

Maximum-Likelihood-Methode (ML-Methode)

- Weitere geläufige Schätzmethode: **Maximum-Likelihood-Methode**
- **Vor** Erläuterung der Methode: einleitendes Beispiel

Beispiel: ML-Methode durch Intuition (?)

Ein „fairer“ Würfel sei auf einer unbekanntem Anzahl $r \in \{0, 1, 2, 3, 4, 5, 6\}$ von Seiten rot lackiert, auf den übrigen Seiten andersfarbig.

Der Würfel wird 100-mal geworfen und es wird festgestellt, wie oft eine rote Seite (oben) zu sehen war.

- ▶ Angenommen, es war 34-mal eine rote Seite zu sehen; wie würden Sie die Anzahl der rot lackierten Seiten auf dem Würfel schätzen?
- ▶ Angenommen, es war 99-mal eine rote Seite zu sehen; wie würden Sie nun die Anzahl der rot lackierten Seiten auf dem Würfel schätzen?

Welche Überlegungen haben Sie insbesondere zu dem zweiten Schätzwert geführt?

Erläuterung Beispiel I

- Bei der Bearbeitung des obigen Beispiels wendet man (zumindest im 2. Fall) vermutlich intuitiv die Maximum-Likelihood-Methode an!
- Prinzipielle Idee der Maximum-Likelihood-Methode:

Wähle denjenigen der möglichen Parameter als Schätzung aus, bei dem die beobachtete Stichprobenrealisation am plausibelsten ist!
- Im Beispiel interessiert die (unbekannte) Anzahl der roten Seiten.
- Kenntnis der Anzahl der roten Seiten ist (Würfel ist „fair“!) gleichbedeutend mit der Kenntnis der Wahrscheinlichkeit, dass eine rote Seite oben liegt; offensichtlich ist diese Wahrscheinlichkeit nämlich $\frac{r}{6}$, wenn $r \in \{0, \dots, 6\}$ die Anzahl der roten Seiten bezeichnet.
- Interessierender Umweltausschnitt kann also durch die Zufallsvariable Y beschrieben werden, die den Wert 1 annimmt, falls bei einem Würfelwurf eine rote Seite oben liegt, 0 sonst.
- Y ist dann offensichtlich $B(1, p)$ -verteilt mit unbekanntem Parameter $p \in \{0, \frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, 1\}$, die 2. Grundannahme ist also erfüllt mit

$$W = \left\{ B(1, p) \mid p \in \left\{ 0, \frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, 1 \right\} \right\} .$$

Erläuterung Beispiel II

- 100-maliges Werfen des Würfels und jeweiliges Notieren einer 1, falls eine rote Seite oben liegt, einer 0 sonst, führt offensichtlich zu einer Realisation x_1, \dots, x_n einer einfachen Stichprobe X_1, \dots, X_n vom Umfang $n = 100$ zu Y , denn X_1, \dots, X_n sind als Resultat wiederholter Würfelwürfe offensichtlich unabhängig identisch verteilt wie Y .
- Wiederum (vgl. Taschengeldbeispiel) ist es aber nützlich, sich schon *vorher* Gedanken über die Verteilung der Anzahl der (insgesamt geworfenen) Würfe mit oberliegender roten Seite zu machen!
- Aus Veranstaltung „Deskriptive Statistik und Wahrscheinlichkeitsrechnung“ bekannt: Für die Zufallsvariable Z , die die Anzahl der roten Seiten bei 100-maligem Werfen beschreibt, also für

$$Z = \sum_{i=1}^{100} X_i = X_1 + \dots + X_{100} ,$$

gilt $Z \sim B(100, p)$, falls $Y \sim B(1, p)$.

- Ziel: Aus Stichprobe X_1, \dots, X_{100} bzw. der Realisation x_1, \dots, x_{100} (über die Stichprobenfunktion Z bzw. deren Realisation $z = x_1 + \dots + x_{100}$) auf unbekanntem Parameter p und damit die Anzahl der roten Seiten r schließen.

