g(T_1(\mathbf{x}), \boldsymbol{\theta}) := \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{n\mu^2}{2\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i\right)\right)$$

ist  $p(\mathbf{x}, \boldsymbol{\theta}) = g(T_1(\mathbf{x}), \boldsymbol{\theta})h(\mathbf{x})$  und somit ist  $T_1(\mathbf{X})$  suffizient für  $\boldsymbol{\theta}$ . Dann ist auch der zufällige Vektor  $T_2$ , definiert durch

$$T_2(\mathbf{X}) := \begin{pmatrix} \bar{X} \\ s^2(\mathbf{X}) \end{pmatrix}$$

suffizient, denn  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  und  $s^2(\mathbf{X}) = \frac{1}{n-1} \sum_{i=1}^n X_i^2 - (\bar{X})^2$  nach Aufgabe 2.2.

## 2.3 Exponentielle Familien

**Definition 2.8.** Eine Familie von Verteilungen  $\{\mathbb{P}_\theta, \theta \in \Theta\}$  mit  $\Theta \subset \mathbb{R}$  heißt eine *einparametrische exponentielle* Familie, falls Funktionen  $c, d : \Theta \rightarrow \mathbb{R}$  und  $T, S : \mathbb{R}^n \rightarrow \mathbb{R}$  und eine Menge  $A \subset \mathbb{R}^n$  existieren, so dass die Dichte oder Wahrscheinlichkeitsfunktion  $p(\mathbf{x}, \theta)$  von  $\mathbb{P}_\theta$  als

$$p(\mathbf{x}, \theta) = \exp\left(c(\theta) \cdot T(\mathbf{x}) + d(\theta) + S(\mathbf{x})\right) \cdot \mathbf{1}_A(\mathbf{x}) \quad (2.5)$$

dargestellt werden kann.

Es ist wesentlich, dass  $A$  hierbei unabhängig von  $\theta$  ist. Die Funktion  $d(\theta)$  kann als Normierung aufgefasst werden. An dieser Stelle soll betont werden, dass die Verteilung eines mehrdimensionalen Zufallsvektors durchaus zu einer *einparametrischen* exponentiellen Familie gehören kann. Diese wird allerdings nur von einem eindimensionalen Parameter aufgespannt.  $K$ -parametrische exponentielle Familien werden in Definition 2.14 vorgestellt.

Die Nützlichkeit dieser Darstellung von Verteilungsklassen erschließt sich bereits durch folgende Beobachtung:  $T(\mathbf{X})$  ist stets suffiziente Statistik für  $\theta$ : Dies folgt aus dem Faktorisierungssatz 2.7 mit

$$g(t, \theta) = \exp(c(\theta)t + d(\theta)) \quad \text{und} \quad h(\mathbf{x}) = \exp(S(\mathbf{x})) \cdot \mathbf{1}_A(\mathbf{x}).$$

$T$  heißt *natürliche suffiziente Statistik*. Eine Vielzahl von Verteilungen lassen sich als exponentielle Familien schreiben. Wir stellen die Normalverteilung in verschiedenen Varianten vor, die Binomialverteilung und es folgen die Poisson-Verteilung, die Gamma- und die Beta-Verteilung. Die Verteilung einer Stichprobe, welche aus i.i.d. Zufallsvariablen einer exponentiellen Verteilung entsteht, bildet wieder eine exponentielle Familie, wie in Bemerkung 2.10 gezeigt wird. Die beiden folgenden Beispiele illustrieren die Normalverteilung als einparametrische exponentielle Familie. Da die Normalverteilung durch zwei Parameter beschrieben wird, muss jeweils einer festgehalten werden, um eine einparametrische Familie zu erhalten. Die Normalverteilung als zweiparametrische exponentielle Familie wird in Beispiel 2.17 vorgestellt.

Ist  $c(\theta) = \theta$  in Darstellung (2.5), so spricht man von einer *natürlichen* exponentiellen Familie. Jede exponentielle Familie hat eine Darstellung als natürlich exponentielle Familie, was man stets durch eine Reparametrisierung erreichen kann: Mit  $\eta := c(\theta)$  erhält man die Darstellung

$$p_0(\mathbf{x}, \eta) = \exp\left(\eta \cdot T(\mathbf{x}) + d_0(\eta) + S(\mathbf{x})\right) \cdot \mathbf{1}_A(\mathbf{x}) \quad (2.6)$$

mit *Normierungskonstante*

$$d_0(\eta) := -\ln\left(\int_{\mathbb{R}^n} \exp(\eta T(\mathbf{x}) + S(\mathbf{x})) d\mathbf{x}\right).$$

was äquivalent ist zu  $\int p_0(\mathbf{x}, \eta) d\mathbf{x} = 1$ .

**Bemerkung 2.9.** Ist  $c : \Theta \rightarrow \mathbb{R}$  eine injektive Funktion, so ist die Normierungskonstante einfacher zu bestimmen, denn in diesem Fall folgt  $d_0(\eta) = d(c^{-1}(\eta))$ . Ferner, falls  $\eta = c(\theta)$  mit  $\theta \in \Theta$ , so folgt  $d_0(\eta) < \infty$  da  $p_0(\mathbf{x}, \eta)$  Dichtefunktion ist.

**Beispiel 2.11.** (*Normalverteilung mit bekanntem  $\sigma$ .*) In Analogie zu dem Messmodell aus Beispiel 2.2 und den dortigen Annahmen (i)-(vi) betrachte wir ein festes  $\sigma_0$  und das statistische Modell

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

Dann ist  $\mathcal{P}$  eine einparametrische exponentielle Familie, denn die Dichte lässt sich schreiben als

$$\begin{aligned} p(x, \mu) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(x - \mu)^2\right) \\ &= \exp\left[\frac{\mu}{\sigma_0^2} \cdot x + \frac{-\mu^2}{2\sigma_0^2} - \left(\frac{x^2}{2\sigma_0^2} + \ln\left(\sqrt{2\pi\sigma_0^2}\right)\right)\right]. \end{aligned} \quad (2.7)$$

Mit  $c(\mu) := \frac{\mu}{\sigma_0^2}$ ,  $T(x) := x$ ,  $d(\mu) := \frac{-\mu^2}{2\sigma_0^2}$  und  $S(x) := -\left(\frac{x^2}{2\sigma_0^2} + \ln\left(\sqrt{2\pi\sigma_0^2}\right)\right)$  sowie  $A := \mathbb{R}$  erhält man die die Gestalt (2.5).

**Beispiel 2.12.** (*Normalverteilung mit bekanntem  $\mu$ .*) Anders als in dem vorigen Beispiel nehmen wir nun an, dass der Erwartungswert der Normalverteilung bekannt ist, etwa  $\mu_0$ . Dies führt zu dem statistischen Modell

$$\mathcal{P} = \{P_{\sigma^2} = \mathcal{N}(\mu_0, \sigma^2) : \sigma > 0\}.$$

Die entsprechende Dichte hat, analog zu Gleichung (2.7), die Gestalt

$$p(x, \sigma^2) = \exp\left(-\frac{1}{2\sigma^2}(x - \mu_0)^2 - \ln\left(\sqrt{2\pi\sigma^2}\right)\right).$$

Mit der Wahl von  $c(\sigma^2) := -\frac{1}{2\sigma^2}$ ,  $T(x) := (x - \mu_0)^2$ ,  $d(\sigma^2) := -\ln\left(\sqrt{2\pi\sigma^2}\right)$  und  $S(x) := 0$ , sowie  $A := \mathbb{R}$  erhält man eine Darstellung in der Form (2.5) und somit ist  $\mathcal{P}$  ebenfalls eine eine exponentielle Familie.

**Beispiel 2.13.** (*Binomialverteilung.*) Nicht nur stetige Verteilungen lassen sich als exponentielle Familien beschreiben, auch für diskrete Verteilungen ist dies möglich. So ist zum Beispiel die Binomialverteilung eine exponentielle Familie: Denn, die Wahrscheinlichkeitsfunktion einer  $\text{Bin}(n, \theta)$ -Verteilung ist für  $k \in \{0, \dots, n\}$

$$p(x, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} = \exp\left(x \cdot \ln\left(\frac{\theta}{1 - \theta}\right) + n \cdot \ln(1 - \theta) + \ln\binom{n}{x}\right).$$

Mit der Wahl von  $c(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$ ,  $T(x) = x$ ,  $d(\theta) = n \ln(1-\theta)$ , und  $S(x) = \ln\binom{n}{x}$ , sowie  $A = \{0, 1, \dots, n\}$  ergibt sich die Darstellung (2.5). Die Familie der Binomialverteilungen, gegeben durch ihre Wahrscheinlichkeitsfunktionen  $\{p(x, \theta) : \theta \in (0, 1)\}$ , ist also eine exponentielle Familie.

Es sei daran erinnert, dass unabhängige und identisch verteilte Zufallsvariablen als i.i.d. bezeichnet werden.

**Bemerkung 2.10.** Die i.i.d. Kombination einer exponentiellen Familie ist eine exponentielle Familie. Insbesondere trifft dies auf die obigen Beispiele 2.11-2.13 zu. Die  $m$  Zufallsvektoren  $\mathbf{X}_1, \dots, \mathbf{X}_m$  seien i.i.d. mit  $\mathbf{X}_i \in \mathbb{R}^n$ .  $P_\theta$  sei eine einparametrische exponentielle Familie und  $\mathbf{X}_i \sim P_\theta$ ,  $i = 1, \dots, m$ . Setze  $\mathbf{X} := (\mathbf{X}_1, \dots, \mathbf{X}_m)^\top \in \mathbb{R}^{n \times m}$ . Die Dichte bzw. Wahrscheinlichkeitsfunktion von  $\mathbf{X}$  ist

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}, \theta) &= \prod_{i=1}^m p_{\mathbf{X}_i}(\mathbf{x}_i, \theta) \\ &= \prod_{i=1}^m \exp\left(c(\theta)T(\mathbf{x}_i) + d(\theta) + S(\mathbf{x}_i)\right) \cdot \mathbf{1}_A(\mathbf{x}_i) \\ &= \exp\left(c(\theta) \sum_{i=1}^m T(\mathbf{x}_i) + m \cdot d(\theta) + \sum_{i=1}^m S(\mathbf{x}_i)\right) \cdot \mathbf{1}_{A^m}(\mathbf{x}_1, \dots, \mathbf{x}_m), \end{aligned}$$

mit  $A^m := \{(\mathbf{x}_1, \dots, \mathbf{x}_m) : \mathbf{x}_i \in A \forall 1 \leq i \leq m\}$ . Durch die Wahl der suffizienter Statistik  $T'(\mathbf{x}) := \sum_{i=1}^m T(\mathbf{x}_i)$ , sowie  $c'(\theta) := c(\theta)$ ,  $d'(\theta) := m \cdot d(\theta)$ ,  $A' := A^m$  und  $S'(\mathbf{x}) = \sum_{i=1}^m S(\mathbf{x}_i)$  erhält man eine Darstellung als exponentiellen Familie gemäß (2.5).

Somit gehört die Verteilung von  $\mathbf{X}$  wieder einer einparametrischen exponentiellen Familie mit suffizienter Statistik  $T'(\mathbf{x}) := \sum_{i=1}^m T(\mathbf{x}_i)$  an.

**Beispiel 2.14.** (i.i.d. Normalverteilung mit bekanntem  $\sigma$ .) Als Beispiel obiger Bemerkung betrachten wir  $\mathbf{X} = (X_1, \dots, X_m)^\top$ , wobei  $X_1, \dots, X_m$  i.i.d. seien mit  $X_i \sim N(\mu, \sigma_0^2)$  und bekanntem  $\sigma_0$  (vergleiche Beispiel 2.11). Dann ist  $T(\mathbf{X}) := \sum_{i=1}^m X_i$  und somit auch das arithmetische Mittel  $\bar{X}$  suffiziente Statistik für  $\mu$  und  $\mathbf{X} \sim P_{\mathbf{X}}$  eine einparametrische exponentielle Familie.

Weitere Beispiele für einparametrische exponentielle Familien sind in Tabelle 2.1 dargestellt.

Das folgende Resultat beschreibt die Verteilung der kanonischen Statistik einer einparametrischen exponentiellen Familie.

Verteilungsfamilie	$c(\theta)$	$T(x)$	$A$
Poiss( $\theta$ )	$\log(\theta)$	$x$	$\{0, 1, 2, \dots\}$
$\text{Gamma}(p, \lambda)$ , $p$ bekannt	$-\lambda$	$x$	$\mathbb{R}^+$
$\text{Gamma}(p, \lambda)$ , $\lambda$ bekannt	$p - 1$	$\ln x$	$\mathbb{R}^+$
$\text{Beta}(r, s)$ , $r$ bekannt	$s - 1$	$\ln(1 - x)$	$[0, 1]$
$\text{Beta}(r, s)$ , $s$ bekannt	$r - 1$	$\ln(x)$	$[0, 1]$

Tabelle 2.1: Einparametrische exponentielle Familien.  $c$ ,  $T$  und  $A$  aus Darstellung (2.5) sind in der Tabelle angegeben,  $d$  ergibt sich durch Normierung.

**Satz 2.11.** Sei  $\mathbb{P}_\theta$  eine einparametrische exponentielle Familie mit Darstellung (2.5) und kanonischer Statistik  $T(\mathbf{X})$ . Gilt  $\mathbf{X} \sim \mathbb{P}_\theta$ , so ist  $T(\mathbf{X}) \sim Q_\theta$  wobei  $Q_\theta$  eine einparametrische exponentielle Familie mit Dichte bzw. Wahrscheinlichkeitsfunktion

$$q(t, \theta) = \exp\left(c(\theta) \cdot t + d(\theta) + S^*(t)\right) \cdot 1_{A^*}(t);$$

hierbei ist  $A^* := \{T(\mathbf{x}) : \mathbf{x} \in A\}$  und  $S^*(t) := \ln \sum_{\{\mathbf{x}: T(\mathbf{x})=t\}} \exp(S(\mathbf{x}))$ .

*Beweis.* Wir beweisen den diskreten Fall, der stetige Fall ist Teil von Aufgabe 2.3. Da  $\mathbf{X}$  eine diskrete Zufallsvariable ist mit Wahrscheinlichkeitsfunktion  $p(\mathbf{x}, \theta)$ , ist  $T(\mathbf{X})$  ebenfalls diskret mit Wahrscheinlichkeitsfunktion

$$\begin{aligned} q(t, \theta) &= P_\theta(T(\mathbf{x}) = t) = \sum_{\{\mathbf{x}: T(\mathbf{x})=t\}} p(\mathbf{x}, \theta) \\ &= \sum_{\mathbf{x}: T(\mathbf{x})=t} \left[ \exp\left(c(\theta) \cdot T(\mathbf{x}) + d(\theta) + S(\mathbf{x})\right) \cdot 1_A(\mathbf{x}) \right] \\ &= \exp\left(c(\theta)t + d(\theta)\right) \left( \sum_{\{\mathbf{x}: T(\mathbf{x})=t\}} e^{S(\mathbf{x})} \right) \cdot 1_{A^*}(t). \end{aligned}$$

Damit ist die Verteilung von  $T$  eine exponentielle Familie nach Darstellung (2.5).  $\square$

**Beispiel 2.15.** (*Verteilung des arithmetischen Mittels.*) Bsp: Verteilung des arithmetischen Mittels im Normalverteilungsfall

In einer natürlichen exponentiellen Familie gilt  $c(\theta) = \theta$  und man hat die einfachere Darstellung (2.6). Die folgende Aussage bestimmt die momentenerzeugende Funktion von  $T(\mathbf{X})$  in einer natürlichen einparametrischen exponentiellen Familie.

**Satz 2.12.** *Betrachtet man eine natürliche einparametrische exponentielle Familie mit Dichte oder Wahrscheinlichkeitsfunktion  $p_0(\mathbf{x}, \eta)$  aus (2.6) und ist  $\mathbf{X} \sim p_0$ , so gilt für die momentenerzeugende Funktion von  $T(\mathbf{X})$ :*

$$\Psi(s) = \mathbb{E}(e^{s \cdot T(\mathbf{X})}) = \exp(d_0(\eta) - d_0(\eta + s)) < \infty,$$

für alle  $\eta, \eta + s \in H$  mit  $H := \{\eta : d_0(\eta) < \infty\}$ .

*Beweis.* Wir führen den Beweis für den Fall, in welchem  $p_0$  eine Dichte ist. Der diskrete Fall folgt analog. Es gilt

$$\begin{aligned} \Psi(s) &= \mathbb{E}(\exp(s \cdot T(\mathbf{X}))) \\ &= \int_A \exp\left((\eta + s)T(\mathbf{x}) + d_0(\eta) + S(\mathbf{x})\right) d\mathbf{x} \\ &= \exp\left(d_0(\eta) - d_0(\eta + s)\right) \int_A \exp\left((\eta + s)T(\mathbf{x}) + d_0(\eta + s) + S(\mathbf{x})\right) d\mathbf{x} \\ &= \exp\left(d_0(\eta) - d_0(\eta + s)\right) \int_A p_0(\mathbf{x}, \eta + s) d\mathbf{x}. \end{aligned}$$

Da  $\eta + s \in H$  ist  $p_0$  eine Dichte und das Integral gleich eins. Da weiterhin  $d_0(\eta)$  und  $d_0(\eta + s)$  endlich sind, da  $\eta, \eta + s \in H$ , folgt die Behauptung.  $\square$

**Bemerkung 2.13.** *Erwartungswert und Varianz der suffizienten Statistik in exponentiellen Familien.* Aus der momentenerzeugenden Funktion kann man die Momente von  $T(\mathbf{X})$  bestimmen. Es sei daran erinnert, dass jede exponentielle Familie eine natürliche Darstellung der Form (2.6) hat. Daraus bestimmt sich

$$\mathbb{E}(T(\mathbf{X})) = \Psi'(0) = \Psi(0) \left( -d'_0(\eta + s) \Big|_{s=0} \right) = -d'_0(\eta).$$

Weiterhin ist  $\mathbb{E}(T(\mathbf{X})^2) = (d'_0(\eta))^2 - d''_0(\eta)$  und damit

$$\text{Var}(T(\mathbf{X})) = -d''_0(\eta).$$

**Beispiel 2.16.** (*Momente der Rayleigh-Verteilung.*) Seien  $X_1, \dots, X_n$  i.i.d. und Rayleighverteilt, d.h.  $X_i$  hat Dichte  $x\theta^{-2} \exp(-x^2/2\theta^2)$  für  $x > 0$  und unbekanntem  $\theta > 0$ , siehe Bemerkung 1.12. Die Rayleigh-Verteilung ist eine exponentielle Familie, denn  $\mathbf{X} = (X_1, \dots, X_n)^\top$  hat Dichte

$$\begin{aligned} p(\mathbf{x}, \theta) &= \exp\left(-\sum_{i=1}^n \frac{x_i^2}{2\theta^2}\right) \cdot \prod_{i=1}^n \frac{x_i}{\theta^2} \\ &= \exp\left(-\frac{1}{2\theta^2} \sum_{i=1}^n x_i^2 - n \ln \theta^2 + \sum_{i=1}^n \ln x_i\right), \end{aligned}$$

und durch die Wahl von  $c(\theta) := -\frac{1}{2\theta^2}$ ,  $d(\theta) = -n \ln(\theta^2)$ ,  $A := \mathbb{R}^+$ , natürlicher suffizienter Statistik  $T(\mathbf{X}) = \sum_{i=1}^n X_i^2$  und  $S(\mathbf{x}) := \sum_{i=1}^n \log x_i$  erhält man Darstellung (2.5). Die Transformation auf eine natürliche Familie erfolgt mit  $\eta := c(\theta) < 0$ . Das bedeutet

$$c^{-1}(\eta) = \sqrt{-\frac{1}{2\eta}} \quad \text{und} \quad d_0(\eta) = d(c^{-1}(\eta)) = n \ln(-2\eta).$$

Nach Satz 2.12 hat  $T(\mathbf{X})$  die momentenerzeugende Funktion  $\Psi(s) = \exp\{d_0(\eta) - d_0(\eta + s)\}$ . Aus Bemerkung 2.13 bestimmt sich nun leicht Erwartungswert und Varianz:

$$\mathbb{E}(T(\mathbf{X})) = \mathbb{E}\left(\sum_{i=1}^n X_i^2\right) = -d'_0(\eta) = -\frac{n}{\eta} = 2n\theta^2,$$

was mit den Ergebnissen aus Aufgabe 1.13 übereinstimmt. Die Berechnung der Varianz erfolgt in Aufgabe 2.4.

**Definition 2.14.** Eine Familie von Verteilungen  $\{\mathbb{P}_\theta : \theta \in \Theta\}$  mit  $\Theta \subset \mathbb{R}^K$  heißt *K-parametrische exponentielle Familie*, falls Funktionen  $c_i, d : \Theta \rightarrow \mathbb{R}$ ,  $T_i : \mathbb{R}^n \rightarrow \mathbb{R}$  und  $S : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $i = 1, \dots, K$  sowie eine Menge  $A \subset \mathbb{R}^n$  existieren, so dass die Dichte oder Wahrscheinlichkeitsfunktion  $p(\mathbf{x}, \theta)$  von  $\mathbb{P}_\theta$  als

$$p(\mathbf{x}, \theta) = \exp\left(\sum_{i=1}^K c_i(\theta) T_i(\mathbf{x}) + d(\theta) + S(\mathbf{x})\right) \cdot \mathbf{1}_A(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n \quad (2.8)$$

dargestellt werden kann.

In Analogie zu den einparametrischen Familien ist die Statistik

$$\mathbf{T}(\mathbf{x}) := (T_1(\mathbf{x}), \dots, T_K(\mathbf{x}))^\top$$

suffizient, sie wird als *natürliche suffiziente Statistik* bezeichnet.

**Beispiel 2.17.** (*Die Normalverteilung ist eine zweiparametrische exponentielle Familie.*)

Die Familie der (eindimensionalen) Normalverteilungen gegeben durch  $\mathbb{P}_{\boldsymbol{\theta}} = \mathcal{N}(\mu, \sigma^2)$  mit  $\boldsymbol{\theta} = (\mu, \sigma^2)^\top \in \Theta$  und  $\Theta = \{(\mu, \sigma^2)^\top : \mu \in \mathbb{R}, \sigma > 0\}$  ist eine zweiparametrische exponentielle Familie mit  $n = 1$ , denn die Dichte von  $\mathbb{P}_{\boldsymbol{\theta}}$  hat die Gestalt

$$p(x, \boldsymbol{\theta}) = \exp\left(\frac{\mu}{\sigma^2} x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)\right)$$

und durch die Wahl von  $c_1(\boldsymbol{\theta}) := \mu/\sigma^2$ ,  $T_1(x) := x$ ,  $c_2(\boldsymbol{\theta}) := -1/2\sigma^2$ ,  $T_2(x) := x^2$ ,  $S(\mathbf{x}) := 0$ ,  $A = \mathbb{R}$  und der entsprechenden Normierung  $d(\boldsymbol{\theta}) := -1/2(\mu^2\sigma^{-2} + \ln(2\pi\sigma^2))$  erhält die Dichte die Gestalt (2.8).

**Beispiel 2.18.** (*i.i.d. Normalverteilung als exponentielle Familie.*) Seien  $X_1, \dots, X_n$  i.i.d. und  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ . Dann ist die Verteilung von  $\mathbf{X} = (X_1, \dots, X_n)^\top$  darstellbar als zweiparametrische exponentielle Familie: Mit den Resultaten aus Bemerkung 2.10 führt die Darstellung der Normalverteilung aus Beispiel 2.17 unmittelbar zu einer exponentiellen Familie. Damit ist

$$T(\mathbf{X}) = \left(\sum_{i=1}^n T_1(X_i), \sum_{i=1}^n T_2(X_i)\right)^\top = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)^\top$$

suffizient für  $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ . Dies wurde in Beispiel 2.10 bereits auf elementarem Wege gezeigt.

