Kapitel 2

Schätzer und Konfidenzintervalle

Bisher haben wir eine mathematische Theorie entwickelt, die es uns erlaubt, gewisse zufällige Phänomene zu modellieren. Zum Beispiel modellieren wir die Anzahl der radioaktiven Zerfälle in einer bestimmten Menge eines bestimmten Stoffs als Poisson-verteilt. Allerdings liefert unsere bisherige Theorie keinen Anhaltspunkt, wie der Parameter λ in der Poisson-Verteilung zu wählen ist, damit wir hiermit wirklich radioaktiven Zerfall beschreiben; genauer gesagt wird der Parameter λ wohl proportional zur Masse der beobachteten Probe sein und darüber hinaus vom radioaktiven Element, dessen Zerfall beschrieben werden soll, abhängen. In der Praxis wird man daher häufig den Parameter aus gewissen Beobachtungen "schätzen".

Hier ist ein weiteres Beispiel:

Beispiel 2.0.1. Betrachten wir wieder das Werfen einer Reißzwecke aus Beispiel 1.2.1(b). Wir hatten dort als Grundraum den Raum $\Omega = \{F, S\}$ gewählt, wobei F das Ergebnis "Die Reißzwecke landet auf der flachen Seite (mit der Spitze senkrecht nach oben)" und S das Ergebnis "Die Reißzwecke landet mit der Spitze schräg nach unten" bezeichnet. Wir haben ein mathematisches Modell für das Werfen einer Reißzwecke, sobald wir ein Wahrscheinlichkeitsmaß auf Ω spezifizieren. Wir müssen also $p = \mathbb{P}(\{F\})$ bestimmen (denn damit ist auch $\mathbb{P}(\{S\}) = 1 - p$ eindeutig festgelegt).

Um einen guten Wert für p zu wählen, kann man wie folgt vorgehen: Man wirft eine Reißzwecke "oft" (z.B. 1000 mal), zählt wie oft sie auf der flachen Seite landet (z.B. k mal) und nimmt dann $k/1000$ als Nährungswert für p.

Dies ist ein typisches Beispiel eines *Schätzproblemes*: Man möchte einen Näherungswert für einen Parameter einer Verteilung bestimmen. In diesem Kapitel werden wir solche Probleme genauer untersuchen. Insbesondere wollen wir allgemeine Prinzipien zur Konstruktion von Schätzern kennenlernen und die Güte einiger Schätzer beurteilen.

2.1 Zufallsstichproben und Schätzer

Die Beobachtungen, derer man sich in der Statistik bedient, werden mathematisch wie folgt modelliert.

Definition 2.1.1. Es sei P eine Wahrscheinlichkeitsverteilung auf einer abzählbaren Menge M. Es seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit Verteilung P. Dann nennt man $X_1, \ldots, X_n$ eine *Zufallsstichprobe* vom Umfang n zur Verteilung P. Die Werte $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$ heißen *Realisierung der Zufallsstichprobe.* Die Menge M^n aller potentiell möglichen Realisierungen einer Stichprobe nennt man *Stichprobenraum.*

Ein Wort zur Verwendung des Symbols P für eine Verteilung. Da jede Verteilung ein Wahrscheinlichkeitsmaß ist, verwenden einige Autoren das gleiche Symbol für Wahrscheinlichkeitsmaße und Verteilungen. Allerdings ist für eine Zufallsvariable X ihre Verteilung, die wir auch mit $\mathbb{P}_X$ bezeichnet hatten, in der Regel nicht gleich dem Wahrscheinlichkeitsmaß $\mathbb{P}$ des Wahrscheinlichkeitsraums, auf dem X definiert ist. Deshalb verwenden wir mit P ein ähnliches, nicht jedoch das gleiche Symbol.

Typischerweise ist die Verteilung P nicht (vollständig) bekannt und es sollen aus einer Realisierung der Zufallsstichprobe Rückschlüsse auf die Verteilung gezogen werden. Als ersten Schritt kann man einige Kenngrößen der Stichprobe berechnen.