Erläuterung Beispiel III

- Im Beispiel: Umsetzung der ML-Methode besonders einfach, da Menge W der möglichen Verteilungen (aus Verteilungsannahme) **endlich**.
- „Plausibilität“ einer Stichprobenrealisation kann hier direkt anhand der Eintrittswahrscheinlichkeit der Realisation gemessen und für alle möglichen Parameter p bestimmt werden.
- Wahrscheinlichkeit (abhängig von p), dass Z Wert z annimmt:

$$P\{Z = z|p\} = \binom{100}{z} \cdot p^z \cdot (1 - p)^{100-z}$$

- Für die erste Realisation $z = 34$ von Z :

r	0	1	2	3	4	5	6
p	0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1
$P\{Z = 34 p\}$	0	$1.2 \cdot 10^{-5}$	$8.31 \cdot 10^{-2}$	$4.58 \cdot 10^{-4}$	$1.94 \cdot 10^{-11}$	$5.17 \cdot 10^{-28}$	0

- Für die zweite Realisation $z = 99$ von Z :

r	0	1	2	3	4	5	6
p	0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1
$P\{Z = 99 p\}$	0	$7.65 \cdot 10^{-76}$	$3.88 \cdot 10^{-46}$	$7.89 \cdot 10^{-29}$	$1.23 \cdot 10^{-16}$	$2.41 \cdot 10^{-7}$	0

Bemerkungen zum Beispiel

- Die angegebenen Wahrscheinlichkeiten für Z fassen jeweils mehrere mögliche Stichprobenrealisationen zusammen (da für den Wert von Z irrelevant ist, *welche* der Stichprobenzufallsvariablen X_i den Wert 0 bzw. 1 angenommen haben), für die ML-Schätzung ist aber eigentlich die Wahrscheinlichkeit einer einzelnen Stichprobenrealisation maßgeblich. Die Wahrscheinlichkeit einer einzelnen Stichprobenrealisation erhält man, indem der Faktor $\binom{100}{z}$ entfernt wird; dieser ist jedoch in jeder der beiden Tabellen konstant und beeinflusst daher die Bestimmung des Maximums nicht.
- Eher untypisch am Beispiel (aber umso geeigneter zur Erklärung der Methode!) ist die Tatsache, dass W eine endliche Menge von Verteilungen ist. In der Praxis wird man in der Regel unendlich viele Möglichkeiten für die Wahl des Parameters haben, z.B. bei Alternativverteilungen $p \in [0, 1]$. Dies ändert zwar *nichts* am Prinzip der Schätzung, wohl aber an den zur Bestimmung der „maximalen Plausibilität“ nötigen (mathematischen) Techniken.
- Dass die „Plausibilität“ hier genauer einer Wahrscheinlichkeit entspricht, hängt an der diskreten Verteilung von Y . Ist Y eine stetige Zufallsvariable, übernehmen Dichtefunktionswerte die Messung der „Plausibilität“.

Maximum-Likelihood-Methode (im Detail)

Schritte zur ML-Schätzung

Die Durchführung einer ML-Schätzung besteht aus folgenden Schritten:

- 1 Aufstellung der sog. **Likelihood-Funktion** $L(\theta)$, die *in Abhängigkeit des (unbekannten) Parametervektors* θ die Plausibilität der beobachteten Stichprobenrealisation misst.
- 2 Suche des (eines) Parameters bzw. Parametervektors $\hat{\theta}$, der den (zu der beobachteten Stichprobenrealisation) maximal möglichen Wert der Likelihoodfunktion liefert.

Es ist also *jeder* Parameter(vektor) $\hat{\theta}$ ein ML-Schätzer, für den gilt:

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$$

- Je nach Anwendungssituation unterscheidet sich die Vorgehensweise in beiden Schritten erheblich.
- Wir setzen bei der Durchführung von ML-Schätzungen **stets** voraus, dass eine **einfache (Zufalls-)Stichprobe** vorliegt!

