Die Maximum-Likelihood-Methode

7

In Kap. 5 haben wir Schätzer bezüglich gewisser Gütekriterien untersucht. Beispielsweise haben wir gesehen, dass im Kontext des Bernoullimodells die relative Häufigkeit $\hat{p}$ ein sinnvoller Schätzer für die Erfolgswahrscheinlichkeit p ist, denn $\hat{p}$ ist konsistent, erwartungstreu und asymptotisch normalverteilt, vgl. Beispiel 4.4. Das Vorgehen war, zunächst intuitiv einen Schätzer als sinnvoll zu behaupten, um dann seine Güte zu prüfen. Nun betrachten wir ein allgemeines Schätzverfahren – die Maximum-Likelihood-Methode –, welches uns selbst Schätzer vorschlägt. Die funktionale Struktur des gewonnenen Schätzers hängt dann zwar vom Modell ab, aber trotzdem werden wir die Gutartigkeit des Verfahrens in einem recht allgemeingültigen Rahmen behaupten können.

7.1 Definition, Beispiele und Eigenschaften

Wir betrachten hier durchweg parametrische Modelle. Ein Modell ist also gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ mit Bildraum $\mathcal{X}$ und eine konkrete Familie $(\nu_\vartheta)_{\vartheta \in \Theta}$ von Verteilungen auf $\mathcal{X}$. Unter jeder Verteilung ν_ϑ besitzt $\mathfrak{X}$ entweder eine Dichte f_ϑ oder Gewichte g_ϑ, anhand derer wir die Verteilung ν_ϑ im Folgenden beschreiben. Um nicht immer zwischen dem Fall von Dichten oder Gewichten unterscheiden zu müssen, bezeichnen wir hier auch Gewichte mit f_ϑ.

M. Messer und G. Schneider, *Statistik,* https://doi.org/10.1007/978-3-662-59339-4_7

Definition 7.1 (Likelihood-Funktion)
Es sei ein parametrisches Modell gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ *mit Bildraum* $\mathcal{X}$ *und eine Familie gemeinsamer Dichten bzw. Gewichte* $(f_\vartheta)_{\vartheta\in\Theta}$. *Dann ist die Likelihood-Funktion eine Abbildung* $L : \mathcal{X} \times \Theta \to [0, \infty)$ *gegeben durch*

$$L(\mathbf{x}, \vartheta) := f_\vartheta(\mathbf{x}). \tag{7.1}$$

Weiter ist die Log-Likelihood-Funktion eine Abbildung $\ell : \mathcal{X} \times \Theta \to [-\infty, \infty)$ *via*

$$\ell(\mathbf{x}, \vartheta) := \log(L(\mathbf{x}, \vartheta)), \tag{7.2}$$

mit der Konvention $\log(0) := -\infty$.

Sind die Komponenten von $\mathfrak{X}$ unter sämtlichen Verteilungen ν_ϑ unabhängig, dann faktorisieren die Dichten bzw. Gewichte in (7.1): Für $\mathbf{x} = (x_1, \ldots, x_n)^t \in \mathcal{X}$ gilt für alle $\vartheta \in \Theta$, dass

$$L(\mathbf{x}, \vartheta) = \prod_{i=1}^{n} f_\vartheta^{(i)}(x_i),$$

wobei $f_\vartheta^{(i)}$ die entsprechende Randdichte der i-ten Komponente X_i von $\mathfrak{X}$ bezeichne, mit $i = 1, \ldots, n$. Sind die X_i unter allen $\vartheta \in \Theta$ zusätzlich identisch verteilt, dann ist

$$L(\mathbf{x}, \vartheta) = \prod_{i=1}^{n} f_\vartheta^{(1)}(x_i). \tag{7.3}$$

Im Falle der Unabhängigkeit liegt die Bedeutung des Übergangs zur Log-Likelihood-Funktion (7.2) darin, dass sie das Produkt über die Funktionalgleichung des Logarithmus in eine Summe überführt:

$$\ell(\mathbf{x}, \vartheta) = \log\left[\prod_{i=1}^{n} f_\vartheta^{(i)}(x_i)\right] = \sum_{i=1}^{n} \log f_\vartheta^{(i)}(x_i).$$

Konvention und Bedeutung der Likelihood-Funktion: Wir indizieren den wahren unbekannten Parameter in diesem Kapitel fortan mit 0, d. h., wir schreiben ϑ_0 anstatt ϑ. Denn wir müssen nun eine Unterscheidung vornehmen: Zum einen gibt es die wahre unbekannte Verteilung ν_{ϑ_0} von $\mathfrak{X}$ bzw. die wahre Dichte oder die wahren Gewichte f_{ϑ_0}. Zum anderen setzen wir bei Betrachtung der Likelihood-Funktion L den Vektor $\mathfrak{X}$, bzw. $\mathbf{x}$, einfach in sämtliche Dichten bzw. Gewichte f_ϑ ein.

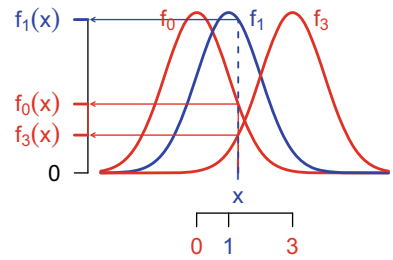

Abb. 7.1 Idee der Maximum-Likelihood-Methode

Grund dafür ist die Idee der Maximum-Likelihood-Methode: Für festgehaltenes $\mathbf{x} \in \mathcal{X}$ suche man aus allen Parametern $\vartheta \in \Theta$ nach demjenigen Kandidaten $\hat{\vartheta}(\mathbf{x})$, für den die Likelihood-Funktion $L(\mathbf{x}, \hat{\vartheta}(\mathbf{x}))$ maximal wird. Beispielsweise sehen wir in Abb. 7.1 eine Familie bestehend aus drei Verteilungen $N(\vartheta, 1)$ mit $\vartheta \in \Theta = \{0, 1, 3\}$ und ein einzelnes Datum $\mathbf{x} \in \mathcal{X} = \mathbb{R}$. Für dieses $\mathbf{x}$ ist die Dichte $f_\vartheta(\mathbf{x})$ für $\vartheta = 1$ größer als die beiden anderen. Der Wert von $\hat{\vartheta}(\mathbf{x}) = 1$ liefert also unter allen $\vartheta \in \Theta = \{0, 1, 3\}$ den maximalen Wert $L(\mathbf{x}, \hat{\vartheta}(\mathbf{x})) = \max_{\vartheta \in \Theta} L(\mathbf{x}, \vartheta)$ der Likelihood Funktionen. Formal schreiben wir

Definition 7.2 (Maximum-Likelihood-Schätzer)
Es sei ein parametrisches Modell gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ *und eine Familie gemeinsamer Dichten bzw. Gewichte* $(f_\vartheta)_{\vartheta \in \Theta}$. *Ein Schätzer* $\hat{\vartheta}$ *für* ϑ *heißt ein Maximum-Likelihood-Schätzer (engl. maximum likelihood estimator, kurz MLE), wenn gilt*

$$L(\mathbf{x}, \hat{\vartheta}(\mathbf{x})) = \max_{\vartheta \in \Theta} L(\mathbf{x}, \vartheta),$$

für alle $\mathbf{x}$ *aus dem Bildraum von* $\mathfrak{X}$.

Bedeutung: Ein MLE sucht denjenigen Parameter $\vartheta \in \Theta$, für den die Beobachtung $\mathbf{x} = (x_1, \ldots, x_n)^t$ die größte gemeinsame Dichte bzw. das größte Gewicht besitzt.

In der Praxis verwendet man zur Bestimmung eines MLE häufig die Log-Likelihood-Funktion, denn da der Logarithmus streng monoton wächst, gilt für alle $\mathbf{x}$ aus dem Bildraum von $\mathfrak{X}$

$$\arg\max_{\vartheta \in \Theta} L(\mathbf{x}, \vartheta) = \arg\max_{\vartheta \in \Theta} \ell(\mathbf{x}, \vartheta), \tag{7.4}$$

d. h., es ist gleichbedeutend, die Maximierer der Likelihood- oder der Log-Likelihood-Funktion zu bestimmen. Ist der MLE eindeutig, so besteht (7.4) für jedes $\mathbf{x}$ aus genau einem Element – nämlich $\hat{\vartheta}(\mathbf{x})$ –, also dem MLE $\hat{\vartheta}$ ausgewertet bei $\mathbf{x}$.