Definition 2.1.2. Es sei $X_1, \dots, X_n$ eine Zufallsstichprobe zur Verteilung P. Dann heißt

$$\bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k$$

das Stichprobenmittel und

$$S^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X})^2$$

die Stichprobenvarianz. Weiter heißt $S := \sqrt{S^2}$ Stichprobenstandardabweichung. Die konkreten, auf einer Realisierung der Stichprobe basierenden, Werte werden häufig mit kleinen Buchstaben bezeichnet:

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \qquad \text{und} \qquad s^2 = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2 .$$

Beispiel 2.1.3. Der Besitzer eines Restaurants zählt jeden Tag die Anzahl seiner Gäste. Er erhält folgende Tabelle:

Tag	1	2	3	4
Anzahl der Gäste	44	37	49	52

In diesem Fall sind weder die Verteilung P (von der wir zumindest unterstellen, dass es sie gibt) noch die Zufallsstichprobe $X_1, \dots, X_4$ konkret bekannt. Die obige Tabelle gibt lediglich eine Realisierung $x_1, \dots, x_4$ der Stichprobe an. Für diese Realisierung haben wir

$$\bar{x} = \frac{1}{4}\big(44 + 37 + 49 + 52\big) = 45,5$$

und

$$s^2 = \frac{1}{3}\big(1,5^2 + 8,5^2 + 3,5^2 + 6,5^2\big) = \frac{516}{12} = 43,$$

was einer Standardabweichung von etwa $6,56$ entspricht.

Lemma 2.1.4. *Es sei $X_1, \dots, X_n$ eine Zufallsstichprobe zur Verteilung P. Wir nehmen an, dass $\mu = \mathbb{E}X_1$ und $\sigma^2 = \mathsf{Var}X_1$ existieren. Dann gilt*

(1) $\mathbb{E}\bar{X} = \mu$.

(2) $\mathsf{Var}\bar{X} = \sigma^2/n$

(3) $\mathbb{E}S^2 = \sigma^2$.

Beweis. (1) Es ist
$$\mathbb{E}\bar{X} = \frac{1}{n}\sum_{k=1}^n \mathbb{E}X_k = \frac{1}{n}\sum_{k=1}^n \mu = \mu.$$
(2) Aufgrund der Unabhängigkeit der $X_1, \ldots, X_n$ gilt
$$\mathsf{Var}\bar{X} = \frac{1}{n^2}\sum_{k=1}^n \mathsf{Var}X_k = \frac{\sigma^2}{n}.$$
(3) Wir haben
$$\begin{aligned}
\mathbb{E}S^2 &= \frac{1}{n-1}\sum_{k=1}^n \mathbb{E}(X_k - \bar{X})^2 \\
&= \frac{1}{n-1}\sum_{k=1}^n \mathbb{E}\big((X_k-\mu)^2 + 2(X_k-\mu)(\mu-\bar{X}) + (\mu-\bar{X})^2\big) \\
&= \frac{1}{n-1}\Big(n\sigma^2 + \frac{2}{n}\sum_{k=1}^n\sum_{j=1}^n \mathbb{E}(X_k-\mu)(\mu-X_j) + \sigma^2\Big) \\
&= \frac{1}{n-1}\Big(n\sigma^2 - \frac{2}{n}\sum_{k=1}^n \mathbb{E}(X_k-\mu)^2 + \sigma^2\Big) \\
&= \frac{1}{n-1}\big(n\sigma^2 - 2\sigma^2 + \sigma^2\big) \\
&= \sigma^2.
\end{aligned}$$
Hier haben wir in der dritten Gleichheit die Definition von $\mathsf{Var}X$ eingesetzt und verwendet, dass die Varianz von $\bar{X} = \sigma^2/n$ ist. In der vierten Gleichheit haben wir verwendet, dass $(X_k - \mu)$ und $(X_j - \mu)$ für $k \neq j$ unabhängig, also unkorreliert, sind. □

2.2 Parametrische Modelle

Wir disktieren nun sogenannte *parametrische Modelle*. Hierbei ist die Verteilung P, die wir näher untersuchen wollen, zwar unbekannt, es ist aber bekannt (oder wir nehmen es zumindest an), dass P zu einer bestimmten Familie $\{P_\theta : \theta \in \Theta\}$ von Verteilungen gehört, wobei die Parametermenge Θ für gewöhnlich eine geeignete Teilmenge von $\mathbb{R}^d$ ist.