1. Schritt: Aufstellen der Likelihoodfunktion

- „Plausibilität“ oder „Likelihood“ der Stichprobenrealisation wird gemessen
 - ▶ mit Hilfe der **Wahrscheinlichkeit**, die Stichprobenrealisation (x_1, \dots, x_n) zu erhalten, d.h. dem Wahrscheinlichkeitsfunktionswert

$$L(\theta) := p_{X_1, \dots, X_n}(x_1, \dots, x_n | \theta) ,$$

falls Y diskrete Zufallsvariable ist,

- ▶ mit Hilfe der **gemeinsamen Dichtefunktion** ausgewertet an der Stichprobenrealisation (x_1, \dots, x_n) ,

$$L(\theta) := f_{X_1, \dots, X_n}(x_1, \dots, x_n | \theta) ,$$

falls Y stetige Zufallsvariable ist.

- Bei Vorliegen einer einfachen Stichprobe lässt sich die Likelihoodfunktion für diskrete Zufallsvariablen Y **immer** darstellen als

$$L(\theta) = p_{X_1, \dots, X_n}(x_1, \dots, x_n | \theta)$$

$$\stackrel{X_i \text{ unabhängig}}{=} \prod_{i=1}^n p_{X_i}(x_i | \theta)$$

$$\stackrel{X_i \text{ verteilt wie } Y}{=} \prod_{i=1}^n p_Y(x_i | \theta) .$$

- Analog erhält man bei Vorliegen einer einfachen Stichprobe für stetige Zufallsvariablen Y **immer** die Darstellung

$$\begin{aligned}
 L(\theta) &= f_{X_1, \dots, X_n}(x_1, \dots, x_n | \theta) \\
 &\stackrel{X_i \text{ unabhängig}}{=} \prod_{i=1}^n f_{X_i}(x_i | \theta) \\
 &\stackrel{X_i \text{ verteilt wie } Y}{=} \prod_{i=1}^n f_Y(x_i | \theta) .
 \end{aligned}$$

für die Likelihoodfunktion.

- Ist der Parameterraum Θ endlich, kann im Prinzip $L(\theta)$ für alle $\theta \in \Theta$ berechnet werden und eines der θ als ML-Schätzwert $\hat{\theta}$ gewählt werden, für das $L(\theta)$ maximal war.
Für diese (einfache) Situation wird Schritt 2 nicht weiter konkretisiert.
- Ist der Parameterraum Θ ein Kontinuum (z.B. ein Intervall in \mathbb{R}^K), müssen für den 2. Schritt i.d.R. Maximierungsverfahren aus der Analysis angewendet werden.

2. Schritt: Maximieren der Likelihoodfunktion

(falls Θ ein Intervall in \mathbb{R}^K ist)

- Wichtige Eigenschaft des Maximierungsproblems aus Schritt 2:
Wichtig ist nicht der **Wert** des Maximums $L(\hat{\theta})$ der Likelihoodfunktion, sondern die **Stelle** $\hat{\theta}$, an der dieser Wert angenommen wird!
- Aus Gründen (zum Teil ganz erheblich) vereinfachter Berechnung:
 - ▶ Bilden der **logarithmierten** Likelihoodfunktion (Log-Likelihoodfunktion) $\ln L(\theta)$.
 - ▶ Maximieren der Log-Likelihoodfunktion $\ln L(\theta)$ **statt** Maximierung der Likelihoodfunktion.
- Diese Änderung des Verfahrens ändert nichts an den Ergebnissen, denn
 - ▶ $\ln : \mathbb{R}_{++} \rightarrow \mathbb{R}$ ist eine streng monoton wachsende Abbildung,
 - ▶ es genügt, die Likelihoodfunktion in den Bereichen zu untersuchen, in denen sie *positive* Werte annimmt, da nur dort das Maximum angenommen werden kann. Dort ist auch die log-Likelihoodfunktion definiert.