Wir betrachten nun die Maximum-Likelihood-Methode in drei Beispielen. In allen Beispielen wird ein Modell mit unabhängigen und identisch verteilten Komponenten $X_1, \ldots, X_n$ formuliert. Insbesondere faktorisiert die Likelihood-Funktion nach (7.3). Unsere Aufgabe: Suche zu gegebenem $\mathbf{x} = (x_1, \ldots, x_n)^t$ nach denjenigen Parametern $\vartheta \in \Theta$, die die (Log-) Likelihood-Funktion maximieren.

Beispiel 7.3 (MLE bei der Bernoulli-Verteilung)
Es seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim ber(p_0)$ und $p_0 \in \Theta = [0,1]$. Sei $\mathbf{x} = (x_1, \ldots, x_n)^t \in \{0,1\}^n$. Dann finden wir die Likelihood-Funktion als

$$L(\mathbf{x}, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}.$$

Hier nach p zu differenzieren, wäre unangenehm. Daher berechnen wir die Log-Likelihood-Funktion

$$\begin{aligned}\ell(\mathbf{x}, p) &= \sum_{i=1}^{n}(x_i \log(p) + (1-x_i)\log(1-p)) \\ &= \log(p) n\bar{x}_n + \log(1-p)(n - n\bar{x}_n).\end{aligned}$$

Für $p \in (0,1)$ suchen wir die Nullstellen der ersten Ableitung von ℓ

$$\frac{\partial}{\partial p}\ell(\mathbf{x}, p) = \frac{1}{p} n\bar{x}_n - \frac{1}{1-p}(n - n\bar{x}_n) \stackrel{!}{=} 0.$$

Das ist äquivalent zu $0 = (1-p)\bar{x}_n - p(1-\bar{x}_n) = \bar{x}_n - p$, d. h., der MLE für p_0 ergibt sich als die relative Häufigkeit $\hat{p}(\mathbf{x}) = \bar{x}_n$. Denn offenbar ist $\hat{p}(\mathbf{x})$ das eindeutige Maximum von $L(\mathbf{x}, p)$, da $L(\mathbf{x}, p)$ als Funktion von p stetig und nichtnegativ ist mit $L(\mathbf{x}, 0) = L(\mathbf{x}, 1) = 0$. Letzteres gilt zumindest, wenn $\mathbf{x}$ ungleich $(0, \ldots, 0)^t$ bzw. $(1, \ldots, 1)^t$ ist. Für diese beiden langweiligen Fälle gilt aber $\hat{p}(\mathbf{x}) = 0$ bzw. 1 und $L(\mathbf{x}, p) = (1-p)^n$ bzw. p^n, d. h., $L(\mathbf{x}, p)$ fällt bzw. steigt streng monoton, also gilt auch hier, dass $\hat{p}(\mathbf{x})$ Maximierer ist.

Beispiel 7.4 (MLE bei der Exponentialverteilung)
Es seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim exp(\lambda_0)$ und $\lambda_0 \in \Theta := (0, \infty)$. Sei $\mathbf{x} = (x_1, \ldots, x_n)^t \in (\mathbb{R}^+)^n$. Die Likelihood-Funktion ergibt sich als

$$L(\mathbf{x}, \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

und weiter die Log-Likelihood-Funktion als

$$\ell(\mathbf{x}, \lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i .$$

Wir suchen wieder Nullstellen der ersten Ableitung

$$\frac{\partial}{\partial \lambda} \ell(\mathbf{x}, \lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \stackrel{!}{=} 0, \tag{7.5}$$

d. h., der MLE für λ_0 ergibt sich als $\hat{\lambda}(\mathbf{x}) = 1/\bar{x}_n$, denn $\ell''(\mathbf{x}, \lambda) = -n/\lambda^2 < 0$ für alle $\lambda > 0$.

Im nächsten Beispiel ist die Likelihood-Funktion nicht differenzierbar.

Beispiel 7.5 (MLE bei der uniformen Verteilung)
Es seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim unif(0, b_0]$ und $b_0 \in \Theta = (0, \infty)$. Sei $\mathbf{x} = (x_1, \ldots, x_n)^t \in (\mathbb{R}^+)^n$. Mit $x_{(n)} = \max\{x_1, \ldots, x_n\}$ ergibt sich die Likelihood-Funktion als

$$L(\mathbf{x}, b) = \prod_{i=1}^{n} f_b(x_i) = \prod_{i=1}^{n} \frac{1}{b} \mathbb{1}_{(0,b]}(x_i) = \frac{1}{b^n} \mathbb{1}_{[x_{(n)}, \infty)}(b).$$

Dabei haben wir in der letzten Gleichung ausgenutzt, dass alle Indikatorfunktionen den Wert 1 annehmen genau dann, wenn alle x_i kleiner b sind (linke Seite), was aber äquivalent dazu ist, dass b größer ist als jedes x_i (rechte Seite). Damit ergibt sich der MLE für b_0 als die maximale Beobachtung $\hat{b}(\mathbf{x}) = x_{(n)}$. Denn die Likelihood-Funktion L ist überhaupt erst echt größer null, wenn b mindestens $x_{(n)}$ ist, aber in diesem Fall fällt es mit wachsendem b, vgl. Abb. 7.2.

Wir diskutieren nun Eigenschaften von MLEs im Zusammenhang der Beispiele.

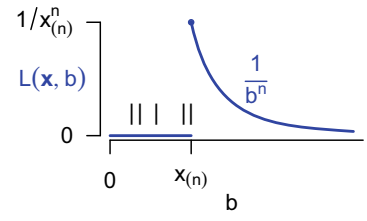

Abb. 7.2 Likelihood-Funktion bei der uniformen Verteilung

1. Ein MLE muss nicht eindeutig sein. Sei zum Beispiel $n = 1$ und eine Familie bestehend aus zwei uniformen Verteilungen $unif[\vartheta, \vartheta + 1]$ mit $\vartheta \in \{0, 0.5\}$ gegeben. Dann nehmen beide assoziierten Dichten bei zum Beispiel $x = 0.9$ den gleichen Wert 1 an und maximieren damit die Likelihood. In den diskutierten Beispielen 7.3–7.5 dagegen waren die MLEs eindeutig.
2. Ein MLE ist i. Allg. nicht erwartungstreu. Zum Beispiel 7.5 wissen wir, dass $[(n + 1)/n]x_{(n)}$ erwartungstreu für b_0 ist, vgl. Beispiel 5.9. Bezüglich Beispiel 7.4 bemerken wir, dass die Funktion $h : \mathbb{R}^+ \to \mathbb{R}^+$ via $h(x) = 1/x$ konvex ist, sodass nach der Jensen-Ungleichung (Lemma 2.8) folgt, dass

$$\mathbb{E}_{\lambda_0}\left[\hat{\lambda}(\mathfrak{X})\right] = \mathbb{E}_{\lambda_0}\left[\frac{1}{\bar{X}}\right] > \frac{1}{\mathbb{E}_{\lambda_0}[\bar{X}]} = \frac{1}{\mathbb{E}_{\lambda_0}[X_1]} = \frac{1}{1/\lambda_0} = \lambda_0,$$

   mit echter Ungleichheit, da $1/\bar{X}$ unter keiner $exp(\lambda)$-Verteilung mit Wahrscheinlichkeit 1 konstant ist.
3. Einen MLE kann man darstellen als Funktion einer suffizienten Statistik S. Dies sehen wir anhand des Satzes von Neyman und Fisher (Satz 5.13) für den Fall eines Modells mit Gewichten sofort: Wegen der Faktorisierung der Likelihood-Funktion der Gestalt

$$L(\mathbf{x}, \vartheta) = f_\vartheta(\mathbf{x}) = r(S(\mathbf{x}), \vartheta) \cdot h(\mathbf{x})$$

