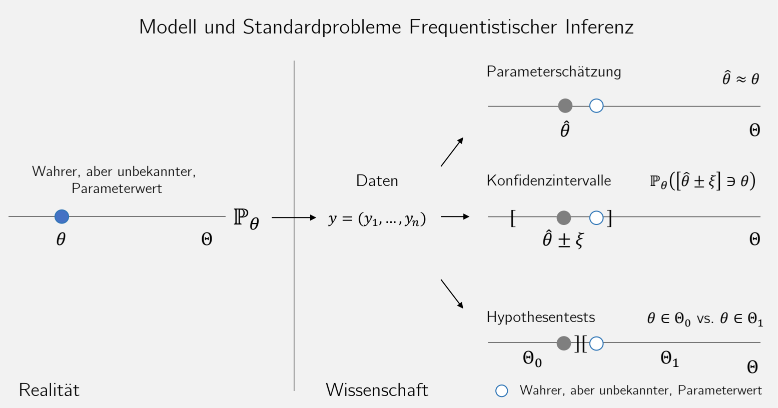
Wahrscheinlichkeitstheorie und Frequentistische Inferenz

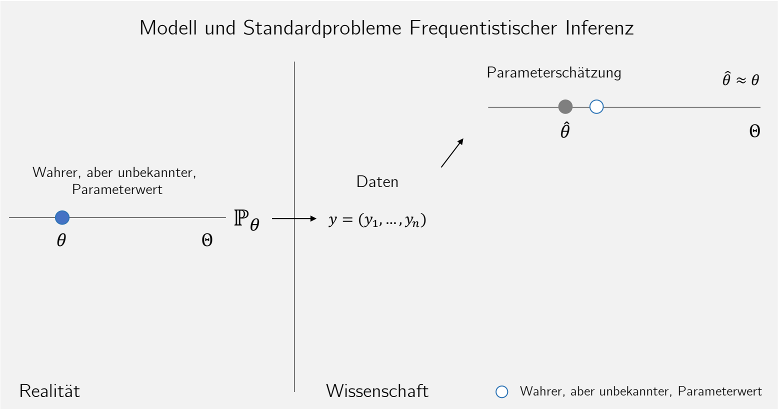
BSc Psychologie WiSe 2022/23

Prof. Dr. Dirk Ostwald

(10) Parameterschätzung



Modell und Standardprobleme Frequentistischer Inferenz
Wahrer, aber unbekannter, Parameterwert
θ
Θ
$\mathbb{P}_\theta$
Daten
$y = (y_1, \ldots, y_n)$
Parameterschätzung
$\hat{\theta} \approx \theta$
$\hat{\theta}$
Θ
Konfidenzintervalle
$\mathbb{P}_\theta([\hat{\theta} \pm \xi] \ni \theta)$
$\hat{\theta} \pm \xi$
Θ
Hypothesentests
$\theta \in \Theta_0$ vs. $\theta \in \Theta_1$
Θ_0
Θ_1
Θ
Realität
Wissenschaft
Wahrer, aber unbekannter, Parameterwert


Modell und Standardprobleme Frequentistischer Inferenz
Parameterschätzung
$\hat{\theta} \approx \theta$
$\hat{\theta}$
Θ
Wahrer, aber unbekannter, Parameterwert
Daten
$\mathbb{P}_\theta$
$y = (y_1, \ldots, y_n)$
θ
Θ
Realität
Wissenschaft
Wahrer, aber unbekannter, Parameterwert

Standardannahmen Frequentistischer Inferenz

$\mathcal{M}$ sei ein statistisches Modell mit Stichprobe $\upsilon_1, ..., \upsilon_n \sim p_\theta$. **Es wird angenommen, dass ein konkret vorliegender Datensatz $y = (y_1, ..., y_n) \in \mathbb{R}^n$ eine der möglichen Realisierungen von $\upsilon_1, ..., \upsilon_n \sim p_\theta$ ist.** Aus Frequentistischer Sicht kann man eine Studie unter gleichen Umständen unendlich oft wiederholen und zu jedem Datensatz Schätzer oder Statistiken auswerten, z.B. das Stichprobenmittel:

$$\text{Datensatz (1)} : y^{(1)} = \left(y_1^{(1)}, y_2^{(1)}, ..., y_n^{(1)}\right) \text{ mit } \bar{y}_n^{(1)} = \frac{1}{n}\sum\nolimits_{i=1}^n y_i^{(1)}$$

$$\text{Datensatz (2)} : y^{(2)} = \left(y_1^{(2)}, y_2^{(2)}, ..., y_n^{(2)}\right) \text{ mit } \bar{y}_n^{(2)} = \frac{1}{n}\sum\nolimits_{i=1}^n y_i^{(2)}$$

$$\text{Datensatz (3)} : y^{(3)} = \left(y_1^{(3)}, y_2^{(3)}, ..., y_n^{(3)}\right) \text{ mit } \bar{y}_n^{(3)} = \frac{1}{n}\sum\nolimits_{i=1}^n y_i^{(3)}$$

$$\text{Datensatz (4)} : y^{(4)} = \left(y_1^{(4)}, y_2^{(4)}, ..., y_n^{(4)}\right) \text{ mit } \bar{y}_n^{(4)} = \frac{1}{n}\sum\nolimits_{i=1}^n y_i^{(4)}$$

$$\text{Datensatz (5)} : y^{(5)} = ...$$

Um die Qualität statistischer Methoden zu beurteilen betrachtet die Frequentistische Statistik deshalb die Wahrscheinlichkeitsverteilungen von Schätzern und Statistiken unter Annahme von $\upsilon_1, ..., \upsilon_n \sim p_\theta$. Was zum Beispiel ist die Verteilung der $\bar{y}_n^{(1)}$, $\bar{y}_n^{(2)}$, $\bar{y}_n^{(3)}$, $\bar{y}_n^{(4)}$, ... also die Verteilung der Zufallsvariable $\bar{\upsilon}_n$?

Wenn eine statistische Methode im Sinne der Frequentistischen Standardannahmen "gut" ist, dann heißt das also, dass sie bei häufiger Anwendung "im Mittel gut" ist. Im Einzelfall, also im Normalfall nur eines vorliegenden Datensatzes, kann sie auch "schlecht" sein.

Grundbegriffe

Maximum-Likelihood Schätzer

Schätzereigenschaften bei endlichen Stichproben

Asymptotische Schätzereigenschaften

Eigenschaften von Maximum-Likelihood Schätzern

Selbstkontrollfragen

Grundbegriffe

Maximum-Likelihood Schätzer

Schätzereigenschaften bei endlichen Stichproben

Asymptotische Schätzereigenschaften

Eigenschaften von Maximum-Likelihood Schätzern

Selbstkontrollfragen

Definition (Parameterpunktschätzer)

$\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\})$ sei ein statistisches Modell, $(\Theta, \mathcal{S})$ sei ein Messraum und $\hat{\theta} : \mathcal{Y} \to \Theta$ sei eine Abbildung. Dann nennen wir $\hat{\theta}$ einen *Parameterpunktschätzer* für θ.

Bemerkungen

- Parameterpunktschätzer nennt man auch einfach *Parameterschätzer*.
- Parameterpunktschätzer sind Schätzer mit $\tau := \mathrm{id}_\Theta$
- Parameterschätzer nehmen Zahlwerte in Θ an.
- Notationstechnisch wird oft nicht zwischen $\hat{\theta}$ und $\hat{\theta}(y)$ unterschieden.

Prinzipien zur Gewinnung von Parameterschätzern

Die Definition eines Parameterschätzers macht keine Aussage darüber, wie man Parameterschätzer findet. Zur Gewinnung von Parameterschätzern in statistischen Modellen haben sich deshalb verschiedene Prinzipien etabliert. Populäre Prinzipien zur Gewinnung von Parameterschätzern sind

- Momentenmethode ($\approx$ est. 1890)
- Maximum-Likelihood Methode ($\approx$ est. 1920)
- M-, Z-, W-Schätzung ($\approx$ est. 1960)

Perse garantiert keine der obengenannten Methoden, dass die mit ihrer Hilfe generierten Parameterschätzer in einem wohldefinierten Sinn gute Schätzer sind.

Die Eigenschaften von durch die Maximum-Likelihood Methode generierten Schätzern sind generell wünschenswert. Wir betrachten also in der Folge nur die Maximum-Likelihood Methode genauer. Mithilfe der Maximum-Likelihood Methode generierte Parameterpunktschätzer nennen wir *Maximum-Likelihood (ML) Schätzer*.