Beispiel 2.2.1. (a) Wenn wir annehmen, dass eine gewisse Größe Poisson-verteilt ist, so kann man $P_\theta = Pois(\theta)$ für $\theta \in \Theta = (0, \infty) \subseteq \mathbb{R}$ betrachten.

(b) Im Falle des Restaurant Besitzers aus Beispiel 2.1.3 ist es schwierig eine geeignete Familie von Verteilungen zu finden. Angesichts fehlender Alternativen scheint die Annahme, dass die zugrunde liegende Verteilung eine Binomialverteilung ist, vertretbar: Es gibt eine gewisse Anzahl n potentieller Gäste, die unabhängig voneinander jeden Tag entscheiden, ob sie essen gehen oder nicht. Jeder Gast besucht an jedem Tag mit Wahrscheinlichkeit p das Restaurant. Wir können also die Familie $\{Bin(n,p) \mid (n,p) \in \Theta\}$ verwenden, wobei $\Theta = \mathbb{N} \times [0, 1]$ ist.

In dieser Situation will man nun entweder den Parameter θ oder die Zahl $g(\theta)$ für eine Funktion $g : \Theta \to \mathbb{R}^l$ schätzen. In Beispiel 2.2.1(b) könnte zum Beispiel der Erwartungswert

$g_1(n,p) = np$ oder auch $g_2(n,p) = p$ interessieren. In der allgemeinen Theorie betrachten wir nur die Schätzung von $g(\theta)$, denn die Schätzung von θ ist hierin als der Spezialfall, dass g die Identität ($g(\theta) = \theta$ für alle $\theta \in \Theta$) ist, enthalten.

Definition 2.2.2. Es sei eine Zufallsstichprobe $X_1, \dots, X_n$ zu einer Verteilung $P \in \{P_\theta : \theta \in \Theta\}$ gegeben, wobei $\Theta \subseteq \mathbb{R}^d$. Weiter sei $g : \Theta \to \mathbb{R}^l$ eine Funktion. Sei schließlich eine Abbildung $T : M^n \to \mathbb{R}^l$, also vom Stichprobenraum M^n in den Raum $\mathbb{R}^l$, der den Wertebereich von g umfasst, gegeben. Dann heißt $T(X_1, \dots, X_n)$ *Schätzer* für $g(\theta)$.

Häufig unterscheiden wir nicht zwischen der Funktion $T : M^n \to \mathbb{R}^m$ und dem Schätzer $T(X_1, \dots, X_n)$.

Wir geben einige Beispiele von Schätzern.

Beispiel 2.2.3. Im Beispiel 2.0.1 hatten wir das Problem betrachtet, den Parameter $p \in [0,1]$ einer Bernoulli-Verteilung, also einer Binomialverteilung $Bin(1,p)$ mit $n = 1$, zu schätzen. Dazu sei $X_1, \dots, X_n$ eine Zufallsstichprobe zur Verteilung $Bin(1,p)$ (wobei wir in Beispiel 2.0.1 den Wert $X_j = 1$ mit dem Elementarereignis P, und den Wert $X_j = 0$ mit dem Elementarereignis S identifizieren wollen).

Wir hatten in Beispiel 2.0.1 bereits den Schätzer T_1, gegeben durch $T_1(X_1, \dots, X_n) = \bar{X}$, betrachtet. Unsere Definition lässt aber auch andere Schätzer zu. Zum Beispiel:

$$\begin{aligned} T_2(X_1, \dots, X_n) &= \frac{2}{5} \\ T_3(X_1, \dots, X_n) &= \frac{1}{2}(X_1 + X_n) \end{aligned}$$

Das Beispiel zeigt, dass wir weitere Kriterien benötigen, um die Güte eines Schätzers zu beurteilen. Es scheint klar, dass T_1 der “beste” Schätzer für p ist. Der Schätzer T_2 berücksichtigt die Stichprobe überhaupt nicht. Der Schätzer T_3 berücksichtigt zwar die Stichprobe, jedoch nicht alle verfügbaren Informationen.