**Beispiel 2.19.** (*Lineare Regression*) Bei der linearen Regression beobachtet man Paare von Daten  $(x_1, Y_1), \dots, (x_n, Y_n)$ . Man vermutet einen *linearen* Einfluss der Größen  $x_i$  auf  $Y_i$  und möchte diesen bestimmen. Die Beobachtungen  $x_1, \dots, x_n$  werden als konstant angesehen. Diese Methodik wird in Kapitel ?? wesentlich vertieft, für Beispiele sei jetzt schon dorthin verwiesen. Wir gehen von folgendem Modell aus:

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i,$$

für  $i = 1, \dots, n$ . Hierbei sind  $\beta_1, \beta_2 \in \mathbb{R}$  nicht zufällige, aber unbekannte Konstanten und  $\epsilon_1, \dots, \epsilon_n$  und i.i.d. mit  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$  (vergleiche mit dem Messmodell, Beispiel 2.2). Setze

$\mathbf{Y} = (Y_1, \dots, Y_n)^\top$  und  $\boldsymbol{\theta} = (\beta_1, \beta_2, \sigma^2)^\top$ . Die Dichte von  $\mathbf{Y}$  ist

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\theta}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^n \exp\left(-\frac{(y_i - \beta_1 - \beta_2 x_i)^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2 - \frac{n\beta_1^2}{2\sigma^2} - \frac{\beta_2^2}{2\sigma^2} \sum_{i=1}^n x_i^2 \right. \\ &\quad \left. + \frac{\beta_1}{\sigma^2} \sum_{i=1}^n y_i + \frac{\beta_2}{\sigma^2} \sum_{i=1}^n x_i y_i - \frac{\beta_1 \beta_2}{\sigma^2} \sum_{i=1}^n x_i^2 - \frac{n}{2} \ln(2\pi\sigma^2)\right) \\ &= \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2 + \frac{\beta_1}{\sigma^2} \sum_{i=1}^n y_i + \frac{\beta_2}{\sigma^2} \sum_{i=1}^n x_i y_i \right. \\ &\quad \left. - \frac{n\beta_1^2}{2\sigma^2} - \left(\frac{\beta_2^2}{2\sigma^2} + \frac{\beta_1 \beta_2}{\sigma^2}\right) \sum_{i=1}^n x_i^2 - \frac{n}{2} \ln(2\pi\sigma^2)\right). \end{aligned}$$

Dies ist eine dreiparametrische exponentielle Familie. Sei zunächst  $\mathbb{R}^+ := \{x \in \mathbb{R} : x > 0\}$ . In der Tat, setzt man  $T_1(\mathbf{y}) := \sum_{i=1}^n y_i$ ,  $T_2(\mathbf{y}) := \sum_{i=1}^n y_i^2$ ,  $T_3(\mathbf{y}) := \sum_{i=1}^n x_i y_i$  sowie  $c_1(\boldsymbol{\theta}) := \beta_1/\sigma^2$ ,  $c_2(\boldsymbol{\theta}) := -(2\sigma^2)^{-1}$ ,  $c_3(\boldsymbol{\theta}) := \beta_2/\sigma^2$ , so erhält man, mit entsprechender Wahl von  $d$  und  $S \equiv 0$ ,  $A := \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$  eine Darstellung der Form (2.8). Damit ist die Statistik

$$T(\mathbf{Y}) := \left( \sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2, \sum_{i=1}^n x_i Y_i \right)^\top$$

suffizient für  $\boldsymbol{\theta} = (\beta_1, \beta_2, \sigma^2)^\top$ .

## 2.4 Bayesianische Modelle

Bis jetzt haben wir angenommen, dass keine weiteren Informationen bezüglich der Parameter vorliegen außer den Daten. In den Anwendungen gibt es aber Situationen, in denen sich weitere Informationen beziehungsweise Annahmen gewinnbringend verwenden lassen. Wir stellen zwei Beispiele vor.

**Beispiel 2.20.** (*Qualitätssicherung unter Vorinformation.*) Wir betrachten die Situation von Beispiel 2.1. Allerdings nehmen wir an, dass bereits in der Vergangenheit Ladungen angenommen wurden, welches eine Vorinformation darstellt, die genutzt werden sollte. Es handele sich um  $K$  Lieferungen mit jeweils (der Einfachheit halber)  $N$  Teilen. Mit  $H_i$  sei die Anzahl derjenigen Lieferungen mit  $i$  defekten Teilen bezeichnet. Definieren wir die empirischen Häufigkeiten

$$\pi_i := \frac{H_i}{K},$$

so induzieren  $\pi_1, \dots, \pi_N$  ein Wahrscheinlichkeitsmaß, welches die Vorinformation summiert. Daher kann der Anteil  $\theta$  der defekten Teile pro Ladung kann als zufällig betrachtet werden und die Vorinformation liefert  $\mathbb{P}(\theta = \frac{i}{N}) = \pi_i$ . Dies bezeichnet man als die *a priori Verteilung* von  $\theta$ .

Es kommt eine neue Lieferung vom Umfang  $N$  an, welche untersucht werden soll.  $\theta$  bezeichne den (zufälligen) Anteil der defekten Teile in der Lieferung. Wir nehmen nun an, dass  $\theta$  nach  $\pi$  verteilt ist, d.h.  $\mathbb{P}(\theta = \frac{i}{N}) = \pi_i$ . Untersucht werde eine Stichprobe vom Umfang  $n$ , und  $X$  bezeichne den Anteil defekter Teile der Stichprobe. Wie in Beispiel 2.1 ist die bedingte Verteilung von  $X$  gegeben  $\theta$  eine hypergeometrische Verteilung, d.h. nach Gleichung (2.1) ist

$$P\left(X = k \mid \theta = \frac{i}{N}\right) = \frac{\binom{i}{k} \binom{N-i}{n-k}}{\binom{N}{n}},$$

welches eine *Hypergeo*( $i, N, n$ )-Verteilung ist. Für die gemeinsame Verteilung von  $(X, \theta)$  erhalten wir

$$P\left(X = k, \theta = \frac{i}{N}\right) = P\left(\theta = \frac{i}{N}\right) \cdot P\left(X = k \mid \theta = \frac{i}{N}\right) = \pi_i \frac{\binom{i}{k} \binom{N-i}{n-k}}{\binom{N}{n}}.$$

Schließlich ergibt sich für die Wahrscheinlichkeit, dass  $k$  Teile der Stichprobe defekt sind, unter Nutzung der Vorinformation, dass

$$P(X = k) = \sum_{i=1}^N \pi_i \frac{\binom{i}{k} \binom{N-i}{n-k}}{\binom{N}{n}}.$$

Dies ist eine gewichtete Form der bedingten Verteilungen von  $X$ . Wenn etwa für ein festes  $\theta_0 = \frac{i_0}{N}$  gilt, dass  $\pi_{i_0} = 1$  und 0 sonst, so erhält man wieder die ungewichtete Darstellung (2.1).

### Beispiel 2.21. (*Operational Risk.*)

Diese Vorgehensweise nennt man einen *Bayesianischer Ansatz*: Man nimmt an, dass der Wert des unbekanntem Parameters eine Realisierung einer Zufallsvariable mit gegebener *a priori Verteilung* (prior) ist. Die *a priori Verteilung* summiert die Annahmen über den wahren Wert des Parameters *bevor die Daten erhoben worden sind*, etwa wenn Vorinformationen oder subjektive Einschätzungen vorliegen. Man spricht daher auch von *subjektiver Inferenz*.

**Definition 2.15.** Ein *Bayesianisches Modell* für Daten  $\mathbf{X}$  und Parameter  $\theta$  ist spezifiziert durch

- (i) eine *a priori* Verteilung  $\pi$ , so dass  $\boldsymbol{\theta} \sim \pi$
- (ii) eine reguläre Verteilung  $\mathbb{P}_{\boldsymbol{\theta}}$ , so dass  $\mathbf{X}|\boldsymbol{\theta} \sim \mathbb{P}_{\boldsymbol{\theta}}$ .

Der zentrale Punkt der Bayesianischen Statistik ist, dass man das Vorwissen – gegeben durch die *a priori* Verteilung – nach Erhebung der Daten  $\mathbf{X}$  an das neu gewonnene Wissen über  $\boldsymbol{\theta}$  unpasst. Dies geschieht durch Bestimmung der bedingten Verteilung von  $\boldsymbol{\theta}$  gegeben der Daten  $\mathbf{X}$ . Diese Verteilung wird als *a posteriori* Verteilung bezeichnet. Sie ist gegeben durch Dichte oder Wahrscheinlichkeitsfunktion  $p(\boldsymbol{\theta} | \mathbf{X})$  und kann mit Hilfe des Satzes von Bayes (Satz 1.1) bestimmt werden:

$$p(\boldsymbol{\theta} | \mathbf{x}) = \frac{\pi(\boldsymbol{\theta}) \cdot p(\mathbf{x} | \boldsymbol{\theta})}{m(\mathbf{x})},$$

wobei  $m(\mathbf{x})$  die unbedingte Verteilung oder *marginale Verteilung* von  $\mathbf{X}$  bezeichnet. Ist  $\boldsymbol{\theta}$  diskret mit Werten  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_T$ , so ist

$$m(\mathbf{x}) = \sum_{i=1}^T \pi(\boldsymbol{\theta}_i) \cdot p(\mathbf{x} | \boldsymbol{\theta}_i).$$

Ist  $\boldsymbol{\theta}$  hingegen eine stetige Zufallsvariable, so ist

$$m(\mathbf{x}) = \int \pi(\boldsymbol{\theta}) \cdot p(\mathbf{x} | \boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Wie man sieht, ist  $m$  bereits durch  $\pi$  und  $p$  bestimmt. Oft beschreibt man deswegen  $p(\boldsymbol{\theta}|\mathbf{x})$  nur bis auf Proportionalität. Die Normierung, in diesem Falle  $m$ , bestimmt sich durch die Bedingung, dass  $p(\boldsymbol{\theta}|\mathbf{x})$  sich zu eins summiert bzw. integriert, siehe Aufgabe 2.5. Wir schreiben kurz

$$p(\boldsymbol{\theta} | \mathbf{x}) \propto \pi(\boldsymbol{\theta}) \cdot p(\mathbf{x} | \boldsymbol{\theta}).$$

**Beispiel 2.22.** (*Konjugierte Familie der Bernoulli-Verteilung.*) Diese Beispiel betrachtet Bernoulli-Zufallsvariablen mit zufälligem Parameter  $\theta$ .  $\theta$  hat eine Beta-Verteilung als *a priori*-Verteilung. Dies führt zu einer Beta-Verteilung als *a posteriori*-Verteilung: Seien  $X_1, \dots, X_n$  i.i.d. Bernoulli, d.h.  $X_i \in \{0, 1\}$  mit  $\mathbb{P}(X_i = 1 | \theta) = \theta$ . Sei  $\theta \sim \pi$ . Setze  $S := \sum_{i=1}^n X_i$ . Dann ist die *a posteriori*-Verteilung gegeben durch

$$\mathbb{P}(\theta | \mathbf{X}) = \frac{\pi(\theta) \theta^S (1 - \theta)^{n-S}}{\int \pi(t) t^S (1 - t)^{n-S} dt}.$$

Die *a posteriori* Verteilung hängt nur von dem beobachteten Wert der beobachteten suffizienten Statistik  $S$  ab. Wählen wir für die *a priori*-Verteilung eine Beta( $a, b$ )-Verteilung, vorgestellt in Definition 1.14, so ist

$$\pi(\theta) = \frac{1}{B(a, b)} \theta^{a-1} (1 - \theta)^{b-1}.$$

Betrachten wir die Beobachtung  $\{S = s\}$ , so ist die a posteriori-Verteilung gerade

$$\mathbb{P}(\theta | \mathbf{X}) \propto \theta^{a+s-1}(1-\theta)^{n-s+b-1}.$$

Wir erhalten demnach die Dichte einer  $Beta(a+s, b+n-s)$ -Verteilung. Damit ist die a priori Verteilung aus der gleichen Klasse wie die a posteriori Verteilung.

Falls die a posteriori Verteilung zur selben Klasse von Verteilungen wie die a priori Verteilung gehört, dann spricht man von einer *konjugierten Familie*. Für exponentielle Familien können wir leicht konjugierte Familien angeben.

**Lemma 2.16.** Sei  $\mathbf{X} = X_1, \dots, X_n$  i.i.d. und Stichprobe einer  $K$ -parametrischen exponentiellen Familie mit Dichte oder Wahrscheinlichkeitsfunktion

$$p(\mathbf{x} | \boldsymbol{\theta}) = \exp \left( \sum_{j=1}^K c_j(\boldsymbol{\theta}) \cdot \sum_{i=1}^n T_j(\mathbf{x}_i) + \sum_{i=1}^n S(\mathbf{x}_i) + nd(\boldsymbol{\theta}) \right) \cdot \mathbf{1}_{A^n}(\mathbf{x}). \quad (2.9)$$

Durch die  $(K+1)$ -parametrische exponentielle a priori Verteilung

$$\pi(\boldsymbol{\theta}) \propto \exp \left( \sum_{j=1}^K c_j(\boldsymbol{\theta}) t_j + t_{K+1} d(\boldsymbol{\theta}) \right)$$

ist eine konjugierte Familie gegeben. Für die a posteriori Verteilung gilt

$$p(\boldsymbol{\theta} | \mathbf{x}) \propto \pi \left( \boldsymbol{\theta} \mid t_1 + \sum_{i=1}^n T_1(\mathbf{x}_i), \dots, t_K + \sum_{i=1}^n T_K(\mathbf{x}_i), t_{K+1} + n \right).$$

*Beweis.* Mit der gewählten a priori Verteilung gilt

$$\begin{aligned} p(\boldsymbol{\theta} | \mathbf{x}) &\propto p(\mathbf{x}, \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta} | t_1, \dots, t_{K+1}) \\ &\propto \exp \left( \sum_{j=1}^K c_j(\boldsymbol{\theta}) \left( \sum_{i=1}^n T_j(\mathbf{x}_i) + t_j \right) + (t_{K+1} + n) d(\boldsymbol{\theta}) \right) \\ &\propto \pi \left( \boldsymbol{\theta} \mid t_1 + \sum_{i=1}^n T_1(\mathbf{x}_i), \dots, t_K + \sum_{i=1}^n T_K(\mathbf{x}_i), t_{K+1} + n \right) \end{aligned}$$

und das ist die Behauptung.  $\square$

**Beispiel 2.23.** (*Konjugierte Familie der Normalverteilung bei bekannter Varianz.*) Seien  $X_1, \dots, X_n$  i.i.d. mit  $X_i \sim N(\mu, \sigma^2)$ . Die Varianz  $\sigma^2$  sei bekannt und der Erwartungswert  $\mu =: \theta$  unbekannt. Da für die Dichte einer Normalverteilung

$$p(x | \theta) \propto \exp\left(\frac{\theta x}{\sigma^2} - \frac{\theta^2}{2\sigma^2}\right),$$

erhalten wir wie in Beispiel 2.11 eine einparametrische exponentielle Familie mit  $T_1(x) = x$  und  $c_1(\theta) = \theta/\sigma^2$  wie in Gleichung 2.9. Die konjugierte zweiparametrische exponentielle Familie erhält man nach Lemma 2.16 durch die folgende a priori Verteilung mit Parameter  $(t_1, t_2)^\top$ :

$$\pi(\theta) \propto \exp\left(\frac{\theta}{\sigma^2} t_1 - \frac{\theta^2}{2\sigma^2} t_2\right).$$

Diese Dichte kann man als eine Normalverteilungsdichte (von  $\theta$ ) identifizieren:

$$\begin{aligned} \pi(\theta) &\propto \exp\left(-\frac{t_2}{2\sigma^2} \left(\theta^2 - \frac{2\sigma^2 \theta t_1}{t_2} + \left(\frac{t_1}{t_2}\right)^2\right)\right) \\ &= \exp\left(-\frac{t_2}{2\sigma^2} \left(\theta - \frac{t_1}{t_2}\right)^2\right); \end{aligned} \quad (2.10)$$

für  $t_2 > 0$  ist die eine  $\mathcal{N}(t_1/t_2, \sigma^2/t_2)$ -Verteilung. Damit ist die Frage nach der konjugierten Familie zunächst gelöst. Ein natürlichere Darstellung geht allerdings direkt von einer normalverteilten a priori-Verteilung aus, welche nun noch bestimmt werden soll. Sei dazu die a priori-Verteilung  $\pi$  eine  $\mathcal{N}(\eta, \tau^2)$ -Verteilung mit  $\tau^2 > 0$ ,  $\eta \in \mathbb{R}$ . Dies ergibt folgende Reparametrisierung:  $t_2 = \frac{\sigma^2}{\tau^2}$  und  $t_1 = \eta \frac{\sigma^2}{\tau^2}$ . Nach Lemma 2.16 ist die a posteriori Verteilung gegeben durch

$$p(\theta | \mathbf{x}) \propto \pi\left(\theta \mid t_1 + \sum_{i=1}^n T_1(x_i), t_2 + n\right).$$

Unter Verwendung der suffizienten Statistik lässt sich dies wie folgt ausdrücken: Da  $T_1(x) = x$  ist mit  $s := \sum_{i=1}^n x_i$  ist nach (2.10)

$$p(\theta | \mathbf{x}) \propto \phi\left(\theta \mid \frac{t_1 + s}{t_2 + n}, \frac{\sigma^2}{t_2 + n}\right),$$

wobei  $\phi(\theta | a, b^2)$  die Dichte einer  $\mathcal{N}(a, b^2)$ -Verteilung ist. Setzen wir die Reparametrisierung ein, so ergibt sich

$$\frac{t_1 + s}{t_2 + n} = w\bar{X} + (1 - w)\eta \quad \text{und} \quad \frac{\sigma^2}{t_2 + n} = \frac{\sigma^2}{\frac{\sigma^2}{\tau^2} + n}.$$

Der linke Ausdruck ist die a posteriori Erwartung, der rechte die a posteriori Varianz. Damit stellt sich die a posteriori Erwartung als gewichtetes Mittel der Stichprobenmittels  $\bar{X}$  und der a priori Erwartung dar. Außerdem gilt, dass  $w \rightarrow 1$  für  $n \rightarrow \infty$ , der Einfluss der a priori-Verteilung wird für zunehmende Stichprobengrößen immer geringer.

**Bemerkung 2.17.** *Nicht-informative a priori-Verteilung.* Falls man keine Vorinformation über den Parameter  $\boldsymbol{\theta}$  hat, dann verwendet man eine so genannte nicht-informative a priori Verteilung. Hierunter haben alle möglichen Parameter die gleiche Wahrscheinlichkeit (oder Dichte):

$$\pi(\boldsymbol{\theta}) \propto 1. \quad (2.11)$$

Ist der Parameterraum  $\Theta = \mathbb{R}^n$  und damit unbeschränkt, so gibt es keine nicht-informative a priori-Verteilung, denn die Dichte in Gleichung (2.11) integriert sich zu  $\int_{\mathbb{R}^n} d\boldsymbol{\theta} = \infty$ . Trotzdem kann man Gleichung (2.11) in manchen derartigen Fällen verwenden, falls nämlich die resultierende a posteriori Verteilung eine wohldefinierte Dichte bleibt. Man spricht auch von einem *improper non informative prior*, eine nicht wohldefinierte, nicht-informative a priori-Verteilung. Unter (2.11) gilt zunächst

$$p(\boldsymbol{\theta} | \mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} \propto p(\mathbf{x} | \boldsymbol{\theta}).$$

$p(\mathbf{x}|\boldsymbol{\theta})$  ist die so genannte *Likelihoodfunktion*  $L(\boldsymbol{\theta}; x_1, \dots, x_n)$ . Sie gibt an, welche Wahrscheinlichkeit (Likelihood) jeder Parameter  $\boldsymbol{\theta}$  unter der Beobachtung  $\{\mathbf{X} = \mathbf{x}\}$  hat. Die Likelihoodfunktion bildet die Grundlage der Maximum-Likelihood Schätzung, welche in Kapitel 3.3 ausführlich behandelt wird. Vorgreifend führt obige Beobachtung bereits zu einer Reihe von interessanten Konsequenzen:

- (i) Die a posteriori Verteilung ist proportional zur Likelihoodfunktion falls man eine nicht-informative a priori Verteilung wählt.
- (ii) Der Modus der a posteriori Verteilung ist der Maximum Likelihood Schätzer (im Gegensatz zum Erwartungswert), siehe das folgende Kapitel 3.3 zu Maximum-Likelihood Schätzern.
- (iii) Im nichtinformativen Fall ist die Likelihoodfunktion  $L : \mathbb{R}^n \rightarrow H$  eine Statistik mit Werten im Funktionenraum  $H := \{h : \Theta \rightarrow \mathbb{R}\}$  von Funktionen  $(x_1, \dots, x_n) \mapsto h(\cdot)$ . Weiterhin ist  $L$  ist suffizient für  $\boldsymbol{\theta}$  und eine Funktion jeder anderen suffizienten Statistik. Als unmittelbare Konsequenz folgt, dass, kennt man  $L$  nicht, so verliert man Information über  $\boldsymbol{\theta}$  in den Daten.

## 2.5 Aufgaben

**Aufgabe 2.1.** Sei  $(N_t)_{t \geq 0}$  ein Poisson-Prozess mit Intensität  $\lambda$  und Sprungzeitpunkten  $\tau_1, \tau_2, \dots$ . Definiere die Zwischenankunftszeiten  $X_i := \tau_i - \tau_{i-1}$ , wobei  $\tau_0 = 0$  gesetzt werde. Dann sind  $X_1, X_2, \dots$  unabhängig und  $X_i \sim \text{Exp}(\lambda)$ .

**Aufgabe 2.2.** Zeigen Sie, dass

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2.$$

**Aufgabe 2.3.** Bestimmen Sie die Verteilung der kanonischen Statistik einer einparametrischen exponentiellen Familie, falls die Familie durch eine Dichte darstellbar ist; siehe Satz 2.11.

**Aufgabe 2.4.** Seien  $X_1, \dots, X_n$  i.i.d. und Rayleigh-verteilt, d.h.  $X_i$  besitzt die Dichte  $x\sigma^{-2} \exp(-x^2/2\sigma^2)$ . Die natürliche suffiziente Statistik ist  $T(\mathbf{X}) = \sum_{i=1}^n X_i^2$ . Zeigen Sie, dass  $\mathbb{E}(T(\mathbf{X})) = 2n\sigma^2$  und  $\text{Var}(T(\mathbf{X})) = 4n\sigma^4$ .

**Aufgabe 2.5.** Bestimmen Sie die Normierungskonstante der a posteriori Verteilung

Vorläufig

## 3 Schatzmethoden

Gegeben sei ein statistisches Modell  $\mathcal{P}$  nach Definition 2.2. Dies ist eine Familie von Verteilungen  $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$  welche man als mögliche Verteilungen für eine Beobachtung  $\mathbf{X}$  betrachtet. Der Parameter  $\theta$  ist unbekannt. Mitunter ist man nicht daran interessiert, den vollständigen Parameter  $\theta$  zu schätzen; jede zu schätzende Größe kann man aber als Funktion  $q : \Theta \rightarrow \mathbb{R}$  auffassen, welches an den folgenden beiden Beispielen illustriert wird.