Grundbegriffe

Maximum-Likelihood Schätzer

Schätzereigenschaften bei endlichen Stichproben

Asymptotische Schätzereigenschaften

Eigenschaften von Maximum-Likelihood Schätzern

Selbstkontrollfragen

Definition (Likelihood-Funktion und Log-Likelihood-Funktion)

$\mathcal{M}$ sei ein parametrisches statistisches Produktmodell mit WMF oder WDF p_θ, also $\upsilon_1, ..., \upsilon_n \sim p_\theta$. Dann ist die *Likelihood Funktion* definiert als

$$L_n : \Theta \to [0, \infty[, \theta \mapsto L_n(\theta) := \prod_{i=1}^n p_\theta(y_i) \tag{1}$$

und die *Log-Likelihood-Funktion* ist definiert als

$$\ell_n : \Theta \to \mathbb{R}, \theta \mapsto \ell_n(\theta) := \ln L_n(\theta). \tag{2}$$

Bemerkungen

- L_n ist eine Funktion des Parameters eines statistischen Modells.
- Werte von L_n sind die gemeinsamen Wahrscheinlichkeitsmassen bzw. -dichten von Datenwerten $y_1, ..., y_n$.
- Generell gibt es keinen Grund anzunehmen, dass L_n über Θ zu 1 integriert.
- Die Likelihood-Funktion ist also keine WMF oder WDF.
- Die Log-Likelihood-Funktion ist die logarithmierte Likelihood-Funktion.

Maximum-Likelihood Schätzer

Definition (Maximum-Likelihood Schätzer)

$\mathcal{M}$ sei ein parametrisches statistisches Produktmodell mit Parameter $\theta \in \Theta$. Ein *Maximum-Likelihood Schätzer* von θ ist definiert als

$$\hat{\theta}_n^{\text{ML}} : \mathcal{Y} \to \Theta, y \mapsto \hat{\theta}_n^{\text{ML}}(y) := \arg\max_{\theta \in \Theta} L_n(\theta) = \arg\max_{\theta \in \Theta} \ell_n(\theta). \tag{3}$$

Bemerkungen

- $L_n(\theta) := \prod_{i=1}^n p_\theta(y_i)$ hängt von $y := (y_1, ..., y_n)$ ab, also hängt auch $\hat{\theta}_n^{\text{ML}}(y)$ von y ab.
- Weil $\ln$ monoton steigend ist, entspricht eine Maximumstelle von ℓ_n einer Maximumstelle von L_n.
- Das Arbeiten mit der Log-Likelihood-Funktion ist oft einfacher als mit der Likelihood Funktion.
- Multiplikation von L_n mit einer positiven Konstante, die nicht von θ abhängt, verändert einen Maximum-Likelihood Schätzer nicht, konstante additive Terme in der Log-Likelihood können also vernachlässigt werden.
- Maximum-Likelihood Schätzung ist ein Optimierungsproblem

Vorgehen zur Gewinnung von Maximum-Likelihood Schätzern

(1) Formulierung der Log-Likelihood-Funktion.

(2) Auswertung der Ableitung der Log-Likelihood-Funktion und Nullsetzen.

(3) Auflösen nach potentiellen Maximumstellen.

Dabei nutzt man typischerweise

- Methoden der analytischen Optimierung in klassischen Beispielen
- Methoden der numerischen Optimierung im Anwendungskontext.

Maximum-Likelihood Schätzer

Beispiel (Bernoullimodell)

$\mathcal{M}$ sei das Bernoullimodell, also $\upsilon_1, ..., \upsilon_n \sim \text{Bern}(\mu)$.

(1) Formulierung der Log-Likelihood-Funktion

Es gilt

$$L_n :]0,1[\to]0,1[, \mu \mapsto L_n(\mu) := \prod_{i=1}^n \mu^{y_i}(1-\mu)^{1-y_i} = \mu^{\sum_{i=1}^n y_i}(1-\mu)^{n-\sum_{i=1}^n y_i}. \tag{4}$$

Logarithmieren ergibt

$$\ell_n :]0,1[\to \mathbb{R}, \mu \mapsto \ell_n(\mu) = \ln \mu \sum_{i=1}^n y_i + \ln(1-\mu)\left(n - \sum_{i=1}^n y_i\right). \tag{5}$$

Beispiel (Bernoullimodell)

(2) Auswertung der Ableitung der Log-Likelihood-Funktion und Nullsetzen

Es gilt

$$
\begin{aligned}
\frac{d}{d\mu}\ell_n(\mu) &= \frac{d}{d\mu}\left(\ln\mu\sum_{i=1}^n y_i + \ln(1-\mu)\left(n-\sum_{i=1}^n y_i\right)\right) \\
&= \frac{d}{d\mu}\ln\mu\sum_{i=1}^n y_i + \frac{d}{d\mu}\ln(1-\mu)\left(n-\sum_{i=1}^n y_i\right) \\
&= \frac{1}{\mu}\sum_{i=1}^n y_i - \frac{1}{1-\mu}\left(n-\sum_{i=1}^n y_i\right).
\end{aligned} \tag{6}
$$

Die sogenannte *Maximum-Likelihood Gleichung* ergibt sich in diesem Beispiel also zu

$$
\frac{1}{\hat{\mu}_n^{\mathsf{ML}}}\sum_{i=1}^n y_i - \frac{1}{1-\hat{\mu}_n^{\mathsf{ML}}}\left(n-\sum_{i=1}^n y_i\right) = 0. \tag{7}
$$

Beispiel (Bernoullimodell)

(3) Auflösen nach potentiellen Maximumstellen

Es gilt

$$
\begin{aligned}
\frac{1}{\hat{\mu}_n^{\mathsf{ML}}}\sum_{i=1}^n y_i - \frac{1}{1-\hat{\mu}_n^{\mathsf{ML}}}\left(n - \sum_{i=1}^n y_i\right) &= 0 \\
\Leftrightarrow \hat{\mu}_n^{\mathsf{ML}}(1-\hat{\mu}_n^{\mathsf{ML}})\left(\frac{1}{\hat{\mu}_n^{\mathsf{ML}}}\sum_{i=1}^n y_i - \frac{1}{1-\hat{\mu}_n^{\mathsf{ML}}}\left(n - \sum_{i=1}^n y_i\right)\right) &= 0 \\
\Leftrightarrow \sum_{i=1}^n y_i - \hat{\mu}_n^{\mathsf{ML}}\sum_{i=1}^n y_i - n\hat{\mu}_n^{\mathsf{ML}} + \hat{\mu}_n^{\mathsf{ML}}\sum_{i=1}^n y_i &= 0 \\
\Leftrightarrow n\hat{\mu}_n^{\mathsf{ML}} &= \sum_{i=1}^n y_i \\
\Leftrightarrow \hat{\mu}_n^{\mathsf{ML}} &= \frac{1}{n}\sum_{i=1}^n y_i .
\end{aligned}
\tag{8}
$$

Beispiel (Bernoullimodell)

$\hat{\mu}_n^{\mathsf{ML}} = \frac{1}{n}\sum_{i=1}^n y_i$ ist also ein potentieller Maximum-Likelihood Schätzer von μ. Dies kann durch Betrachten der zweiten Ableitung von ℓ_n verifiziert werden, worauf hier verzichtet werden soll.

$$\hat{\mu}_n^{\mathsf{ML}} : \{0,1\}^n \to [0,1], y \mapsto \hat{\mu}_n^{\mathsf{ML}}(y) := \frac{1}{n}\sum_{i=1}^n y_i \tag{9}$$

ist also ein Maximum-Likelihood Schätzer von μ im Bernoullimodell.

Beispiel (Normalverteilungsmodell)

$\mathcal{M}$ sei das Normalverteilungsmodell, also $\upsilon_1, ..., \upsilon_n \sim N\left(\mu, \sigma^2\right)$

(1) Formulierung der Log-Likelihood-Funktion

Es gilt

$$
\begin{aligned}
L_n : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}_{>0}, (\mu, \sigma^2) \mapsto L_n(\mu, \sigma^2) &:= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) \\
&= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right).
\end{aligned} \tag{10}
$$

Logarithmieren ergibt

$$
\ell_n : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}, (\mu, \sigma^2) \mapsto \ell_n(\mu, \sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2. \tag{11}
$$

Beispiel (Normalverteilungsmodell)

(2) Auswertung der Ableitung der Log-Likelihood-Funktion und Nullsetzen

Es ergibt sich

$$\frac{d}{d\mu}\ell_n(\mu) = -\frac{d}{d\mu}\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2 = -\frac{1}{2\sigma^2}\sum_{i=1}^n \frac{d}{d\mu}(y_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (y_i-\mu). \tag{12}$$

Weiterhin ergibt sich

$$\frac{d}{d\sigma^2}\ell_n(\sigma^2) = -\frac{n}{2}\frac{d}{d\sigma^2}\ln\sigma^2 - \frac{d}{d\sigma^2}\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (y_i-\mu)^2. \tag{13}$$

Die Maximum-Likelihood Gleichungen haben also die Form

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{\mu}_n^{\mathsf{ML}}) &= 0 \\ -\frac{n}{2\hat{\sigma}_n^{2^{\mathsf{ML}}}} + \frac{1}{2\hat{\sigma}_n^{4^{\mathsf{ML}}}}\sum_{i=1}^n (y_i-\mu)^2 &= 0 \end{aligned} \tag{14}$$

Beispiel (Normalverteilungsmodell)