Um weitere Eigenschaften eines Schätzers zu definieren führen wir folgende Notation ein. Gegeben eine parametrisierte Familie $\{P_\theta : \theta \in \Theta\}$ und eine Zufallsstichprobe $X_1, \dots, X_n$ schreiben wir $\mathbb{P}_\theta(A)$ respektive $\mathbb{E}_\theta Y$ für die Wahrscheinlichkeit des Ereignisses A respektive den Erwartungswert der Zufallsvariablen Y unter der Annahme, dass die zugrunde liegende Verteilung der X_j gerade P_θ ist.

Definition 2.2.4. Es seien $X_1, \dots, X_n, X_{n+1}, ...$ unabhängige Zufallsvariablen mit $X_k \sim P \in \{P_\theta : \theta \in \Theta\}$. Weiter sei $T_n := T_n(X_1, \dots, X_n)$ eine Folge von Schätzern für $g(\theta) \in \mathbb{R}$.

(a) Der Schätzer T_n heißt *erwartungstreu*, falls $\mathbb{E}_\theta T_n = g(\theta)$ für alle $\theta \in \Theta$.

(b) Die Folge T_n heißt *asymptotisch erwartungstreu*, falls für alle $\theta \in \Theta$ stets $\mathbb{E}_\theta T_n \to g(\theta)$ für $n \to \infty$.

(c) Die Folge T_n heißt *schwach konsistent*, falls $T_n \to g(\theta)$ in θ- Wahrscheinlichkeit für $n \to \infty$ für alle $\theta \in \Theta$, d.h. für alle $\epsilon > 0$ gilt

$$\mathbb{P}_\theta\big(|T_n - g(\theta)| \geq \varepsilon\big) \to 0.$$

Beispiel 2.2.5. Wir betrachten wieder die Schätzer T_1, T_2 und T_3 für $\theta = p$ aus Beispiel 2.2.3. Dann sind T_1 und T_3 erwartungstreu, T_2 jedoch nicht. Es gilt nämlich

$$\mathbb{E}_p T_1(X_1, \dots, X_n) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}_p X_k = p$$

$$\mathbb{E}_p T_2(X_1,\dots,X_n) = \frac{2}{5} \neq p \qquad \text{für alle } p \in [0,1] \setminus \{\tfrac{2}{5}\}$$
$$\mathbb{E}_p T_3(X_1,\dots,X_n) = \frac{1}{2}\big(\mathbb{E}_p X_1 + \mathbb{E}_p X_k\big) = \frac{1}{2}(p+p) = p$$

Die Folge der Schätzer T_1 ist schwach konsistent, da

$$T_1(X_1,\dots,X_n) = \tfrac{1}{n}\sum_{i=1}^n X_i \overset{n\to\infty}{\longrightarrow} \mathbb{E}X_1 = p$$

in p-Wahrscheinlichkeit nach dem schwachen Gesetz der großen Zahlen. Für $p = \frac{1}{2}$ und $\varepsilon = 1/4$ erhalten wir

$$\mathbb{P}_{\frac{1}{2}}(|T_3 - \frac{1}{2}| \geq 1/4) = \mathbb{P}_{\frac{1}{2}}(X_1 = 0, X_n = 0 \text{ oder } X_1 = 1, X_n = 1) = \frac{1}{2} \not\to 0$$

sodass T_3 nicht schwach konsistent ist. Ebenso ist T_2 nicht schwach konsistent.

Wir stellen nun zwei allgemeine Verfahren zur Konstruktion von Schätzern vor.

Momentenmethode

Es sei $X_1,\dots,X_n$ eine Zufallsstichprobe zur Verteilung $P \in \{P_\theta : \theta \in \Theta\}$. Für $r \in \mathbb{N}$ ist $m_r(\theta) := \mathbb{E}_\theta X_1^r$ das *r-te Moment der Verteilung P_θ*. Wir können auch die empirischen r-ten Momente $\hat{m}_r := \frac{1}{n}\sum_{k=1}^n X_k^r$ betrachten. Es folgt aus dem schwachen Gesetz der großen Zahlen, dass $\hat{m}_r \to m_r(\theta)$ in θ-Wahrscheinlichkeit für $n \to \infty$.