**Beispiel 3.1.** (*Qualitätssicherung aus Beispiel 2.1.*) Eine Ladung von  $N$  Teilen soll auf ihre Qualität untersucht werden. Die Ladung enthält defekte und nicht defekte Teile. Mit  $\theta$  sei der Anteil der defekten Teile bezeichnet. Man interessiert sich etwa für die Anzahl der defekten Teile und schätzt

$$q(\theta) = N \cdot \theta.$$

**Beispiel 3.2.** (*Messmodell aus Beispiel 2.2.*)  $n$  Messungen einer physikalischen Konstante  $\mu$  werden vorgenommen. Die Messergebnisse seien  $X_1, \dots, X_n$  und man nimmt an, dass  $X_i = \mu + \epsilon_i$ .  $\epsilon_i$  bezeichnet den Messfehler, und in Beispiel 2.2 wurden eine Reihe von möglichen Annahmen an die Messfehler vorgestellt. Unter Annahmen (i)-(v) sind die  $X_i$  i.i.d.  $\mathcal{N}(\mu, \sigma^2)$ -verteilt und  $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ . Gesucht ist die physikalische Konstante  $\mu$ , weswegen man  $q(\boldsymbol{\theta}) = \mu$  schätzt. Unter den Annahmen (i)-(iv) sind die  $\epsilon_i$  symmetrisch um Null verteilt und besitzen die unbekannte Dichte  $f$ . D.h.  $\boldsymbol{\theta} = (\mu, f)$  und man schätzt wieder  $q(\boldsymbol{\theta}) = \mu$ .

Um  $q(\boldsymbol{\theta})$  zu schätzen, wählt man eine Statistik  $T$  und wertet sie an den beobachteten Datenpunkten  $\mathbf{x} = (x_1, \dots, x_n)^\top$  aus. Falls der wahre unbekannte Wert für  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$  wäre, dann schätzt man die unbekannte Größe  $q(\boldsymbol{\theta}_0)$  durch die bekannte Größe  $T(\mathbf{x})$ . Oft verwenden wir auch die Notation  $T(\mathbf{X})$  für den zufälligen Schätzwert, ohne uns bereits auf die beobachteten Daten  $\mathbf{x}$  festzulegen.

**Beispiel 3.3.** (*Messmodell aus Beispiel 3.2.*) In dem Messmodell aus Beispiel 3.2 ist ein Schätzer für den unbekannt Parameter  $\mu$  aus den Daten  $\mathbf{X} = (X_1, \dots, X_n)^\top$

$$T(\mathbf{X}) := \frac{1}{n} \sum_{i=1}^n X_i.$$

Dann ist  $T(\mathbf{X})$  eine Zufallsvariable. Darüber hinaus ist  $T$  arithmetisches Mittel der Daten, und, wie im vorigen Kapitel gesehen, oft eine suffiziente Statistik. Liegt die Beobachtung  $\{\mathbf{X} = \mathbf{x}\}$  vor, so nimmt der Schätzer den Wert  $T(\mathbf{x})$  an.

Wie wählt man vernünftige Schätzer für  $q(\theta)$ ? In diesem Kapitel stellen wir vier Methoden hierfür vor:

- Substitutionsprinzip
- Momentenmethode
- Kleinste Quadrate
- Maximum Likelihood

Die Momentenmethode ist hierbei ein Spezialfall des Substitutionsprinzips, wie wir in Kürze sehen werden. Im Folgenden werden *Schätzungen* immer mit einem  $\hat{\cdot}$  bezeichnet: Insbesondere steht  $\hat{\theta}$  sowohl für die Zufallsvariable  $\hat{\theta}(\mathbf{X})$  Zufallsvariable, als auch für  $\hat{\theta}(\mathbf{x})$ , den Wert der Zufallsvariable falls  $\mathbf{X} = \mathbf{x}$  beobachtet wird. Wir sprechen auch vom *Schätzer*  $\hat{\theta}(\mathbf{x})$  mit *Schätzwert*  $\hat{\theta}(\mathbf{x})$ .

## 3.1 Substitutionsprinzip

Die Idee des Substitutionsprinzips ist die unbekannt Parameter in Beziehung zu Größen zu setzen, welche sich leicht schätzen lassen. Dieses allgemeine Prinzip erläutern wir in zwei wichtigen Fällen: Die Schätzung von Häufigkeiten durch relative Häufigkeiten, welche zur Häufigkeitssubstitution führt sowie die Schätzung von Momenten durch empirische Momente, welche zur Momentensubstitution führt.

### 3.1.1 Häufigkeitssubstitution

In diskreten Modellen lassen sich die Wahrscheinlichkeiten der Elementarereignisse unter geringen Voraussetzungen durch relative Häufigkeiten schätzen.

**Beispiel 3.4.** (*Relative Häufigkeiten.*) Die Zufallsvariablen  $X_1, \dots, X_n$  seien i.i.d. und jeweils Multinomialverteilt mit Klassen  $\nu_1, \dots, \nu_K$ , siehe Abschnitt 1.2. Demnach ist  $X_i \in \{\nu_1, \dots, \nu_K\}$  und es gelte  $p_k := \mathbb{P}(X_1 = \nu_k)$  für  $k \in \{1, \dots, K\}$ . Wir möchten einen Schätzer für  $p_1, \dots, p_K$  bestimmen unter Berücksichtigung der Eigenschaften  $\sum_{k=1}^K p_k = 1$  und  $p_k \in [0, 1]$ . Ein intuitiver Schätzer für  $p_k$  ist die *relative Häufigkeit*  $\hat{p}_k = \hat{p}_k(\mathbf{X})$  der Klasse  $k$ ,  $k \in \{1, \dots, K\}$ . Sie ist gegeben durch die Anzahl der Beobachtung in Klasse  $k$  geteilt durch die Gesamtzahl der Beobachtungen:

$$\hat{p}_k(\mathbf{X}) := \frac{1}{n} \sum_{i=1}^n 1_{\{X_i = \nu_k\}}.$$

Ein Datenbeispiel illustriert die Bestimmung der  $\hat{p}_k$ : Man klassifiziere Arbeitnehmer eines Betriebes in Stellenkategorien 1-5, wobei  $N_k$  Arbeitnehmer in Stellenkategorie  $k$  beschäftigt werden:

$k$	1	2	3	4	5
$N_k$	23	84	289	217	95
$\hat{p}_k$	0.03	0.12	0.41	0.31	0.13

Die relativen Häufigkeiten erhält man durch  $\hat{p}_k = N_k/n$  mit Gesamtzahl  $n = \sum_{k=1}^5 N_k = 708$  Beobachtungen. Man beachte, dass stets  $\hat{p}_k \in [0, 1]$  und  $\sum_{k=1}^5 \hat{p}_k = 1$ . Allgemeiner schätzt man die Funktion  $q(p_1, \dots, p_k)$  durch  $q(\hat{p}_1, \dots, \hat{p}_k)$  geschätzt, d.h. man substituiert die Wahrscheinlichkeiten  $p_1, \dots, p_K$  durch ihre Schätzer  $\hat{p}_1, \dots, \hat{p}_K$ . Sind beispielsweise in Kategorie 4 und 5 Facharbeiter beschäftigt und in Kategorie 2 und 3 Angestellte so wird die Anteilsdifferenz  $q(p_1, \dots, p_5) := (p_4 + p_5) - (p_2 + p_3)$  zwischen Facharbeitern und Angestellten durch

$$q(\hat{p}_1, \dots, \hat{p}_5) = (\hat{p}_4 + \hat{p}_5) - (\hat{p}_2 + \hat{p}_3) = (0.31 + 0.13) - (0.12 + 0.41) = -0.009$$

geschätzt.

Das im Beispiel verwendete Prinzip kann man auch allgemeiner formulieren. Sei dazu

$$F_n(x) := \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \leq x\}}$$

die *empirische Verteilungsfunktion*. Möchte man ein Funktional

$$q := \int f(y) dF(y)$$

schätzen, so ersetzt man  $F$  durch den (nichtparametrischen) Schätzer  $F_n$  und erhält als möglichen Schätzer

$$\hat{q} := \int f(y) dF_n(y) = \frac{1}{n} \sum_{i=1}^n f(X_i)$$

In der Tat, in Beispiel 3.4 ist  $p_k = \int 1_{\{y=\nu_k\}} dF(y)$  und somit  $\hat{p}_k = 1/n \sum 1_{\{X_i=\nu_k\}}$ .

Manchmal erhält man durch die Parametrisierungen Probleme mit der Eindeutigkeit, wie im folgenden illustriert wird. Falls  $p_1, \dots, p_k$  nicht frei wählbar sind, sondern stetige Funktionen eines  $m$ -dimensionalen Parameters  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)^T$  sind, und es gibt

$$q(\boldsymbol{\theta}) = h(p_1(\boldsymbol{\theta}), \dots, p_k(\boldsymbol{\theta}))$$

mit stetiger Funktion  $h$ , definiert auf

$$I_k = \left\{ (p_1, \dots, p_k) : p_i \geq 0 \forall i, \sum_{i=1}^k p_i = 1 \right\},$$

so schätzt man  $q$  durch  $\hat{q}(\theta) = h(\hat{p}_1, \dots, \hat{p}_k)$ .

**Beispiel 3.5.** (*Genotypen.*) Als Anwendungsbeispiel von Beispiel 3.4 betrachten wir ein Gen mit zwei Ausprägungen  $A$  und  $B$ . Sei  $\theta := \mathbb{P}(\text{Gen hat Ausprägung } A)$ , dann gibt es drei Genotypen mit den folgenden Häufigkeiten

	Typ 1	Typ 2	Typ 3
Häufigkeiten	$p_1 = \theta^2$	$p_2 = 2\theta(1 - \theta)$	$p_3 = (1 - \theta)^2$
„M=mother“	$M = A$	$M = A, F = B$	$M = B$
„F=father“	$F = A$	$M = B, F = A$	$F = B$

Die so erhaltenen Häufigkeiten werden in der Genetik als *Hardy-Weinberg Gleichgewicht* bezeichnet. Wesentlich hierbei ist, dass der Zusammenhang von  $p_1, p_2$  und  $p_3$  nun durch zwei Gleichungen bestimmt ist: Erstens,  $p_1 + p_2 + p_3 = 1$  und zweitens, durch die gemeinsame Abhängigkeit von  $\theta$ , wie oben erläutert. Dies wird in der Schätzung wie folgt berücksichtigt: Es werde eine Stichprobe vom Umfang  $n$  beobachtet. Sei  $N_i$  die Anzahl der Personen mit Genotyp  $i$  in der Stichprobe. Dann ist  $(N_1, N_2, N_3)$  Multinomial-verteilt,  $(N_1, N_2, N_3) \sim M(n, p_1, p_2, p_3)$  mit  $n = N_1 + N_2 + N_3$ . Man könnte zwei Substitutionen betrachten, etwa

$$\theta = \sqrt{p_1} \rightarrow \hat{\theta} = \sqrt{\hat{p}_1} = \sqrt{\frac{N_1}{n}}$$

und

$$\theta = 1 - \sqrt{p_3} \rightarrow \hat{\theta} = 1 - \sqrt{\frac{N_3}{n}}.$$

Die Parametrisierungen sind demnach unterschiedlich, obwohl in etwa die gleichen Werte herauskommen können.

### 3.1.2 Momentenmethode

Als einen Spezialfall des im vorigen Abschnittes formulierten Substitutionsprinzips erhält man die *Momentenmethode*. Betrachtet sei eine Stichprobe von i.i.d. Zufallsvariablen  $X_1, \dots, X_n$  mit Verteilung  $\mathbb{P}_\theta$ . Mit  $\mathbb{E}_\theta$  sei der Erwartungswert bezüglich der Verteilung  $\mathbb{P}_\theta$  bezeichnet und weiterhin seien

$$m_k(\theta) := \mathbb{E}_\theta(X^k), \quad k = 1, \dots, r$$

seien die ersten  $r$  Momente der generischen<sup>1</sup> Zufallsvariable  $X := X_1$  bezeichnet. Nach dem Substitutionsprinzip schätzt man die unbekanntenen Momente durch das  $k$ -te Stichprobenmoment

$$\hat{m}_k := \int x^k F_n(dx) = \frac{1}{n} \sum_{i=1}^n X_i^k.$$

Um  $q(\boldsymbol{\theta})$  zu schätzen, muss man einen Bezug zwischen  $\boldsymbol{\theta}$  und den Momenten herstellen.

Lässt sich  $q(\boldsymbol{\theta})$  als

$$q(\boldsymbol{\theta}) = g(m_1(\boldsymbol{\theta}), \dots, m_r(\boldsymbol{\theta})) \quad (3.1)$$

mit einer stetigen Funktion  $g$  darstellen, so schätzt man in der *Momentenmethode*  $q(\boldsymbol{\theta})$  durch

$$T(\mathbf{X}) = g(\hat{m}_1, \dots, \hat{m}_r).$$

Wir illustrieren die Momentenmethode in einer Reihe von Beispielen.

**Beispiel 3.6.** (*Normalverteilung.*) Die Daten einer Stichprobe  $X_1, \dots, X_n$  seien i.i.d. mit  $X_i \sim \mathcal{N}(\mu, \sigma^2)$  wie in Beispielen 2.2 und 2.18. Dann ist das erste Moment  $m_1 = \mu$  und somit  $\hat{\mu} = \hat{m}_1 = \bar{X}$ . Weiterhin gilt  $\sigma^2 = m_2 - (m_1)^2$ . Man schätzt also die Varianz mittels  $g(m_1, m_2) = m_2 - (m_1)^2$  und als Schätzer von  $\sigma^2$  ergibt sich

$$\hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^n X_i^2 - \left( \frac{1}{n} \sum_{i=1}^n X_i \right)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Man beachte, dass die Schätzer konsistent aber nicht erwartungstreu ist – im Gegensatz zur Stichprobenvarianz  $s^2(\mathbf{X})$  aus Beispiel 1.1, siehe auch Aufgabe 1.3.

Die Momentenmethode führt nicht zwingend zu einem eindeutigen Schätzer, denn typischerweise gibt es viele Darstellungen der Form (3.1), wie folgende Beispiele zeigen.

**Beispiel 3.7.** (*Bernoulli-Verteilung.*) Seien  $X_1, \dots, X_n$  i.i.d. Bernoulli-verteilt, siehe Beispiel 1.3. D.h.  $X_i \in \{0, 1\}$  und  $\mathbb{P}(X_i = 1) = \theta$ . In diesem Falle ist  $m_1(\theta) = P(X_i = 1) = \theta$  und somit  $\hat{\theta} = \bar{X}$  Momentenschätzer für  $\theta$ . Allerdings ist auch  $m_2(\theta) = \theta$  und erstaunlicherweise  $\hat{m}_2 = \hat{m}_1$ , da  $X_i \in \{0, 1\}$ . Für die Varianz gilt  $\text{Var}(X_1) = \theta(1 - \theta)$  und somit ist  $\bar{X}(1 - \bar{X})$  Momentenschätzer für  $\text{Var}(X_i)$ .

Dies muss allerdings nicht immer so sein:

<sup>1</sup>Da die  $X_1, \dots, X_n$  identisch verteilt sind, ist somit auch  $\mathbb{E}_\theta(X_i^k) = m_k(\theta)$ .

**Beispiel 3.8.** (*Poisson-Verteilung.*) Für eine Poisson-verteilte Zufallsvariable  $X$  gilt nach Aufgabe 1.5, dass  $\mathbb{E}(X) = \text{Var}(X) = \lambda$ . Damit erhält man aus der Momentenmethode zwei Schätzer:

$$\hat{\lambda}_1 := \bar{X} = \hat{m}_1$$

und

$$\hat{\lambda}_2 := \hat{m}_2 - (\hat{m}_1)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2.$$

Allerdings gilt typischerweise  $\hat{\lambda}_1 \neq \hat{\lambda}_2$ .

Dass die Momentenmethode nicht immer zu sinnvollen Ergebnissen führt, zeigt folgendes Beispiel.

**Beispiel 3.9.** (*Gleichverteilung.*) Eine Population mit  $\theta$  Mitgliedern, bezeichnet mit  $1, \dots, \theta$  werde betrachtet. Von dieser Population werde  $n$ -mal mit Wiederholung gezogen. Mit  $X_i$  werde die gezogene Nummer des  $i$ -ten Zuges bezeichnet. Dann gilt  $\mathbb{P}(X_i = r) = \frac{1}{\theta}$  für  $r = 1, \dots, \theta$  und  $i = 1, \dots, n$ . Ferner folgt

$$m_1(\theta) = \mathbb{E}(X_i) = \sum_{r=1}^{\theta} r \cdot \mathbb{P}(X_i = r) = \frac{1}{\theta} \sum_{r=1}^{\theta} r = \frac{1}{\theta} \cdot \frac{\theta(\theta+1)}{2} = \frac{\theta+1}{2}.$$

Schätzt man  $\theta$  durch Momentenmethode, so erhält man, dass  $\theta = 2m_1(\theta) - 1$  und somit ist der Momentenschätzer von  $\theta$  gerade

$$\hat{\theta} = 2\bar{X} - 1.$$

Allerdings ist dies *kein* sinnvoller Schätzer, wenn  $\max\{X_i\} > 2\bar{X} - 1 = \hat{\theta}$  ist, da natürlich  $\theta \geq \max\{X_i\}$  gilt.

**Bemerkung 3.1.** Die wesentlichen Merkmale der Momentenmethode sollen noch einmal kurz zusammengefasst werden.

- Der Momentenschätzer braucht nicht eindeutig zu sein.
- Substitutionsprinzipien ergeben im Allgemeinen einfach zu berechnende Schätzer. Daher werden sie häufig als erste bzw. vorläufige Schätzung benutzt.
- Falls der Stichprobenumfang groß ist ( $n \rightarrow \infty$ ), dann sind die Schätzungen nahe dem wahren Parameterwert. Diese Konsistenz wird in Abschnitt 4.4.1 genauer vorgestellt und diskutiert.

## 3.2 Methode der Kleinsten Quadrate

### 3.2.1 Allgemeine und lineare Regressionsmodelle

Die lineare Regression und in diesem Zusammenhang die Methode der kleinsten Quadrate ist eine überraschend alte Methode, die bereits Gauß für astronomische Messungen verwendete, siehe Gauß (1809). Die erhaltenen Formeln werden in der Numerik oft auch als verallgemeinerte Inverse benutzt. Das Anpassen der Regressionsgeraden an die Daten verwendete Prinzip der Minimierung eines quadratischen Abstandes hat in vielen Bereichen Anwendung, so etwa auch in der Kalibrierung in der Finanzmathematik.

Regressionsprobleme untersuchen die Abhängigkeit der *Zielvariablen* (Response, endogene Variable) von anderen Variablen (Kovariablen, unabhängige Variablen, exogene Variablen).

Der Begriff Regression geht hierbei auf Experimente zur Schätzung der Körpergröße von Söhnen basierend auf der Körpergröße ihrer Väter zurück.

**Definition 3.2.** Eine *allgemeine Regression* ist gegeben durch einen zu bestimmenden  $r$ -dimensionalen Parametervektor  $\boldsymbol{\theta} \in \Theta$  und bekannte, zufällige Funktionen  $g_1, \dots, g_n : \Omega \times \Theta \rightarrow \mathbb{R}$ . Das zugehörige *Modell* ist

$$Y_i = g_i(\boldsymbol{\theta}) + \epsilon_i \quad i = 1, \dots, n.$$

Weiterhin gilt, dass die Zufallsvariablen  $\epsilon_1, \dots, \epsilon_n$  folgende Bedingungen erfüllen:

- (i)  $\mathbb{E}(\epsilon_i) = 0$  für alle  $i = 1, \dots, n$ .
- (ii)  $\text{Var}(\epsilon_i) = \sigma^2 > 0$  für alle  $i = 1, \dots, n$ .  $\sigma^2$  ist unbekannt.
- (iii)  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$  für alle  $1 \leq i \neq j \leq n$ .

Die Variablen  $\epsilon_i$  stellen wie in Beispiel 2.2 Abweichungen von der systematischen Beziehung  $Y_i = g_i(\boldsymbol{\theta})$  dar. Die Bedingung (i) veranschaulicht, dass die Regression keinen systematischen Fehler macht. Die Bedingung (ii) verlangt eine homogene Fehlervarianz, was man als *homoskedastisch* bezeichnet.

Die Bedingungen (i)-(iii) gelten, falls etwa  $\epsilon_1, \dots, \epsilon_n$  i.i.d. mit verschwindendem Erwartungswert und  $\text{Var}(\epsilon_i) > 0$ . Ein wichtiger Spezialfall ist durch die zusätzliche Normalverteilungsannahme  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$  gegeben. An dieser Stelle sei noch einmal auf die Analogie zu den Annahmen des Messmodells aus Beispiel 2.2 verwiesen.

**Beispiel 3.10.** (*Messmodell aus Beispiel 2.2.*)  $n$  Messungen einer physikalischen Konstante  $\theta$  werden vorgenommen. Die Messergebnisse seien  $Y_1, \dots, Y_n$ . Variiert der Messfehler additiv um  $\theta$ , so erhält man

$$Y_i = \theta + \epsilon_i, \quad i = 1, \dots, n.$$

In diesem Falle ist  $r = 1$  und  $g_i(\boldsymbol{\theta}) = \theta$ .

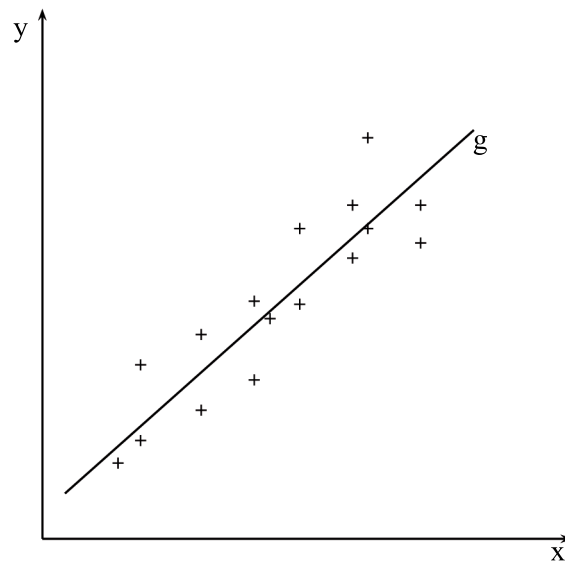


Abbildung 3.1: Eine einfache lineare Regression wie in Beispiel ?? vorgestellt. Beobachtet werden Paare  $(x_i, Y_i)$ ,  $i = 1, \dots, n$ , welche in der Abbildung durch Kreuze gekennzeichnet sind. Die den Daten angepasste Regressionsgerade  $g : x \rightarrow \hat{\theta}_1 + \hat{\theta}_2 x$  mit geschätzten Parametern  $\hat{\theta}_1$  und  $\hat{\theta}_2$  ist ebenfalls dargestellt.

**Beispiel 3.11.** (*Einfache lineare Regression.*) Die einfache lineare Regression wurde bereits in Beispiel 2.19 im Kontext von exponentiellen Familien betrachtet, welches wir an dieser Stelle wieder aufgreifen. Man beobachtet Paare von Daten  $(x_1, Y_1), \dots, (x_n, Y_n)$ .  $x_1, \dots, x_n$  werden als deterministische, bekannte Größen betrachtet und es wird angenommen, dass für  $i = 1, \dots, n$

$$Y_i = \theta_1 + \theta_2 x_i + \epsilon_i.$$

$\mathbf{Y}$  heißt *Zielvariable* mit Beobachtung  $Y_i$  und  $X_i$  heißt *Kovariable*. Wir verwenden  $g_i(\theta_1, \theta_2) = \theta_1 + \theta_2 x_i$ . In Abbildung 3.1 werden die Beobachtungen zusammen mit der geschätzten Regressionsgeraden  $x \rightarrow \hat{\theta}_1 + \hat{\theta}_2 x$  im Falle einer einfachen linearen Regression gezeigt.

### 3.2.2 Methode der kleinsten Quadrate

Bei dieser Methode schätzt man den unbekannt Parameter  $\boldsymbol{\theta}$  durch dasjenige  $\hat{\boldsymbol{\theta}}$ , welches den Abstand von  $\mathbb{E}(\mathbf{Y})$  und den beobachteten Daten  $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$  minimiert. Der Abstand wird als hierbei durch einen *quadratischen* Abstand  $Q$  gemessen.