(3) Auflösen nach potentiellen Maximumstellen

Es ergibt sich

$$\sum_{i=1}^{n}(y_i - \hat{\mu}_n^{\mathsf{ML}}) = 0 \Leftrightarrow \sum_{i=1}^{n} y_i = n\hat{\mu}_n^{\mathsf{ML}} \Leftrightarrow \hat{\mu}_n^{\mathsf{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{15}$$

Also ist $\hat{\mu}_n^{\mathsf{ML}} = n^{-1}\sum_{i=1}^{n} y_i$ ein potentieller Maximum-Likelihood Schätzer von μ. Einsetzen ergibt dann

$$\begin{aligned}
-\frac{n}{2\hat{\sigma}_n^{2\,\mathsf{ML}}} + \frac{1}{2\hat{\sigma}_n^{4\,\mathsf{ML}}}\sum_{i=1}^{n}(y_i - \hat{\mu}_n^{\mathsf{ML}})^2 &= 0 \\
\Leftrightarrow -n\hat{\sigma}_n^{2\,\mathsf{ML}} + \sum_{i=1}^{n}(y_i - \hat{\mu}_n^{\mathsf{ML}})^2 &= 0 \\
\Leftrightarrow \hat{\sigma}_n^{2\,\mathsf{ML}} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_n^{\mathsf{ML}})^2.
\end{aligned} \tag{16}$$

$\hat{\sigma}_n^{2\,\mathsf{ML}} = n^{-1}\sum_{i=1}^{n}(y_i - \hat{\mu}_n^{\mathsf{ML}})^2$ ist also ein potentieller Maximum-Likelihood Schätzer von σ^2.

Maximum-Likelihood Schätzer

Beispiel (Normalverteilungsmodell)

Beide potentiellen Maximum-Likelihood Schätzer können durch Betrachten der zweiten Ableitung von ℓ_n verifiziert werden, worauf hier verzichtet werden soll.

Also sind

$$\hat{\mu}_n^{\text{ML}} : \mathbb{R}^n \to \mathbb{R}, y \mapsto \hat{\mu}_n^{\text{ML}}(y) := \frac{1}{n}\sum_{i=1}^n y_i \tag{17}$$

und

$$\hat{\sigma}_n^{2^{\text{ML}}} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}, y \mapsto \hat{\sigma}_n^{2^{\text{ML}}}(y) := \frac{1}{n}\sum_{i=1}^n \left(y_i - \hat{\mu}_n^{\text{ML}}\right)^2. \tag{18}$$

die Maximum-Likelihood Schätzer von μ und σ^2 im Normalverteilungsmodell. $\hat{\mu}_n^{\text{ML}}$ ist identisch mit dem Stichprobenmittel $\bar{\upsilon}_n$, $\hat{\sigma}_n^{2^{\text{ML}}}$ ist dagegen nicht identisch mit der Stichprobenvarianz S_n^2.

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

Pre-BDI

BDI-II Fragebogen

Name | Alter | Geschlecht m / w | Datum

Anleitung: Dieser Fragebogen enthält 21 Gruppen von Aussagen. Bitte lesen Sie jede dieser Gruppen von Aussagen sorgfältig durch und suchen Sie sich dann in jeder Gruppe **eine Aussage** heraus, die am besten beschreibt, wie Sie sich **in den letzten zwei Wochen, einschließlich heute, gefühlt haben.** Kreuzen Sie die Zahl neben der Aussage an, die Sie sich herausgesucht haben (0, 1, 2 oder 3). Falls in einer Gruppe mehrere Aussagen gleichermaßen auf Sie zutreffen, kreuzen Sie die Aussage mit der höheren Zahl an. Achten Sie bitte darauf, dass Sie in jeder Gruppe nicht mehr als eine Aussage ankreuzen, das gilt auch für Gruppe 16 (Veränderungen der Schlafgewohnheiten) oder Gruppe 18 (Veränderungen des Appetits).

1.) Traurigkeit
- 0 Ich bin nicht traurig.
- 1 Ich bin oft traurig.
- 2 Ich bin ständig traurig.
- 3 Ich bin so traurig oder unglücklich, dass ich es nicht aushalte.

2.) Pessimismus
- 0 Ich sehe nicht mutlos in die Zukunft.
- 1 Ich sehe mutloser in die Zukunft als sonst.
- 2 Ich bin mutlos und erwarte nicht, dass meine Situation besser wird.
- 3 Ich glaube, dass meine Zukunft hoffnungslos ist und nur noch schlechter wird.

3.) Versagensgefühle
- 0 Ich fühle mich nicht als Versager.
- 1 Ich habe häufiger Versagensgefühle.
- 2 Wenn ich zurückblicke, sehe ich eine Menge Fehlschläge.
- 3 Ich habe das Gefühl, als Mensch ein völliger Versager zu sein.

4.) Verlust von Freude
- 0 Ich kann die Dinge genauso gut genießen wie früher.
- 1 Ich kann die Dinge nicht mehr so genießen wie früher.
- 2 Dinge, die mir früher Freude gemacht haben, kann ich kaum mehr genießen.
- 3 Dinge, die mir früher Freude gemacht haben, kann ich überhaupt nicht mehr genießen.

5.) Schuldgefühle
- 0 Ich habe keine besonderen Schuldgefühle.
- 1 Ich habe oft Schuldgefühle wegen Dingen, die ich getan habe oder hätte tun sollen.
- 2 Ich habe die meiste Zeit Schuldgefühle.
- 3 Ich habe ständig Schuldgefühle.

6.) Bestrafungsgefühle
- 0 Ich habe nicht das Gefühl, für etwas bestraft zu sein.
- 1 Ich habe das Gefühl, vielleicht bestraft zu werden.
- 2 Ich erwarte, bestraft zu werden.
- 3 Ich habe das Gefühl, bestraft zu sein.

7.) Selbstablehnung
- 0 Ich halte von mir genauso viel wie immer.
- 1 Ich habe Vertrauen in mich verloren.
- 2 Ich bin von mir enttäuscht.
- 3 Ich lehne mich völlig ab.

8.) Selbstvorwürfe
- 0 Ich kritisiere oder tadle mich nicht mehr als sonst.
- 1 Ich bin mir gegenüber kritischer als sonst.
- 2 Ich kritisiere mich für all meine Mängel.
- 3 Ich gebe mir die Schuld für alles Schlimme, was passiert.

9.) Selbstmordgedanken
- 0 Ich denke nicht daran, mir etwas anzutun.
- 1 Ich denke manchmal an Selbstmord, aber ich würde es nicht tun.
- 2 Ich möchte mich am liebsten umbringen.
- 3 Ich würde mich umbringen, wenn ich die Gelegenheit dazu hätte.

10.) Weinen
- 0 Ich weine nicht öfter als früher.
- 1 Ich weine jetzt mehr als früher.
- 2 Ich weine beim geringsten Anlass.
- 3 Ich möchte gern weinen, aber ich kann nicht.

PEARSON PsychCorp

Summe Seite 1:

Bitte wenden

Post-BDI

⇒ Pre-Post BDI Score Reduktion

Maximum-Likelihood Schätzer

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

```
fname = file.path(getwd(), "10_Parameterschätzung.csv")
D     = read.table(fname, sep = ",", header = T)
```

i	BDI.Reduktion
1	-1
2	3
3	-2
4	9
5	3
6	-2
7	4
8	5
9	5
10	1
11	9
12	4

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

Für die Pre-Post BDI Score Reduktion υ_i der $i$ten von n Patient:innen legen wir das Modell

$$\upsilon_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, ..., n \tag{19}$$

zugrunde. Dabei wird die Pre-Post BDI Reduktion υ_i der $i$ten Patient:in also mithilfe einer über die Gruppe von Patient:innen identischen Pre-Post BDI Score Reduktion $\mu \in \mathbb{R}$ und einer Patient:innen-spezifischen normalverteilten Pre-Post BDI Score Reduktionsabweichung ε_i erklärt

Wie gezeigt ist dieses Modell äquivalent zum Normalverteilungsmodell

$$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2). \tag{20}$$

Die Standardprobleme der Frequentistischen Inferenz führen in diesem Szenario auf folgende Fragen:

(1) Was sind sinnvolle Tipps für die wahren, aber unbekannten, Parameterwerte μ und σ^2?
(2) Wie hoch ist im frequentistischen Sinn die mit diesen Tipps assoziierte Unsicherheit?
(3) Entscheiden wir uns sinnvollerweise für die Hypothese, dass gilt $\mu \neq 0$?