- Maximierung von $\ln L(\theta)$ kann *oft* (aber nicht immer!) auf die aus der Mathematik bekannte Art und Weise erfolgen:

- 1 Bilden der ersten Ableitung $\frac{\partial \ln L}{\partial \theta}$ der log-Likelihoodfunktion.

(Bei mehrdimensionalen Parametervektoren: Bilden der partiellen Ableitungen

$$\frac{\partial \ln L}{\partial \theta_1}, \dots, \frac{\partial \ln L}{\partial \theta_K}$$

der log-Likelihoodfunktion.)

- 2 Nullsetzen der ersten Ableitung, um „Kandidaten“ für Maximumstellen von $\ln L(\theta)$ zu finden:

$$\frac{\partial \ln L}{\partial \theta} \stackrel{!}{=} 0 \quad \rightsquigarrow \quad \hat{\theta}$$

(Bei mehrdimensionalen Parametervektoren: Lösen des Gleichungssystems

$$\frac{\partial \ln L}{\partial \theta_1} \stackrel{!}{=} 0, \quad \dots, \quad \frac{\partial \ln L}{\partial \theta_K} \stackrel{!}{=} 0$$

um „Kandidaten“ $\hat{\theta}$ für Maximumstellen von $\ln L(\theta)$ zu finden.)

- 3 Überprüfung anhand des Vorzeichens der 2. Ableitung $\frac{\partial^2 \ln L}{(\partial \theta)^2}$ (bzw. der Definitheit der Hessematrix), ob tatsächlich eine Maximumstelle vorliegt:

$$\frac{\partial^2 \ln L}{(\partial \theta)^2}(\hat{\theta}) \stackrel{?}{<} 0$$

- Auf die Überprüfung der 2. Ableitung bzw. der Hessematrix verzichten wir häufig, um nicht durch mathematische Schwierigkeiten von den statistischen abzulenken.
- Durch den Übergang von der Likelihoodfunktion zur log-Likelihoodfunktion erhält man gegenüber den Darstellungen aus Folie 39 und 40 im diskreten Fall nun

$$\ln L(\theta) = \ln \left(\prod_{i=1}^n p_Y(x_i|\theta) \right) = \sum_{i=1}^n \ln (p_Y(x_i|\theta))$$

und im stetigen Fall

$$\ln L(\theta) = \ln \left(\prod_{i=1}^n f_Y(x_i|\theta) \right) = \sum_{i=1}^n \ln (f_Y(x_i|\theta)) .$$

- Die wesentliche Vereinfachung beim Übergang zur log-Likelihoodfunktion ergibt sich meist dadurch, dass die Summen in den obigen Darstellungen deutlich leichter abzuleiten sind als die Produkte in den Darstellungen der Likelihoodfunktion auf Folie 39 und Folie 40.
- Falls „Standardverfahren“ keine Maximumsstelle liefert \rightsquigarrow „Gehirn einschalten“

Beispiel: ML-Schätzung für Exponentialverteilung

Erinnerung: $f_Y(y|\lambda) = \lambda e^{-\lambda y}$ für $y > 0$, $\lambda > 0$

- ① Aufstellen der Likelihoodfunktion (im Fall $x_i > 0$ für alle i):

$$L(\lambda) = \prod_{i=1}^n f_Y(x_i|\lambda) = \prod_{i=1}^n (\lambda e^{-\lambda x_i})$$

- ② Aufstellen der log-Likelihoodfunktion (im Fall $x_i > 0$ für alle i):

$$\ln L(\lambda) = \sum_{i=1}^n \ln(\lambda e^{-\lambda x_i}) = \sum_{i=1}^n (\ln \lambda + (-\lambda x_i)) = n \cdot \ln \lambda - \lambda \cdot \sum_{i=1}^n x_i$$

- ③ Ableiten und Nullsetzen der log-Likelihoodfunktion:

$$\frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i \stackrel{!}{=} 0$$

liefert

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}}$$

als ML-Schätzer (2. Ableitung $\frac{\partial^2 \ln L}{(\partial \lambda)^2}(\frac{1}{\bar{x}}) = -\frac{n}{\bar{x}^2} < 0$).