   genügt es zur Maximierung der Likelihood-Funktion, den Term $r(S(\mathbf{x}), \vartheta)$ zu maximieren, der nur von der suffizienten Statistik $S(\mathbf{x})$ abhängt. Dies gilt im Modell mit Dichten ganz analog (vgl. zum Beispiel Pruscha 2000).
4. Eine weitere schöne Eigenschaft von MLEs ist ihre Invarianz gegen Umparametrisierung. Es sei ein statistisches Modell gegeben durch einen Zufallsvektor $\mathfrak{X}$ mit Bildraum $\mathcal{X}$ und eine Familie gemeinsamer Dichten/Gewichte $(f_\vartheta)_{\vartheta\in\Theta}$, mit $\Theta \subseteq \mathbb{R}^d$. Sei $\Delta \subseteq \mathbb{R}^k$. Unter einer Umparametrisierung des Modells verstehen wir eine bijektive Abbildung $h : \Theta \to \Delta$. Wir setzen $\varphi := h(\vartheta)$ und finden die umparametrisierte Familie $(\tilde{f}_\varphi)_{\varphi\in\Delta}$ gegeben durch

$$\tilde{f}_\varphi := f_{h^{-1}(\varphi)} = f_\vartheta. \tag{7.6}$$

   Beide betrachteten Familien sind gleich, aber durch verschiedene Parameter beschrieben. Bezüglich der Familie $(f_\vartheta)_{\vartheta\in\Theta}$ sei nun $\hat{\vartheta}$ ein MLE für ϑ. Dann findet sich hinsichtlich der umparametrisierten Familie $(\tilde{f}_\varphi)_{\varphi\in\Delta}$ ein MLE für φ durch die Komposition

$$\hat{\varphi} := h \circ \hat{\vartheta}, \tag{7.7}$$

   d. h. für alle $\mathbf{x} \in \mathcal{X}$ gilt $\hat{\varphi}(\mathbf{x}) = h(\hat{\vartheta}(\mathbf{x}))$. Denn für die Likelihood-Funktion $\tilde{L}$ bezüglich der neuen Parametrisierung gilt für alle $\mathbf{x} \in \mathcal{X}$ nach (7.6) und (7.7), dass $\tilde{L}(\mathbf{x}, \hat{\varphi}(\mathbf{x})) = L(\mathbf{x}, \hat{\vartheta}(\mathbf{x})) \geq L(\mathbf{x}, \vartheta) = \tilde{L}(\mathbf{x}, \varphi)$, und dies ist gerade die definierende Eigenschaft eines MLE für φ.

Wir betrachten Beispiel 7.4, mit $X_1, \ldots, X_n$ unabhängigen und identisch verteilten Zufallsvariablen, mit $X_1 \sim exp(\lambda)$ und $\lambda \in \Theta = (0, \infty)$. Es sei $\mathbf{x} = (x_1, \ldots, x_n)^t \in (\mathbb{R}^+)^n$. Wir hatten die Familie der gemeinsamen Dichten beschrieben durch

$$\left(\prod_{i=1}^{n} \lambda e^{(-\lambda x_i)} \right)_{\lambda \in (0,\infty)}$$

und in dieser Parametrisierung den MLE als $\hat{\lambda}(\mathbf{x}) = 1/\bar{x}_n$ gefunden. Sei nun $h : (0, \infty) \to (0, \infty)$ via $h(\lambda) = 1/\lambda =: \varphi$. Also ist $\lambda = 1/\varphi$, und die umparametrisierte Familie von Dichten ergibt sich als

$$\left(\prod_{i=1}^{n} \frac{1}{\varphi} \exp\left(\frac{1}{\varphi} x_i \right) \right)_{\varphi \in (0,\infty)} .$$

Hier finden wir den MLE von φ als $\hat{\varphi}(\mathbf{x}) = h(\hat{\lambda}(\mathbf{x})) = \bar{x}_n$. Wir erkennen den Charme dieser Parametrisierung: Der MLE ist der Mittelwert und damit erwartungstreu, stark konsistent und asymptotisch normal!

5. In allen Beispielen war der MLE stark konsistent. Für asymptotische Aussagen denken wir dabei wieder an eine Folge $X_1, X_2, \ldots$ von Zufallvariablen, und für $n = 1, 2, \ldots$ betrachten wir das Modell, welches aus der Einschränkung auf die ersten n Zufallsvariablen hervorgeht. Für die Beispiele 7.3 und 7.4 ist die Konsistenz eine direkte Folgerung aus dem Starken Gesetz der großen Zahlen. Für Beispiel 7.5 wenden wir das Borel-Cantelli-Lemma an (siehe Feller 1968): Es sei $\varepsilon \in (0, b_0)$, dann gilt

$$\sum_{i=1}^{\infty} \mathbb{P}_{b_0}(X_i \in [b_0 - \varepsilon, b_0]) = \sum_{i=1}^{\infty} \frac{\varepsilon}{b_0} = \infty,$$

und, da die $(X_i)_i$ unabhängig sind, folgt nach Borel-Cantelli

$$\mathbb{P}_{b_0}\left(\lim_{n\to\infty} X_{(n)} \in [b_0 - \varepsilon, b_0] \right) \geq \mathbb{P}_{b_0}\left(\{X_i \in [b_0 - \varepsilon, b_0]\} \text{ für unendlich viele i} \right) = 1.$$

Das ist die definierende Eigenschaft der starken Konsistenz, weil ε beliebig klein gewählt werden kann. Wir werden in Abschn. 7.2 allgemeine Bedingungen kennenlernen, die gewisse Konsistenzaussagen im Zusammenhang mit MLEs ermöglichen, siehe auch Satz 7.9.

6. Zudem werden sich unter gewissen Voraussetzungen auch Aussagen über asymptotische Normalität machen lassen, siehe ebenfalls Satz 7.9. In Beispiel 7.3 erkennen wir asymptotische Normalität des MLE schon direkt, denn nach dem Zentralen Grenzwertsatz gilt für $n \to \infty$

$$\sqrt{n}(\hat{p}_n(\mathfrak{X}_n) - p_0) \xrightarrow{d_{p_0}} N\left(0, p_0(1 - p_0)\right),$$

wobei wir für die Fälle $p_0 \in \{0, 1\}$ eine $N(0, 0)$-verteilte Zufallsvariable als eine Konstante mit Wert 0 verstehen. In Beispiel 7.4 dagegen konvergiert das Maximum $X_{(n)} = \max(X_1, \ldots, X_n)$ unabhängiger und identisch uniformverteilter Zufallsvariablen unter geeigneter Reskalierung gegen eine Verteilung, die nicht die Normalverteilung ist, siehe Ferguson (1996). Dass dort oben genannter Satz 7.9 nicht greift, liegt im Wesentlichen daran, dass die Likelihood-Funktion am wahren Parameter nicht differenzierbar ist, siehe Abb. 7.2.

7.2 Konsistenz und asymptotische Normalität

Wie wir in den Beispielen 7.3–7.5 gesehen haben, kann ein MLE unter gewissen Bedingungen konsistent und asymptotisch normalverteilt sein. Asymptotische Normalität schreibt sich als

$$\sqrt{n}(\hat{\vartheta}_n(\mathfrak{X}_n) - \vartheta_0) \xrightarrow{d_{\vartheta_0}} N(0, \sigma^2_{ML}),$$

wobei die Anzahl n an Beobachtungen gegen unendlich streben soll. Das Beispiel der uniformen Verteilung aber zeigte, dass die asymptotische Normalverteilung nicht immer zutrifft. Der wesentliche Grund dafür ist, dass die Likelihood-Funktion dort nicht hinreichend glatt ist, vgl. Abb. 7.2. Daher benötigt man sogenannte *Regularitätsbedingungen* an die zugrunde liegende Familie von Dichten bzw. Gewichten, um asymptotische Aussagen machen zu können. Teile dieser Regularitätsbedingungen werden bereits benötigt, um die asymptotische Varianz σ^2_{ML} des MLE zu formulieren. Diese wird über den Begriff der sogenannten *Fisher-Information* formuliert, in welche Information über die zugrunde liegende Familie von Dichten bzw. Gewichten eingeht.

Um die folgenden recht technischen Begrifflichkeiten nicht zu überladen, betrachten wir den Fall des eindimensionalen Parameterraums $\Theta \subseteq \mathbb{R}^1$.