Maximum-Likelihood Schätzer

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

```
# Einlesen und Auswahl der Daten
fname = file.path(getwd(), "10_Parameterschätzung.csv")
D     = read.table(fname, sep = ",", header = T)
y     = D$BDI.Reduktion

# ML Schätzung des Erwartungswertparameters
mu_hat   = mean(y)                             # mean(y) berechnet das Stichprobenmittel
print(mu_hat)                                  # Ausgabe
```

```
> [1] 3.17
```

```
# ML Schätzung des Varianzparameters
n          = length(y)                         # Anzahl der Datenpunkte
sigsqr_hat  = ((n-1)/n)*var(y)                 # var(y) berechnet die Stichprobenvarianz
print(sigsqr_hat)                              # Ausgabe
```

```
> [1] 12.6
```

Basierend auf dem Prinzip der Maximum-Likelihood Schätzung sind also

$$\hat{\mu}_{12}^{\text{ML}} = 3.17, \text{ und } \hat{\sigma}_{12}^{2^{\text{ML}}} = 12.6 \qquad (21)$$

sinnvolle Tipps für μ und σ^2 basierend auf den vorliegenden 12 Datenpunkten.

Grundbegriffe

Maximum-Likelihood Schätzer

Schätzereigenschaften bei endlichen Stichproben

Asymptotische Schätzereigenschaften

Eigenschaften von Maximum-Likelihood Schätzern

Selbstkontrollfragen

Vorbemerkungen zu Frequentistischen Schätzereigenschaften

Wir gehen von einem statistischem parametrischem Produktmodell $\mathcal{M} := \{\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\}\}$ mit n-dimensionalen Stichprobenraum (z.B. $\mathcal{Y} := \mathbb{R}^n$), d-dimensionalen Parameteraum $\Theta \subset \mathbb{R}^d$ und gegebener WMF oder WDF p_θ für alle $\theta \in \Theta$ aus. $\upsilon := \upsilon_1, ..., \upsilon_n \sim p_\theta$ bezeichnet die zu $\mathcal{M}$ gehörende Stichprobe unabhängig und identisch verteilter Zufallsvariablen, es gilt also $\upsilon_1 \sim p_\theta$ und $\upsilon_i \sim p_\theta$ für alle $i = 1, ..., n$.

Für einen Messraum (Σ, S) sei $\hat{\tau} : \mathcal{Y} \to \Sigma$ ein Schätzer von $\tau : \Theta \to \Sigma$. Wir betrachten Erwartungswerts-Varianz-, und Standardabweichungsschätzer, also Schätzer für

$$\tau : \Theta \to \Sigma,\ \theta \mapsto \tau(\theta) \text{ mit } \tau(\theta) := \mathbb{E}_\theta(\upsilon_1), \tau(\theta) := \mathbb{V}_\theta(\upsilon_1), \text{ und } \tau(\theta) := \mathbb{S}_\theta(\upsilon_1) \tag{22}$$

respektive, sowie Parameterschätzer, also Schätzer für

$$\tau : \Theta \to \Sigma, \tau(\theta) := \theta. \tag{23}$$

In der Folge führen wir *Frequentistische Schätzereigenschaften* ein. Frequentistische Schätzereigenschaften betrachten die Verteilung der Schätzwerte $\hat{\tau}(y_1, ..., y_n)$ in Abhängigkeit von der Verteilung der Datennwerte $\upsilon := \upsilon_1, ..., \upsilon_n$. Weil die Stichprobenwerte zufällig sind, sind auch die Schätzwerte zufällig; ein Schätzer $\hat{\tau}$ ist also wie oben gesehen eine Zufallsvariable. Wir unterscheiden zwischen *Schätzereigenschaften bei endlichen Stichproben*, d.h. Eigenschaften von $\hat{\tau}_n$ für ein fixes $n \in \mathbb{N}$ (z.B. $n = 12$) und *Asymptotischen Schätzereigenschaften*, d.h. Eigenschaften von $\hat{\tau}_n$ für unendlich groß werdende Stichproben mit $n \to \infty$.

Schätzereigenschaften bei endlichen Stichproben

Wir betrachten in der Folge zwei Aspekte von Schätzereigenschaften bei endlichen Stichproben:

(1) Erwartungstreue
(2) Varianz und Standardfehler

Intuitiv haben diese folgende Bedeutungen: Ein Schätzer $\hat{\tau}_n$ heißt *erwartungstreu*, wenn sein Erwartungswert dem wahren, aber unbekannten, Wert $\tau(\theta)$ für alle $\theta \in \Theta$ gleicht. Die *Varianz* eines Schätzers $\hat{\tau}_n$ ist die Varianz der Zufallsvariable $\hat{\tau}_n(\upsilon)$. Der *Standardfehler* eines Schätzers $\hat{\tau}_n$ ist die Standardabweichung der Zufallsvariable $\hat{\tau}_n(\upsilon)$.

Für folgende Schätzereigenschaften bei endlichen Stichproben verweisen wir auf den Appendix:

(1) Mittlerer Quadratischer Fehler
(2) Cramér-Rao Ungleichung

Intuitiv haben diese folgende Bedeutungen: Der *mittlere quadratische Fehler* von $\hat{\tau}_n$ ist der Erwartungswert der quadrierten Abweichung von $\hat{\tau}_n(\upsilon)$ von $\tau(\theta)$ über Stichproben vom Umfang n. Die *Cramér-Rao-Ungleichung* gibt eine untere Schranke für die Varianz erwartungstreuer Schätzer an. Ein erwartungstreuer Schätzer mit Varianz gleich der in der Cramér-Rao-Ungleichung gegebenen unteren Schranke hat die kleinstmögliche Varianz aller erwartungstreuen Schätzer und ist in diesem Sinne ein optimaler Schätzer.

Definition (Fehler, Systematischer Fehler, und Erwartungstreue)

$\upsilon := \upsilon_1, ..., \upsilon_n \sim p_\theta$ sei eine Stichprobe und $\hat{\tau}_n$ sei ein Schätzer für τ.

- Der *Fehler* von $\hat{\tau}_n$ ist definiert als
$$\hat{\tau}_n(\upsilon) - \tau(\theta). \tag{24}$$
- Der *systematische Fehler (Bias)* von $\hat{\tau}_n$ ist definiert als
$$\mathrm{B}(\hat{\tau}_n) := \mathbb{E}_\theta(\hat{\tau}_n(\upsilon)) - \tau(\theta). \tag{25}$$
- $\hat{\tau}_n$ heißt *erwartungstreu (unbiased)*, wenn
$$\mathrm{B}(\hat{\tau}_n) = 0 \Leftrightarrow \mathbb{E}_\theta(\hat{\tau}_n(\upsilon)) = \tau(\theta) \text{ für alle } \theta \in \Theta, n \in \mathbb{N}. \tag{26}$$
Andernfalls heißt $\hat{\tau}_n$ *verzerrt (biased)*.

Bemerkungen

- Der Fehler hängt von einer Realisation der Stichprobe ab.
- Der systematische Fehler ist der erwartete Fehler über viele Stichprobenrealisationen.
- Ein Parameterschätzer ist erwartungstreu, wenn $\mathbb{E}_\theta(\hat{\theta}_n(\upsilon)) = \theta$.

Theorem (Erwartungstreue von Stichprobenmittel und Stichprobenvarianz)

$\upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines statistischen parametrischen Produktmodells $\mathcal{M}$.

- Das Stichprobenmittel

$$\bar{\upsilon}_n := \frac{1}{n}\sum_{i=1}^{n} \upsilon_i \tag{27}$$

  ist ein erwartungstreuer Schätzer des Erwartungswerts $\mathbb{E}_\theta(\upsilon_1)$.

- Die Stichprobenvarianz

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n} (\upsilon_i - \bar{\upsilon}_n)^2 \tag{28}$$

  ist ein erwartungstreuer Schätzer der Varianz $\mathbb{V}_\theta(\upsilon_1)$.

Beweis

Um die Notation zu vereinfachen, definieren wir $\mathbb{E} := \mathbb{E}_\theta$ und $\mathbb{V} := \mathbb{V}_\theta$. Mit der Linearität von Erwartungswerten ergibt sich dann

$$\mathbb{E}(\bar{\upsilon}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \upsilon_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\upsilon_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\upsilon_1\right) = \frac{1}{n}n\mathbb{E}\left(\upsilon_1\right) = \mathbb{E}\left(\upsilon_1\right).$$

Dies zeigt die Erwartungstreue des Stichprobenmittels als Schätzer des Erwartungswertes.