Bemerkungen

- Häufiger wird die Abhängigkeit der Likelihoodfunktion von der Stichprobenrealisation auch durch Schreibweisen der Art $L(\theta; x_1, \dots, x_n)$ oder $L(x_1, \dots, x_n | \theta)$ ausgedrückt.
- Vorsicht geboten, falls Bereich positiver Dichte bzw. der Träger der Verteilung von Y von Parametern abhängt!
 Im Beispiel: Bereich positiver Dichte \mathbb{R}_{++} *unabhängig* vom Verteilungsparameter λ , Maximierungsproblem unter Vernachlässigung des Falls „*mindestens ein x_i kleiner oder gleich 0*“ betrachtet, da dieser Fall **für keinen der möglichen Parameter** mit positiver Wahrscheinlichkeit eintritt. Dieses „Vernachlässigen“ ist nicht immer unschädlich!
- Bei diskreten Zufallsvariablen mit „wenig“ verschiedenen Ausprägungen oft Angabe der absoluten Häufigkeiten für die einzelnen Ausprägungen in der Stichprobe statt Angabe der Stichprobenrealisation x_1, \dots, x_n selbst.
 Beispiel: Bei Stichprobe vom Umfang 25 zu alternativverteilter Zufallsvariablen Y häufiger Angabe von „18 Erfolge in der Stichprobe der Länge 25“ als Angabe der Stichprobenrealisation

0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1 .

Beispiel: ML-Schätzung für Alternativverteilungen I

- Verteilungsannahme: $Y \sim B(1, p)$ für $p \in \Theta = [0, 1]$ mit

$$p_Y(y|p) = \left\{ \begin{array}{ll} p & \text{falls } y = 1 \\ 1 - p & \text{falls } y = 0 \end{array} \right\} = p^y \cdot (1 - p)^{1-y} \text{ für } y \in \{0, 1\} .$$

- 1 Aufstellen der Likelihoodfunktion:

$$L(p) = \prod_{i=1}^n p_Y(x_i|p) = \prod_{i=1}^n (p^{x_i} \cdot (1 - p)^{1-x_i}) = p^{\sum_{i=1}^n x_i} \cdot (1 - p)^{n - \sum_{i=1}^n x_i}$$

bzw. — wenn $n_1 := \sum_{i=1}^n x_i$ die Anzahl der „Einsen“ (Erfolge) in der Stichprobe angibt —

$$L(p) = p^{n_1} \cdot (1 - p)^{n - n_1}$$

- 2 Aufstellen der log-Likelihoodfunktion:

$$\ln L(p) = n_1 \ln(p) + (n - n_1) \ln(1 - p)$$

Beispiel: ML-Schätzung für Alternativverteilungen II

- ④ Ableiten und Nullsetzen der log-Likelihoodfunktion:

$$\begin{aligned} \frac{\partial \ln L}{\partial p} &= \frac{n_1}{p} - \frac{n - n_1}{1 - p} \stackrel{!}{=} 0 \\ \Leftrightarrow n_1 - n_1 p &= np - n_1 p \\ \Rightarrow \hat{p} &= \frac{n_1}{n} \end{aligned}$$

Die 2. Ableitung $\frac{\partial^2 \ln L}{(\partial p)^2} = -\frac{n_1}{p^2} - \frac{n - n_1}{(1 - p)^2}$ ist negativ für $0 < p < 1$, der Anteil der Erfolge in der Stichprobe $\hat{p} = n_1/n$ ist also der ML-Schätzer.

Bemerkung:

Die Bestimmung des ML-Schätzers \hat{p} ist so nur zulässig, wenn $0 < p < 1$ gilt (sonst Multiplikation einer Gleichung mit 0). Allerdings gilt

- ▶ für $p = 0$ offensichtlich stets $n_1 = 0$, damit ist $\hat{p} = \frac{n_1}{n} = 0$ für $p = 0$ ebenfalls der ML-Schätzer, denn für $n_1 = 0$ ist $L(p) = (1 - p)^n$ maximal für $p = 0$;
- ▶ für $p = 1$ offensichtlich stets $n_1 = n$, damit ist $\hat{p} = \frac{n}{n} = 1$ für $p = 1$ ebenfalls der ML-Schätzer, denn für $n_1 = n$ ist $L(p) = p^n$ maximal für $p = 1$.