Ist man nun an einem Parameter $\vartheta = g(\theta) \in \mathbb{R}^l$ interessiert, der sich als Funktion der ersten l Momente ausdrücken lässt, etwa $\vartheta = f(m_1(\theta),\dots,m_l(\theta))$, so liegt es nahe, als Schätzer für ϑ gerade

$$\hat{\vartheta} = f(\hat{m}_1,\dots,\hat{m}_l)$$

zu verwenden.

Beispiel 2.2.6. Wir wollen $(\mathbb{E}X_1, \mathsf{Var}X_1)$ mit Hilfe der Momentenmethode aus der Zufallsstichprobe $X_1,\dots,X_n$ schätzen.

Es gilt $(\mathbb{E}X, \mathsf{Var}X) = f(\mathbb{E}X, \mathbb{E}X^2)$ mit $f(x_1,x_2) = (x_1, x_2 - x_1^2)$. Somit liefert die Momentenmethode den Schätzer $(\bar{X}, \frac{1}{n}\sum_{k=1}^n X_k^2 - \bar{X}^2)$.

Es gilt

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^n X_k^2 - \Big(\frac{1}{n}\sum_{k=1}^n X_k\Big)^2 &= \frac{1}{n}\sum_{k=1}^n X_k^2 - \frac{1}{n}\sum_{k=1}^n X_k\bar{X} \\ &= \frac{1}{n}\sum_{k=1}^n X_k(X_k - \bar{X}) \\ &= \frac{1}{n}\sum_{k=1}^n (X_k - \bar{X})^2 = \frac{n-1}{n}S^2.\end{aligned}$$

Hier haben wir verwendet, dass $\sum_{k=1}^n \bar{X}(X_k - \bar{X}) = \bar{X}\cdot(\sum_{k=1}^n X_k - n\bar{X}) = 0$ ist. Beachte, dass dieser Schätzer *nicht* erwartungstreu ist, denn aus Lemma 2.1.4 folgt $\mathbb{E}\frac{n-1}{n}S_n = \frac{n-1}{n}\sigma^2$.

Wir geben noch einige Beispiele in denen der Parameter θ selbst geschätzt werden soll:

Beispiel 2.2.7. Ist $P_\lambda = Pois(\lambda)$, so ist $\lambda = m_1(\lambda)$. Somit liefert die Momentenmethode $\hat{\lambda} = \bar{X}$.

Beispiel 2.2.8. (Taxiproblem)

In einer großen Stadt gibt es N Taxis, die – gut zu erkennen – außen die Nummern $1, \ldots, N$ tragen. Es stellt sich die Frage, wie man N schätzen kann, wenn man die Nummern $x_1, \ldots, x_n$ von n vorbeifahrenden Taxis notiert.

Hierzu sei U_N die Gleichverteilung auf den Zahlen $1, 2, \ldots, N$, also $U_N(\{k\}) = N^{-1}$ für $k = 1, \ldots, N$. Ist $X \sim U_N$, so ist

$$\mathbb{E}X = \sum_{k=1}^{N} k\frac{1}{N} = \frac{1}{N}\frac{N(N+1)}{2} = \frac{N+1}{2},$$

also $N = 2m_1(N) - 1$. Somit liefert die Momentenmethode als Schätzer für den Parameter N gerade $\hat{N} = 2\bar{X} - 1$.

Hat man also die Nummern 242, 681, 44 und 512 notiert, so liefert die Momentenmethode $\hat{N} = 2 \cdot 369,75 - 1 = 738,5$. Beachte, dass es passieren kann, dass $\hat{N}$ kleiner als die größte beobachtete Zahl ist (etwa wenn man die Taxis mit den Nummern 22, 4 und 121 beobachtet; dann ist $\hat{N} = 97$).

Maximum-Likelihood-Methode

Die grundlegende Idee bei der Konstruktion von Schätzern mit der Maximum-Likelihood-Methode ist folgende:

Die beste Schätzung für einen Parameter θ ist diejenge,
bei der die beobachtete Stichprobe die höchste Wahrscheinlichkeit hat.