**Definition 3.3.** Definiere den Quadratischen Abstand

$$Q(\boldsymbol{\theta}) := \sum_{i=1}^n (Y_i - g_i(\boldsymbol{\theta}))^2, \quad (3.2)$$

für jedes  $\boldsymbol{\theta} \in \Theta$ . Ein Wert  $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{Y})$  heißt *Kleinste-Quadrate Schätzer* (KQS) für  $\boldsymbol{\theta}$ , falls  $Q(\hat{\boldsymbol{\theta}}) \leq Q(\tilde{\boldsymbol{\theta}})$  für alle  $\tilde{\boldsymbol{\theta}} \in \Theta$ .

Ein KQS wird auch als Least Squares Estimator bezeichnet. Sind die Funktionen  $g_i$  in  $\boldsymbol{\theta}$  differenzierbar, und ist das Bild von  $(g_1, \dots, g_n)$  abgeschlossen, so ist  $\hat{\boldsymbol{\theta}}$  wohldefiniert. Falls darüber hinaus  $\Theta$  offen ist, so muss  $\hat{\boldsymbol{\theta}}$  notwendigerweise die so genannten *Normalgleichungen*

$$\frac{\partial}{\partial \theta_j} Q(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0, \quad j = 1, \dots, r$$

erfüllen. Nach Definition von  $Q$ , Gleichung (3.2), sind sie äquivalent zu

$$\sum_{i=1}^n \left( [y_i - g_i(\boldsymbol{\theta})] \cdot \frac{\partial}{\partial \theta_j} g_i(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right) = 0 \quad j = 1, \dots, r. \quad (3.3)$$

**Bemerkung 3.4.** Sind  $g_i(\theta_1, \dots, \theta_r)$  linear in  $\theta_1, \dots, \theta_r$ , so erhält man ein lineares Gleichungssystem, welches man explizit lösen kann.

Die Kleinst-Quadrate-Methode soll nun an den obigen Beispielen illustriert werden.

**Beispiel 3.12.** (*Messmodell.*) Sei wie in Beispiel 3.10 ein lineares Modell gegeben durch

$$Y_i = \theta + \epsilon_i, \quad i = 1, \dots, n.$$

Dann ist  $g_i(\theta) = \theta$  und somit  $\frac{\partial}{\partial \theta} g_i(\theta) = 1$  für alle  $i = 1, \dots, n$ . Die Normalgleichungen (3.3) ergeben

$$\sum_{i=1}^n (g_i(\hat{\theta}) - \hat{\theta}) = 0.$$

Hieraus folgt unmittelbar, dass  $\hat{\theta} = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ , das arithmetische Mittel der Beobachtungen. Der durch die Momentenmethode in Beispiel 3.6 erhaltene Schätzer ist also gleich dem Schätzer, welcher aus der Kleinst-Quadrate-Methode errechnet wird. Nach Beispiel 2.18 ist  $\bar{Y}$  darüber hinaus suffiziente Statistik für  $\theta$ .

**Beispiel 3.13.** (*Einfache lineare Regression.*) In Fortsetzung von Beispiel 3.11 betrachten wir ein lineares Modell gegeben durch

$$Y_i = \theta_1 + \theta_2 x_i + \epsilon_i, \quad i = 1, \dots, n.$$

In diesem Fall ist  $g_i(\boldsymbol{\theta}) = \theta_1 + \theta_2 x_i$  und  $\frac{\partial g_i}{\partial \theta_1}(\boldsymbol{\theta}) = 1$ ,  $\frac{\partial g_i}{\partial \theta_2}(\boldsymbol{\theta}) = x_i$ . Die Normalgleichungen (3.3) erhalten damit folgende Gestalt:

$$\sum_{i=1}^n (Y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i) \cdot 1 = 0 \quad (3.4)$$

$$\sum_{i=1}^n (Y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i) \cdot x_i = 0 \quad (3.5)$$

Aus Gleichung (3.4) erhält man mit  $\bar{Y} := \frac{1}{n} \sum_{i=1}^n Y_i$  und  $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$ , dass

$$\hat{\theta}_1 = \bar{Y} - \hat{\theta}_2 \bar{x}.$$

Setzt man dies in (3.5) ein, so ergibt sich

$$\begin{aligned} \sum_{i=1}^n x_i y_i - (\bar{Y} - \hat{\theta}_2 \bar{x}) \sum_{i=1}^n x_i - \hat{\theta}_2 \sum_{i=1}^n x_i^2 &= 0 \\ \Leftrightarrow \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{Y} \bar{x} &= \hat{\theta}_2 \left( \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \right). \end{aligned}$$

Damit erhält man die Schätzer der *einfachen linearen Regression*

$$\hat{\theta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\theta}_1 = \bar{Y} - \hat{\theta}_2 \bar{x}.$$

Die Gerade  $x \mapsto \hat{\theta}_1 + \hat{\theta}_2 x$  heißt *Regressionsgerade*. Sie minimiert die Summe der quadratischen Abstände zwischen  $(x_i, Y_i)$  und  $(x_i, \theta_1 + \theta_2 x_i)$ . Der Erwartungswert von  $Y_i$ , gegeben durch  $\mathbb{E}(Y_i) = \theta_1 + \theta_2 x_i$  wird durch

$$\hat{Y}_i := \hat{\theta}_1 + \hat{\theta}_2 x_i, \quad i = 1, \dots, n$$

geschätzt. Die Regressionsgerade zusammen mit  $Y_i$  und  $\hat{Y}_i$  werden in Abbildung 3.2 illustriert.

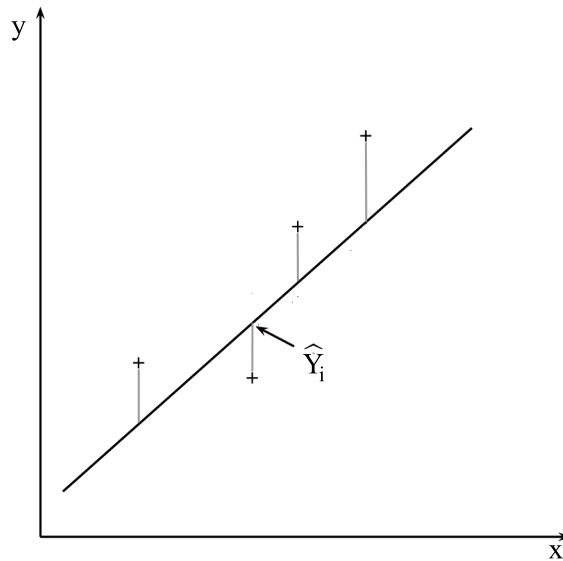


Abbildung 3.2: Illustration der Regressionsgeraden  $g : x \mapsto \hat{\theta}_1 + \hat{\theta}_2 x$  und der Erwartung eines Datenpunktes  $\hat{Y}_i = \hat{\theta}_1 + \hat{\theta}_2 x_i$ ; siehe auch Abbildung 3.1.

### 3.2.3 Gewichtete Kleinste-Quadrate-Schätzer

In praktischen Anwendungen kann es durchaus nützlich sein, in allgemeinen Regressionsmodellen die Annahme (ii) aus Definition 3.2,  $\text{Var}(\epsilon_i) = \sigma^2$ , abzuschwächen. Dies hatten wir als homoskedastisch bezeichnet. Ist die Varianz der Fehler abhängig von  $i$ , so heißt das Modell heteroskedastisch.

Eine allgemeine Regression heißt *heteroskedastisch*, falls

$$\text{Var}(\epsilon_i) = \sigma^2 \cdot w_i$$

mit  $w_i > 0$ ,  $i = 1, \dots, n$ . Man nennt die  $w_i$  auch Gewichte und nimmt an, dass sie *bekannt* sind.

Da die Gewichte bekannt sind, kann man durch eine Reparametrisierung eine homoskedastische allgemeine Regression erhalten: Setze

$$Z_i := \frac{Y_i}{\sqrt{w_i}}$$

für  $i = 1, \dots, n$ . Mit  $g_i^*(\boldsymbol{\theta}) := g_i(\boldsymbol{\theta})w_i^{-1/2}$  und  $\epsilon_i^* := \epsilon_i w_i^{-1/2}$  erhält man

$$Z_i = g_i^*(\boldsymbol{\theta}) + \epsilon_i^*.$$

Dies ist eine homoskedastische allgemeine Regression, denn  $\mathbb{E}(\epsilon_i^*) = 0$ ,  $\text{Cov}(\epsilon_i^*, \epsilon_j^*) = 0$  und

$$\text{Var}(\epsilon_i^*) = \frac{1}{w_i} \cdot \text{Var}(\epsilon_i) = \frac{1}{w_i} w_i \sigma^2 = \sigma^2.$$

Als Schätzer in dem heteroskedastischen Modell erhält man den *gewichteten Kleinst-Quadrate Schätzer*  $\hat{\boldsymbol{\theta}}^w$ .  $\hat{\boldsymbol{\theta}}^w$  minimiert

$$\sum_{i=1}^n (Z_i - g_i^*(\boldsymbol{\theta}))^2 = \sum_{i=1}^n \frac{1}{w_i} (Y_i - g_i(\boldsymbol{\theta}))^2.$$

Im Kontext der einfachen linearen Regression wird in Aufgabe 3.1  $\hat{\boldsymbol{\theta}}^w$  bestimmt.

### 3.3 Maximum-Likelihood-Schätzung

Sicherlich die wichtigste und flexibelste Methode zur Bestimmung von Schätzern ist die Maximum-Likelihood-Methode. Es werde ein reguläres statistisches Modell  $\mathcal{P}$  gegeben durch eine Familie von Dichten oder Wahrscheinlichkeitsfunktionen  $\{p(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k\}$  betrachtet. Die Funktion

$$L(\boldsymbol{\theta}, \mathbf{x}) := p(\mathbf{x}, \boldsymbol{\theta})$$

heißt *Likelihoodfunktion*. Falls die Beobachtung  $\mathbf{X}$  eine diskrete Zufallsvariable ist, dann gibt  $L(\boldsymbol{\theta}, \mathbf{x})$  die Wahrscheinlichkeit an, die konkrete Beobachtung  $\{\mathbf{X} = \mathbf{x}\}$  unter dem Parameter  $\boldsymbol{\theta}$  zu erhalten. Mit anderen Worten, man kann  $L(\boldsymbol{\theta}, \mathbf{x})$  als Maß dafür interpretieren, wie wahrscheinlich (likely) der Parameter  $\boldsymbol{\theta}$  ist, falls  $\mathbf{x}$  beobachtet wird. Im stetigen Fall kann man diese Interpretation ebenfalls erlangen, indem man das Ereignis  $\mathbf{X}$  liegt in einer  $\epsilon$ -Umgebung von  $\mathbf{x}$  betrachtet und  $\epsilon$  gegen Null gehen lässt.

Die *Maximum Likelihood Methode* besteht darin, denjenigen Wert  $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{X})$  zu finden, unter den beobachtenden Daten die höchste Wahrscheinlichkeit erlangt.

**Definition 3.5.** Für das reguläre statistische Modell  $\mathcal{P}$  und Beobachtung  $\{\mathbf{X} = \mathbf{x}\}$  heißt  $\hat{\boldsymbol{\theta}}(\mathbf{X})$  *Maximum-Likelihood-Schätzer* (MLS), falls gilt, dass

$$L(\hat{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{x}) = \max \{L(\boldsymbol{\theta}, \mathbf{x}) : \boldsymbol{\theta} \in \Theta\}$$

Falls der MLS  $\hat{\boldsymbol{\theta}}(\mathbf{X})$  existiert, dann schätzen wir  $q(\boldsymbol{\theta})$  durch  $q(\hat{\boldsymbol{\theta}}(\mathbf{X}))$ . In diesem Fall heißt

$$q(\hat{\boldsymbol{\theta}}(\mathbf{X}))$$

der *Maximum Likelihood Schätzer* von  $q(\boldsymbol{\theta})$ . Dieser wird auch als MLE oder Maximum-Likelihood-Estimate bezeichnet.

Ist die Likelihood-Funktion differenzierbar in  $\boldsymbol{\theta}$ , so sind *mögliche* Kandidaten für den MLS durch die Bedingung

$$\frac{\partial}{\partial \theta_i} L(\boldsymbol{\theta}, \mathbf{x}) = 0, \quad i = 1, \dots, k$$

gegeben. Darüberhinaus ist die zweite Ableitung zu überprüfen, um festzustellen, ob es sich tatsächlich um ein Maximum handelt. Weitere Maxima könnten auch auf dem Rand des Parameterraums angenommen werden.

Für die praktische Anwendung ist es äußerst nützlich den Logarithmus der Likelihood-Funktion zu betrachten. Da der Logarithmus eine streng monoton wachsende Funktion ist, bleibt die Maximalität unter dieser Transformation erhalten. Durch

$$l(\boldsymbol{\theta}, \mathbf{x}) := \ln L(\boldsymbol{\theta}, \mathbf{x})$$

ist die *Log-Likelihood-Funktion*  $l$  definiert. Falls  $\Theta$  offen,  $l$  differenzierbar in  $\boldsymbol{\theta}$  für festes  $\mathbf{x}$  und  $\hat{\boldsymbol{\theta}}(\mathbf{x})$  existiert, so muß der Maximum-Likelihood-Schätzer  $\hat{\boldsymbol{\theta}}$  die *Log-Likelihood-Gleichung* erfüllen:

$$\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{0}. \quad (3.6)$$

Des weiteren sind hinreichende Bedingungen, etwa an die zweite Ableitung, zu überprüfen um zu verifizieren, dass  $\hat{\boldsymbol{\theta}}$  auch tatsächlich Maximalstelle ist.

**Bemerkung 3.6.** *Konkavität der Likelihood-Funktion.* Nicht immer muss man die zweite Ableitung bemühen, um Maximalität zu zeigen: Falls etwa  $L$  konkav ist, so ist eine Lösung von  $\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{x}) = 0$  für  $\boldsymbol{\theta} \in \mathbb{R}$  stets Maximum-Likelihood-Schätzer für  $\boldsymbol{\theta}$  – Gleiches gilt natürlich ebenso für  $l$ . In Abbildung 3.3 wird dies an einer konkaven Funktion illustriert. Hierbei ist eine Funktion  $f: \mathbb{R} \rightarrow \mathbb{R}$  konkav, falls  $f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y)$  für alle  $\lambda \in (0, 1)$ . Angewendet etwa auf die Log-Likelihood-Funktion  $l$  heisst das: Ist  $l$  zweimal differenzierbar in  $\boldsymbol{\theta}$ , so ist  $l$  konkav in  $\boldsymbol{\theta}$  genau dann, wenn  $\frac{d^2}{d\boldsymbol{\theta}^2} l(\boldsymbol{\theta}, \mathbf{x}) \leq 0$ .

**Beispiel 3.14.** (*Log-Likelihood-Funktion unter Unabhängigkeit.*) Sind die  $X_1, \dots, X_n$  unabhängig und gilt  $X_i$  hat Dichte oder Wahrscheinlichkeitsfunktion  $p_i(x, \boldsymbol{\theta})$ , so ist die Log-Likelihood-Funktion gegeben durch

$$l(\boldsymbol{\theta}, \mathbf{x}) = \ln \prod_{i=1}^n p_i(x_i, \boldsymbol{\theta}) = \sum_{i=1}^n \ln p_i(x_i, \boldsymbol{\theta}).$$

**Bemerkung 3.7.** Maximum-Likelihood-Schätzer müssen nicht notwendigerweise existieren und sind auch nicht immer eindeutig.

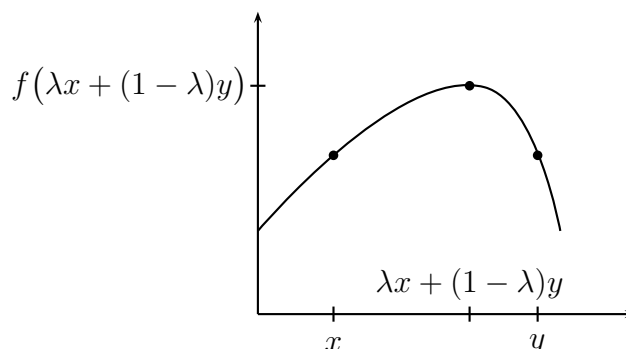


Abbildung 3.3: Ist die Funktion  $L$  konkav, so ist das Verschwinden der ersten Ableitung auch hinreichend für ein Maximum von  $L$ .

### 3.3.1 Maximum-Likelihood in eindimensionalen Modellen

In diesem Abschnitt nehmen wir an, dass  $\theta \in \mathbb{R}$  ein eindimensionaler Parameter ist. Wir beginnen mit zwei Beispielen

**Beispiel 3.15.** (*Normalverteilungsfall,  $\sigma$  bekannt.*) (Siehe Beispiel 2.11) Sei  $X$  normalverteilt,  $X \sim \mathcal{N}(\theta, \sigma^2)$  und die Varianz von  $X$ ,  $\sigma^2$ , sei bekannt. Mit der Dichte der Normalverteilung, gegeben in (1.4) erhält man die Likelihoodfunktion

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

Diese ist in Abbildung 3.4 dargestellt. Nach Beispiel 3.14 kann man dies leicht auf die i.i.d.-Situation übertragen: Seien  $X_1, \dots, X_n$  i.i.d. mit  $X_i \sim \mathcal{N}(\theta, \sigma^2)$ . Nach wie vor sei  $\sigma^2$  bekannt. Dann gilt für die Likelihood-Funktion

$$L(\theta, \mathbf{x}) \propto \exp\left(-\sum_{i=1}^n \frac{(x_i - \theta)^2}{2\sigma^2}\right).$$

Daraus erhält man die Log-Likelihood-Funktion mit einer geeigneten Konstanten  $c \in \mathbb{R}$

$$l(\theta, \mathbf{x}) = c - \sum_{i=1}^n \frac{(x_i - \theta)^2}{2\sigma^2}.$$

Die Log-Likelihood-Gleichung 3.6 ergibt direkt, dass

$$\hat{\theta} = \bar{X}.$$

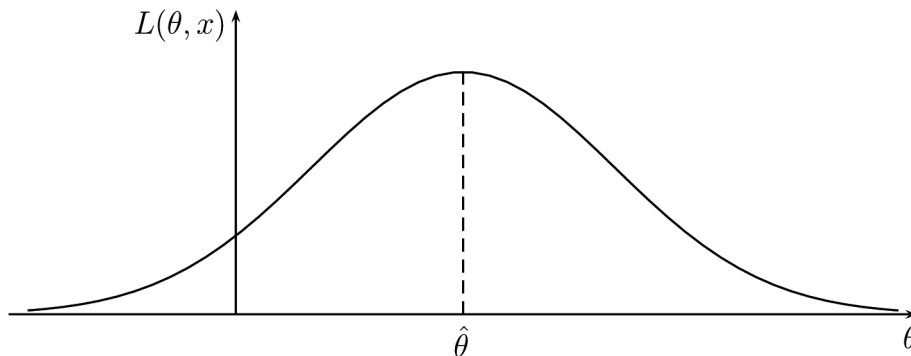


Abbildung 3.4: Die Likelihood-Funktion  $L$  aus Beispiel 3.15. Der Maximum-Likelihood-Schätzer  $\hat{\theta}$  maximiert die Likelihood-Funktion  $L(\theta, x)$  für festes  $x$ .

Die zweite Ableitung von  $l$  nach  $\theta$  ist negativ und somit ist das gefundene  $\hat{\theta}$  Maximalstelle.

Die verschiedenen Schätzmethoden für den Normalverteilungsfall, etwa die Momentenmethode in Beispiel 3.6, oder die Kleinst-Quadrate-Methode in Beispiel ?? ergeben also den gleichen Schätzer wie die Maximum-Likelihood-Methode.

**Beispiel 3.16.** (*Gleichverteilung.*) (Fortsetzung von Beispiel 3.9) Eine Population mit  $\theta$  Mitgliedern, bezeichnet mit  $1, \dots, \theta$  werde betrachtet. Von dieser Population werde  $n$ -mal mit Wiederholung gezogen. Mit  $X_i$  werde die gezogene Nummer des  $i$ -ten Zuges bezeichnet und das Maximum der Beobachtungen durch  $x_{(n)} := \max\{x_1, \dots, x_n\}$ . Es gilt, dass  $\mathbb{P}(X_i = r) = \theta^{-1} 1_{\{r \in \{1, \dots, \theta\}\}}$ . Nach Beispiel 3.14 ist die Likelihoodfunktion

$$\begin{aligned}
 L(\theta; \mathbf{x}) &= \prod_{i=1}^n \theta^{-1} 1_{\{x_i \in \{1, \dots, \theta\}\}} = \theta^{-n} 1_{\{x_{(n)} \leq \theta\}} \\
 &= \begin{cases} 0 & \text{für } \theta \in \{1, \dots, x_{(n)} - 1\} \\ \max\{x_1, \dots, x_n\}^{-n} & \text{für } \theta = x_{(n)} \\ \theta^{-n} & \text{für } \theta > x_{(n)}. \end{cases}
 \end{aligned} \tag{3.7}$$

Damit ergibt sich  $\hat{\theta} = X_{(n)}$  als Kleinst-Quadrate-Schätzer. Die Likelihood-Funktion ist in Abbildung 3.5 dargestellt.

**Beispiel 3.17.** (*Genotypen.*) Wie in Beispiel 3.5 werde eine Population mit drei Genotypen bezeichnet durch  $1, 2, 3$  betrachtet. Sei mit  $p(i, \theta)$  die Wahrscheinlichkeit für Genotyp  $i$  für gegebenes  $\theta \in (0, 1)$ . Wir hatten gesehen, dass in dem so genannten Hardy-

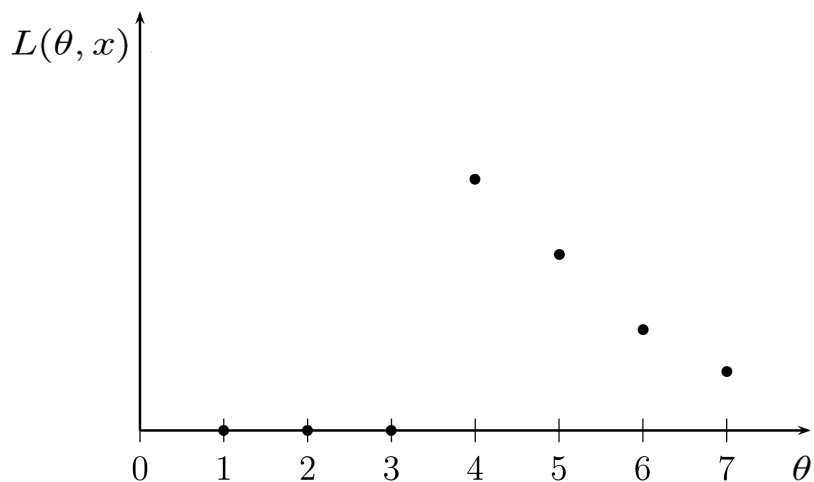


Abbildung 3.5: Die Likelihoodfunktion für eine Population mit  $\theta$  Mitgliedern, wie in Gleichung 3.7 berechnet. Die Darstellung ist für  $x_{(n)} = 4$ .