$\Rightarrow \hat{p} = \frac{n_1}{n}$ ist ML-Schätzer für Verteilungsannahme $Y \sim B(1, p)$, $p \in [0, 1]$.

Beispiel: ML-Schätzung für Poissonverteilungen I

- Verteilungsannahme: $Y \sim \text{Pois}(\lambda)$ für $\lambda \in \Theta = \mathbb{R}_{++}$ mit

$$p_Y(k|\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

für $k \in \mathbb{N}_0$.

- Aufstellen der Likelihoodfunktion:

$$L(\lambda) = \prod_{i=1}^n p_Y(x_i|\lambda) = \prod_{i=1}^n \left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right)$$

(falls alle $x_i \in \mathbb{N}_0$)

- Aufstellen der log-Likelihoodfunktion:

$$\ln L(\lambda) = \sum_{i=1}^n (x_i \ln(\lambda) - \ln(x_i!) - \lambda) = \left(\sum_{i=1}^n x_i \right) \ln(\lambda) - \left(\sum_{i=1}^n \ln(x_i!) \right) - n\lambda$$

Beispiel: ML-Schätzung für Poissonverteilungen II

- 3 Ableiten und Nullsetzen der log-Likelihoodfunktion:

$$\begin{aligned}\frac{\partial \ln L}{\partial \lambda} &= \frac{\sum_{i=1}^n x_i}{\lambda} - n \stackrel{!}{=} 0 \\ \Rightarrow \hat{\lambda} &= \frac{\sum_{i=1}^n x_i}{n} = \bar{x}\end{aligned}$$

mit $\frac{\partial^2 \ln L}{(\partial \lambda)^2} = -\frac{\sum_{i=1}^n x_i}{\lambda^2} < 0$ für alle $\lambda > 0$, $\hat{\lambda} = \bar{x}$ ist also der ML-Schätzer für λ .

- Aus Wahrscheinlichkeitsrechnung bekannt: $Y \sim \text{Pois}(\lambda) \Rightarrow E(Y) = \lambda$, also ergibt sich (hier) auch für den Schätzer nach der Momentenmethode offensichtlich $\hat{\lambda} = \bar{X}$.
- Wird (ähnlich zur Anzahl n_1 der Erfolge in einer Stichprobe zu einer alternativverteilten Grundgesamtheit) statt der (expliziten) Stichprobenrealisation x_1, \dots, x_n eine „Häufigkeitsverteilung“ der in der Stichprobe aufgetretenen Werte angegeben, kann \bar{x} mit der aus der deskriptiven Statistik bekannten „Formel“ ausgerechnet werden.

Beispiel: ML-Schätzung bei diskreter Gleichverteilung

- Verteilungsannahme: für ein (unbekanntes) $M \in \mathbb{N}$ nimmt Y die Werte $\{1, \dots, M\}$ mit der gleichen Wahrscheinlichkeit von jeweils $1/M$ an, d.h.:

$$p_Y(k|M) = \begin{cases} \frac{1}{M} & \text{falls } k \in \{1, \dots, M\} \\ 0 & \text{falls } k \notin \{1, \dots, M\} \end{cases}$$

- Aufstellen der Likelihoodfunktion:

$$\begin{aligned} L(M) &= \prod_{i=1}^n p_Y(x_i|M) = \begin{cases} \frac{1}{M^n} & \text{falls } x_i \in \{1, \dots, M\} \text{ für alle } i \\ 0 & \text{falls } x_i \notin \{1, \dots, M\} \text{ für mindestens ein } i \end{cases} \\ &= \begin{cases} \frac{1}{M^n} & \text{falls } \max\{x_1, \dots, x_n\} \leq M \\ 0 & \text{falls } \max\{x_1, \dots, x_n\} > M \end{cases} \quad (\text{gegeben } x_i \in \mathbb{N} \text{ für alle } i) \end{aligned}$$