Formal geht man wie folgt vor:

Definition 2.2.9. Es sei $X_1, \ldots X_n$ eine Zufallsstichprobe zur Verteilung $P \in \{P_\theta : \theta \in \Theta\}$. Ferner sei $f(x, \theta) = P_\theta(\{x\})$ die zugehörige Zähldichte. Dann heißt $L : \mathbb{R}^n \times \Theta \to [0, 1]$, gegeben durch

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

die zugehörige *Likelihood-Funktion*. Es sei nun $\hat{\theta} : \mathbb{R}^n \to \Theta$ eine Funktion mit

$$L(x_1, \ldots, x_n; \theta) \leq L(x_1, \ldots, x_n; \hat{\theta}(x_1, \ldots, x_n))$$

für alle $x_1, \ldots, x_n$ und alle $\theta \in \Theta$. Dann heißt $\hat{\theta}(X_1, \ldots, X_n)$ *Maximum-Likelihood-Schätzer* für θ.

Bemerkung 2.2.10. (a) Weder existiert ein Maximum-Likelihood-Schätzer immer, noch ist er eindeutig bestimmt. In vielen Beispielen gibt es jedoch einen eindeutigen Maximum-Likelihood-Schätzer und in der Regel handelt es sich hierbei auch um einen "guten" Schätzer.

(b) Oft ist es einfacher statt der Likelihood-Funktion L die sogenannte log-Likelihood-Funktion $\log L$ zu maximieren. Wegen der Monotonie von Logarithmus und Exponentialfunktion nimmt diese an der selben Stelle wie L ihre Maxima an.

Um Maximum-Likelihood-Schätzer zu bestimmen, muss man also Funktionen maximieren. Die folgende Proposition hilft dabei:

Proposition 2.2.11. *Sei $f : (a,b) \to \mathbb{R}$ eine stetige Funktion mit $-\infty \le a < b \le \infty$. Falls es eine Stelle $x_0 \in (a,b)$ gibt mit*

$$f(x_0) \ge \max\big\{\lim_{x\to a} f(x), \lim_{x\to b} f(x)\big\},$$

dann hat f eine Maximalstelle $x_1 \in (a,b)$.

Die Voraussetzungen der Proposition sind insbesondere dann erfüllt, wenn

$$\lim_{x\to a} f(x) = \lim_{x\to b} f(x) = -\infty.$$

Sind die Voraussetzungen der Proposition erfüllt, ist f darüber hinaus differenzierbar und hat die Ableitung von f nur eine Nullstelle, dann muss diese Nullstelle also bereits die Maximalstelle von f sein.

Beispiel 2.2.12. Wir bestimmen zunächst einen Maximum-Likelihood-Schätzer für die Erfolgswahrscheinlichkeit in einem Bernoulli-Experiment. In diesem Fall ist der Parameter $\theta = p \in [0,1]$ und für $x \in \{0,1\}$ ist $f(x,\theta) = p^x(1-p)^{1-x}$. Es ergibt sich für die Likelihood-Funktion

$$L(x_1,\ldots,x_n;p) = \prod_{k=1}^n p^{x_k}(1-p)^{1-x_k}.$$

Hier ist es vorteilhaft, zur log-Likelihood-Funktion überzugehen. Wir erhalten

$$\log L(x_1,\ldots,x_n;p) = \sum_{k=1}^n x_k \log(p) + (1-x_k)\log(1-p) = n\bar{x}\log(p) + n(1-\bar{x})\log(1-p).$$

Um das Maximum zu bestimmen berechnen wir die Nullstellen der Ableitung (Beachte, falls $\bar{x} \in (0,1)$, dann hat $\log L$ nach Proposition 2.2.11 ein Maximum, denn die Grenzwerte bei 0 und 1 sind jeweils $-\infty$; ist $\bar{x} = 0$ bzw. $\bar{x} = 1$, so ist die Ableitung der log-Likelihood-Funktion überall negativ bzw. positiv, was impliziert, dass 0 bzw. 1 die Maximalstelle ist). Es ist

$$0 \stackrel{!}{=} \frac{d}{dp}\log L(x_1,\ldots,x_n;p) = \frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} \quad \Leftrightarrow \quad p = \bar{x}.$$

Demnach ist der Maximum-Likelihood-Schätzer $\hat{p} = \bar{x}$. In diesem Fall stimmt also der Maximum-Likelihood-Schätzer mit dem Schätzer, den man aus der Momentenmethode erhält, überein.