Um die Erwartungstreue der Stichprobenvarianz zu zeigen, halten wir zunächst fest, dass

$$\mathbb{V}(\bar{\upsilon}_n) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^n \upsilon_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}\left(\upsilon_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}\left(\upsilon_1\right) = \frac{1}{n^2}n\mathbb{V}\left(\upsilon_1\right) = \frac{\mathbb{V}\left(\upsilon_1\right)}{n}.$$

Weiterhin gilt wie im Folgenden gezeigt mit $\mu := \mathbb{E}(\upsilon_1)$

$$\sum_{i=1}^n \left(\upsilon_i - \bar{\upsilon}_n\right)^2 = \sum_{i=1}^n (\upsilon_i - \mu)^2 - n(\bar{\upsilon}_n - \mu)^2.$$

$$
\begin{aligned}
\sum_{i=1}^{n}(\upsilon_i - \bar{\upsilon}_n)^2 & = \sum_{i=1}^{n}(\upsilon_i - \mu - \bar{\upsilon}_n + \mu)^2 \\
& = \sum_{i=1}^{n}((\upsilon_i - \mu) - (\bar{\upsilon}_n - \mu))^2 \\
& = \sum_{i=1}^{n}(\upsilon_i - \mu)^2 - 2(\bar{\upsilon}_n - \mu)\left(\sum_{i=1}^{n}(\upsilon_i - \mu)\right) + \sum_{i=1}^{n}(\bar{\upsilon}_n - \mu)^2 \\
& = \sum_{i=1}^{n}(\upsilon_i - \mu)^2 - 2(\bar{\upsilon}_n - \mu)\left(\sum_{i=1}^{n}\upsilon_i - n\mu\right) + n(\bar{\upsilon}_n - \mu)^2 \qquad (29) \\
& = \sum_{i=1}^{n}(\upsilon_i - \mu)^2 - 2(\bar{\upsilon}_n - \mu)\left(n\left(\frac{1}{n}\sum_{i=1}^{n}\upsilon_i\right) - n\mu\right) + n(\bar{\upsilon}_n - \mu)^2 \\
& = \sum_{i=1}^{n}(\upsilon_i - \mu)^2 - 2n(\bar{\upsilon}_n - \mu)^2 + n(\bar{\upsilon}_n - \mu)^2 \\
& = \sum_{i=1}^{n}(\upsilon_i - \mu)^2 - n(\bar{\upsilon}_n - \mu)^2.
\end{aligned}
$$

Es ergibt sich also

$$
\begin{aligned}
\mathbb{E}\left((n-1)S_n^2\right) &= \mathbb{E}\left(\sum_{i=1}^{n}(\upsilon_i - \bar{\upsilon}_n)^2\right) \\
&= \mathbb{E}\left(\sum_{i=1}^{n}(\upsilon_i - \mu)^2 - n(\bar{\upsilon}_n - \mu)^2\right) \\
&= \sum_{i=1}^{n}\mathbb{E}\left((\upsilon_i - \mu)^2\right) - n\mathbb{E}\left((\bar{\upsilon}_n - \mu)^2\right) \\
&= n\mathbb{V}(\upsilon_1) - n\mathbb{V}(\bar{\upsilon}_n) \\
&= n\mathbb{V}(\upsilon_1) - n\frac{\mathbb{V}(\upsilon_1)}{n} \\
&= n\mathbb{V}(\upsilon_1) - \mathbb{V}(\upsilon_1) \\
&= (n-1)\mathbb{V}(\upsilon_1)
\end{aligned}
$$

Schließlich ergibt sich

$$
\mathbb{E}(S_n^2) = \mathbb{E}\left(\frac{1}{n-1}(n-1)S_n^2\right) = \frac{1}{n-1}\mathbb{E}\left((n-1)S_n^2\right) = \frac{1}{n-1}(n-1)\mathbb{V}(\upsilon_1) = \mathbb{V}(\upsilon_1)
$$

und damit die Erwartungstreue der Stichprobenvarianz als Schätzer der Varianz.

Theorem (Verzerrtheit der Stichprobenstandardabweichung)

$\upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines statistischen parametrischen Produktmodells $\mathcal{M}$. Dann ist die Stichprobenstandardabweichung

$$S_n := \sqrt{S_n^2} \tag{30}$$

ein verzerrter Schätzer der Standardabweichung $\mathbb{S}_\theta(\upsilon_1)$.

Beweis

Wir halten zunächst fest, dass $\sqrt{\cdot}$ eine strikt konkave Funktion und $\sigma^2 > 0$ ist. Dann aber gilt mit der Jensenschen Ungleichung $\mathbb{E}(f(\xi)) < f(\mathbb{E}(\xi))$ für strikt konkave Funktionen, dass

$$\mathbb{E}(S_n) = \mathbb{E}\left(\sqrt{S_n^2}\right) < \sqrt{\mathbb{E}(S_n^2)} = \sqrt{\mathbb{V}_\theta(\upsilon_1)} = \mathbb{S}_\theta(\upsilon_1). \tag{31}$$

$\square$

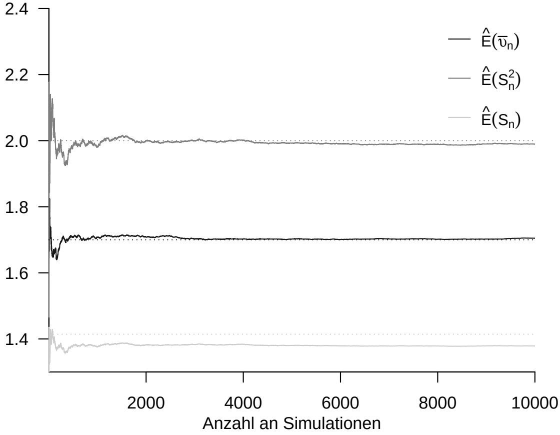
Bemerkung

- Nichtlineare Transformationen von erwartungstreuen Schätzern liefern oft verzerrte Schätzer.

Simulation von $\upsilon_1, ..., \upsilon_{12} \sim N(\mu, \sigma^2)$ mit $n = 12$, $\mu = 1.7$, $\sigma^2 = 2$, $\sigma \approx 1.41$

```
# Modellformulierung
mu      = 1.7                      # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr  = 2                        # wahrer, aber unbekannter, Varianzparameter
n       = 12                       # Stichprobengroesse n
nsim    = 1e4                      # Anzahl der Simulationen
y_bar   = rep(NaN,nsim)            # Stichprobenmittelarray
s_sqr   = rep(NaN,nsim)            # Stichprobenvarianzarray
s       = rep(NaN,nsim)            # Stichprobenstandardabweichungarray

# Simulationsiterationen
for(sim in 1:nsim){

    # Stichprobenrealisation von \ups_1,...,\ups_{12}
    y           = rnorm(n,mu,sqrt(sigsqr))

    # Erwartungswert-, Varianz-, Standardabweichungschaetzer
    y_bar[sim] = mean(y)           # Stichprobenmittel
    s_sqr[sim] = var(y)            # Stichprobenvarianz
    s[sim]     = sd(y)             # Stichprobenstandardabweichung
}

# Erwartungswertschaetzung
E_hat_y_bar = cumsum(y_bar)/(1:nsim) # \mathbb{E}(\bar{\ups}_n) Schaetzungen
E_hat_s_sqr = cumsum(s_sqr)/(1:nsim) # \mathbb{E}(S^2) Schaetzungen
E_hat_s     = cumsum(s)    /(1:nsim) # \mathbb{E}(S) Schaetzungen
```

Simulation von $\upsilon_1, ..., \upsilon_{12} \sim N(\mu, \sigma^2)$ mit $n = 12$, $\mu = 1.7$, $\sigma^2 = 2$, $\sigma \approx 1.41$



Definition (Varianz und Standardfehler)

$\upsilon := \upsilon_1, ..., \upsilon_n \sim p_\theta$ sei eine Stichprobe und $\hat{\tau}_n$ sei ein Schätzer von τ.

- Die *Varianz* von $\hat{\tau}_n$ ist definiert als

$$\mathbb{V}_\theta(\hat{\tau}_n) := \mathbb{E}_\theta\left(\left(\hat{\tau}_n(\upsilon) - \mathbb{E}_\theta(\hat{\tau}_n(\upsilon))\right)^2\right). \quad (32)$$

- Der *Standardfehler* von $\hat{\tau}_n$ ist definiert als

$$\mathsf{SE}(\hat{\tau}_n) := \sqrt{\mathbb{V}_\theta(\hat{\tau}_n)} \quad (33)$$

Bemerkungen

- Die Varianz eines Schätzers $\hat{\tau}_n$ ist die Varianz der Zufallsvariable $\hat{\tau}_n(\upsilon)$.
- Der Standardfehler eines Schätzers $\hat{\tau}_n$ ist die Standardabweichung von $\hat{\tau}_n(\upsilon)$.

Theorem (Standardfehler des Stichprobenmittels)

$\upsilon := \upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells. Dann ist der *Standardfehler des Stichprobenmittels* gegeben durch

$$\mathrm{SE}(\bar{\upsilon}_n) = \frac{\mathbb{S}_\theta(\upsilon_1)}{\sqrt{n}}. \tag{34}$$

Der Standardfehler des Stichprobenmittels heißt auch *Standardfehler des Mittelwertes*.

Beweis

Per definitionem und mit $\mathbb{V}_\theta(\bar{\upsilon}_n) = \mathbb{V}_\theta(\upsilon_1)/n$, ergibt sich

$$\mathrm{SE}(\bar{\upsilon}_n) = \sqrt{\mathbb{V}_\theta(\bar{\upsilon}_n)} = \sqrt{\frac{\mathbb{V}_\theta(\upsilon_1)}{n}} = \frac{\mathbb{S}_\theta(\upsilon_1)}{\sqrt{n}}. \tag{35}$$

Bemerkungen

- Der Standardfehler des Mittelwerts beschreibt die Variabilität des Stichprobenmittels.
- Da $\mathbb{S}_\theta(\upsilon_1)$ unbekannt ist, ist auch $\mathrm{SE}(\bar{\upsilon}_n)$ unbekannt.
- Ein verzerrter Schätzer für den Standardfehler des Stichprobenmittels ist gegeben durch $\hat{\mathrm{SE}}(\bar{\upsilon}_n) = \frac{s_n}{\sqrt{n}}$.