- Maximieren der Likelihoodfunktion:

Offensichtlich ist $L(M)$ für $\max\{x_1, \dots, x_n\} \leq M$ streng monoton fallend in M , M muss also **unter Einhaltung der Bedingung** $\max\{x_1, \dots, x_n\} \leq M$ möglichst klein gewählt werden. Damit erhält man den ML-Schätzer als $\hat{M} = \max\{x_1, \dots, x_n\}$.

Beurteilung von Schätzfunktionen

- *Bisher:* Zwei Methoden zur Konstruktion von Schätzfunktionen bekannt.

- *Problem:*

Wie kann Güte/Qualität dieser Methoden bzw. der resultierenden Schätzfunktionen beurteilt werden?

- *Lösung:*

Zu gegebener Schätzfunktion $\hat{\theta}$ für θ : Untersuchung des **zufälligen** Schätzfehlers $\hat{\theta} - \theta$ (bzw. dessen Verteilung)

- Naheliegende Forderung für „gute“ Schätzfunktionen:

Verteilung des Schätzfehler sollte möglichst „dicht“ um 0 konzentriert sein (d.h. Verteilung von $\hat{\theta}$ sollte möglichst „dicht“ um θ konzentriert sein)

- Aber:

- ▶ Was bedeutet das?
- ▶ Wie vergleicht man zwei Schätzfunktionen $\hat{\theta}$ und $\tilde{\theta}$? Wann ist Schätzfunktion $\hat{\theta}$ „besser“ als $\tilde{\theta}$ (und was bedeutet „besser“)?
- ▶ Was ist zu beachten, wenn Verteilung des Schätz**fehlers** noch vom zu schätzenden Parameter abhängt?

Bias, Erwartungstreue

- Eine offensichtlich gute Eigenschaft von Schätzfunktionen ist, wenn der zu schätzende (wahre) Parameter zumindest *im Mittel* getroffen wird, d.h. der *erwartete* Schätzfehler gleich Null ist:

Definition 3.4 (Bias, Erwartungstreue)

Seien W eine parametrische Verteilungsannahme mit Parameterraum Θ , $\hat{\theta}$ eine Schätzfunktion für θ . Dann heißt

- 1 der erwartete Schätzfehler

$$\text{Bias}(\hat{\theta}) := E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$$

die **Verzerrung** oder der **Bias** von $\hat{\theta}$,

- 2 die Schätzfunktion $\hat{\theta}$ **erwartungstreu für** θ oder auch **unverzerrt für** θ , falls $\text{Bias}(\hat{\theta}) = 0$ bzw. $E(\hat{\theta}) = \theta$ für alle $\theta \in \Theta$ gilt.
- 3 Ist allgemeiner $g : \Theta \rightarrow \mathbb{R}$ eine (messbare) Abbildung, so betrachtet man auch Schätzfunktionen $\widehat{g(\theta)}$ für $g(\theta)$ und nennt diese **erwartungstreu für** $g(\theta)$, wenn $E(\widehat{g(\theta)} - g(\theta)) = 0$ bzw. $E(\widehat{g(\theta)}) = g(\theta)$ für alle $\theta \in \Theta$ gilt.