7.2.1 Die Fisher-Information

Definition 7.6 (Reguläres Modell und Fisher-Information)
Es sei ein parametrisches Modell gegeben durch einen Zufallsvektor $\mathfrak{X}_n = (X_1, \ldots, X_n)^t$ mit Bildraum $\mathcal{X} \subseteq \mathbb{R}^n$ und eine Familie gemeinsamer Dichten bzw. Gewichte $(f_\vartheta)_{\vartheta \in \Theta}$, und $\Theta \subseteq \mathbb{R}$ sei offen. Das Modell heißt regulär, wenn folgende Bedingungen erfüllt sind:

1. *Glattheit: Für alle* $\mathbf{x} \in \mathcal{X}$ *ist die Abbildung* $\vartheta \mapsto \ell(\mathbf{x}, \vartheta)$ *zweimal stetig differenzierbar.*
2. *Vertauschbarkeit: Für alle* $\vartheta \in \Theta$ *gilt für* $j = 1, 2$

$$\int_{\mathcal{X}} \frac{\partial^j}{\partial \vartheta^j} L(\mathbf{x}, \vartheta) \mathrm{d}\mathbf{x} = \frac{\partial^j}{\partial \vartheta^j} \int_{\mathcal{X}} L(\mathbf{x}, \vartheta) \mathrm{d}\mathbf{x} \quad \left(= \frac{\partial^j}{\partial \vartheta^j} \cdot 1 = 0 \right). \tag{7.8}$$

3. *Für alle* $\vartheta_0 \in \Theta$ *ist*

$$I_n(\vartheta_0) := \mathbb{V}ar_{\vartheta_0} \left(\frac{\partial}{\partial \vartheta} \ell(\mathfrak{X}_n, \vartheta_0) \right) \in (0, \infty).$$

Insbesondere heißt die Funktion $I_n : \Theta \to (0, \infty)$ *die Fisher-Informationsfunktion bezüglich* $(f_\vartheta)_{\vartheta \in \Theta}$*, die Auswertung* $I_n(\vartheta_0)$ *bezeichnen wir als Fisher-Information.*

Die Terme $(\partial/\partial\vartheta)\ell(\mathbf{x}, \vartheta)$ und $(\partial^j/\partial\vartheta^j)L(\mathbf{x}, \vartheta)$ sind so zu lesen, dass wir die (Log-) Likelihood-Funktion erst nach ϑ ableiten (evtl. $(j = 2)$-mal) und dann an der Stelle ϑ auswerten. Insbesondere schreibt sich der Ausdruck in 3. als

$$I_n(\vartheta_0) = \mathbb{V}ar_{\vartheta_0} \left(\frac{\partial}{\partial \vartheta} \ell(\mathfrak{X}_n, \vartheta) \Big|_{\vartheta = \vartheta_0} \right),$$

d. h. die abgeleitete Log-Likelihood-Funktion wird am wahren zugrunde liegenden Parameter ϑ_0 betrachtet und durch Einsetzen von $\mathfrak{X}_n$ eine Zufallsvariable, deren Varianz dann unter dem Parameter ϑ_0 bestimmt wird. Zudem verschwindet der Term in Bedingung 2, da wir über den kompletten Bildraum $\mathcal{X}$ integrieren und $L(\mathbf{x}, \vartheta) = f_\vartheta(x)$ für jedes $\vartheta \in \Theta$ eine Dichte ist, deren Integral konstant 1 ist. Im Falle von Gewichten ist hier und im Folgenden die Integration durch Summation zu ersetzen.

Wir werden sehen, dass die asymptotische Varianz σ^2_{ML} der inversen Fisher-Information $I_n^{-1}(\vartheta_0)$ entspricht (Satz 7.9). Bevor wir die Bedeutung der Fisher-Information genauer diskutieren, stellen wir sie anders dar. In einem regulären Modell heißen

$$S(\mathbf{x}, \vartheta) := \frac{\partial}{\partial \vartheta} \ell(\mathbf{x}, \vartheta) \qquad \text{die Scorefunktion und} \tag{7.9}$$

$$J(\mathbf{x}, \vartheta) := -\frac{\partial}{\partial \vartheta} S(\mathbf{x}, \vartheta) = -\frac{\partial^2}{\partial \vartheta^2} \ell(\mathbf{x}, \vartheta) \qquad \text{die Informationsfunktion.} \tag{7.10}$$

Wir bemerken, dass die Fisher-Information die Varianz der Scorefunktion bei ϑ_0 ist, denn per definitionem gilt $I_n(\vartheta_0) = \mathbb{V}ar_{\vartheta_0}(S(\mathfrak{X}_n, \vartheta_0))$.

Zudem löst ein MLE $\hat{\vartheta}_n$ notwendigerweise die *Scoregleichung,* d. h., es gilt für alle $\mathbf{x}$ aus dem Bildraum von $\mathfrak{X}_n$, dass

$$S(\mathbf{x}, \hat{\vartheta}_n(\mathbf{x})) = 0,$$

denn $\hat{\vartheta}_n(\mathbf{x})$ ist ja eine Maximalstelle der Log-Likelihood-Funktion, also eine Nullstelle ihrer Ableitung.

Weiter besagt das nächste Lemma, dass die Fisher-Information dem Erwartungswert der Informationsfunktion entspricht.

Lemma 7.7 (Darstellung der Fisher-Information)
Es sei ein reguläres Modell gegeben durch einen Zufallsvektor $\mathfrak{X}_n = (X_1, \ldots, X_n)^t$ *mit Bildraum* $\mathcal{X} \subseteq \mathbb{R}^n$ *und eine Familie gemeinsamer Dichten bzw. Gewichte* $(f_\vartheta)_{\vartheta \in \Theta}$. *Dann gilt für alle* $\vartheta_0 \in \Theta$

$$i): \ \mathbb{E}_{\vartheta_0}[S(\mathfrak{X}_n, \vartheta_0)] = 0, \tag{7.11}$$

$$ii): \ \mathbb{E}_{\vartheta_0}[J(\mathfrak{X}_n, \vartheta_0)] = I_n(\vartheta_0). \tag{7.12}$$

Beweis

i) Es bezeichne die Hochstellung $'$ die partielle Ableitung nach ϑ, d. h. den Operator $\partial/\partial\vartheta$. Dann gilt

$$\begin{aligned}
\mathbb{E}_{\vartheta_0}[S(\mathfrak{X}_n, \vartheta_0)] &= \int_{\mathcal{X}} \frac{\partial}{\partial\vartheta} \ell(\mathbf{x}, \vartheta_0) L(\mathbf{x}, \vartheta_0) \mathrm{d}\mathbf{x} \\
&\overset{(*)}{=} \int_{\mathcal{X}} \frac{L'(\mathbf{x}, \vartheta_0)}{L(\mathbf{x}, \vartheta_0)} L(\mathbf{x}, \vartheta_0) \mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}} \frac{\partial}{\partial\vartheta} L(\mathbf{x}, \vartheta_0) \mathrm{d}\mathbf{x} \\
&\overset{(7.8)}{=} \frac{\partial}{\partial\vartheta} \int_{\mathcal{X}} L(\mathbf{x}, \vartheta_0) \mathrm{d}\mathbf{x} = 0.
\end{aligned}$$

Dabei haben wir in der vorletzten Gleichung lediglich die Vertauschbarkeitseigenschaft ausgenutzt. Und in $(*)$ wurde die Kettenregel auf den Logarithmus angewendet.
Wir bemerken, dass i. Allg. $\mathbb{E}_{\vartheta_0}[S(\mathfrak{X}_n, \vartheta)] \neq 0$ ist – hier setzen wir die Beobachtungen, deren Verteilung die Dichte f_{ϑ_0} besitzt, explizit in eine falsche Dichte f_ϑ ein –, da sich dann die Likelihood-Funktion auf der rechten Seite von $(*)$ nicht wegkürzt, denn dort tauchen $L(\mathbf{x}, \vartheta)$ und $L(\mathbf{x}, \vartheta_0)$ auf, welche i. Allg. ungleich sind. Dass die Auswertung am wahren Parameter $\mathbb{E}_{\vartheta_0}[S(\mathfrak{X}_n, \vartheta_0)] = 0$ liefert, kann man so interpretieren, dass

die wahre Dichte f_{ϑ_0} die Beobachtung $\mathfrak{X}_n$ erwartungsgemäß am besten beschreibt, im Sinne der Maximierung der Log-Likelihood-Funktion.