Weinberg-Gleichgewicht gilt, dass

$$p(1, \theta) = \theta^2, \quad p(2, \theta) = 2\theta(1 - \theta), \quad p(3, \theta) = (1 - \theta)^2.$$

für ein  $\theta \in (0, 1)$ . In einer Untersuchung werden drei nicht verwandte Personen typisiert.  $X_i$  bezeichne den Typ der  $i$ -ten Person. Die Untersuchung ergebe die Beobachtung  $\mathbf{x}_0 = (1, 2, 1)^\top$ . Dann ist die Likelihood-Funktion gegeben durch

$$L(\theta, \mathbf{x}_0) = p(1, \theta) \cdot p(2, \theta) \cdot p(1, \theta) = 2\theta^5(1 - \theta)$$

und somit ist die Log-Likelihood-Funktion

$$l(\theta, \mathbf{x}_0) = 5 \ln(\theta) + \ln(1 - \theta) + \ln(2).$$

Aus der notwendigen Bedingung für eine Maximalstelle, (3.6), folgt

$$\frac{\partial l(\theta, \mathbf{x})}{\partial \theta} = \frac{5}{\theta} - \frac{1}{1 - \theta} = 0$$

und somit  $\hat{\theta} = \frac{5}{6}$ . Um Maximalität nachzuweisen, überprüfen wir die zweite Ableitung. Da

$$\frac{\partial^2 l(\theta, \mathbf{x})}{\partial \theta^2} = -\frac{5}{\theta^2} - \frac{1}{(1 - \theta)^2} < 0$$

für alle  $\theta \in (0, 1)$ , ist  $\hat{\theta} = \frac{5}{6}$  Maximalstelle von  $L(\theta, \mathbf{x})$  und somit ein Maximum-Likelihood-Schätzer für  $\theta$  unter der Beobachtung  $\mathbf{X} = (1, 2, 1)$ . Die Situation mit  $n$  Beobachtungen wird in Beispiel 3.20 untersucht.

**Beispiel 3.18.** (*Warteschlange.*) (Siehe Beispiel 2.7) Sei  $X$  die Anzahl der Kunden, welche an einem Schalter in  $n$  Stunden ankommen. Wir nehmen an, dass die Anzahl der ankommenden Kunden einem Poissonprozess folgt und bezeichnen die Intensität (beziehungsweise die erwartete Anzahl Kunden pro Stunde) mit  $\lambda$ . Dann gilt  $X \sim \text{Pois}(n\lambda)$ . Mit der Wahrscheinlichkeitsfunktion einer Poisson-Verteilung, gegeben in Gleichung (1.3) erhält man die Likelihoodfunktion

$$L(\lambda, x) = \frac{e^{-\lambda n} (\lambda n)^x}{x!}$$

für  $x = 0, 1, \dots$ . Damit ist die Log-Likelihood-Funktion  $l(x, \lambda) = -\lambda n + x \ln(\lambda n) - \ln x!$  und für den Maximum-Likelihood-Schätzer folgt aus der Log-Likelihood-Gleichung (3.6)

$$\frac{\partial l(\lambda, x)}{\partial \lambda} = -n + \frac{x \cdot n}{\lambda \cdot n} = 0.$$

Somit ist  $\hat{\lambda} = x/n$ . Die zweite Ableitung ist  $-x/\lambda^2$ , welche für  $x > 0$  negativ ist, und somit ist für  $x > 0$  das arithmetische Mittel

$$\hat{\lambda} = \frac{x}{n}$$

der Maximum-Likelihood-Schätzer für  $\lambda$ . Gilt allerdings  $x = 0$ , so existiert kein MLS für  $\lambda$ .

Wieder bezeichnen wir für das betrachtete reguläre statistische Modell  $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$  mit  $E_\theta(T(\mathbf{X}))$  den Erwartungswert von  $T(\mathbf{X})$  bezüglich der Verteilung  $\mathbb{P}_\theta$ . Weiterhin wird das Bild von  $c$  mit  $c(\Theta) := \{c(\theta) : \theta \in \Theta\}$  bezeichnet.

**Satz 3.8** (MLS für eindimensionale exponentielle Familien). *Betrachtet werde das statistische Modell  $\mathcal{P}$  gegeben durch die Dichte oder Wahrscheinlichkeitsfunktion*

$$p(\mathbf{x}, \theta) = \exp\{c(\theta)T(\mathbf{x}) + d(\theta) + S(\mathbf{x})\}1_A(x), \quad \theta \in \Theta.$$

*Sei  $C$  das Innere von  $c(\Theta)$  und weiterhin sei  $c$  injektiv. Falls*

$$\mathbb{E}_\theta(T(\mathbf{X})) = T(\mathbf{x})$$

*eine Lösung  $\hat{\theta}(\mathbf{x})$  besitzt mit  $c(\hat{\theta}(\mathbf{x})) \in C$ , dann ist  $\hat{\theta}(\mathbf{x})$  der eindeutige Maximum-Likelihood-Schätzer von  $\theta$ .*

*Beweis.* Betrachte zunächst die natürliche exponentielle Familie:

$$\begin{aligned} p(\mathbf{x}, \eta) &= \exp\{\eta \cdot T(\mathbf{x}) + d_0(\eta) + S(\mathbf{x})\}, \quad \mathbf{x} \in A \\ \Rightarrow \frac{\partial}{\partial \eta} l(\eta, \mathbf{x}) &= T(\mathbf{x}) + d'_0(\eta), \quad \frac{\partial^2}{\partial \eta^2} l(\eta, \mathbf{x}) = d''_0(\eta) \end{aligned}$$

Ist  $\eta$  innerer Punkt von  $\Theta$ , so gilt nach Bemerkung 2.13, dass

$$\begin{aligned} \mathbb{E}_\eta(T(\mathbf{X})) &= -d'_0(\eta), \\ \text{Var}_\eta(T(\mathbf{X})) &= -d''_0(\eta) > 0. \end{aligned}$$

Damit ist  $d''_0(\eta) < 0$ . Somit ist  $l$  konkav und die Likelihoodgleichung ist äquivalent zu  $\mathbb{E}_\eta(T(\mathbf{X})) = T(\mathbf{x})$ . Existiert eine Lösung für  $\mathbb{E}_\eta(T(\mathbf{x})) = T(\mathbf{x})$ , so muß diese Lösung der MLE sein. Eindeutigkeit folgt aus der strikten Konkavität von  $l$  (da  $d''_0(\eta) > 0$ ).

Den allgemeinen Fall behandeln wir wie folgt. Es gilt, dass

$$\{l(\theta, \mathbf{x}) = c(\theta)T(\mathbf{x}) + d(\theta) + S(\mathbf{x}), \theta \in \Theta\} \subset \{\eta \cdot T(\mathbf{x}) + d_0(\eta) + S(\mathbf{x}), \eta \in H\}$$

mit  $H = \{\eta : d_0(\eta) < \infty\}$  und  $\eta = c(\theta)$ . Falls  $\hat{\theta}$  Lösung von  $\mathbb{E}_\theta(T(\mathbf{X})) = T(\mathbf{x})$  ist, dann maximiert  $c(\hat{\theta})$  die Gleichung  $\eta \cdot T(\mathbf{x}) + d_0(\eta) + S(\mathbf{x})$  für alle  $\eta \in H$  und  $\hat{\eta} = c(\hat{\theta})$ . Dies folgt wegen der Eindeutigkeit von  $\hat{\eta}$  und da  $c : \Theta \rightarrow \mathbb{R}$  injektiv. Also maximiert  $\hat{\theta}$  die Log-Likelihood-Funktion  $l(\cdot, \mathbf{x})$ , da ein Maximum über einer kleineren Menge genommen wird.  $\square$

**Beispiel 3.19.** (*Normalverteilungsfall,  $\sigma$  bekannt.*) (Siehe Beispiel 3.15) Seien  $X_1, \dots, X_n$  i.i.d. mit  $X_i \sim \mathcal{N}(\theta, \sigma^2)$  und die Varianz  $\sigma^2$  sei bekannt. Nach Beispiel 2.18 ist die Verteilung von  $\mathbf{X} = (X_1, \dots, X_n)^\top$  eine exponentielle Familie mit kanonischer Statistik  $T(\mathbf{X}) = \sum_{i=1}^n X_i$ . Für  $T(\mathbf{X})$  gilt, dass

$$\mathbb{E}_\theta(T(\mathbf{X})) = n\theta$$

und somit ist die Bedingung  $\mathbb{E}_\theta(T(\mathbf{X})) = T(\mathbf{x})$  äquivalent zu

$$\theta = \frac{1}{n} \sum_{i=1}^n x_i.$$

Da  $c(\theta) = \theta/\sigma^2$  nach Beispiel 2.11 ist  $c$  injektiv und das Bild von  $c$  ist  $\mathbb{R}$ . Damit liegt  $\hat{\theta} := \bar{X}$  im Inneren des Bild von  $c$ . Mit Satz 3.8 folgt somit, dass  $\hat{\theta} = \bar{X}$  eindeutiger MLS ist.

**Beispiel 3.20.** (*Genotypen.*) Wir setzen Beispiel 3.17 fort. Dort wurde eine Population mit Genotypen 1, 2, 3 betrachtet und für den unbekannt Parameter  $\theta \in (0, 1)$  folgte, dass

$$p(1, \theta) = \theta^2, \quad p(2, \theta) = 2\theta(1 - \theta), \quad p(3, \theta) = (1 - \theta)^2. \quad (3.8)$$

Es werde eine Stichprobe  $X_1, \dots, X_n$  untersucht, wobei  $X_1, \dots, X_n$  i.i.d. mit  $X_i \in \{1, 2, 3\}$  seien und  $X_i$  habe Wahrscheinlichkeitsfunktion  $p(\cdot, \theta)$  aus Gleichung (3.8). Mit  $N_i$ ,  $i = 1, 2, 3$  werde die Anzahl der Beobachtungen mit Wert  $i$  bezeichnet. Dann ist

$$\mathbb{E}(N_1) = n \cdot P(X_1 = 1) = n \cdot p(1, \theta) = n\theta^2$$

und

$$\mathbb{E}(N_2) = n \cdot p(2, \theta) = 2n\theta(1 - \theta).$$

Weiterhin ist  $\mathbb{E}(N_1 + N_2 + N_3) = n$ . Betrachtet man eine Beobachtung  $\mathbf{x}$ , für welche sich  $n_1, n_2, n_3$  Elemente in den Gruppen 1, 2, 3 ergeben, so ist die Likelihood-Funktion gegeben durch

$$\begin{aligned} L(\theta, \mathbf{x}) &= \theta^{2n_1} (2\theta(1 - \theta))^{n_2} ((1 - \theta)^2)^{n_3} = 2^{n_2} \theta^{2n_1 + n_2} (1 - \theta)^{2n_3 + n_2} \\ &= \left( \frac{\theta}{1 - \theta} \right)^{2n_1 + n_2} (1 - \theta)^{2n} 2^{n_2}. \end{aligned}$$

Damit liegt eine eindimensionale exponentielle Familie vor mit  $T(\mathbf{X}) = 2N_1 + N_2$  und  $c(\theta) = \ln \frac{\theta}{1 - \theta}$ . Weiterhin ist

$$\mathbb{E}_\theta(T(\mathbf{X})) = \mathbb{E}_\theta(2N_1 + N_2) = 2n\theta^2 + 2n\theta(1 - \theta) = 2n\theta.$$

Damit ist  $\mathbb{E}_\theta(T(\mathbf{X})) = T(\mathbf{x})$  äquivalent zu  $2n\theta = 2n_1 + n_2$  und somit ist

$$\hat{\theta} = \frac{2n_1 + n_2}{2n}$$

nach Satz 3.8 der eindeutige MLS für  $\theta$ , denn  $c$  ist injektiv und darüber hinaus liegt  $c(\hat{\theta})$  im Inneren des Bildes von  $c$ .

**Bemerkung 3.9.** *Der MLS ist auch Momentenschätzer.* Da nach Satz 3.8  $\mathbb{E}_\theta(T(\mathbf{X})) = T(\mathbf{x})$  für den eindeutigen MLS in einer eindimensionalen exponentiellen Familie gilt, ist dieser auch der Momentenschätzer.

### 3.3.2 Maximum Likelihood in mehrdimensionalen Modellen

In diesem Abschnitt wird die Verallgemeinerung der Maximum-Likelihood-Methode auf den Fall vorgestellt, in welchem der Parameterraum  $\Theta$   $k$ -dimensional ist. Betrachte hierzu das reguläre statistische Modell  $\mathcal{P}$  gegeben durch eine Familie von Dichten oder Wahrscheinlichkeitsfunktionen  $\{p(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k\}$ . Wir nehmen an, dass  $\Theta$  offen ist. Falls die partiellen Ableitungen der Log-Likelihood-Funktion existiert und der MLS  $\hat{\boldsymbol{\theta}}$  existiert, so löst er die Log-Likelihood-Gleichung, (3.6),

$$\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{0}.$$

Wieder bezeichnen wir für das betrachtete reguläre statistische Modell  $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$  mit  $E_\theta(T(\mathbf{X}))$  den Erwartungswert von  $T(\mathbf{X})$  bezüglich der Verteilung  $\mathbb{P}_\theta$ . Weiterhin wird das Bild von  $c$  mit  $c(\Theta) := \{c(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$  bezeichnet. Der folgende Satz gibt Kriterien für einen eindeutigen MLS in  $K$ -parametrischen exponentiellen Familien.

**Satz 3.10.** *Betrachtet werde das statistische Modell  $\mathcal{P}$  gegeben durch die Dichte oder Wahrscheinlichkeitsfunktion aus einer  $K$ -parametrischen exponentiellen Familie,*

$$p(\mathbf{x}, \boldsymbol{\theta}) = \exp \left( \sum_{i=1}^K c_i(\boldsymbol{\theta}) T_i(\mathbf{x}) + d(\boldsymbol{\theta}) + S(\mathbf{x}) \right) 1_A(\mathbf{x}), \quad \boldsymbol{\theta} \in \Theta. \quad (3.9)$$

*Sei  $C$  das Innere von  $c(\Theta)$  und  $c_1, \dots, c_K$  injektiv. Falls*

$$\mathbb{E}_\theta(T_i(\mathbf{X})) = T_i(\mathbf{x}), \quad i = 1, \dots, K$$

*eine Lösung  $\hat{\boldsymbol{\theta}}(\mathbf{x})$  besitzt mit  $(c_1(\hat{\boldsymbol{\theta}}(\mathbf{x})), \dots, c_K(\hat{\boldsymbol{\theta}}(\mathbf{x})))^\top \in C$ , dann ist  $\hat{\boldsymbol{\theta}}(\mathbf{x})$  der eindeutige Maximum-Likelihood-Schätzer von  $\boldsymbol{\theta}$ .*

Der Beweis des Satzes ist ähnlich dem eindimensionalen Fall und ist Gegenstand von Aufgabe 3.2. In Verallgemeinerung von Beispiel 3.15 betrachten wir nun die allgemeine Situation der MLS von normalverteilten Beobachtungen.

**Beispiel 3.21.** *(MLS für Normalverteilung,  $\mu$  und  $\sigma$  unbekannt.)* Seien  $X_1, \dots, X_n$  i.i.d. mit  $X_i \sim \mathcal{N}(\mu, \sigma^2)$  und sowohl  $\mu$  als auch  $\sigma^2$  unbekannt. Setze  $\boldsymbol{\theta} := (\mu, \sigma^2)^\top$  und  $\Theta := \mathbb{R} \times \mathbb{R}^+$  mit  $\mathbb{R}^+ := \{x \in \mathbb{R} : x > 0\}$ . Nach Beispiel 2.17 führt die Darstellung der Normalverteilung als exponentielle Familie gemäß Gleichung (3.9) zu  $c_1(\boldsymbol{\theta}) = \mu/\sigma^2$  und

$c_2(\boldsymbol{\theta}) = -1/2\sigma^2$ . Damit ist  $C = \mathbb{R} \times \mathbb{R}^-$  mit  $\mathbb{R}^- := \{x \in \mathbb{R} : x < 0\}$ . Weiterhin sind

$$T_1(\mathbf{x}) = \sum_{i=1}^n x_i, \quad T_2(\mathbf{x}) = \sum_{i=1}^n x_i^2.$$

Daraus ergeben sich die folgenden beiden Gleichungen. Zunächst ist  $\mathbb{E}_{\boldsymbol{\theta}}(T_1(\mathbf{X})) = n\mu$ . Damit ist  $\mathbb{E}_{\boldsymbol{\theta}}(T_1(\mathbf{X})) = T_1(\mathbf{x})$  äquivalent zu

$$n\mu = \sum_{i=1}^n x_i,$$

woraus  $\hat{\mu} = \hat{\theta}_1 = \bar{X}$  folgt. Weiterhin ist

$$\mathbb{E}_{\boldsymbol{\theta}}(T_2(\mathbf{X})) = \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta}}(X_i^2) = n(\sigma^2 + \mu^2).$$

Damit ist  $\mathbb{E}_{\boldsymbol{\theta}}(T_2(\mathbf{X})) = T_2(\mathbf{x})$  äquivalent zu  $n(\sigma^2 + \mu^2) = \sum_{i=1}^n x_i^2$ . Wir erhalten

$$\hat{\sigma}^2 = \hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2,$$

falls  $n \geq 2$ .

Mit Satz 3.10 folgt, dass

$$\hat{\boldsymbol{\theta}} = \left( \bar{X}, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right)^\top$$

eindeutiger Maximum-Likelihood-Schätzer für  $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$  ist.

### 3.3.3 Numerische Bestimmung des MLS

Nicht immer lässt sich der Maximum-Likelihood-Schätzer direkt ausrechnen, sondern es sind mitunter numerische Methoden notwendig, um den MLS zu bestimmen, wie in folgendem Beispiel.

**Beispiel 3.22.** (*Diskret beobachtete Überlebenszeiten.*) Man untersucht gewisse Bauteile auf ihre Lebensdauer. Nimmt man an, dass die Bauteile ermüdungsfrei arbeiten, so bietet sich eine Exponentialverteilung zur Modellierung der Lebensdauer an, vergleiche auch

Aufgabe 1.6. Seien  $X_1, \dots, X_n$  i.i.d. und  $X_i \sim \text{Exp}(\theta)$  die Überlebenszeiten von  $n$  beobachteten Bauteilen. Allerdings werden die Bauteile nicht permanent untersucht, sondern nur zu den Zeitpunkten  $a_1 < a_2 < \dots < a_k$ . Setze  $a_0 := 0$  und  $a_{k+1} := a_k + 1$  (das Bauteil überlebt alle Inspektionen). Man beobachtet

$$Y_j := \begin{cases} a_l & \text{falls } a_{l-1} < X_j \leq a_l, \quad l = 1, \dots, k \\ a_{k+1} & \text{falls } X_j > a_k. \end{cases}$$

Sei  $N_i$  die Anzahl der  $Y_1, \dots, Y_n$  welche den Wert  $a_i$  annehmen,  $i = 1, \dots, k+1$ . Dann ist der Vektor  $(N_1, \dots, N_{k+1})^\top$  Multinomial-verteilt. Darüber hinaus ist er suffizient für  $\theta$ . Zur Berechnung der Likelihood-Funktion  $L$  setzen wir

$$p_j(\theta) := P(Y = a_j) = P(a_{j-1} < X \leq a_j) = e^{-\theta a_{j-1}} - e^{-\theta a_j}$$

für  $j = 1, \dots, k$  und

$$p_{k+1}(\theta) := P(Y = a_{k+1}) = P(X > a_k) = e^{-\theta a_k}$$

Dann ist

$$L(\theta, n_1, \dots, n_{k+1}) = \frac{n!}{n_1! \dots n_{k+1}!} \prod_{j=1}^{k+1} p_j(\theta)^{n_j}$$

und die Log-Likelihood-Funktion

$$l(\theta, n_1, \dots, n_{k+1}) = \sum_{j=1}^{k+1} n_j \log p_j(\theta) + c$$

mit von  $\theta$  unabhängigem  $c = c(n_1, \dots, n_{k+1})$ . Damit folgt aus der Log-Likelihood-Gleichung (3.6), dass der MLS  $\hat{\theta}$  folgende Gleichung lösen muss:

$$0 = \sum_{j=1}^{k+1} n_j \frac{\frac{\partial}{\partial \theta} p_j(\theta)}{p_j(\theta)} = \sum_{j=1}^k n_j \frac{a_j e^{-a_j \theta} - a_{j-1} e^{-a_{j-1} \theta}}{e^{-a_{j-1} \theta} - e^{-a_j \theta}} + n_{k+1} \frac{-a_k e^{-a_k \theta}}{e^{-a_k \theta}}$$

Falls  $a_j \neq b_j + d$  für alle  $j = 1, \dots, k$  kann (??) nicht mehr explizit gelöst werden und die Bestimmung von  $\hat{\theta}$  muss numerisch erfolgen.

Zur numerischen Bestimmung des MLS stellen wir kurz die *Newton-Methode* und deren Variante, die *Fischer-Scoring-Methode* vor. Hierbei möchte man die Log-Likelihood-Gleichung 3.6 lösen. Zunächst einmal lässt sich diese als nichtlineares Gleichungssystem der Form

$$\mathbf{h}(\boldsymbol{\theta}) = \begin{bmatrix} h_1(\theta_1, \dots, \theta_k) \\ \vdots \\ h_k(\theta_1, \dots, \theta_k) \end{bmatrix} = \mathbf{0} \quad (3.10)$$

schreiben. Sei  $\hat{\boldsymbol{\theta}}$  die Lösung von (3.10) und  $\boldsymbol{\theta}_0$  nahe bei  $\hat{\boldsymbol{\theta}}$ . Dann gilt mit der Taylorentwicklung 1. Ordnung um  $\boldsymbol{\theta}_0$

$$\mathbf{0} = \mathbf{h}(\hat{\boldsymbol{\theta}}) \approx \mathbf{h}(\boldsymbol{\theta}_0) + D\mathbf{h}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

mit

$$D\mathbf{h}(\boldsymbol{\theta}_0) = \begin{pmatrix} \left. \frac{\partial h_1}{\partial \theta_1} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} & \cdots & \left. \frac{\partial h_1}{\partial \theta_k} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ \vdots & & \vdots \\ \left. \frac{\partial h_k}{\partial \theta_1} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} & \cdots & \left. \frac{\partial h_k}{\partial \theta_k} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \end{pmatrix}$$

Aber die Gleichung  $\mathbf{h}(\boldsymbol{\theta}_0) + D\mathbf{h}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0}$  wird gelöst von

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - (D\mathbf{h}(\boldsymbol{\theta}_0))^{-1} \mathbf{h}(\boldsymbol{\theta}_0).$$

Dies wird nun in einem iterativen Verfahren eingesetzt: Sei  $\boldsymbol{\theta}_0$  ein Startwert und

$$\hat{\boldsymbol{\theta}}_{i+1} := \boldsymbol{\theta}_i - (D\mathbf{h}(\boldsymbol{\theta}_i))^{-1} \mathbf{h}(\boldsymbol{\theta}_i).$$

Man iteriert diesen Algorithmus so lange bis  $\|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i\|$  unter eine vorgegebene Schranke fällt und setzt dann  $\hat{\boldsymbol{\theta}} := \boldsymbol{\theta}_{i+1}$ .

Allgemeine Konvergenzaussagen sind vorhanden. In der Statistik wird im Allgemeinen  $D\mathbf{h}(\boldsymbol{\theta})$  von den Daten  $\mathbf{X}$  abhängen, d.h. man erhält eine zufällige Matrix. In der *Fisher-Scoring-Methode* wird deswegen  $\mathbb{E}_{\boldsymbol{\theta}}(D\mathbf{h}(\boldsymbol{\theta}_i, \mathbf{X}))$  an Stelle von  $D\mathbf{h}(\boldsymbol{\theta}_i, \mathbf{X})$  verwendet.

## 3.4 Vergleich der Maximum Likelihood Methode mit anderen Schätzverfahren

In diesem Abschnitt halten wir kurz einige Beobachtungen fest, die den MLS in andere Schätzmethoden einordnen.