Beispiel (Standardfehler des Bernoulli Parameter Maximum-Likelihood Schätzes)

Es sei $\upsilon := \upsilon_1, ..., \upsilon_n \sim \text{Bern}(\mu)$ und $\hat{\mu}_n^{\text{ML}}$ der Maximum-Likelihood Schätzer für μ. Dann ist

$$\text{SE}\left(\hat{\mu}_n^{\text{ML}}\right) = \sqrt{\frac{\mu(1-\mu)}{n}}. \tag{36}$$

<u>Beweis</u>

Es gilt

$$\begin{aligned} \text{SE}\left(\hat{\mu}_n^{\text{ML}}\right) = \sqrt{\mathbb{V}_\mu\left(\hat{\mu}_n^{\text{ML}}\right)} & = \sqrt{\mathbb{V}_\mu\left(\frac{1}{n}\sum_{i=1}^n \upsilon_i\right)} \\ & = \sqrt{\frac{1}{n^2}\sum_{i=1}^n \mathbb{V}_\mu(\upsilon_i)} = \sqrt{\frac{n\mu(1-\mu)}{n^2}} = \sqrt{\frac{\mu(1-\mu)}{n}}, \end{aligned} \tag{37}$$

wobei die dritte Gleichung mit der Unabhängigkeit der $\upsilon_i, i = 1, ..., n$ und die vierte Gleichung mit der Varianz $\mathbb{V}_\mu(\upsilon_1) = \mathbb{V}_\mu(\upsilon_i) = \mu(1-\mu), i = 1, ..., n$ der Bernoulli Stichprobenvariablen folgt.

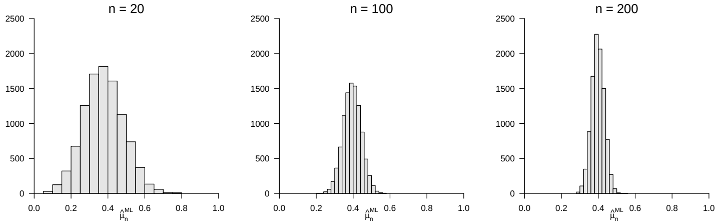
Bemerkung

- Ein Schätzer für den Standardfehler $\text{SE}\left(\hat{\mu}_n^{\text{ML}}\right)$ ist $\hat{\text{SE}}\left(\hat{\mu}_n^{\text{ML}}\right) = \sqrt{\frac{\hat{\mu}_n^{\text{ML}}(1-\hat{\mu}_n^{\text{ML}})}{n}}$

Simulation von $\upsilon_1, ..., \upsilon_n \sim \text{Bern}(\mu)$ mit $\mu = 0.4$

```
# Modellformulierung
mu          = 0.4                                 # wahrer, aber unbekannter, Parameterwert
n_all       = c(20,100,200)                       # Stichprobengrößen n
ns          = 1e4                                 # Anzahl der Simulationen
mu_hat_ML   = matrix(rep(NaN, length(n_all)*ns),  # Maximum-Likelihood Schätzearray
                     nrow = length(n_all))

# Stichprobengroessenniterationen
for(i in seq_along(n_all)){

    # Simulationsiterationen
    for(s in 1:ns){
        y               = rbinom(n_all[i],1,mu)   # Stichprobenrealisation von y_1,...,y_n
        mu_hat_ML[i,s]  = mean(y)                 # Stichprobenmittel
    }
}
```

Die Varianz bzw. der Standardfehler von $\hat{\mu}_n^{\text{ML}}$ hängen von n ab.



Vorbemerkungen zu Asymptotischen Schätzereigenschaften

Dieser Abschnitt ist eine Kurzeinführung in die *Asymptotische Statistik (AS)*.

Die AS befasst sich mit dem Verhalten von Statistiken bei großen Stichproben.

Methoden der AS werden benutzt, um

- qualitative Schätzereigenschaften zu studieren und
- Schätzereigenschaften für große Stichprobengrößen zu approximieren.

Moderne Stichproben sind üblicherweise groß.

Die Methoden der AS sind also praktisch einsetzbar und gerechtfertigt.

Vaart (1998) gibt eine ausführliche Einführung in die AS.

Asymptotische Schätzereigenschaften

Wir betrachten im Folgenden drei asymptotische Schätzereigenschaften:

(1) Asymptotische Erwartungstreue
(2) Konsistenz
(3) Asymptotische Normalverteilung

Intuitiv haben diese folgende Bedeutungen: Ein Schätzer $\hat{\tau}_n$ für τ heißt *asymptotisch erwartungstreu*, wenn der Erwartungswert von $\hat{\tau}_n$ für große Stichprobengrößen $n \to \infty$ gleich dem wahren, aber unbekannten, Wert $\tau(\theta)$ ist. Ein Schätzer $\hat{\tau}_n$ für τ heißt *konsistent*, wenn für große Stichprobengrößen $n \to \infty$ die Wahrscheinlichkeit dafür, dass $\hat{\tau}_n(\upsilon)$ vom wahren, aber unbekannten, Wert $\tau(\theta)$ abweicht beliebig klein wird. Ein Schätzer $\hat{\tau}_n$ für τ heißt *asymptotisch normalverteilt*, wenn für große Stichprobengrößen $n \to \infty$, die Verteilung von $\hat{\tau}_n$ durch eine Normalverteilung gegeben ist.

Für folgende asymptotische Schätzereigenschaften verweisen wir auf den Appendix:

(1) Asymptotische Effizienz

Intuitiv hat diese die folgende Bedeutung: Ein Schätzer $\hat{\tau}_n$ für τ heißt *asymptotisch effizient*, wenn für große Stichprobengrößen $n \to \infty$ die Verteilung von $\hat{\tau}_n$ durch eine Normalverteilung mit Erwartungswertparameter $\tau(\theta)$ und Varianzparameter gleich der Cramér-Rao-Schranke gegeben ist.

Definition (Asymptotische Erwartungstreue)

$\upsilon := \upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells $\mathcal{M}$ und $\hat{\tau}_n$ sei ein Schätzer für τ. $\hat{\tau}_n$ heißt *asymptotisch erwartungstreu*, wenn

$$\lim_{n \to \infty} \mathbb{E}_\theta(\hat{\tau}_n(\upsilon)) = \tau(\theta) \text{ für alle } \theta \in \Theta. \tag{38}$$

Bemerkungen

- Asymptotisch erwartungstreue Schätzer sind für "unendlich große" Stichproben erwartungstreu.
- Erwartungstreue Schätzer sind immer auch asymptotisch erwartungstreu.

Theorem (Asymptotische Erwartungstreue des Varianzparameterschätzers)

Es sei $\upsilon := \upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$. Dann ist der Maximum-Likelihood Schätzer des Varianzparameters,

$$\hat{\sigma}_n^{2^{\text{ML}}} := \frac{1}{n}\sum_{i=1}^n (\upsilon_i - \bar{\upsilon}_n)^2 \tag{39}$$

asymptotisch erwartungstreu.

Beweis

Mit der Erwartungstreue der Stichprobenvarianz ergibt sich

$$\mathbb{E}_{\mu,\sigma^2}\left(\hat{\sigma}_n^{2^{\text{ML}}}\right) = \mathbb{E}_{\mu,\sigma^2}\left(\frac{1}{n}\sum_{i=1}^n \left(\upsilon_i - \bar{\upsilon}_n\right)^2\right) = \frac{1}{n}\mathbb{E}_{\mu,\sigma^2}\left(\sum_{i=1}^n \left(\upsilon_i - \bar{\upsilon}_n\right)^2\right) = \frac{n-1}{n}\sigma^2$$

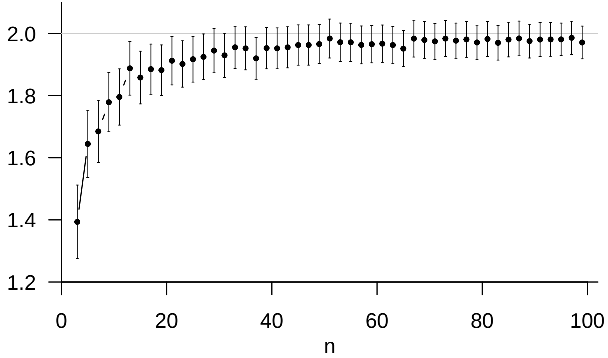
Also gilt $\mathbb{E}_{\mu,\sigma^2}\left(\hat{\sigma}_n^{2^{\text{ML}}}\right) \neq \sigma^2$. $\hat{\sigma}_n^{2^{\text{ML}}}$ ist also ein verzerrter Schätzer von σ^2. Allerdings gilt $(n-1)/n \to 1$ für $n \to \infty$, so dass

$$\lim_{n\to\infty} \mathbb{E}_{\mu,\sigma^2}\left(\hat{\sigma}_n^{2^{\text{ML}}}\right) = \lim_{n\to\infty} \frac{n-1}{n}\sigma^2 = \sigma^2 \lim_{n\to\infty} \frac{n-1}{n} = \sigma^2. \tag{40}$$