Bemerkungen

- Obwohl Definition 3.4 auch für mehrdimensionale Parameterräume Θ geeignet ist („0“ entspricht dann ggf. dem Nullvektor), betrachten wir zur Vereinfachung im Folgenden meist nur noch **eindimensionale** Parameterräume $\Theta \subseteq \mathbb{R}$.
- Ist beispielsweise W als Verteilungsannahme für Y die Menge aller Alternativverteilungen $B(1, p)$ mit Parameter $p \in \Theta = [0, 1]$, so ist der ML-Schätzer $\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ bei Vorliegen einer Zufallsstichprobe X_1, \dots, X_n zu Y erwartungstreu für p , denn es gilt:

$$\begin{aligned}
 E(\hat{p}) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \stackrel{E \text{ linear}}{=} \frac{1}{n} \sum_{i=1}^n E(X_i) \\
 &\stackrel{F_{X_i} = F_Y}{=} \frac{1}{n} \sum_{i=1}^n E(Y) \\
 &= \frac{1}{n} \cdot n \cdot p = p \text{ für alle } p \in [0, 1]
 \end{aligned}$$

- Allgemeiner gilt, dass \bar{X} bei Vorliegen einer Zufallsstichprobe stets erwartungstreu für $E(Y)$ ist, denn es gilt analog zu oben:

$$\begin{aligned}
 E(\bar{X}) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \stackrel{E \text{ linear}}{=} \frac{1}{n} \sum_{i=1}^n E(X_i) \\
 &\stackrel{F_{X_i}=F_Y}{=} \frac{1}{n} \sum_{i=1}^n E(Y) \\
 &= \frac{1}{n} \cdot n \cdot E(Y) = E(Y)
 \end{aligned}$$

- Genauso ist klar, dass man für beliebiges k mit dem k -ten empirischen Moment $\overline{X^k}$ bei Vorliegen einer Zufallsstichprobe stets erwartungstreu Schätzer für das k -te theoretische Moment $E(Y^k)$ erhält, denn es gilt:

$$E(\overline{X^k}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i^k\right) = \frac{1}{n} \sum_{i=1}^n E(X_i^k) = \frac{1}{n} \sum_{i=1}^n E(Y^k) = E(Y^k)$$

- Der nach der Methode der Momente erhaltene Schätzer

$$\widehat{\sigma^2} = \overline{X^2} - \bar{X}^2 \stackrel{\text{Verschiebungssatz}}{=} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

für den Parameter σ^2 einer normalverteilten Zufallsvariable ist **nicht** erwartungstreu für σ^2 .

Bezeichnet $\sigma^2 := \text{Var}(Y)$ nämlich die (unbekannte) Varianz der Zufallsvariablen Y , so kann gezeigt werden, dass für $\widehat{\sigma^2}$ generell

$$E(\widehat{\sigma^2}) = \frac{n-1}{n} \sigma^2$$

gilt. Einen erwartungstreuen Schätzer für σ^2 erhält man folglich mit der sogenannten **Stichprobenvarianz**

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

denn es gilt offensichtlich

$$E(S^2) = E\left(\frac{n}{n-1} \widehat{\sigma^2}\right) = \frac{n}{n-1} E(\widehat{\sigma^2}) = \frac{n}{n-1} \cdot \frac{n-1}{n} \cdot \sigma^2 = \sigma^2.$$

Vergleich von Schätzfunktionen

- Beim Vergleich von Schätzfunktionen: **oft** Beschränkung auf erwartungstreue Schätzfunktionen
- In der Regel: viele erwartungstreue Schätzfunktionen denkbar.
- Für die Schätzung von $\mu := E(Y)$ beispielsweise alle *gewichteten* Mittel

$$\hat{\mu}_{w_1, \dots, w_n} := \sum_{i=1}^n w_i \cdot X_i$$

mit der Eigenschaft $\sum_{i=1}^n w_i = 1$ erwartungstreu für μ , denn es gilt dann offensichtlich

$$E(\hat{\mu}_{w_1, \dots, w_n}) = E\left(\sum_{i=1}^n w_i \cdot X_i\right) = \sum_{i=1}^n w_i E(X_i) = E(Y) \cdot \sum_{i=1}^n w_i = E(Y) = \mu.$$

- Problem: Welche Schätzfunktion ist „die beste“?
- Übliche Auswahl (bei Beschränkung auf erwartungstreue Schätzfunktionen!): Schätzfunktionen mit geringerer **Streuung (Varianz)** bevorzugen.