ii) Zunächst gilt aufgrund der Vertauschbarkeitseigenschaft

$$\mathbb{E}_{\vartheta_0}\left[\frac{L''(\mathfrak{X}_n,\vartheta_0)}{L(\mathfrak{X}_n,\vartheta_0)}\right] = \int_{\mathcal{X}} \frac{L''(\mathbf{x},\vartheta_0)}{L(\mathbf{x},\vartheta_0)} L(\mathbf{x},\vartheta_0)\mathrm{d}\mathbf{x} = \int_{\mathcal{X}} \frac{\partial^2}{\partial\vartheta^2} L(\mathbf{x},\vartheta_0)\mathrm{d}\mathbf{x} \overset{(7.8)}{=} 0. \tag{7.13}$$

Damit folgt unter Ausnutzung der Kettenregel $(**)$ und der Quotientenregel $(***)$, dass

$$\begin{aligned}
\mathbb{E}_{\vartheta_0}\left[J(\mathfrak{X}_n,\vartheta_0)\right] &= -\mathbb{E}_{\vartheta_0}\left[\ell''(\mathfrak{X}_n,\vartheta_0)\right] \\
&\overset{(**)}{=} -\mathbb{E}_{\vartheta_0}\left[\left(\frac{L'(\mathfrak{X}_n,\vartheta_0)}{L(\mathfrak{X}_n,\vartheta_0)}\right)'\right] \\
&\overset{(***)}{=} -\mathbb{E}_{\vartheta_0}\left[\frac{L(\mathfrak{X}_n,\vartheta_0)L''(\mathfrak{X}_n,\vartheta_0) - L'(\mathfrak{X}_n,\vartheta_0)^2}{L(\mathfrak{X}_n,\vartheta_0)^2}\right] \\
&\overset{(7.13)}{=} \mathbb{E}_{\vartheta_0}\left[\left(\frac{L'(\mathfrak{X}_n,\vartheta_0)}{L(\mathfrak{X}_n,\vartheta_0)}\right)^2\right] \\
&\overset{(**)}{=} \mathbb{E}_{\vartheta_0}\left[S(\mathfrak{X}_n,\vartheta_0)^2\right] \overset{(7.11)}{=} \mathbb{V}ar_{\vartheta_0}(S(\mathfrak{X}_n,\vartheta_0)) = I_n(\vartheta_0).
\end{aligned}$$

Zur Interpretation der Fisher-Information stellen wir fest, dass diese die erwartete (negative) Krümmung der Log-Likelihood-Funktion am wahren Parameter ist. Ist die Krümmung stark negativ, hat die Likelihood-Funktion im Mittel einen prominenten Gipfel am wahren Parameter. Daher schwankt die Maximalstelle über verschiedene Realisierungen von $\mathfrak{X}_n$ wenig, sodass der Maximierer eine kleine Varianz hat.

In Abb. 7.3 ist die erwartete Log-Likelihood-Funktion $\mathbb{E}_{\vartheta_0}[\ell(\mathfrak{X}_n,\vartheta)]$ durch die mittlere Log-Likelihood-Funktion (rot) approximiert, wobei der punktweise Mittelwert aus den 20 schwarzen Realisierungen von Log-Likelihood-Funktionen $\ell(\mathbf{x},\vartheta)$ gebildet wurde. Die Funktionen sind gegen den Parameter ϑ aufgetragen. Die betrachteten Realisierungen $\mathbf{x} \in \mathbb{R}^n$ beziehen sich jeweils auf $n = 20$ Ziehungen aus der Exponentialverteilung zum wahren Parameter $\lambda_0 = 2$ (A) und $\lambda_0 = 4$ (B) und sind nicht dargestellt. Wir erkennen, dass die Log-Likelihood-Funktionen am wahren Parameter in Abb. 7.3a viel stärker gekrümmt sind als in Abb. 7.3b. Das Maximum ist in Abb. 7.3a prominenter und weniger variabel, und folglich schwanken die aus den schwarzen Log-Likelihood-Funktionen resultierenden MLEs (vertikalen Striche) viel weniger als in Abb. 7.3b. In der Tat werden wir die inverse Fisher-Information bei der Exponentialverteilung als $I_1(\lambda_0)^{-1} = \lambda_0^2$ errechnen, siehe Beispiel 7.10. Auch das besagt also, dass die asymptotische Varianz des MLE in Abb. 7.3a kleiner ist als in Abb. 7.3b.

Das folgende Lemma besagt, dass die Fisher-Information die schöne Eigenschaft der Additivität besitzt, falls die Beobachtungen als unabhängig und identisch verteilt angenommen werden. Es sei $I_1(\vartheta_0) = \mathbb{V}ar_{\vartheta_0}(\ell'(X_1,\vartheta_0))$ die Fisher-Information der

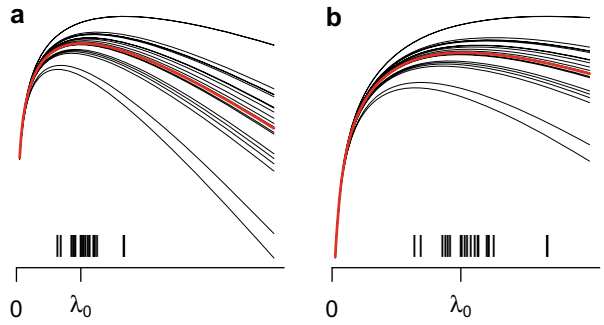

Abb. 7.3 Zur Interpretation der Fisher-Information. Realisierungen der Log-Likelihood-Funktion für jeweils 20 Ziehungen aus der Exponentialverteilung mit Parameter $\lambda_0 = 2$ (**a**) und $\lambda_0 = 4$ (**b**). Rot: Mittlere Log-Likelihood-Funktion. Die MLEs (vertikale Striche) schwanken in **a** weniger als in **b**

Individualbeobachtung. Wir betrachten hier also genau genommen das Modell, das durch Restriktion von $\mathfrak{X}_n = (X_1, \ldots, X_n)^t$ auf die erste Komponente X_1 hervorgeht und bilden diesbezüglich die Likelihood-Funktion.

Lemma 7.8 (Additivität der Fisher-Information)
Es sei ein reguläres Modell gegeben durch einen Zufallsvektor $\mathfrak{X}_n = (X_1, \ldots, X_n)^t$ und eine Familie gemeinsamer Dichten bzw. Gewichte $(f_\vartheta)_{\vartheta\in\Theta}$. Unter allen Dichten bzw. Gewichten seien die Komponenten von $\mathfrak{X}_n$ unabhängig und identisch verteilt. Dann gilt für alle $\vartheta_0 \in \Theta$

$$I_n(\vartheta_0) = nI_1(\vartheta_0).$$

Beweis Der Grund ist die Funktionalgleichung des Logarithmus. Für $\mathbf{x}_n = (x_1, \ldots, x_n)^t$ aus dem Bildraum $\mathcal{X}$ von $\mathfrak{X}_n$ faktorisieren die Dichten wegen der Unabhängigkeit. Es gilt für alle $\vartheta_0 \in \Theta$

$$S(\mathbf{x}_n, \vartheta_0) \overset{(7.3)}{=} \frac{\partial}{\partial\vartheta} \log\left(\prod_{i=1}^n f_\vartheta^{(1)}(x_i)\right)\Bigg|_{\vartheta=\vartheta_0} = \frac{\partial}{\partial\vartheta}\sum_{i=1}^n \log\left(f_\vartheta^{(1)}(x_i)\right)\Bigg|_{\vartheta=\vartheta_0}$$
$$=: \sum_{i=1}^n S_1(x_i, \vartheta_0).$$

Damit folgt

$$I_n(\vartheta_0) = \mathbb{V}ar_{\vartheta_0}\left(S(\mathfrak{X}_n, \vartheta_0)\right) = \sum_{i=1}^{n} \mathbb{V}ar_{\vartheta_0}\left(S_1(X_i, \vartheta_0)\right) = nI_1(\vartheta_0).$$

7.2.2 Konsistenz und asymptotische Normalität

Im Hinblick auf asymptotische Aussagen ziehen wir uns im Folgenden auf den Fall eines regulären Modells zurück, in dem die Zufallsvariablen insbesondere unabhängig und identisch verteilt sind. Wir formulieren Bedingungen, die sowohl die Konsistenz als auch die asymptotische Normalität eines Schätzers sichern, welcher die ML-Gleichung löst, d. h. Nullstelle der Scorefunktion ist. Wir fordern die Bedingungen aus Definition 7.6 des regulären Modells und noch etwas mehr.