Simulation von $\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$ mit $\mu = 1$, $\sigma^2 = 2$

```
# Modellformulierung
mu        = 1                       # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr    = 2                       # wahrer, aber unbekannter, Varianzparameter
n         = seq(1,100, by = 2)      # Stichprobengroessen
ns        = 1e3                     # Anzahl Simulation pro Stichprobengroesse
sigsqr_ml = matrix(                 # \hat{\sigma^2}^{ML} Array
            rep(NaN, length(n)*ns),
            ncol = length(n))

# Stichprobengroesseniterationen
for(i in seq_along(n)){

    # Simulationsiterationen
    for(s in 1:ns){

        # Stichprobenrealisation
        y               = rnorm(n[i], mu, sqrt(sigsqr))

        # \hat{\sigma^2}^{ML}
        sigsqr_ml[s,i]  = ((n[i]-1)/n[i])*var(y)
    }
}
E_sigsqr_ml = colMeans(sigsqr_ml)   # Erwartungswertschaetzung
```

Simulation $\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$ mit $\mu = 1$, $\sigma^2 = 2$



Definition (Konsistenz)

$\upsilon := \upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells $\mathcal{M}$ und $\hat{\tau}_n$ sei ein Schätzer von τ. Eine Folge von Schätzern $\hat{\tau}_1, \hat{\tau}_2, ...$ wird dann eine *konsistente Folge von Schätzern* genannt, wenn für jedes $\epsilon > 0$ und jedes $\theta \in \Theta$ gilt, dass

$$\lim_{n\to\infty} \mathbb{P}_\theta \left(|\hat{\tau}_n(\upsilon) - \tau(\theta)| \geq \epsilon\right) = 0.$$

Wenn $\hat{\tau}_1, \hat{\tau}_2, ...$ eine konsistente Folge von Schätzern ist, dann heißt $\hat{\tau}_n$ *konsistenter Schätzer*.

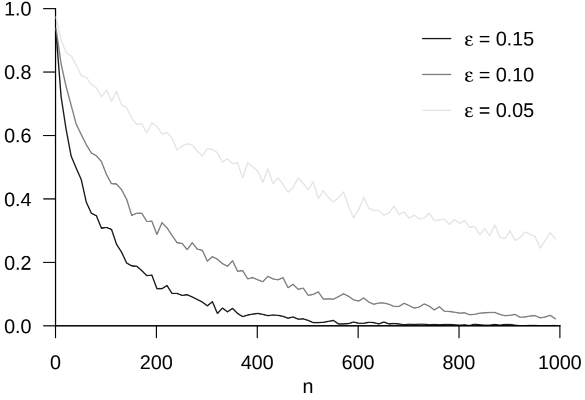
Bemerkungen

- Für $n \to \infty$ wird die Wahrscheinlichkeit, dass $\hat{\tau}_n(\upsilon)$ beliebig nah bei $\tau(\theta)$ liegt, groß.
- Für $n \to \infty$ wird die Wahrscheinlichkeit, dass $\hat{\tau}_n(\upsilon)$ von $\tau(\theta)$ abweicht, klein.
- Diese Eigenschaften gelten für alle möglichen wahren, aber unbekannten, Parameterwerte.
- Die Konvergenz ist *Konvergenz in Wahrscheinlichkeit*.
- Konsistenz von Schätzern kann direkt oder mit Kriterien nachgewiesen werden.

Simulation der Konsistenz von $\bar{\upsilon}_n$ bei $\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$ mit $\mu = 1$, $\sigma^2 = 2$

```
# Modellformulierung
mu      = 1                                               # w.a.u \mu Wert
sigsqr  = 2                                               # w.a.u. \sigma^2 Wert
n       = seq(1,1e3,by = 10)                              # Stichprobengroesse n
eps     = c(0.15, 0.10, 0.05)                             # \epsilon Werte
ne      = length(eps)                                     # Anzahl \epsilon Werte
nn      = length(n)                                       # Anzahl Stichprobengroessen
ns      = 1e3                                             # Anzahl Simulationen
E       = array(rep(NaN,nn,ne,ns),                        # Ereignisindikatorarray
                dim = c(nn,ne,ns))

# Simulation
for(e in seq_along(eps)){                                 # \epsilon Iterationen
    for(i in seq_along(n)){                               # n Iterationen
        for(s in 1:ns){                                   # Simulationsiterationen

            # Stichprobenrealisationen
            y   = rnorm(n[i], mu, sqrt(sigsqr))
            if(abs(mean(y) - mu) >= eps[e]){              # |y_bar - \mu)| \ge \epsilon
                E[i,e,s] = 1
            } else {                                      # |y_bar - \mu)| < \epsilon
                E[i,e,s] = 0
            }
        }
    }
}

# Schaetzung von \mathbb{P}(|\hat{\tau}_n(\ups)-\tau(\theta)| \ge \epsilon)
P_hat       = apply(E, c(1,2), mean)
```

Simulation der Konsistenz von $\bar{\upsilon}_n$ bei $\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$ mit $\mu = 1$, $\sigma^2 = 2$



Definition (Asymptotische Normalität)

$\upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells $\mathcal{M}$ und $\hat{\theta}_n$ sei ein Parameterschätzer für θ. Weiterhin sei $\tilde{\theta} \sim N(\mu, \sigma^2)$ eine normalverteilte Zufallsvariable mit Erwartungswertparameter μ und Varianzparameter σ^2. Wenn $\hat{\theta}_n$ in Verteilung gegen $\tilde{\theta}$ konvergiert, dann heißt $\hat{\theta}_n$ *asymptotisch normalverteilt* und wir schreiben

$$\hat{\theta}_n \overset{a}{\sim} N(\mu, \sigma^2). \tag{41}$$

Bemerkung

- Konvergenz in Verteilung heißt $\lim_{n\to\infty} P_n(\hat{\theta}_n) = P(\tilde{\theta})$.

Grundbegriffe

Maximum-Likelihood Schätzer

Schätzereigenschaften bei endlichen Stichproben

Asymptotische Schätzereigenschaften

Eigenschaften von Maximum-Likelihood Schätzern

Selbstkontrollfragen

Theorem (Eigenschaften von Maximum-Likelihood Schätzern)

$\upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells $\mathcal{M}$ und $\hat{\theta}_n^{\text{ML}}$ sei ein Maximum-Likelihood Schätzer für θ. Dann gilt, dass $\hat{\theta}_n^{\text{ML}}$

(1) nicht notwendigerweise erwartungstreu, aber

(2) konsistent,

(3) asymptotisch normalverteilt und

(4) asymptotisch erwartungstreu ist.

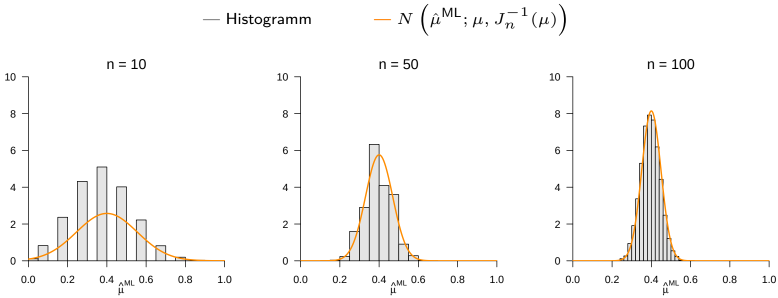
Bemerkungen

- Maximum-Likelihood Schätzer sind überdies asymptotisch effizient.
- Für einen Beweis verweisen wir auf Held and Sabanés Bové (2014), Abschnitt 3.4.