- (i) Maximum Likelihood für diskrete Zufallsvariablen  $\mathbf{X}$  entspricht dem Substitutionsprinzip.
- (ii) Der Kleinste-Quadrate-Schätzer einer allgemeinen Regression unter Normalverteilungsannahme aus Abschnitt 3.2 kann als spezieller Maximum-Likelihood-Schätzer betrachtet werden: Für  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^\top$  und

$$Y_i = g_i(\boldsymbol{\theta}) + \epsilon_i, \quad i = 1, \dots, n$$

mit i.i.d.  $\epsilon_1, \dots, \epsilon_n$  und  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$  ist die Likelihood-Funktion gegeben durch

$$L(\boldsymbol{\theta}, \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - g_i(\boldsymbol{\theta}))^2\right) \quad (3.11)$$

Für alle  $\sigma^2 > 0$  ist (3.11) maximal genau dann, wenn

$$\sum_{i=1}^n (x_i - g_i(\theta_1, \dots, \theta_r))^2$$

minimal ist. Damit entspricht der Kleinste-Quadrate Schätzer dem Maximum-Likelihood-Schätzer.

- (iii) Für Bayes-Modelle mit endlichem Parameterraum  $\Theta$  und a priori Verteilung für  $\boldsymbol{\theta}$  die Gleichverteilung, ist der Maximum-Likelihood-Schätzer  $\hat{\boldsymbol{\theta}}$  derjenige Wert von  $\boldsymbol{\theta}$ , der die höchste a posteriori Wahrscheinlichkeit besitzt. Falls  $\Theta = [a, b]$  und  $\theta$  gleichverteilt, also  $\theta \sim U(a, b)$ , dann ist der Maximum-Likelihood-Schätzer  $\hat{\theta}$  der Modus der a posteriori Dichte.

## 3.5 Aufgaben

**Aufgabe 3.1.** *Gewichtete einfache lineare Regression.* Finden Sie eine Formel für den Kleinste-Quadrate-Schätzer  $\hat{\boldsymbol{\theta}}^w$  im Modell

$$Y_i = \theta_1 + \theta_2 x_i + \epsilon_i,$$

wobei  $\epsilon_1, \dots, \epsilon_n$  unabhängig seien mit  $\epsilon_i \sim N(0, \sigma^2 w_i)$ .

**Aufgabe 3.2.** Beweisen Sie die Aussage von Satz 3.10.

# 4 Vergleich von Schätzern: Optimalitätstheorie

## 4.1 Schätzkriterien

In diesem Abschnitt betrachten wir stets das statistische Modell  $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ . Wie kann man das Verhalten eines Schätzers  $T := T(\mathbf{X})$  für  $q(\theta)$  messen? Ein erster Ansatz wäre, den Fehler  $F := |T(\mathbf{X}) - q(\theta)|$  zu betrachten. Dieser erweist sich aus folgenden zwei Gründen als ungeeignet:

1.  $F$  hängt vom unbekanntem Parameter  $\theta$  ab.
2.  $F$  ist zufällig und kann erst nach der Datenerhebung zur Beurteilung herangezogen werden.

Es gilt vielmehr ein Kriterium zu finden, welches bereits vor der Datenerhebung zur Beurteilung eines Schätzers herangezogen werden kann. Daher mißt man die Qualität des Schätzers  $T(\mathbf{X})$  mit der Streuung des Schätzers um das gesuchte  $q(\theta)$ . Hierfür kommen zum Beispiel die folgenden beiden Maße in Frage.

**Definition 4.1.** Sei  $T(\mathbf{X})$  ein Schätzer für  $q(\theta)$ . Dann heißt

$$R(\theta, T) := \mathbb{E}_\theta \left( (T(\mathbf{X}) - q(\theta))^2 \right).$$

*mittlerer quadratischer Fehler (MQF)* von  $T$ . Weiterhin heißt

$$b(\theta, T) := \mathbb{E}_\theta (T(\mathbf{X})) - q(\theta)$$

*Verzerrung (bias)* des Schätzer  $T$ . Gilt  $b(\theta, T) = 0$ , so heißt  $T$  *unverzerrt*.

Als Alternative zu dem MQF kann man auch den *mittleren betraglichen Fehler*  $\mathbb{E}_\theta (|T(\mathbf{X}) - q(\theta)|)$  betrachten, was wir aber hier nicht eingehend vertiefen werden. Für den mittleren quadratischen Fehler erhält man unmittelbar die folgende wichtige Zerlegung in Varianz des Schätzers und Bias:

$$\begin{aligned} R(\theta, T) &= \mathbb{E}_\theta \left( (T(\mathbf{X}) - q(\theta))^2 \right) \\ &= \mathbb{E}_\theta \left( (T(\mathbf{X}) - \mathbb{E}_\theta(T(\mathbf{X})) + \mathbb{E}_\theta(T(\mathbf{X}) - q(\theta)))^2 \right) \\ &= \text{Var}_\theta(T(\mathbf{X})) + b^2(\theta, T). \end{aligned}$$

Man erkennt, dass der MQF sowohl von  $\boldsymbol{\theta}$  als auch von der Wahl des Schätzers  $T$  abhängt. Die Varianz  $\text{Var}_{\boldsymbol{\theta}}(T(\mathbf{X}))$  ist ein Maß der Präzision des Schätzers  $T(\mathbf{X})$ .

**Beispiel 4.1.** (MQF für die Normalverteilung.) Seien  $X_1, \dots, X_n$  i.i.d. mit  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ . Wie bereits in Beispiel 3.21 gezeigt, ist der MLS für  $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$  gegeben durch  $\hat{\mu} = \bar{X}$  und

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

Ferner gilt nach Beispiel 2.15, dass  $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ . Somit folgt, dass für  $q(\boldsymbol{\theta}) = \mu$

$$b(\boldsymbol{\theta}, \bar{X}) = \mathbb{E}_{\boldsymbol{\theta}}(\bar{X}) - q(\boldsymbol{\theta}) = \mu - \mu = 0,$$

d.h. das arithmetische Mittel  $\hat{\mu}$  ist unverzerrt, und für den mittleren quadratischen Fehler gilt

$$R(\boldsymbol{\theta}, \bar{X}) = \text{Var}_{\boldsymbol{\theta}}(\bar{X}) = \frac{\sigma^2}{n} \xrightarrow{n \rightarrow \infty} 0.$$

Damit verschwindet der MQF mit steigender Stichprobenzahl ( $n \rightarrow \infty$ ). Für den Schätzer  $\hat{\sigma}^2$  gilt

$$S := \frac{n\hat{\sigma}^2(\mathbf{X})}{\sigma^2} = \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi_{n-1}^2$$

nach Lemma ???. Damit folgt  $\mathbb{E}(S) = n - 1$  und  $\text{Var}(S) = 2(n - 1)$  nach Bemerkung 1.7. Für  $q(\boldsymbol{\theta}) = \sigma^2$  gilt somit, dass

$$b(\boldsymbol{\theta}, \hat{\sigma}^2(\mathbf{X})) = \frac{\sigma^2}{n} \mathbb{E}_{\boldsymbol{\theta}} \left( \frac{n\hat{\sigma}^2(\mathbf{X})}{\sigma^2} \right) - \sigma^2 = \frac{\sigma^2 \cdot (n - 1)}{n} - \sigma^2 = -\frac{\sigma^2}{n} \xrightarrow{n \rightarrow \infty} 0,$$

also ist  $\sigma^2(\mathbf{X})$  nicht unverzerrt. Allerdings ist  $\hat{\sigma}^2(\mathbf{X})$  asymptotisch unverzerrt. Dieses Manko behebt man durch Verwendung der Stichprobenvarianz  $s^2(\mathbf{X})$ , wie bereits in Aufgabe 1.3 besprochen. Als MQF erhält man

$$R(\boldsymbol{\theta}, \hat{\sigma}^2(\mathbf{X})) = \left( \frac{\sigma^2}{n} \right)^2 \text{Var}_{\boldsymbol{\theta}} \left( \frac{n\hat{\sigma}^2}{\sigma^2} \right) + \frac{\sigma^4}{n^2} = \frac{\sigma^4}{n^2} (2(n - 1) + 1) = \frac{\sigma^4(2n - 1)}{n^2} \xrightarrow{n \rightarrow \infty} 0.$$

**Bemerkung 4.2.** Im Allgemeinen ist es oft nicht möglich Verzerrung und mittleren quadratischen Fehler eines Schätzers zu berechnen und man muss sich mit Approximationen behelfen. Darüber hinaus ist der Vergleich des MQF zweier Schätzern nicht einfach, da häufig die Situation entsteht, dass in verschiedenen Teilen des Parameterraums  $\Theta$  jeweils unterschiedliche Schätzer besser sind. Eine solche Situation ist in Abbildung 4.1 dargestellt.

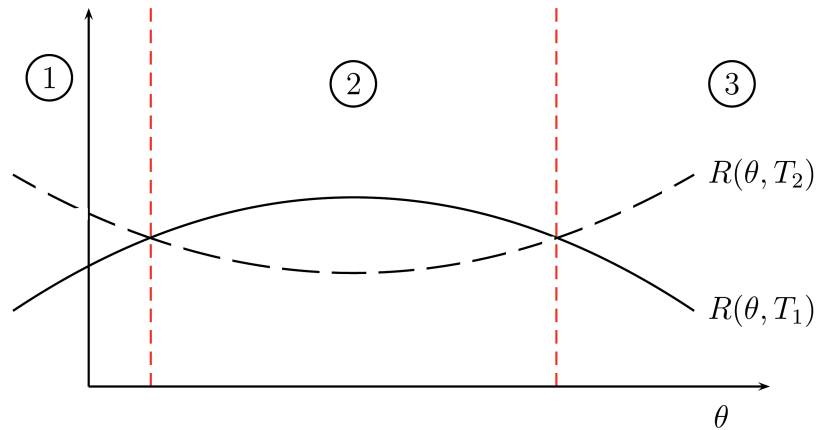


Abbildung 4.1: Vergleich des mittleren quadratischen Fehlers zweier Schätzer. In den Bereichen 1 und 3 hat Schätzer  $T_1$  einen geringeren MQF als Schätzer  $T_2$ , während die Umgekehrung in Bereich 2 der Fall ist.

**Beispiel 4.2.** (*Vergleich von Mittelwertschätzern anhand des MQF.*) In diesem Beispiel sollen die beiden Schätzer  $T_1(\mathbf{X}) = \bar{X}$  und  $T_2(\mathbf{X}) = a\bar{X}$ , mit einem  $a \in (0, 1)$  zur Schätzung des Mittelwertes im Normalverteilungsfall untersucht werden. Seien dazu  $X_1, \dots, X_n$  i.i.d. mit  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ . Wie im Beispiel 3.21 betrachten wir  $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ , d.h. Mittelwert und Varianz sind unbekannt und untersuchen  $q(\boldsymbol{\theta}) := \mu$ . Dann ist  $b(\boldsymbol{\theta}, T_1) = 0$  sowie  $R(\boldsymbol{\theta}, T_1) = \sigma^2/n$  nach Beispiel 4.1. Weiterhin ist

$$b(\boldsymbol{\theta}, T_2) = \mathbb{E}_\mu(T_2(\mathbf{X})) - \mu = a\mu - \mu = (a - 1)\mu,$$

und damit ergibt sich der MQF

$$R(\boldsymbol{\theta}, T_2) = \text{Var}_\mu(a\bar{X}) + ((a - 1)\mu)^2 = \frac{a^2\sigma^2}{n} + (a - 1)^2\mu^2.$$

Daran liest man ab, dass für  $|\mu|$  groß genug folgt, dass  $R(\boldsymbol{\theta}, T_1) < R(\boldsymbol{\theta}, T_2)$ , also Schätzer  $T_1$  ist besser als Schätzer  $T_2$ . Ist umgekehrt  $|\mu|$  nahe genug bei Null, so folgt dass  $R(\boldsymbol{\theta}, T_1) > R(\boldsymbol{\theta}, T_2)$  und somit ist in diesem Falle  $T_2$  ist besser als  $T_1$ . Damit liegt die Situation aus Bemerkung 4.2 vor. Zur Verdeutlichung ist die konkrete Situation in Abbildung 4.2 dargestellt.

**Definition 4.3.** Betrachtet werde ein statistisches Modell  $\{\mathbb{P}_\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta\}$  und sei  $S$  ein Schätzer. Gibt es einen weiteren Schätzer  $T$  mit  $R(\boldsymbol{\theta}, T) \leq R(\boldsymbol{\theta}, S)$  für alle  $\boldsymbol{\theta} \in \Theta$  und  $R(\boldsymbol{\theta}, T) < R(\boldsymbol{\theta}, S)$  für mindestens ein  $\boldsymbol{\theta} \in \Theta$  so heißt  $S$  *unzulässig*.

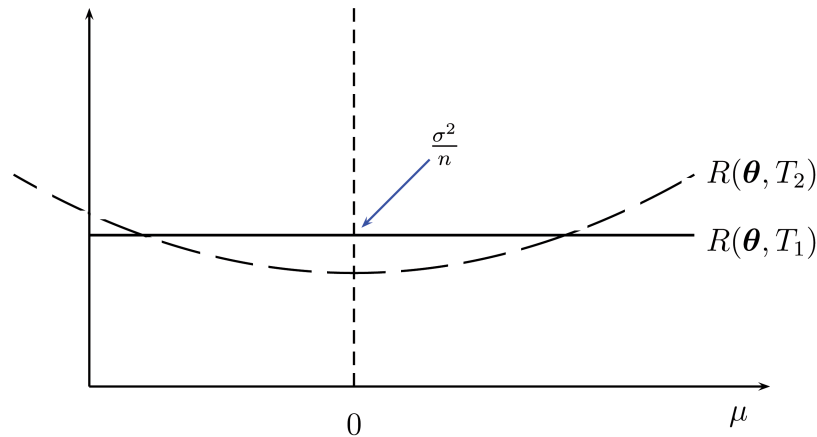


Abbildung 4.2: Vergleich des mittleren quadratischen Fehlers bezüglich  $\mu$  für die Schätzer  $T_1 = \bar{X}$  und  $T_2 = a\bar{X}$  bei normalverteilten Daten.

Für einen unzulässigen Schätzer gibt es also einen weiteren Schätzer, der echt besser ist im Sinne des MQF.

**Beispiel 4.3.** (*Der perfekte Schätzer*) Man ist versucht, zu fragen, ob es einen „besten“ Schätzer  $T$  gibt, für welchen

$$R(\boldsymbol{\theta}, T) \leq R(\boldsymbol{\theta}, S) \quad (4.1)$$

für alle Parameter  $\boldsymbol{\theta} \in \Theta$  und darüber hinaus auch für alle Schätzer  $S$  gilt? Leider ist dies nicht der Fall, wie man leicht sieht: Wählt man  $\boldsymbol{\theta}_0 \in \Theta$  beliebig und betrachtet den Schätzer  $S(\mathbf{X}) := q(\boldsymbol{\theta}_0)$ . Dieser Schätzer nutzt die erhobenen Daten nicht, trifft aber den wahren Parameter perfekt falls gerade  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ . Mit diesem Schätzer gilt, dass

$$R(\boldsymbol{\theta}_0, S) = \text{Var}_{\boldsymbol{\theta}_0}(S(\mathbf{X})) + (\mathbb{E}_{\boldsymbol{\theta}_0}(S(\mathbf{X})) - q(\boldsymbol{\theta}_0))^2 = 0.$$

Für den perfekten Schätzer  $T$  müsste (4.1) erfüllt sein, woraus wegen  $R(\boldsymbol{\theta}_0, T) = 0$  folgte, dass

$$R(\boldsymbol{\theta}, T) = 0$$

für alle  $\boldsymbol{\theta} \in \Theta$ . Das bedeutete, dass  $T(\mathbf{X})$  den gesuchten  $q(\boldsymbol{\theta})$  für alle  $\boldsymbol{\theta} \in \Theta$  perfekt schätzen, was in keinem natürlichen Modellen möglich ist.

Aus diesem Beispiel erkennen wir, dass es nicht sinnvoll ist alle möglichen Schätzer zu betrachten. Man muss die Klasse der zu betrachtenden Schätzer geeignet einschränken. Eine bereits bekannte und wünschenswerte Eigenschaft ist die Unverzerrtheit eines Schätzers. Für alle unverzerrten Schätzer gilt, dass der mittlere quadratische Fehler

$$R(\boldsymbol{\theta}, T) = \text{Var}_{\boldsymbol{\theta}}(T(\mathbf{x})).$$

Betrachtet man nur die Klasse der unverzerrten Schätzer und beurteilt die Qualität eines Schätzer mittels MQF, so wird der zunächst der systematische Fehler (Verzerrung) kontrolliert, bevor die Präzision des Schätzers betrachtet wird.

**Beispiel 4.4.** (*Unverzerrte Schätzer*) Haben  $X_1, \dots, X_n$  den Erwartungswert  $\mu$ , so ist das arithmetische Mittel  $\bar{X}$  ein unverzerrter Schätzer für  $\mu$ , denn

$$\mathbb{E}(\bar{X}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) = \mu.$$

Sind die  $X_i$  darüberhinaus unabhängig mit  $\text{Var}(X_i) = \sigma^2$ , so ist die Stichprobenvarianz ein unverzerrter Schätzer für  $\sigma^2$ , wie in Aufgabe 1.3 gezeigt.

Der Schätzer  $S(\mathbf{X}) := q(\theta_0)$  aus Beispiel 4.3 ist natürlich verzerrt, denn

$$b(\theta, S) - q(\theta) \neq 0$$

für alle  $\Theta \ni \theta$  welche von  $\theta_0$  verschieden sind.

## 4.2 UMVUE Schätzer

Unter den unverzerrten Schätzern kann man häufig einen Schätzer mit folgender Optimalitätseigenschaft finden.

**Definition 4.4.** Ein unverzerrter Schätzer  $T$  für welchen

$$R(\theta, T) = \text{Var}_\theta(T(\mathbf{X})) \leq \text{Var}_\theta(S(\mathbf{X}))$$

für alle unverzerrten Schätzer  $S$  gilt, heißt *UMVUE*-Schätzer.

UMVUE steht für Uniformly Minimum Variance Unbiased Estimator. Allerdings können eine Reihe von Problemen mit unverzerrten Schätzern auftreten:

- Unverzerrte Schätzer müssen nicht zu existieren.
- Ein UMVUE Schätzer muß nicht zulässig zu sein.
- Unverzerrtheit ist nicht invariant unter Transformation, d.h.  $\hat{\theta}$  kann unverzerrt für  $\theta$  sein, aber  $q(\hat{\theta})$  ist typischerweise ein verzerrter Schätzer für  $q(\theta)$ .

Im folgenden betrachte wir die Situation  $q(\theta)$  für  $\theta \in \Theta$  basierend auf einer beobachteten Stichprobe  $\mathbf{X} = (X_1, \dots, X_n) \sim P_\theta$  zu schätzen. Sei  $T(\mathbf{X})$  ein suffizienter Schätzer für  $\theta$ . Falls  $S(\mathbf{X})$  ein weiterer Schätzer für  $q(\theta)$  ist, dann können wir einen besseren (oder zumindest nicht schlechteren) Schätzer unter Mithilfe von  $T(\mathbf{X})$  wie folgt konstruieren:

Da  $T$  suffizient ist, hängt die Verteilung bedingt auf  $T(\mathbf{X})$  nicht von  $\boldsymbol{\theta}$  ab und man setzt  $\mathbb{E}(S(\mathbf{X})|T(\mathbf{X})) := \mathbb{E}_{\boldsymbol{\theta}}(S(\mathbf{X})|T(\mathbf{X}))$  für ein beliebiges  $\boldsymbol{\theta}$ . Definiere

$$T^*(t) := \mathbb{E}(S(\mathbf{X})|T(\mathbf{X})).$$

Im Zusammenhang mit dem folgenden Satz sagt man auch, dass  $T^*$  aus  $S$  mit Hilfe von  $T$  durch *Rao-Blackwellisierung* erzeugt wurde.

**Satz 4.5** (Rao-Blackwell). *Sei  $T(\mathbf{X})$  ein suffizienter Schätzer für  $\boldsymbol{\theta}$  und  $S$  ein Schätzer mit  $\mathbb{E}_{\boldsymbol{\theta}}(|S(\mathbf{X})|) < \infty$  für alle  $\boldsymbol{\theta} \in \Theta$ . Setze  $T^*(\mathbf{X}) := \mathbb{E}(S(\mathbf{X})|T(\mathbf{X}))$ . Dann gilt für alle  $\boldsymbol{\theta} \in \Theta$ , dass*

$$\mathbb{E}_{\boldsymbol{\theta}}\left(\left(T^*(\mathbf{X}) - q(\boldsymbol{\theta})\right)^2\right) \leq \mathbb{E}_{\boldsymbol{\theta}}\left(\left(S(\mathbf{X}) - q(\boldsymbol{\theta})\right)^2\right). \quad (4.2)$$

*Gilt darüber hinaus  $\text{Var}_{\boldsymbol{\theta}}(S) < \infty$ , so erhält man Gleichheit genau dann, wenn  $T^*(\mathbf{X}) = S(\mathbf{X})$   $\mathbb{P}_{\boldsymbol{\theta}}$ -fast sicher für alle  $\boldsymbol{\theta} \in \Theta$ .*

*Beweis.* Es gilt nach den Voraussetzungen, dass

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}(T^*) &= \mathbb{E}_{\boldsymbol{\theta}}(\mathbb{E}(S|T)) = \mathbb{E}_{\boldsymbol{\theta}}(S) \\ b(\boldsymbol{\theta}, T^*) &= \mathbb{E}_{\boldsymbol{\theta}}(T^*) - \boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta}}(S) - \boldsymbol{\theta} = b(\boldsymbol{\theta}, S) \end{aligned}$$

Damit haben  $T^*$  und  $S$  die gleiche Verzerrung. Somit gilt

$$\begin{aligned} (4.2) &\Leftrightarrow \text{Var}_{\boldsymbol{\theta}}(T^*) \leq \text{Var}_{\boldsymbol{\theta}}(S) \\ &\Leftrightarrow \mathbb{E}_{\boldsymbol{\theta}}(\mathbb{E}(S|T) - \mathbb{E}_{\boldsymbol{\theta}}(S))^2 \leq \mathbb{E}_{\boldsymbol{\theta}}(S - \mathbb{E}_{\boldsymbol{\theta}}(S))^2 \\ &\Leftrightarrow \mathbb{E}_{\boldsymbol{\theta}}((\mathbb{E}(S|T))^2) \leq \mathbb{E}_{\boldsymbol{\theta}}(S^2). \end{aligned}$$

Aber mit der Jensen'schen Ungleichung aus Satz 1.5 und der Monotonie des Erwartungswertes, (1.1), gilt

$$\mathbb{E}_{\boldsymbol{\theta}}((\mathbb{E}(S|T))^2) \leq \mathbb{E}_{\boldsymbol{\theta}}(\mathbb{E}(S^2|T)) = \mathbb{E}_{\boldsymbol{\theta}}(S^2).$$

Gleichheit gilt in der Jensen'schen Ungleichung  $\mathbb{E}_{\boldsymbol{\theta}}(\xi|\eta = x)^2 \leq \mathbb{E}_{\boldsymbol{\theta}}(\xi^2|\eta = x)$  genau dann, wenn  $\xi = \mathbb{E}_{\boldsymbol{\theta}}(\xi|\eta)$   $F_{\eta}$ -fast sicher und somit folgt auch der zweite Teil.  $\square$

Um Optimalitätsaussagen machen zu können, braucht man das Konzept der Vollständigkeit nach Lehmann und Scheffé. Optimalität wird im Rahmen des Vollständigkeitskonzeptes

so verifiziert, dass es für eine vorgegebene suffiziente Statistik  $T(\mathbf{X})$  im wesentlichen nur einen von  $T(x)$  abhängenden, erwartungstreuen Schätzer gibt. Das ist gleichbedeutend mit

$$\mathbb{E}_\theta(g_1(T)) = \mathbb{E}_\theta(g_2(T)) \text{ für alle } \theta \in \Theta \Rightarrow g_1 = g_2.$$

Dies führt zu folgender Definition.

**Definition 4.6.** Eine Statistik  $T$  heißt *vollständig*, falls aus

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

folgt, dass  $\mathbb{P}_\theta(g(T) = 0) = 1$  für alle  $\theta \in \Theta$ .