Beispiel 2.2.13. Wir bestimmen einen Maximum-Likelihood-Schätzer für die Verteilungen $\{Pois(\lambda) : \lambda > 0\}$. In diesem Falle ist

$$L(x_1,\ldots,x_n;\lambda) = e^{-n\lambda}\frac{\lambda^{\sum_{k=1}^n x_k}}{x_1! \cdot \ldots \cdot x_n!}$$

und daher $\log L = n\bar{x}\log\lambda - n\lambda - \log(x_1! \cdot \ldots \cdot x_n!)$. Durch Differenzieren und Nullsetzten erhält man dass als Kandidaten für das Maximum $\lambda = \bar{x}$. Beachten wir noch, dass, falls $\bar{x} > 0$,

$$\lim_{\lambda\to 0}\log L(x_1,\ldots,x_n;\lambda) = -\infty = \lim_{\lambda\to\infty}\log L(x_1,\ldots,x_n;\lambda)$$

so ergibt sich nach Proposition 2.2.11, dass es sich hierbei in der Tat um die Maximumstelle handeln muss (falls $\bar{x} = 0$, so sehen wir wie im Beispiel 2.2.12, dass $\lambda = 0$ die Maximalstelle ist). Also ist Maximum-Likelihood-Schätzer für λ durch $\hat{\lambda} = \bar{x}$ gegeben.

2.3 Konfidenzintervalle

Interessieren wir uns für einen Parameter einer Verteilung, so liefert ein Schätzer eine einzige Zahl, die den uns unbekannten Parameter approximieren soll. Allerdings ist dieser eine Wert für sich alleine genommen nicht aussagekräftig.

Beispiel 2.3.1. (Qualitätskontrolle)
Eine Maschine produziert Bauteile, die manchmal defekt sind. Der zuständige Werkstattsleiter will überprüfen, wie hoch der Anteil der defekten Teile am Output einer Maschine ist. Hierzu nimmt er an, dass die Bauteile unabhängig voneinander defekt oder funktionsfähig sind. Nun entnimmt er eine Stichprobe von 200 Bauteilen und testet diese. Dabei stellt er fest, dass 3 hiervon defekt sind. Er vermutet daher, dass $1,5\%$ der Bauteile defekt sind.

In obigem Beispiel war die Erfolgswahrscheinlichkeit p in einem Bernoulli-Experiment zu schätzen und wir haben hierzu den Schätzer $\bar{X}$ verwendet. Allerdings können wir nicht mit Sicherheit sagen, wie hoch der Anteil defekter Bauteile ist. So kann dieses Experiment niemals einen Schätzwert zwischen $1,5\%$ und 2% liefern, auch wenn der wahre Wert natürlich dazwischen liegen kann. Theoretisch wäre es auch denkbar, dass die Wahrscheinlichkeit, dass ein Bauteil defekt ist, 50% beträgt und nur zufällig so wenige defekte Bauteile in der Stichprobe sind. Die ist aber wenig realistisch, denn nach dem Gesetz der Großen Zahlen ist die Differenz $|\bar{X} - p|$ mit großer Wahrscheinlichkeit klein, wenn n hinreichend groß ist.

Deshalb wollen wir nun statt einer einzigen Zahl $\bar{X}$ ein Intervall $I = I(X_1, \ldots, X_n)$ angeben, in dem der wahre Parameter mit hoher Wahrscheinlichkeit liegt.