Satz 7.9 (Asymptotisches Verhalten einer Lösung der ML-Gleichung)
Es sei ein reguläres Modell beschrieben durch die Zufallsvariable X_1 mit Dichte $f_\vartheta^{(1)}$, und $f_\vartheta^{(1)}$ sei Mitglied einer Familie $(f_\vartheta^{(1)})_{\vartheta\in\Theta}$. Dabei sei $\Theta \subseteq \mathbb{R}$ ein offenes Intervall. Für die Familie gelten zusätzlich folgende Aussagen:

i. Gleichmäßige Beschränktheit: Es existiert eine reellwertige Funktion $M(x_1)$ und eine Konstante K, sodass für alle $\vartheta_0 \in \Theta$ und für alle x_1 aus dem Bildraum von X_1 gilt

$$\left|\frac{\partial^2}{\partial\vartheta^2}\log L_1(x_1, \vartheta_0)\right| < M(x_1) \quad \text{und} \quad \mathbb{E}_{\vartheta_0}[M(X_1)] < K.$$

Dabei ist $L_1(x_1, \vartheta_0) = f_{\vartheta_0}^{(1)}(x_1)$ die Likelihood der Individualbeobachtung x_1.

ii. Identifizierbarkeit: Falls für $\vartheta, \tilde{\vartheta} \in \Theta$ gilt, dass $f_\vartheta^{(1)} = f_{\tilde{\vartheta}}^{(1)}$, dann ist $\vartheta = \tilde{\vartheta}$.

Seien $X_1, X_2, \ldots$ unabhängig und identisch verteilt wie oben, und für $n = 1, 2, \ldots$ betrachte man das Modell beschrieben durch die ersten n Zufallsvariablen $\mathfrak{X}_n = (X_1, \ldots, X_n)^t$. Dann existiert eine Folge von Schätzern $(\hat{\vartheta}_n)_{n=1,2,\ldots}$ mit folgenden Eigenschaften:

1. Nullstelle der Scorefunktion: Unter allen $\vartheta_0 \in \Theta$ gilt für $n = 1, 2, \ldots$

$$\sum_{i=1}^{n} \frac{\partial}{\partial\vartheta}\log L_1(X_i, \hat{\vartheta}_n(\mathfrak{X}_n)) = 0 \quad \text{mit Wahrscheinlichkeit } 1.$$

2. *Starke Konsistenz: Unter allen $\vartheta_0 \in \Theta$ gilt für $n \to \infty$*

$$\hat{\vartheta}_n(\mathfrak{X}_n) \to \vartheta_0 \quad \text{mit Wahrscheinlichkeit } 1.$$

3. *Asymptotische Normalität: Für $n \to \infty$ gilt*

$$\sqrt{n}(\hat{\vartheta}_n(\mathfrak{X}_n) - \vartheta_0) \xrightarrow{d_{\vartheta_0}} N(0, I_1^{-1}(\vartheta_0)).$$

Dabei ist die asymptotische Varianz $I_1^{-1}(\vartheta_0)$ die inverse Fisher-Information.

Achtung: Der Satz macht eine Aussage über einen Schätzer, der eine Nullstelle der Scoregleichung ist. Im Hinblick auf den MLE sind die Inhalte des Satzes nur dann hilfreich, wenn bekannt ist, dass er die eindeutige Nullstelle der Scoregleichung ist. Für einen Beweis siehe Ferguson (1996). Wir geben zwei Heuristiken an. Dabei vermeiden wir die Diskussion um die Existenz der betrachteten Größen und nehmen an, dass $\hat{\vartheta}_n$ eindeutiger MLE ist. Es bezeichne wieder ϑ_0 den wahren unbekannten Parameter.

Heuristik zur Konsistenz Die mit n skalierte Log-Likelihood-Funktion ist gegeben durch

$$\mathscr{L}_n(\mathfrak{X}_n, \vartheta) := \frac{1}{n} \sum_{i=1}^{n} \log f_\vartheta^{(1)}(X_i).$$

Nach dem Starken Gesetz der großen Zahlen gilt für alle $\vartheta \in \Theta$ für $n \to \infty$

$$\mathscr{L}_n(\mathfrak{X}_n, \vartheta) \longrightarrow \mathbb{E}_{\vartheta_0}[\log f_\vartheta^{(1)}(X_1)] =: \mathscr{L}_\infty(\vartheta) \quad \text{mit Wahrscheinlichkeit } 1.$$

Achtung: Wir hatten angenommen, dass ϑ_0 der wahre zugrunde liegende Parameter ist, und daher wird auch der Erwartungswert unter ϑ_0 gebildet. Da $\log(x) \leq x - 1$, folgt für alle $\vartheta \in \Theta$

$$\begin{aligned}
\mathscr{L}_\infty(\vartheta) - \mathscr{L}_\infty(\vartheta_0) &= \mathbb{E}_{\vartheta_0}[\log f_\vartheta^{(1)}(X_1) - \log f_{\vartheta_0}^{(1)}(X_1)] \\
&= \mathbb{E}_{\vartheta_0}\left[\log \frac{f_\vartheta^{(1)}(X_1)}{f_{\vartheta_0}^{(1)}(X_1)}\right] \\
&\leq \mathbb{E}_{\vartheta_0}^{(1)}\left[\frac{f_\vartheta^{(1)}(X_1)}{f_{\vartheta_0}^{(1)}(X_1)} - 1\right] \\
&= \int \left(\frac{f_\vartheta^{(1)}(x)}{f_{\vartheta_0}^{(1)}(x)} - 1\right) f_{\vartheta_0}^{(1)}(x)\mathrm{d}x
\end{aligned}$$

$$= \int f_{\vartheta}^{(1)}(x)\mathrm{d}x - \int f_{\vartheta_0}^{(1)}(x)\mathrm{d}x$$
$$= 1 - 1 = 0.$$

Der MLE $\hat{\vartheta}_n(\mathfrak{X}_n)$ maximiert $\mathscr{L}_n(\mathfrak{X}_n, \vartheta)$ definitionsgemäß. Da ϑ_0 die Grenzfunktion $\mathscr{L}_\infty(\vartheta)$ maximiert, und da $\mathscr{L}_n(\mathfrak{X}_n, \vartheta) \to \mathscr{L}_\infty(\vartheta)$ mit Wahrscheinlichkeit 1 unter ϑ_0 gilt, scheint die starke Konsistenz von $\hat{\vartheta}_n(\mathfrak{X}_n)$ für ϑ_0 plausibel. Schematisch:

$$\begin{array}{ccc} \hat{\vartheta}_n(\mathfrak{X}_n) \text{ maximiert } \mathscr{L}_n(\mathfrak{X}_n, \vartheta) \\ (\downarrow) \qquad\qquad \downarrow \\ \vartheta_0 \text{ maximiert } \mathscr{L}_\infty(\vartheta) \end{array}$$

Heuristik zur asymptotischen Normalität Nach dem Mittelwertsatz existiert $\hat{\rho}_n = \hat{\rho}_n(\mathfrak{X}_n) \in [\min(\hat{\vartheta}_n(\mathfrak{X}_n), \vartheta_0), \max(\hat{\vartheta}_n(\mathfrak{X}_n), \vartheta_0)]$, sodass

$$\frac{\mathscr{L}_n'(\mathfrak{X}_n, \hat{\vartheta}_n(\mathfrak{X}_n)) - \mathscr{L}_n'(\mathfrak{X}_n, \vartheta_0)}{\hat{\vartheta}_n(\mathfrak{X}_n) - \vartheta_0} = \mathscr{L}_n''(\mathfrak{X}_n, \hat{\rho}_n).$$

Da $\hat{\vartheta}_n(\mathfrak{X}_n)$ die Funktion $\mathscr{L}_n(\mathfrak{X}_n, \vartheta)$ maximiert, folgt $\mathscr{L}_n'(\mathfrak{X}_n, \hat{\vartheta}_n(\mathfrak{X}_n)) = 0$, sodass

$$\sqrt{n}(\hat{\vartheta}_n(\mathfrak{X}_n) - \vartheta_0) = \frac{\sqrt{n}\mathscr{L}_n'(\mathfrak{X}_n, \vartheta_0)}{-\mathscr{L}_n''(\mathfrak{X}_n, \hat{\rho}_n)}. \tag{7.14}$$