Eigentlich ist die Vollständigkeit eher eine Eigenschaft der Familie von betrachteten Verteilungen  $\{P_\theta : \theta \in \Theta\}$  beziehungsweise dem betrachteten statistischen Modell. Sie bedeutet, dass  $\Theta$  hinreichend groß ist, um die Implikation in Definition 4.6 zu erzwingen.

**Beispiel 4.5.** (*Vollständigkeit unter Poissonverteilung.*) Seien  $X_1, \dots, X_n$  i.i.d. mit  $X_i \sim \text{Poiss}(\theta)$ . Setze  $\Theta := \{x \in \mathbb{R} : x > 0\}$ . Nach Tabelle 2.1 und Bemerkung 2.10 ist

$T(\mathbf{X}) = \sum_{i=1}^n X_i$  suffiziente Statistik für  $\theta$ . Zum Beispiel unter Verwendung von Satz 2.11 zeigt man, dass  $T(\mathbf{X}) \sim \text{Poiss}(n\theta)$ . Sei  $g$  eine (messbare) Funktion mit  $\mathbb{E}_\theta(g(T)) = 0$  für alle  $\theta > 0$ . Dies ist gleichbedeutend mit

$$e^{-n\theta} \sum_{i=0}^{\infty} g(i) \frac{(n \cdot \theta)^i}{i!} = 0$$

für alle  $\theta > 0$ . Eine Potenzreihe, die identisch mit 0 ist in einer Umgebung von 0, muß alle Koeffizienten gleich 0 haben. Somit folgt  $g(i) = 0$  für alle  $i = 0, 1, 2, \dots$ , was bedeutet, dass  $T$  vollständig ist.

Für vollständige suffiziente Statistiken haben wir folgenden wichtigen Satz.

**Satz 4.7** (Lehmann-Scheffé). *Falls  $T(\mathbf{X})$  eine vollständige suffiziente Statistik und  $S(\mathbf{X})$  ein unverzerrter Schätzer von  $q(\boldsymbol{\theta})$ , dann ist*

$$T^*(\mathbf{X}) := \mathbb{E}(S(\mathbf{X})|T(\mathbf{X}))$$

*ein UMVUE Schätzer für  $q(\boldsymbol{\theta})$ . Falls weiterhin  $\text{Var}_\theta(T^*(\mathbf{X})) < \infty$  für alle  $\boldsymbol{\theta} \in \Theta$ , so ist  $T^*(\mathbf{X})$  der eindeutige UMVUE Schätzer von  $q(\boldsymbol{\theta})$ .*

*Beweis.* Da  $b(\boldsymbol{\theta}, T^*) = b(\boldsymbol{\theta}, S) = 0$  folgt, dass  $T^*$  ein unverzerrter Schätzer für  $q(\boldsymbol{\theta})$  ist. Nach dem Satz von Rao Blackwell, 4.5, gilt dann  $\text{Var}_{\boldsymbol{\theta}}(T^*) \leq \text{Var}_{\boldsymbol{\theta}}(S)$ . Falls  $\text{Var}_{\boldsymbol{\theta}}(S) < \infty$  gilt strikte Ungleichung, falls  $T^* \neq S$ . Man muß noch zeigen, dass  $T^*$  unabhängig von der Wahl von  $S$  ist. Seien  $T_1^* = \mathbb{E}(S_1|T) = g_1(T)$  und  $T_2^* = \mathbb{E}(S_2|T) = g_2(T)$  zwei unverzerrte Schätzer von  $q(\boldsymbol{\theta})$ , die durch Rao-Blackwellisierung erhalten wurden. Dann gilt

$$\mathbb{E}_{\boldsymbol{\theta}}(g_1(T) - g_2(T)) = \mathbb{E}_{\boldsymbol{\theta}}(T_1^*) - \mathbb{E}_{\boldsymbol{\theta}}(T_2^*) = q(\boldsymbol{\theta}) - q(\boldsymbol{\theta}) = 0.$$

Da  $T$  vollständig ist folgt aus  $\mathbb{E}_{\boldsymbol{\theta}}(g_1(T) - g_2(T)) = 0$  für alle  $\boldsymbol{\theta} \in \Theta$ , dass  $\mathbb{P}_{\boldsymbol{\theta}}(g_1(T) = g_2(T)) = 1$  für alle  $\boldsymbol{\theta} \in \Theta$  und somit hängt  $T^*$  nicht von  $S$  ab.

Wir zeigen noch die Eindeutigkeit. Sei  $T_1^*$  ein unverzerrter Schätzer für  $q(\boldsymbol{\theta})$  mit

$$\mathbb{E}_{\boldsymbol{\theta}}(T_1^* - q(\boldsymbol{\theta}))^2 \leq \mathbb{E}_{\boldsymbol{\theta}}(S(\mathbf{X}) - q(\boldsymbol{\theta}))^2$$

für alle  $S(\mathbf{X})$ . Dann gilt

$$\mathbb{E}_{\boldsymbol{\theta}}((T_1^* - q(\boldsymbol{\theta}))^2) \leq \mathbb{E}_{\boldsymbol{\theta}}((T^* - q(\boldsymbol{\theta}))^2).$$

Da  $\text{Var}_{\boldsymbol{\theta}}(T^*) < \infty$  muß nach dem Satz von Rao-Blackwell, 4.5, gelten dass  $T_1^* = T^*$ .  $\square$

**Bemerkung 4.8.** Man kann den Satz von Lehmann-Scheffé, Satz 4.7, auf zwei Arten für die Bestimmung von UMVUE Schätzern verwenden:

- (i) Falls man eine Statistik der Form  $h(T(\mathbf{X}))$  für eine suffiziente Statistik  $T$  findet mit

$$\mathbb{E}_{\boldsymbol{\theta}}(h(T(\mathbf{X}))) = q(\boldsymbol{\theta}),$$

so ist  $h(T(\mathbf{X}))$  UMVUE-Schätzer: In der Tat, da  $\mathbb{E}(h(T(\mathbf{X}))|T(\mathbf{X})) = h(T(\mathbf{X}))$  kann man Satz 4.7 mit  $S(\mathbf{X}) = h(T(\mathbf{X}))$  anwenden.

- (ii) Findet man einen unverzerrten Schätzer  $S(\mathbf{X})$  für  $q(\boldsymbol{\theta})$ , so ist

$$\mathbb{E}(S(\mathbf{X})|T(\mathbf{X}))$$

der UMVUE-Schätzer für  $q(\boldsymbol{\theta})$ .

Der Nachweis von Vollständigkeit ist oft schwierig, aber für exponentielle Familien hat man folgenden Satz:

**Satz 4.9.** Sei  $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$  eine  $K$ -dimensionale exponentielle Familie und  $c(\Theta)$  enthalte ein offenes Rechteck in  $\mathbb{R}^k$ . Dann ist  $\mathbf{T}(\mathbf{X}) := (T_1(\mathbf{X}), \dots, T_k(\mathbf{X}))^\top$  vollständig und suffizient für  $q(\boldsymbol{\theta})$ .

*Beweis.* Für den Beweis verweisen wir auf Lehmann (1997), Theorem 4.3.1 auf Seite ...  $\square$

**Beispiel 4.6.** (*UMVUE Schätzer für die Normalverteilung*) Seien  $\mathbf{X} := (X_1, \dots, X_n)^\top$  i.i.d. mit  $X_i \sim \mathcal{N}(\mu, \sigma^2)$  und  $\boldsymbol{\theta} := (\mu, \sigma^2)^\top$  unbekannt. In Beispiel 3.21 wurden die Maximum-Likelihood-Schätzer für dieses Modell bestimmt und gesehen, dass für  $C$  aus Satz 4.7  $C = \mathbb{R} \times \mathbb{R}^-$ . Damit enthält  $C$  ein offenes Rechteck. In Beispiel 2.17 hatten wir gezeigt, dass es sich um eine exponentielle Familie handelt mit suffizienter Statistik

$$\mathbf{T}(\mathbf{X}) := \left( \sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right)^\top.$$

Da das arithmetische Mittel  $\bar{X}$  eine Funktion von  $\mathbf{T}(\mathbf{X})$  und weiterhin unverzerrt für  $\mu = \theta_1$  ist, folgt mit Satz 4.7, dass  $\bar{X}$  eindeutiger UMVUE-Schätzer für  $\mu$  ist. Ebenso ist die Stichprobenvarianz

$$s^2(\mathbf{X}) := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

ein unverzerrter Schätzer für  $\sigma^2$  nach Aufgabe 1.3. Weiterhin ist sie suffizient, da sie eine Funktion von  $\mathbf{T}(\mathbf{X})$  ist. Damit ist die Stichprobenvarianz der eindeutige UMVUE-Schätzer für  $\sigma^2$ .

Allerdings ist  $s^2(\mathbf{X})$  nicht UMVUE-Schätzer für  $\sigma^2$ , falls der Mittelwert  $\mu$  bekannt ist, siehe Aufgabe 4.1.

**Beispiel 4.7.** (*UMVUE Schätzer für die Exponentialverteilung*) Wir betrachten  $X_1, \dots, X_n$  i.i.d. mit  $X_i \sim \text{Exp}(\theta)$  und  $\Theta := \mathbb{R}^+$ , vergleiche Beispiel 2.8. Man betrachte einen festen zeitlichen Horizont  $s$  und möchte

$$q(\theta) := P_\theta(X \leq s) = 1 - e^{-\theta s}$$

schätzen. Eine Exponentialverteilung mit Parameter  $\theta$  ist gerade *Gamma*(1,  $\theta$ )-verteilt, siehe Abschnitt 1.2. Aus Tabelle 2.1 entnimmt man, dass deswegen die Exponentialverteilung eine eindimensionale exponentielle Familie ist mit kanonischer Statistik  $T(\mathbf{X}) = \sum_{i=1}^n X_i$  und  $c(\theta) = -\theta$ . Damit ist  $C = \mathbb{R}^-$  und enthält ein offenes Rechteck. Nach Satz 4.9 ist  $T(\mathbf{X})$  suffizient und vollständig für  $\theta$ . Betrachte

$$S(X_1) := \mathbf{1}_{\{X_1 \leq s\}}.$$

Dann ist  $\mathbb{E}_\theta(S(X_1)) = P_\theta(X_1 \leq s) = q(\theta)$  und somit ist  $S(X_1)$  unverzerrt für  $q(\theta)$ . Nach dem Satz von Lehmann-Scheffé, 4.7, ist  $T^* = \mathbb{E}(S(X_1)|T)$  UMVUE-Schätzer für  $q(\theta)$ . Wir berechnen  $T^*$ : Es gilt, dass

$$\mathbb{E}(S(X_1)|T) = P(X_1 \leq s | T) = P\left(\frac{X_1}{T} \leq \frac{s}{T} \mid T\right).$$

Nun ist  $X_1/T$  unabhängig von  $T$  nach ... und damit ist

$$P\left(\frac{X_1}{T} \leq \frac{s}{T} \mid T = t\right) = P\left(\frac{X_1}{T} \leq \frac{s}{t} \mid T = t\right).$$

Nach Bemerkung 1.15 ist  $\frac{X_1}{T} \sim \text{Beta}(1, n-1)$ , da  $X_1 \sim \text{Gamma}(1, \theta)$  und  $X_2 + \dots + X_n$  unabhängig von  $X_1$  sind mit  $X_2 + \dots + X_n \sim \text{Gamma}(n-1, \lambda)$ . Somit folgt

$$\begin{aligned} \mathbb{E}(S(X_1) \mid T = t) &= P\left(\frac{X_1}{T} \leq \frac{s}{t} \mid T = t\right) = \int_0^{s/t} (n-1)(1-u)^{n-2} du \\ &= -(1-u)^{n-1} \Big|_0^{s/t} = 1 - \left(1 - \frac{s}{t}\right)^{n-1} \end{aligned}$$

falls  $s \leq t$ . Ist  $s > t$ , so ist  $S(X_1) = 1$ . Damit erhalten wir den UMVUE-Schätzer für  $q(\theta)$  durch

$$T^* = \mathbb{E}(S|T) = \begin{cases} 1 - \left(1 - \frac{s}{T}\right)^{n-1} & \text{falls } T \geq s \\ 1 & \text{falls } T < s \end{cases}.$$

Zum Vergleich: der Maximum-Likelihood-Schätzer und der Momentenschätzer für  $\theta$  ist  $\hat{\theta} = (\bar{X})^{-1}$ . Damit ist der MLS von  $q(\theta)$  gegeben durch

$$q(\hat{\theta}) = 1 - \exp(-\hat{\theta}t).$$

Da  $T^* \neq \hat{\theta}$  ist  $\hat{\theta}$  allerdings nicht UMVUE. Allerdings ist  $q(\hat{\theta})$  eine Funktion von  $T$  und damit suffizient. Demnach muss  $q(\hat{\theta})$  ein verzerrter Schätzer von  $q(\theta)$  sein.

**Beispiel 4.8.** (*UMVUE Schätzer für die Gleichverteilung*) In diesem Beispiel betrachten wir den Fall einer Gleichverteilung, welche keine exponentielle Familie darstellt. Seien  $\mathbf{X} = (X_1, \dots, X_n)^\top$  i.i.d. mit  $X_i \sim U(0, \theta)$  und  $\Theta = \mathbb{R}^+$ . Definiere  $X_{(1)} := \min\{X_1, \dots, X_n\}$  und  $X_{(n)} := \max\{X_1, \dots, X_n\}$  und entsprechend  $x_{(1)}$  und  $x_{(n)}$ . Dann ist die Dichte von  $\mathbf{X}$  gegeben durch

$$p(\mathbf{x}, \theta) = \begin{cases} \theta^{-n} & \text{falls } 0 \leq x_{(1)} \leq x_{(n)} \leq \theta \\ 0 & \text{sonst.} \end{cases}$$

Unter Anwendung des Faktorisierungssatzes, Satz 2.7, sieht man, dass  $X_{(n)}$  suffizient für  $\theta$  ist. Wir zeigen nun, dass  $X_{(n)}$  auch vollständig ist. Zunächst folgt aus  $X_1 \sim U(0, \theta)$ , dass  $\mathbb{P}_\theta(X_1 \leq t) = t\theta^{-1}1_{\{0 \leq t \leq \theta\}}$  für  $t \leq \theta$  und 1 sonst. Damit ist

$$P(X_{(n)} \leq t) = P(X_1 \leq t, \dots, X_n \leq t) = (P(X_1 \leq t))^n$$

und wir erhalten die Dichte von  $X_{(n)}$ :

$$\frac{d}{dt} P_\theta(X_{(n)} \leq t) = n\theta^{-n}t^{n-1} \text{ für } 0 < t < \theta.$$

Für die Anwendung von Satz 4.7 betrachten wir

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

Damit folgt aus  $\mathbb{E}_\theta(g(X_{(n)})) = 0$ , dass  $g(t) = 0$  Lebesgue-fast sicher für alle  $t \geq 0$ . Damit ist  $X_{(n)}$  vollständig und suffizient. Allerdings ist  $X_{(n)}$  verzerrt, da

$$\mathbb{E}_\theta(X_{(n)}) = \frac{n}{\theta^n} \int_0^\theta t^n dt = \frac{n\theta}{n+1} \neq \theta.$$

Die Statistik

$$M_n := \frac{n+1}{n} X_{(n)}$$

ist demnach unverzerrt für  $\theta$ . Sie ist weiterhin Funktion der vollständigen und suffizienten Statistik  $X_{(n)}$ . Da  $\text{Var}(M_n) < \infty$ , ist nach Satz 4.7  $M_n$  eindeutiger UMVUE-Schätzer für  $\theta$ .

**Bemerkung 4.10** (Andere Ansätze). Es gibt eine Reihe von Alternativen zu UMVUE um Optimalitätseigenschaften von Schätzern zu messen.

- (i) Der *Bayesianische Ansatz* betrachtet man  $\boldsymbol{\theta} \sim \pi$  und vergleicht das Verhalten von

$$\mathbb{E}_\theta(R(\boldsymbol{\theta}, T)) = \int R(\boldsymbol{\theta}, T)\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

für verschieden Schätzer  $T$ . ?).

- (ii) Oder man vergleicht  $\max_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, T)$  für verschieden Schätzer und suche  $T$  so, dass  $\max_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, T)$  minimal ist. Ein solcher Schätzer heißt *Minimax Schätzer*.

## 4.3 Die Informationsungleichung

Die Informationsungleichung gibt eine *untere Schranke* für die Varianz einer Statistik. Im folgenden untersuchen wir ein eindimensionales reguläres statistisches Modell  $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$  und nehmen die so genannten *Cramer-Rao Regularitätsbedingungen* an:

**(CR)** (i) Sei  $\Theta \subset \mathbb{R}$  offen.

- (ii)  $A = \{\mathbf{x} \in \mathbb{R}^k : p(\mathbf{x}, \theta) > 0\}$  hängt nicht von  $\theta$  ab. Die Ableitung  $\frac{\partial}{\partial \theta} \ln p(\mathbf{x}, \theta)$  existiert und ist endlich  $\forall \mathbf{x} \in A, \forall \theta \in \Theta$ .

(iii) Falls  $T$  eine beliebige Statistik mit  $\mathbb{E}_\theta(|T|) < \infty$  für alle  $\theta \in \Theta$  ist, so gilt

$$\frac{\partial}{\partial \theta} \int_{\mathbb{R}^k} T(\mathbf{x}) p(\mathbf{x}, \theta) d\mathbf{x} = \int_{\mathbb{R}^k} \frac{\partial}{\partial \theta} p(\mathbf{x}, \theta) T(\mathbf{x}) d\mathbf{x}.$$

**Bemerkung 4.11.** Falls

$$p(\mathbf{x}, \theta) = \exp\{c(\theta)T(\mathbf{x}) + d(\theta) + S(\mathbf{x})\} \cdot \mathbf{1}_A(\mathbf{x})$$

eine einparametrische exponentielle Familie ist mit  $\frac{\partial}{\partial \theta} c(\theta) \neq 0$  für alle  $\theta \in \Theta$  und stetigem  $c$ , dann ist (CR) erfüllt.

*Beweis.* Satz von der monotonen Konvergenz. □

Im Folgenden möchten wir die Information, die in Daten enthalten ist, möglichst effizient ausnutzen. Dazu benötigen wir ein Konzept für Information.

**Definition 4.12.** Die *Fisher-Information* für einen Parameter  $\theta$  ist gegeben durch

$$I(\theta) := \mathbb{E}_\theta \left( \left( \frac{\partial}{\partial \theta} \ln p(\mathbf{X}, \theta) \right)^2 \right).$$

Für die Fisher-Information gilt

$$I(\theta) = \int_{\mathbb{R}^k} \left( \frac{\partial}{\partial \theta} \ln p(\mathbf{x}, \theta) \right)^2 \cdot p(\mathbf{x}, \theta) d\mathbf{x} = \int_{\mathbb{R}^k} \frac{1}{p(\mathbf{x}, \theta)} \cdot \left( \frac{\partial}{\partial \theta} p(\mathbf{x}, \theta) \right)^2 d\mathbf{x}$$

Man bezeichnet  $\frac{\partial}{\partial \theta} \ln p(\mathbf{X}, \theta)$  auch als *Einfluss-* oder *Score-*funktion. Ihr Erwartungswert verschwindet unter den obigen Regularitätsannahmen (CR), denn es gilt

$$\begin{aligned} \mathbb{E}_\theta \left( \frac{\partial}{\partial \theta} \ln p(\mathbf{X}, \theta) \right) &= \int_{\mathbb{R}^k} \frac{\partial}{\partial \theta} \ln p(\mathbf{x}, \theta) \cdot p(\mathbf{x}, \theta) d\mathbf{x} \\ &= \int_{\mathbb{R}^k} \frac{\partial}{\partial \theta} p(\mathbf{x}, \theta) d\mathbf{x} \\ &= \frac{\partial}{\partial \theta} \left( \int_{\mathbb{R}^k} p(\mathbf{x}, \theta) d\mathbf{x} \right) = 0. \end{aligned} \tag{4.3}$$

Die Fisher Information ist demnach gleich der Varianz der Einflussfunktion,

$$I(\theta) = \text{Var} \left( \frac{\partial}{\partial \theta} \ln p(\mathbf{X}, \theta) \right).$$

Sind  $X_1, \dots, X_n$  i.i.d. so erhalten wir mit  $\mathbf{X} = (X_1, \dots, X_n)^\top$  dass die Fisher Information der Stichprobe gerade  $n$  mal die Fisher Information einer einzelnen Zufallsvariable ist:

$$I(\theta) = \text{Var}_\theta \left[ \left( \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln p(X_i, \theta) \right)^2 \right] = n \mathbb{E}_\theta \left[ \left( \frac{\partial}{\partial \theta} \ln p(X_1, \theta) \right)^2 \right]$$

**Beispiel 4.9.** (*Fisher-Information unter Normalverteilung*) Ist  $X$  normalverteilt mit bekanntem  $\sigma$  so erhält man für die Fisher-Information, dass

$$I(\theta) = \frac{1}{\sigma^4} \mathbb{E}_\theta((X - \theta)^2) = \frac{1}{\sigma^2}. \quad (4.4)$$

Je kleiner die Varianz, umso höher der Informationsgehalt, der einer einzelnen Beobachtung zuzuschreiben ist. Somit ist die Fisher Information für die i.i.d. Stichprobe des Umfanges  $n$  gerade  $n\sigma^{-2}$ .

**Beispiel 4.10.** (*Fisher-Information für die Poisson-Verteilung*) Seien  $X_1, \dots, X_n$  i.i.d. mit  $X_i \sim \text{Poiss}(\theta)$ . Das heißt, die Wahrscheinlichkeitsfunktion von  $X_i$  ist  $p(x, \theta) = e^{-\theta} \frac{\theta^x}{x!}$  für  $x = 1, 2, \dots$ . Da

$$\frac{\partial}{\partial \theta} \ln p(x, \theta) = -1 + \frac{x}{\theta},$$

folgt für die Fisher Information einer Stichprobe von Poisson-verteilten Zufallsvariablen

$$I(\theta) = n \text{Var} \left( \frac{\partial}{\partial \theta} \ln p(X_1, \theta) \right) = n\theta^{-2} \cdot \text{Var}(X_1) = \frac{n\theta}{\theta^2} = \frac{n}{\theta}.$$

**Satz 4.13.** Sei  $T(\mathbf{X})$  eine Statistik mit  $\text{Var}_\theta(T(\mathbf{X})) < \infty$  für alle  $\theta \in \Theta$  und  $\Psi(\theta) := \mathbb{E}_\theta(T(\mathbf{X}))$ . Weiterhin sei (CR) erfüllt und  $0 < I(\theta) < \infty$  für alle  $\theta \in \Theta$ . Dann gilt für alle  $\theta \in \Theta$ , dass  $\Psi(\theta)$  differenzierbar ist und

$$\text{Var}_\theta(T(\mathbf{X})) \geq \frac{(\Psi'(\theta))^2}{I(\theta)}. \quad (4.5)$$

Gleichung (4.5) nennt man die *Informationsungleichung*.