Definition 2.3.2. Es sei $X_1, \ldots, X_n$ eine Zufallsstichprobe zur Verteilung $P \in \{P_\theta, \theta \in \Theta\}$. Weiter sei $g : \Theta \to \mathbb{R}$ und $\alpha \in (0, 1)$. Ein *Konfidenzintervall zum Konfidenzniveau* α oder α*-Konfidenzintervall* für $g(\theta)$ ist ein zufälliges Intervall $I = [a(X_1, \ldots, X_n), b(X_1, \ldots, X_n)]$, wobei $a, b : M^n \to \mathbb{R}$ mit $a(x_1, \ldots, a_n) \leq b(x_1, \ldots, x_n)$ für alle $(x_1, \ldots, x_n) \in M^n$, falls

$$\mathbb{P}_\theta(g(\theta) \in I) \geq \alpha$$

für alle $\theta \in \Theta$ gilt.

Bemerkung 2.3.3. Beachte: Das Konfidenzintervall ist zufällig, nicht der Parameter θ.
Dass ein zufälliges Intervall ein α-Konfidenzintervall ist, bedeutet, dass $g(\theta)$ mit Wahrscheinlichkeit größer α im Intervall liegt, wenn die "wahre Verteilung" Parameter θ hat.

Man kann Konfidenzintervalle beispielsweise mit der Tschebyscheffschen Ungleichung bestimmen. Wenn wir etwa eine Zufallsstichprobe zur Verteilung P betrachten, die Erwartungswert μ und Varianz σ^2 hat, und verwenden wir $\bar{X}$ als Schätzer für μ, so ist nach Lemma 2.1.4

$$\mathbb{E}\bar{X} = \mu \quad \text{und} \quad \mathsf{Var}\bar{X} = \frac{\sigma^2}{n}.$$

Nach der Tschebyscheffschen Ungleichung gilt

$$\mathbb{P}(|\bar{X} - \mu| \geq \delta) \leq \frac{\sigma^2}{n\delta^2}.$$

Ist also σ^2 bekannt (oder zumindest beschänkt) und wählen wir $\delta = \sqrt{\frac{\sigma^2}{n(1-\alpha)}}$, so ist

$$[\bar{X} - \delta, \bar{X} + \delta]$$

ein Konfidenzintervall zum Konfidenzniveau α, denn

$$\mathbb{P}(\mu \in [\bar{X} - \delta, \bar{X} + \delta]) \geq \mathbb{P}(\mu \in (\bar{X} - \delta, \bar{X} + \delta)) \geq 1 - \frac{\sigma^2}{\frac{\sigma^2}{(1-\alpha)}} = \alpha.$$

Beispiel 2.3.4. In Beispiel 2.3.1 bestimmen wir ein 95%-Konfidenzintervall für die Bernoulliwahrscheinlichkeit p. Beachte, dass hier $\sigma^2 = p(1-p) = 1/4 - (p - 1/2)^2 \leq 1/4$. Weiter ist der Stichprobenumfang $n = 200$. Somit können wir

$$\delta = \sqrt{\frac{1}{4 \cdot 200 \cdot 0,05}} \approx 0,1581$$

wählen. Beachte, dass dies im Vergleich zum geschätzen Wert ($\bar{x} = 0,015$) relativ groß ist. Wir erhalten also das (relativ lange) Konfidenzintervall $[-0{,}143,\ 0{,}173]$ für $p = \mu$. Da $p \in [0,1]$, folgt

$$\mathbb{P}_p(\mu \in [\bar{X} - \delta, \bar{X} + \delta]) = \mathbb{P}_p(\mu \in [\max\{0, \bar{X} - \delta\}, \min\{\bar{X} + \delta, 1\}]).$$

Somit ist $[\max\{0, \bar{X} - \delta\}, \min\{\bar{X} + \delta, 1\}]$ ebenfalls α-Konfidenzintervall. Setzen wir die beobachteten Werte ein, so ergibt sich $[0,\ 0{,}173]$ als 95%-Konfidenzintervall.

Will man das Konfidenzintervall weiter verkleinern, so kann man z.B. den Stichprobenumfang erhöhen.

Der Grund, warum mit der Tschebyscheff Ungleichung relativ lange Konfidenzintervalle entstehen, liegt in der Allgemeinheit der Ungleichung. Wenn man spezielle Eigenschaften der Verteilung des Schätzers berücksichtigt, so erhält man kürzere Konfidenzintervalle. Wir kommen später darauf zurück.