Für den Zähler in (7.14) nutzen wir zunächst die Vertauschungsrelation $\mathbb{E}_{\vartheta_0}[\ell_1'(X_1, \vartheta_0)] = 0$. Mit dem Zentralen Grenzwertsatz folgt für $n \to \infty$

$$\begin{aligned} &\sqrt{n}\mathscr{L}_n'(\mathfrak{X}_n, \vartheta_0) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \ell_1'(X_i, \vartheta_0) - \mathbb{E}_{\vartheta_0}[\ell_1'(X_1, \vartheta_0)]\right) \quad \xrightarrow{d_{\vartheta_0}} N\left(0, \mathbb{V}ar_{\vartheta_0}(\ell_1'(X_1, \vartheta_0)\right) \\ &= N\left(0, I_1(\vartheta_0)\right). \end{aligned}$$

Für den Nenner in (7.14) finden wir aufgrund des Starken Gesetzes der großen Zahlen für alle $\vartheta \in \Theta$ für $n \to \infty$

$$\mathscr{L}_n''(\mathfrak{X}_n, \vartheta) = \frac{1}{n}\sum_{i=1}^{n} \ell_1''(X_i, \vartheta) \longrightarrow \mathbb{E}_{\vartheta_0}[\ell_1''(X_1, \vartheta)] \qquad \text{mit Wahrscheinlichkeit 1,}$$

unter ϑ_0. Da $\hat{\rho}_n \in [\min(\hat{\vartheta}_n(\mathfrak{X}_n), \vartheta_0), \max(\hat{\vartheta}_n(\mathfrak{X}_n), \vartheta_0)]$, folgt aufgrund der starken Konsistenz des MLE $\hat{\vartheta}_n(\mathfrak{X}_n)$, dass $\hat{\rho}_n \to \vartheta_0$ mit Wahrscheinlichkeit 1 unter ϑ_0. Gilt nun für $n \to \infty$

$$\mathscr{L}_n''(\mathfrak{X}_n, \hat{\rho}_n) \xrightarrow{\mathbb{P}_{\vartheta_0}} \mathbb{E}_{\vartheta_0}[\ell_1''(X_1, \vartheta_0)] \overset{(7.12)}{=} -I_1(\vartheta_0), \tag{7.15}$$

so folgt mit dem Satz von Slutsky (Satz 2.12) die Konvergenz von (7.14) gegen $N(0, I_1^{-1}(\vartheta_0))$.

Als Anwendungsbeispiel betrachten wir nochmals die Exponentialverteilung.

Beispiel 7.10 (Der MLE bei Exponentialverteilung)
Seien $X_1, X_2 \ldots$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim exp(\lambda_0)$ und $\lambda_0 \in \Theta := (0, \infty)$. Im Modell der ersten n Beobachtungen ist der MLE von λ nach Beispiel 7.4 gegeben durch $\hat{\lambda}(\mathbf{x}_n) = 1/\overline{x}_n$, für $\mathbf{x}_n = (x_1, \ldots, x_n)^t \in (\mathbb{R}^+)^n$. Insbesondere ist der MLE die eindeutige Lösung der Scoregleichung (7.5).

Dann besagt Satz 7.9 (dessen Bedingungen nachzuprüfen sind) zum einen, was wir schon wussten: $\hat{\lambda}_n(\mathfrak{X}_n)$ ist stark konsistent für λ_0. Weiter berechnen wir für die asymptotische Normalität die Fisher-Information: Es gilt für $x_1 \in (0, \infty)$

$$\begin{aligned}
L(x_1, \lambda) &= f_\lambda^{(1)}(x_1) = \lambda e^{-\lambda x_1}, \\
\ell(x_1, \lambda) &= \log f_\lambda(x_1) = \log \lambda - \lambda x_1, \\
S(x_1, \lambda) &= \frac{\partial}{\partial \lambda} \ell(x_1, \lambda) = \frac{1}{\lambda} - x_1, \\
J(x_1, \lambda) &= -\frac{\partial^2}{\partial \lambda^2} \ell(x_1, \lambda) = \frac{1}{\lambda^2}, \\
I_1(\lambda_0) &= \mathbb{E}_{\lambda_0}[J(X_1, \lambda_0)] = \frac{1}{\lambda_0^2}.
\end{aligned}$$

Der Satz liefert dann die asymptotische Normalität des MLE

$$\sqrt{n}(\hat{\lambda}_n(\mathfrak{X}_n) - \lambda_0) \xrightarrow{d_{\lambda_0}} N(0, \lambda_0^2).$$

Um ein asymptotisches Konfidenzintervall zu konstruieren, müssen wir die Varianz λ_0^2 konsistent schätzen. Da $\hat{\lambda}_n(\mathfrak{X}_n) = 1/\bar{X}_n$ stark konsistent für λ_0 ist, nutzen wir sein Quadrat $\hat{\lambda}_n^2(\mathfrak{X}_n)$, und finden mit dem Satz von Slutsky für $n \to \infty$

$$\sqrt{n}\bar{X}_n \left(\frac{1}{\bar{X}_n} - \lambda_0 \right) \xrightarrow{d_{\lambda_0}} N(0,1).$$

Sei $\alpha \in (0,1)$ und q das $1 - \alpha/2$-Quantil der $N(0,1)$-Verteilung. Dann ist

$$I_n(\mathbf{x}_n) = \left[\frac{1}{\bar{x}_n} - q \cdot \frac{1}{\sqrt{n}\bar{x}_n}, \frac{1}{\bar{x}_n} + q \cdot \frac{1}{\sqrt{n}\bar{x}_n} \right]$$

ein asymptotisches $(1 - \alpha)$-Konfidenzintervall für λ_0.

7.3 Dialog: Schätzmethoden

Ein Kommilitone von Ihnen trifft auf einer Medizinerparty eine Gruppe von Doktoranden aus der Hirnforschung. Für dieses Thema hat sich Ihr Kollege schon immer interessiert und verwickelt die Gruppe deshalb in ein Gespräch über ihre Forschungsprojekte. Eine Doktorandin berichtet, dass sie in ihren Studien einzelne Nervenzellen beobachtet. Dabei misst sie die Zeitpunkte, zu denen eine Zelle einen sogenannten *spike* – eine kurze elektrische Entladung – aussendet. Um Unterschiede zwischen verschiedenen Zellen zu finden, würde sie die Zeitreihe dieser Zeitpunkte – ein sogenannter *spike train* – und ihre Eigenschaften gerne zusammenfassen. Doch da tun sich schon einige Schwierigkeiten auf. Die Doktorandin (**D**) zeigt Ihrem Kollegen (**K**) eine Abbildung ihrer Daten auf dem Handy (Abb. 7.4).

D: Schau mal hier. . . Ich denke, das müsste eigentlich ganz einfach sein. Meine Kollegen machen das ständig und sagen, die Verteilung der Wartezeiten zwischen *spikes* kann man mit einer sogenannten Gammaverteilung beschreiben. Du bist doch Mathematiker, kannst du mir das erklären?
K: Na klar! Eine Gammaverteilung könnte geeignet sein, denn sie ist schon mal nur auf den positiven reellen Zahlen definiert.
Das sagt ihr nun erst mal gar nichts.
D: Das klingt ja sehr professionell, aber jetzt mal ehrlich, was heißt das denn?
K: Das heißt, dass gammaverteilte Größen immer größer als null sind – was man ja auch braucht bei Wartezeiten, die sind ja auch nie negativ.
D: Okay, das ist schon mal logisch. Aber da muss doch noch mehr dran sein. . . Zum Beispiel verstehe ich nicht, wie eine einzige Verteilung diese beiden total verschiedenen Datensätze beschreiben kann.
Das ist natürlich eine gute Frage. Ihr Kollege erklärt:
K: Ja, das ist so: Die Gammaverteilung bildet eigentlich eine ganze Familie von Verteilungen, denn sie hat zwei Parameter.
Um das zu erläutern, zeichnet er schnell eine grobe Skizze von Abb. 6.3. Dann fährt er fort:

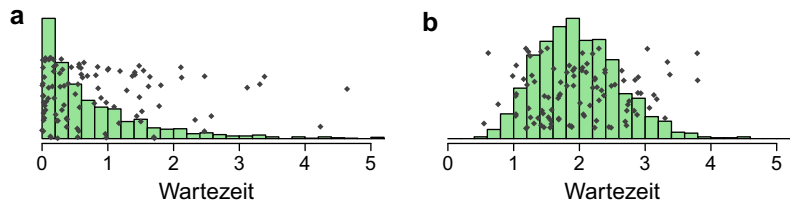

Abb. 7.4 Zwei Verteilungen der Wartezeiten zwischen *spikes*

K: Interessant ist, dass man mit diesen zwei Parametern ganz unterschiedliche Verteilungen mit verschiedener Form und verschiedenem Mittelwert erhalten kann, eben genau wie in deinen echten Daten.
Die Doktorandin ist mit dieser Information sehr glücklich und wähnt ihr Problem schon fast gelöst.
D: Und die Unterschiede in den Parametern zeigen mir dann die Unterschiede zwischen den Zellen, ja? Wenn ich die messen könnte, wäre das ja fantastisch! Kannst du mir vielleicht sagen, wie das geht?
Ihr Kollege erinnert sich an die Maximum-Likelihood-Methode und denkt, dass er ihr damit eigentlich weiterhelfen können müsste.
K: In der Statistik spricht man weniger von einer Messung als von einer Schätzung der Parameter, aber ja, die Frage, wie man das macht, ist genau richtig... Ich glaube, ich hab' da auch schon eine Idee. Letzte Woche haben wir in der Vorlesung eine Methode kennengelernt, die ziemlich allgemein einsetzbar ist – sie heißt Maximum Likelihood. Die müsste hier auch funktionieren.
Die Doktorandin ist begeistert.
D: Ach sehr cool, von Maximum Likelihood habe ich tatsächlich auch schon gehört! Sag mal, könntest du das vielleicht für mich ausrechnen und mir eine Formel schicken, mit der ich die Parameter schätzen kann? Das wäre wirklich toll...
Optimistisch verspricht Ihr Kollege:
K: Na klar! Das müsste schnell gehen, vielleicht kann ich dir die Ergebnisse schon morgen schicken.
Ihr Kommilitone ist voller Elan und glaubt, das Problem seiner neuen Bekannten schnell und leicht in den Griff zu bekommen. Ableiten und Nullsetzen sollte ja zu schaffen sein. Doch schon am nächsten Tag kommen ihm langsam Zweifel. Als er die Likelihood- und Log-Likelihood-Funktion ansieht, bemerkt er, dass die Likelihood-Funktion

$$L(x, (\alpha, \lambda)^t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x)$$

die unangenehme Gammafunktion $\Gamma(\alpha)$ enthält. Diese abzuleiten und nach α aufzulösen, ist ziemlich schwierig. Da er nicht weiß, wie er vorgehen soll, wendet er sich mit dieser Frage an die Betreuerin (**B**) seiner Bachelorarbeit.
K: Gestern Abend habe ich einer Medizinerin versprochen, Schätzer für die Parameter der Gammaverteilung auszurechnen, aber jetzt habe ich gemerkt, dass das mit Maximum Likelihood gar nicht so einfach ist. Kannst Du mir vielleicht helfen?
Seine Betreuerin weiß natürlich, wo das Problem liegt.
B: Da hast du ein mathematisch nicht so leicht zugängliches Beispiel erwischt. Hier sucht man im zweidimensionalen Parameterraum den Maximierer. Das Standardvorgehen, bei dem man die partiellen Ableitungen null setzt, funktioniert aber zumindest beim Formparameter α leider nicht mehr, sodass die Lösung keine geschlossene Form hat. Hier muss uns der Rechner weiterhelfen und die Likelihood numerisch maximieren.

Ihr Kollege ist ein wenig entmutigt und bereut sein vorschnelles Versprechen vom gestrigen Abend.
K: Oh. . . Da muss man ja richtig programmieren. Da habe ich meiner Medizinerkollegin dann wohl zu viel versprochen. Ich dachte, ich kann ihr einfach eine Formel für die beiden Parameterschätzer geben, die sie dann auf alle Datensätze anwenden kann.
B: Leider ist das für den MLE bei der Gammaverteilung nicht möglich. Das ist eigentlich nicht schlimm, da es geeignete Methoden zur mehrdimensionalen Maximierung gibt und sich diese zur Bestimmung von MLEs auch lohnen. Wegen der guten asymptotischen Eigenschaften hat man mit MLEs dann ein mächtiges statistisches Werkzeug an der Hand.
Sie hat aber noch einen anderen Vorschlag:
B: Für deine Anwendung gibt es aber auch eine Alternative, um die Parameter zu schätzen. Und die ist noch um einiges einfacher anwendbar. Sie heißt die Momentenmethode und liefert die Momentenschätzer.
Ihr Kollege schaltet sofort.
K: Moment mal, Momente. . . hat das etwas mit den Momenten der Verteilung zu tun?
B: Ja, genau! Die Momentenmethode ist ganz leicht und tut nichts anderes, als die Momente der Verteilung, also $\mathbb{E}[X^k]$, die ja Funktionen der Parameter sind, mit den empirischen Momenten $(1/n)\sum x_i^k$ zu identifizieren. Damit werden Gleichungssysteme aufgestellt, die man nach den Parametern auflösen kann. Die Gammaverteilung hat erstes Moment (also Erwartungswert) α/λ und Varianz α/λ^2. Statt des zweiten Moments nimmt man wegen des Varianzzerlegungssatzes oft einfach die Varianz, da man ja schon den Erwartungswert mit dem Mittelwert identifiziert. Wir denken also etwa wie folgt:

$$\begin{aligned} \frac{\alpha}{\lambda} &= \mathbb{E}_{(\alpha,\lambda)}[X_1] \approx \bar{x} \\ \frac{\alpha}{\lambda^2} &= \mathbb{V}ar_{(\alpha,\lambda)}[X_1] \approx \frac{1}{n}\sum (x_i - \bar{x})^2 =: \hat{\sigma}^2(x). \end{aligned}$$

Ihr Kollege ist begeistert.
K: Oh krass, und dann muss ich nur noch das einfache Gleichungssystem lösen? Moment mal, dann bekomme ich einfach $\bar{x}/\hat{\sigma}^2$ als Schätzung für λ und $\bar{x}^2/\hat{\sigma}^2$ als Schätzung für α, richtig?
B: Ganz genau, die Momentenschätzer sind $\hat{\alpha}(x) = \bar{x}^2/\hat{\sigma}^2(x)$ und $\hat{\lambda}(x) = \bar{x}/\hat{\sigma}^2(x)$, und schon sind wir fertig!
K: Das ist natürlich perfekt! Das sind schöne einfache Formeln, die ich super weitergeben kann. . .
Auf den zweiten Blick scheint ihm das dann aber doch ein bisschen zu einfach. Etwas skeptisch fragt er:
K: Aber bringt's das? Wie sieht es denn aus mit den Eigenschaften der Momentenschätzer?

B: Nun ja, zum einen kann man sich auf die Eigenschaften der empirischen Momente berufen, die ja Mittelwerte sind, sodass man zum Beispiel die Konsistenz direkt sehen kann. Auch die asymptotische Normalität kann man mithilfe der sogenannten Delta-Methode angehen. Dazu könntest Du zum Beispiel in Bishop et al. (1975); Ferguson (1996) oder van der Vaart (1998) nachlesen.
K: Und wenn die Eigenschaften so schön sind, warum benutzt man dann nicht immer die Momentenmethode?
B: Über MLEs kann man noch etwas bessere Eigenschaften zeigen. Vor allem, dass MLEs in gewisser Weise optimal sind, und zwar in dem Sinne, dass ein MLE zumindest asymptotisch unter allen unverzerrten Schätzern die kleinste Varianz besitzt. Die sogenannte Cramer-Rao-Schranke gibt eine untere Schranke für die asymptotische Varianz eines Schätzers, und diese ist gerade die asymptotische Varianz beim MLE (siehe etwa Ferguson 1996).
Jetzt weiß Ihr Kollege wie er helfen kann. Er ist froh, dass er der Doktorandin doch nicht zu viel versprochen hat, und macht sich gleich an die Arbeit.