*Beweis.* Zunächst ist unter (CR)

$$\begin{aligned} \Psi'(\theta) &= \frac{\partial}{\partial \theta} \mathbb{E}_\theta(T(\mathbf{X})) = \int_{\mathbb{R}^k} \frac{\partial}{\partial \theta} (T(\mathbf{x})p(\mathbf{x}, \theta)) d\mathbf{x} \\ &= \mathbb{E}_\theta \left( T(\mathbf{X}) \frac{\partial}{\partial \theta} \ln p(\mathbf{X}, \theta) \right), \end{aligned}$$

analog zu Gleichung (4.3). Damit erhalten wir

$$\begin{aligned} (\Psi'(\theta))^2 &= \mathbb{E}_\theta \left( T(\mathbf{X}) \frac{\partial}{\partial \theta} \ln p(\mathbf{X}, \theta) \right)^2 \\ &\stackrel{(4.3)}{=} \text{Cov}_\theta \left( T(\mathbf{X}), \frac{\partial}{\partial \theta} \ln p(\mathbf{x}, \theta) \right) \\ &\leq \text{Var}_\theta(T(\mathbf{X})) \cdot \text{Var}_\theta \left( \frac{\partial}{\partial \theta} \ln p(\mathbf{X}, \theta) \right) \end{aligned}$$

mit der Cauchy-Schwarz Ungleichung, Gleichung (1.2). Da der letzte Term gerade die Fisher-Information ist, folgt die Behauptung.  $\square$

**Korollar 4.14.** *Unter den Bedingungen des Satzes 4.13 gilt für unverzerrte Schätzer  $T$  von  $\theta$ , die Cramér-Rao Schranke*

$$\text{Var}_\theta(T(\mathbf{X})) \geq \frac{1}{I(\theta)}.$$

**Korollar 4.15.** *Sei  $\mathbf{X} = (X_1, \dots, X_n)$  mit  $X_1, \dots, X_n$  i.i.d. und  $X_i$  habe Dichte oder Wahrscheinlichkeitsfunktion  $p(x, \theta)$  für  $\theta \in \Theta$  und die Bedingungen des Satzes 4.13 seien erfüllt. Dann gilt*

$$\text{Var}_\theta(T(\mathbf{X})) \geq \frac{(\Psi'(\theta))^2}{n \cdot I_1(\theta)}.$$

Hierbei ist  $I_1(\theta) := \mathbb{E}[(\partial/\partial\theta \ln p(X_1, \theta))^2]$  die Information pro Beobachtung.

## Anwendung der Informationsungleichung

Falls (CR) erfüllt ist und ein unverzerrter Schätzer  $T^*$  für  $\Psi(\theta)$  existiert mit  $\text{Var}_\theta(T^*(X)) = \frac{(\Psi'(\theta))^2}{I(\theta)}$ , dann ist  $T^*$  UMVUE für  $\Psi(\theta)$ .

Überraschenderweise ist die Bedingung, dass die untere Schranke der Informationsungleichung angenommen wird nur in exponentiellen Familien erfüllt, wie folgendes Theorem zeigt. In anderen Verteilungsklassen gibt es also mitunter größere untere Schranken, die Schranke ist dann nicht scharf.

**Satz 4.16.** *Es gelte (CR) und es existiere ein unverzerrter Schätzer  $T^*$  von  $\Psi(\theta)$ , der die untere Schranke annimmt  $\forall \theta$ . Dann ist  $\{P_\theta, \theta \in \Theta\}$  eine eindimensionale exponentielle Familie mit*

$$p(\mathbf{x}, \theta) = \exp \{c(\theta)T^*(\mathbf{x}) + d(\theta) + S(\mathbf{x})\} \cdot \mathbf{1}_A(\mathbf{x}).$$

Umgekehrt, falls  $\{P_\theta, \theta \in \Theta\}$  eine eindimensionale exponentielle Familie ist und  $c(\theta)$  stetige Ableitungen  $c'(\theta) \neq 0 \quad \forall \theta \in \Theta$  besitzt, dann nimmt  $T(\mathbf{X})$  die Informationsschranke an und ist daher UMVUE von  $\mathbb{E}_\theta(T(\mathbf{X}))$ .

*Beweis.* Bickel und Doksum Bickel and Doksum (2001), Seite ... □

#### Bemerkung 4.17.

- UMVUE-Schätzer können auch existieren, wenn (CR) nicht erfüllt wird. Ein Beispiel dafür ist  $X_1, \dots, X_n$  i.i.d. mit  $X_i \sim U(0, \theta)$ , siehe Beispiel 4.8.
- Die Informationsschranke braucht nicht angenommen zu werden, auch wenn UMVUE-Schätzer existieren und (CR) erfüllt ist.

## 4.4 Asymptotische Theorie

Die asymptotische Theorie beschäftigt sich mit dem Verhalten von Schätzern wenn der Stichprobenumfang  $n$  immer größer wird, also  $n \rightarrow \infty$ . Hierzu betrachten wir im folgenden Abschnitt  $X_1, X_2, \dots$  i.i.d. mit glatten Dichten  $p(x, \theta)$  und es gelte  $q(\theta)$  zu schätzen. Hierbei halten wir  $\theta$  typischerweise fest.

### 4.4.1 Konsistenz

**Definition 4.18.** Eine Schätzfolge  $T_n(X_1, \dots, X_n)$ ,  $n = 1, 2, \dots$  für  $q(\theta)$  heißt *konsistent*, falls

$$\mathbb{P}_\theta(|T_n(X_1, \dots, X_n) - q(\theta)| \geq \epsilon) \xrightarrow{n \rightarrow \infty} 0$$

für alle  $\epsilon > 0$ .

Für einen konsistenten Schätzer gilt also

$$T_n \xrightarrow{\mathbb{P}_\theta} q(\theta).$$

Im Gegensatz zur in der Definition eingeführten (schwachen) Konsistenz verlangt die so genannte starke Konsistenz sogar fast sichere Konvergenz.

UMVUE Schätzer sind immer konsistent, MLE's sind in der Regel auch konsistent. Für Beweise verweisen wir hier auf die Literatur. Im Folgenden werden einige Beispiele vorgestellt, in welchen die Konsistenz jeweils mit dem starken Gesetz der großen Zahlen nachgewiesen wird.

**Beispiel 4.11.** (*Multinomialverteilung*) Falls  $\mathbf{N} = (N_1, \dots, N_k)$  Multinomial-verteilt ist,  $\mathbf{N} \sim M(n, p_1, \dots, p_k)$ , dann gilt nach dem schwachen Gesetz der großen Zahlen 1.19, dass

$$\frac{N_i}{n} \xrightarrow{\mathbb{P}} p_i.$$

Insofern ist  $N_i/n$  konsistent für  $p_i$  für  $i = 1, \dots, k$ . Daher ist der Schätzer  $T_n$  gegeben durch

$$T_n := h\left(\frac{N_1}{n}, \dots, \frac{N_k}{n}\right)$$

konsistent für  $q(\theta) = h(p_1, \dots, p_k)$  falls  $h$  stetig ist. Denn nach dem Stetigkeitssatz 1.21 folgt

$$T_n \xrightarrow{\mathbb{P}} h(p_1, \dots, p_k).$$

**Beispiel 4.12.** (*Momentenschätzer.*) Betrachten wir den Momentenschätzer für das  $j$ -te Moment  $\mu^j = \mathbb{E}(X_1^j)$ , wobei wir annehmen, dass  $\mu^j < \infty$ . Mit dem starken Gesetz der großen Zahl 1.20 ist  $\hat{m}_j := \frac{1}{n} \sum_{i=1}^n x_i^j$  konsistent für  $\mu^j$ . Wieder ist  $T_n := g(\hat{m}_1, \dots, \hat{m}_r)$  konsistent für  $q(\theta) := g(m_1(\theta), \dots, m_r(\theta))$  mit  $m_j(\theta) := \mathbb{E}_\theta(X_1^j)$  falls  $g$  stetig ist. Somit ist der Momentenschätzer konsistent für stetige Funktionen der theoretischen Momente.

#### 4.4.2 Asymptotische Normalität und verwandte Eigenschaften

**Definition 4.19.** Eine Schätzfolge  $T_n(X_1, \dots, X_n)$ ,  $n = 1, 2, \dots$  heißt *asymptotisch normalverteilt* falls Folgen  $(\mu_n(\theta))_{n \geq 1}$  und Varianz  $(\sigma_n^2(\theta))_{n \geq 1}$  existieren, so dass

$$\frac{T_n(X_1, \dots, X_n) - \mu_n(\theta)}{\sigma_n(\theta)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Diese Definition bedeutet, dass der (asymptotisch) zentrierte, standardisierte Schätzer in Verteilung gegen eine Standardnormalverteilung konvergiert. Hierbei muß  $\mu_n$  ( $\sigma_n^2$ ) nicht Erwartungswert (Varianz) von  $T_n$  sein. Per Definitionem bedeutet das

$$\lim_{n \rightarrow \infty} P\left(\frac{T_n(X_1, \dots, X_n) - \mu_n(\theta)}{\sigma_n(\theta)} \leq z\right) = \Phi(z), \quad \forall z \in \mathbb{R}$$

wobei  $\Phi(\cdot)$  die Verteilungsfunktion der Standardnormalverteilung ist. Oft gilt  $\mu_n(\theta) = \mathbb{E}_\theta(T_n(\mathbf{X}))$  und  $\sigma_n^2(\theta) = \text{Var}_\theta(T_n(\mathbf{X}))$ .

Asymptotische Normalität wird auch wie folgt ausgedrückt:

$$P(T_n(\mathbf{X}) \leq z) \approx \Phi\left(\frac{z - \mu_n(\theta)}{\sigma_n(\theta)}\right) \quad \text{für } n \text{ groß.} \quad (4.6)$$

Insbesondere wird die Verteilungsfunktion von  $T_n(\mathbf{x})$  an der Stelle  $z$  durch  $\Phi(z - \mu_n(\theta)/\sigma_n(\theta))$  approximiert.

## 4.5 Aufgaben

**Aufgabe 4.1.** Seien  $X_1, \dots, X_n$  i.i.d. mit  $X_i \sim \mathcal{N}(\mu_0, \sigma^2)$  und  $\mu_0 \in \mathbb{R}$  sei bekannt. Zeigen Sie, dass

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

UMVUE Schätzer für  $\sigma^2$  ist.

Vorläufig

# 5 Konfidenzintervalle und Hypothesentest

Vorläufig

## **6 Optimale Tests und Konfidenzintervalle, Likelihood Ratio Tests und verwandte Methoden**

Vorläufig

# **7 Lineare Modelle - Regression und Varianzanalyse (ANOVA)**

Vorläufig

# 8 Verzeichnisse

## Tabellenverzeichnis

2.1	Einparametrische exponentielle Familien. $c$ , $T$ und $A$ aus Darstellung (2.5) sind in der Tabelle angegeben, $d$ ergibt sich durch Normierung. . . . .	36
-----	---	----

## Abbildungsverzeichnis

1.1	Verteilung der Hypergeometrischen Verteilung . . . . .	11
1.2	Dichte der Normalverteilung . . . . .	12
1.3	Dichte der Gamma-Verteilung . . . . .	15
1.4	Dichte der Beta-Verteilung . . . . .	16
2.1	Poisson-Prozess . . . . .	29
3.1	Einfache lineare Regression . . . . .	54
3.2	Einfache lineare Regression . . . . .	57
3.3	Konkave Funktionen und Maxima . . . . .	60
3.4	Likelihood-Funktion für Normalverteilung . . . . .	61
3.5	Likelihoodfunktion einer diskreten Gleichverteilung . . . . .	62
4.1	Nichtidentifizierbarkeit eines besten Schätzers . . . . .	73
4.2	Vergleich von Mittelwertschätzern anhand des MQF . . . . .	74

# Liste der Beispiele

1.1	Mittelwert und Stichprobenvarianz . . . . .	5
1.2	Hypergeometrische Verteilung. . . . .	10
1.3	Bernoulli-Verteilung . . . . .	17
1.4	Fortsetzung. . . . .	17
1.5	Suffiziente Statistik in der Bernoulli-Verteilung . . . . .	17
1.6	Minima und Maxima von gleichverteilten Zufallsvariablen. . . . .	19
2.1	Qualitätssicherung . . . . .	22
2.2	Messmodell . . . . .	23
2.3	Ein nicht identifizierbares Modell . . . . .	25
2.4	Messmodell . . . . .	26
2.5	Qualitätskontrolle, siehe Beispiel 2.1. . . . .	27
2.6	Qualitätskontrolle . . . . .	28
2.7	Warteschlange. . . . .	28
2.8	Warteschlange, Fortsetzung von Beispiel 2.7. . . . .	31
2.9	Titel fehlt . . . . .	32
2.10	Suffiziente Statistiken für die Normalverteilung. . . . .	32
2.11	Normalverteilung mit bekanntem $\sigma$ . . . . .	34
2.12	Normalverteilung mit bekanntem $\mu$ . . . . .	34
2.13	Binomialverteilung. . . . .	34
2.14	i.i.d. Normalverteilung mit bekanntem $\sigma$ . . . . .	35
2.15	Verteilung des arithmetischen Mittels. . . . .	36
2.16	Momente der Rayleigh-Verteilung. . . . .	38
2.17	Die Normalverteilung ist eine zweiparametrische exponentielle Familie. . . . .	39
2.18	i.i.d. Normalverteilung als exponentielle Familie. . . . .	39
2.19	Lineare Regression . . . . .	39
2.20	Qualitätssicherung unter Vorinformation. . . . .	40
2.21	Operational Risk. . . . .	41
2.22	Konjugierte Familie der Bernoulli-Verteilung. . . . .	42
2.23	Konjugierte Familie der Normalverteilung bei bekannter Varianz. . . . .	44
3.1	Qualitätssicherung aus Beispiel 2.1. . . . .	47

3.2	Messmodell aus Beispiel 2.2. . . . .	47
3.3	Messmodell aus Beispiel 3.2. . . . .	47
3.4	Relative Häufigkeiten. . . . .	48
3.5	Genotypen. . . . .	50
3.6	Normalverteilung. . . . .	51
3.7	Bernoulli-Verteilung. . . . .	51
3.8	Poisson-Verteilung. . . . .	52
3.9	Gleichverteilung. . . . .	52
3.10	Messmodell aus Beispiel 2.2. . . . .	53
3.11	Einfache lineare Regression. . . . .	54
3.12	Messmodell. . . . .	55
3.13	Einfache lineare Regression. . . . .	55
3.14	Log-Likelihood-Funktion unter Unabhängigkeit. . . . .	59
3.15	Normalverteilungsfall, $\sigma$ bekannt. . . . .	60
3.16	Gleichverteilung. . . . .	61
3.17	Genotypen. . . . .	61
3.18	Warteschlange. . . . .	63
3.19	Normalverteilungsfall, $\sigma$ bekannt. . . . .	64
3.20	Genotypen. . . . .	65
3.21	MLS für Normalverteilung, $\mu$ und $\sigma$ unbekannt. . . . .	66
3.22	Diskret beobachtete Überlebenszeiten. . . . .	67
4.1	MQF für die Normalverteilung. . . . .	72
4.2	Vergleich von Mittelwertschätzern anhand des MQF. . . . .	73
4.3	Der perfekte Schätzer . . . . .	74
4.4	Unverzerrte Schätzer . . . . .	75
4.5	Vollständigkeit unter Poissonverteilung. . . . .	77
4.6	UMVUE Schätzer für die Normalverteilung . . . . .	79
4.7	UMVUE Schätzer für die Exponentialverteilung . . . . .	79
4.8	UMVUE Schätzer für die Gleichverteilung . . . . .	80
4.9	Fisher-Information unter Normalverteilung . . . . .	83
4.10	Fisher-Information für die Poisson-Verteilung . . . . .	83
4.11	Multinomialverteilung . . . . .	86
4.12	Momentenschätzer. . . . .	86

# Literatur

- Bickel, P. J. and K. A. Doksum (2001). *Mathematical Statistics: Basic Ideas and Selected Topics Vol. I* (2nd ed.). Prentice Hall.
- Gauß, C. F. (1809). *Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium*. Volume 2.
- Georgii, H.-O. (2004). *Stochastik* (2nd ed.).
- Lehmann, E. L. (1997). *Testing Statistical Hypotheses*. Springer, New York.
- Resnick, S. (2003). *A Probability Path* (3rd ed.). Kluwer Academic Publ.
- Rolski, T., H. Schmidli, V. Schmidt, and J. Teugels (1999). *Stochastic Processes for Insurance and Finance*. John Wiley & Sons. New York.
- Schmidt, T. (2007). Coping with copulas. In J. Rank (Ed.), *Copulas: from theory to applications in finance*, pp. 1 – 31. Risk Books.

# Index

- $F_n$ , 49
- $Q(\boldsymbol{\theta})$ , 55
- $\chi^2$ -Verteilung, 12
- $\sigma$ -Algebra, 1
  
- a posteriori Verteilung, 42
- a priori Verteilung, 41
- a priori-Verteilung
  - nicht wohldefiniert, 45
  - nicht-informativ, 45
- arithmetischer Mittelwert, 5
- asymptotisch normalverteilt, 86
- asymptotische Normalität, 86
  
- Bayesianische Schätzer, 81
- Bayesianisches Modell, 41
- bedingte Dichte von Zufallsvektoren, 18
- bedingte Verteilung, 17
- bedingte Wahrscheinlichkeit, 2
- bedingter Erwartungswert, 17
- Bernoulli-Verteilung, 10, 17
- Beta-Funktion, 13
- Beta-Verteilung, 15
- bias (Verzerrung), 71
- Bienaymé, 9
- Binomialverteilung, 9
  
- Cauchy-Schwarz Ungleichung, 8
- Charakteristische Funktion, 9
- CR (Cramer-Rao Regularitätsbed.), 81
- Cramér-Rao, 84
- Cramér-Rao-Schranke, 84
  
- Cramer-Rao
  - Regularitätsbedingungen (CR), 81
  
- Dichte, 4
- diskrete Zufallsvariable, 3
- diskreter Wahrscheinlichkeitsraum, 2
  
- Einfluss-Funktion, 82
- Elementarereignis, 2
- empirische Verteilungsfunktion, 49
- endogene Variable, 53
- Erwartungswert, 7
  - bedingter, 17
- exogene Variable, 53
- Exponentialverteilung, 11, 31
  - Gedächtnislosigkeit, 20
- exponentielle Familie, 33, 38, 82
  - natürliche, 33
- exponentielle Familien
  - MLS, 63, 66
  - Vollständigkeit, 78
  
- F-Verteilung, 13
- Faktorisierungssatz, 30
- Faltungformel, 21
- Familie
  - exponentielle, 33, 38, 82
  - konjugierte, 43
- Fisher-Information, 82
- Fisher-Scoring-Methode, 69
- Funktion
  - Einfluss-, 82

- Likelihood-, 58
- Score-, 82
- Gamma-Verteilung, 14
- Gedächtnislosigkeit, 20
- Gesetz der großen Zahl, 20
- gewichtete Kleinste-Quadrate-Schätzer, 57
- Gleichungen
  - Normalen, 55
- Gleichverteilung, 11, 52
  - diskrete, 52
- Häufigkeit
  - relativ, 48
- Hardy-Weinberg Gleichgewicht, 50
- heteroskedastisch, 57
- homoskedastisch, 53
- Hypergeometrische Verteilung, 10, 22
- Hypothesentest
  - statistischer, 26
- i.i.d., 7, 35
- Identifizierbarkeit, 25
- improper non informative prior, 45
- Information
  - Fisher-, 82
- Informationsungleichung, 83
- Jensensche Ungleichung, 7
- Kleinste-Quadrate-Methode, 54
- Kleinste-Quadrate-Schätzer, 54
  - gewichtete, 57
- konsistent, 85, 86
- Korrelation, 8, 21
- Kovariable, 53, 54
- Kovarianz, 8
- KQS (Kleinste-Quadrate-Schätzer), 54
- Laplacesche Modelle, 10
- Least Squares Estimator, 54
- Likelihoodfunktion, 45, 58
- lineare Regression, 39
- Log-Likelihood-Funktion, 59
- Log-Likelihood-Gleichung, 59
- LSE, 54
- marginale Verteilung, 42
- Maximum Likelihood Methode, 58
- Maximum-Likelihood-Schätzer, 58
  - f.  $K$ -dim. exponentielle Familien, 66
  - f. exponentielle Familien, 63
  - Numerische Bestimmung, 67
- Messmodell, 23, 26, 34, 47, 53
- Methode
  - der kleinsten Quadrate, 54
  - Maximum-Likelihood, 58
- Minimax Schätzer, 81
- Mittelwert, 5, 79
- mittlerer betraglicher Fehler, 71
- mittlerer quadratischer Fehler, 71
- MLE (Maximum-Likelihood-Estimate), 58
- MLS (Maximum-Likelihood-Schätzer), 58
- Modell
  - Bayesianisches, 41
  - nichtparametrisches, 25
  - parametrisches, 25
  - statistisches, 1, 24
- Moment, 8
  - Stichproben-, 51
- Momente, 51
- Momentenerzeugende Funktion, 9
- momentenerzeugende Funktion, 37
- Momentenmethode, 50, 51
- Momentenschätzer, 86
- MQF (mittlerer quadratischer Fehler), 71
- Multinomialverteilung, 10
- Newton-Methode, 68
- Nichtidentifizierbarkeit, 24
- Normalgleichungen, 55

- Normalität
  - asymptotische, 86
- normalverteilt
  - asymptotisch, 86
- Normalverteilung, 11
  - Fisher-Information, 83
- Normierungskonstante, 33
- Nuisance Parameter, 24
- Numerische Bestimmung des MLS, 67
- Operational Risk, 41
- Poisson Prozess, 28
- Poisson-Prozess, 31
- Poisson-Verteilung, 52
  - Fisher-Information, 83
- Poissonverteilung, 10
- Qualitätssicherung, 22, 24, 27, 28, 40
  - Bayesianisch, 40
- Rao-Blackwell
  - Satz von, 76
- Rayleigh-Verteilung, 14, 38
  - Momente, 21
- Regression, 53
  - allgemeine, 53
  - lineare, 39
- Regressionsgerade, 56
- relative Häufigkeit, 48
- Response, 53
- Satz
  - Rao-Blackwell, 76
  - von Bayes, 2
  - Faktorisierungs-, 30
  - Gesetz der großen Zahl, 20
  - Stetigkeits-, 20
  - Substitutions-, 18
- Schätzer
  - Bayesianische, 81
  - konsistenter, 85
  - Maximum-Likelihood, 58
  - Minimax, 81
  - UMVUE, 75, 85
  - unverzerrt, 71, 74, 79, 84
  - unzulässig, 73
- Schranke
  - Cramér-Rao, 84
- schwaches Gesetz der großen Zahl, 20
- Score-Funktion, 82
- Störparameter, 24
- statistisches Modell, 24
- Statistik
  - Definition, 27
  - natürliche suffiziente, 33
- stetige Zufallsvariable, 4
- Stetigkeitssatz, 20
- Stichprobe, 22
- Stichprobenmoment, 51
- Stichprobenvarianz, 5, 20, 72, 75, 79
- Substitutionssatz, 18
- suffizient, 28
- symmetrisch verteilt, 22
- t-Verteilung, 13
- Transformationsatz, 5
- UMVUE-Schätzer, 75
- unabhängige Variablen, 53
- Unabhängigkeit, 3
  - von Zufallsvariablen, 6
- Ungleichung
  - Cauchy-Schwarz, 8
  - Informations-, 83
  - Jensen, 7
- unverzerrt, 75
- unverzerrter Schätzer, 71
- unzulässiger Schätzer, 73
- Variable

- endogene, 53
- exogene, 53
- Ko-, 53
- unabhängige, 53
- Varianz, 8
- Verteilung, 4
  - $\chi^2$ , 12
  - a posteriori, 42
  - a priori, 41
  - bedingte, 17
  - Bernoulli, 10, 17
  - Beta-, 15
  - F-, 13
  - Gamma-, 14
  - Hypergeometrische, 22
  - hypergeometrische, 10
  - marginale, 42
  - Rayleigh, 14, 21, 38
  - t-, 13
- Verteilungsfunktion, 4
  - empirische, 49
- verzerrt, 75
- Verzerrung, 71
- vollständig, 77
- Vollständigkeit
  - exponentielle Familien, 78
- Wahrscheinlichkeit
  - bedingte, 2
- Wahrscheinlichkeitsfunktion, 3
- Wahrscheinlichkeitsmaß, 2
- Wahrscheinlichkeitsraum
  - diskret, 2
- Warteschlange, 28
- Zielvariable, 53, 54
- Zufallsvariable, 3
  - diskret, 3
  - stetig, 4

Vorläufig