Simulation der asymptotische Normalverteilung des Maximum-Likelihood Bernoulliparameterschätzers

```
# Modellformulierung
mu          = 0.4                                                 # w.a.u. Parameterwert
n_all       = c(1e1,5e1,1e2)                                      # Stichprobengroesse n
ns          = 1e4                                                 # Anzahl der Simulationen
mu_hat_ML   = matrix(                                             # ML Schaetzerarray
                        rep(NaN,
                        length(n_all)*ns),
                        nrow = length(n_all))
mu_hat_ML_r = 1e3                                                 # ML Schaetzerraumaufloesung
mu_hat_ML_y = seq(0,1,len = mu_hat_ML_r)                          # ML Schaetzerraum
mu_hat_ML_p = matrix(rep(NaN, length(n_all)*mu_hat_ML_r),         # ML WDF Array
                     nrow = length(n_all))

# Stichprobengroesseniterationen
for(i in seq_along(n_all)){

    # Simulationsiterationen
    for(s in 1:ns){
        y               = rbinom(n_all[i],1,mu)                   # Stichprobenrealisation
        mu_hat_ML[i,s]  = mean(y)                                 # ML Schaetzer
    }

    # WDF der asymptotischen Verteilung
    mu_hat_ML_p[i,] = dnorm(mu_hat_ML_y, mu, sqrt(mu*(1-mu)/n_all[i]))
}
```

Simulation der asymptotische Normalverteilung des Maximum-Likelihood Bernoulliparameterschätzers



Grundbegriffe

Maximum-Likelihood Schätzer

Schätzereigenschaften bei endlichen Stichproben

Asymptotische Schätzereigenschaften

Eigenschaften von Maximum-Likelihood Schätzern

Selbstkontrollfragen

Selbstkontrollfragen

1. Definieren und erläutern Sie den Begriff des Parameterpunktschätzers.
2. Definieren Sie den Begriff der Likelihood-Funktion.
3. Definieren Sie den Begriff der Log-Likelihood-Funktion.
4. Definieren Sie den Begriff des Maximum-Likelihood Schätzes.
5. Erläutern Sie das Vorgehen zur Maximum-Likelihood Schätzung für ein parametrisches Produktmodell.
6. Geben Sie den Maximum-Likelihood Schätzer für den Parameter μ des Bernoullimodells an.
7. Geben Sie den Maximum-Likelihood Schätzer für den Parameter μ des Normalverteilungmodels an.
8. Geben Sie den Maximum-Likelihood Schätzer für den Parameter σ^2 des Normalverteilungmodels an.
9. Definieren und erläutern Sie den Begriff der Erwartungstreue eines Schätzers.
10. Definieren Sie die Begriffe der Varianz und des Standardfehlers eines Schätzers.
11. Erläutern Sie den Begriff des asymptotischen Erwartungstreue eines Schätzers
12. Erläutern Sie den Begriff der Konsistenz eines Schätzers.
13. Erläutern Sie den Begriff der asymptotischen Normalität eines Schätzers.
14. Nennen Sie vier Eigenschaften eines Maximum-Likelihood Schätzers.

Appendix

Appendix | Mittlerer quadratischer Fehler

Definition (Mittlerer quadratischer Fehler)

$\upsilon := \upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells $\mathcal{M}$ und $\hat{\tau}_n$ ein Schätzer für τ. Dann ist der *mittlere quadratischer Fehler (engl. mean squared error)* von $\hat{\tau}_n$ definiert als

$$\text{MQF}(\hat{\tau}_n) := \mathbb{E}_\theta\left(\left(\hat{\tau}_n(\upsilon) - \tau(\theta)\right)^2\right). \tag{42}$$

Bemerkungen

- Der MQF von $\hat{\tau}_n$ ist die erwartete quadrierte Abweichung von $\hat{\tau}_n(\upsilon)$ von $\tau(\theta)$.
- Die Varianz von $\hat{\tau}_n$ ist die erwartete quadrierte Abweichung von $\hat{\tau}_n$ von $\mathbb{E}_\theta(\hat{\tau}_n(\upsilon))$.
- $\mathbb{E}_\theta(\hat{\tau}_n(\upsilon))$ kann mit $\tau(\theta)$ übereinstimmen, muss es aber nicht.

Theorem (Zerlegung des mittleren quadratischen Fehlers)

$\upsilon := \upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells $\mathcal{M}$, $\hat{\tau}_n$ sei ein Schätzer für τ, und $\mathrm{MQF}(\hat{\tau}_n)$ sei der mittlere quadratische Fehler von $\hat{\tau}_n$. Dann gilt

$$\mathrm{MQF}(\hat{\tau}_n) = \mathrm{B}(\hat{\tau}_n)^2 + \mathbb{V}_\theta(\hat{\tau}_n). \tag{43}$$

Bemerkungen

- $\mathrm{MQF} = \mathrm{Bias}^2 + \mathrm{Varianz}$.
- Der MQF kann als Bias-Varianz Abwägungskriterium benutzt werden.
- Kleine Schätzerverzerrungen können gegenüber einer großen Schätzervarianz präferiert werden.

Beweis

Zur Vereinfachung der Notation seien $\tau := \tau(\theta)$, $\hat{\tau}_n := \hat{\tau}_n(\upsilon)$ und $\bar{\tau}_n := \mathbb{E}_\theta(\hat{\tau}_n(\upsilon))$. Dann gilt:

$$
\begin{aligned}
\mathbb{E}_\theta\left((\hat{\tau}_n - \tau)^2\right) &= \mathbb{E}_\theta\left((\hat{\tau}_n - \bar{\tau}_n + \bar{\tau}_n - \tau)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\tau}_n - \bar{\tau}_n)^2 + 2(\hat{\tau}_n - \bar{\tau}_n)(\bar{\tau}_n - \tau) + (\bar{\tau}_n - \tau)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\tau}_n - \bar{\tau}_n)^2\right) + 2\mathbb{E}_\theta\left((\hat{\tau}_n - \bar{\tau}_n)(\bar{\tau}_n - \tau)\right) + \mathbb{E}_\theta\left((\bar{\tau}_n - \tau)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\tau}_n - \bar{\tau}_n)^2\right) + 2\mathbb{E}_\theta\left(\hat{\tau}_n\bar{\tau}_n - \hat{\tau}_n\tau - \bar{\tau}_n\bar{\tau}_n + \bar{\tau}_n\tau\right) + \mathbb{E}_\theta\left((\bar{\tau}_n - \tau)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\tau}_n - \bar{\tau}_n)^2\right) + 2\left(\bar{\tau}_n\bar{\tau}_n - \bar{\tau}_n\tau\right) + \mathbb{E}_\theta\left((\bar{\tau}_n - \tau)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\tau}_n - \bar{\tau}_n)^2\right) + 0 + \mathbb{E}_\theta\left((\bar{\tau}_n - \tau)^2\right) \\
&= \mathbb{E}_\theta\left((\bar{\tau}_n - \tau)^2\right) + \mathbb{E}_\theta\left((\hat{\tau}_n - \bar{\tau}_n)^2\right) \\
&= \mathbb{E}_\theta\left((\mathbb{E}_\theta(\hat{\tau}_n) - \tau)^2\right) + \mathbb{E}_\theta\left((\hat{\tau}_n - \mathbb{E}_\theta(\hat{\tau}_n))^2\right) \\
&= (\mathbb{E}_\theta(\hat{\tau}_n) - \tau)^2 + \mathbb{V}_\theta(\hat{\tau}_n) \\
&= \mathrm{B}(\hat{\tau}_n)^2 + \mathbb{V}_\theta(\hat{\tau}_n).
\end{aligned}
$$

Appendix | Cramér-Rao-Ungleichung

Vorbemerkungen zur Cramér-Rao-Ungleichung

Je kleiner die Varianz eines Schätzers, desto besser. Weil aber Stichproben streuen, kann die Varianz von erwartungstreuen Schätzern nicht beliebig klein sein.

Die **Cramér-Rao-Ungleichung** gibt eine untere Schranke für die Varianz erwartungstreuer Schätzer an. Ein erwartungstreuer Schätzer mit Varianz gleich der in der Cramér-Rao-Ungleichung gegebenen unteren Schranke hat die kleinstmögliche Varianz aller erwartungstreuer Schätzer und ist in diesem Sinne "optimal".

Die Cramér-Rao-Ungleichung basiert auf dem Begriff der **Fisher-Information**. Wir diskutieren deshalb zunächst die Begriffe der **Scorefunktion** und der darauf basierenden **Fisher-Information**.

Die vorgestellten Resultate gelten im Allgemeinen nur unter eine Reihe von Annahmen, den sogenannten **Fisher-Regularitätsbedingungen**:

- Θ ist ein offenes Intervall, d.h. θ liegt nicht an einer Parameterraumgrenze.
- Der Träger von p_θ hängt nicht von θ ab.
- WMFs oder WDF mit unterschiedlichem $\theta \in \Theta$ sind unterschiedlich.
- Die Likelihood-Funktion ist zweimal stetig differenzierbar.
- Integration und Differentiation dürfen vertauscht werden.

Definition (Scorefunktion und Fisher-Information)

$\upsilon := \upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells $\mathcal{M}$ mit eindimensionalem Parameter θ und ℓ_n sei die zugehörige Log-Likelihood-Funktion.

- Die erste Ableitung der Log-Likelihood-Funktion ℓ_n wird *Scorefunktion der Stichprobe* genannt und mit
$$S_n(\theta) := \frac{d}{d\theta}\ell_n(\theta). \tag{44}$$
bezeichnet. Für $n = 1$ schreiben wir $S(\theta) := S_1(\theta)$ und nennen $S(\theta)$ *Scorefunktion einer Zufallsvariable*.
- Die negative zweite Ableitung der Log-Likelihood-Funktion ℓ_n wird *Fisher-Information der Stichprobe* genannt und mit
$$I_n(\theta) := -\frac{d^2}{d\theta^2}\ell_n(\theta). \tag{45}$$
bezeichnet. Für $n = 1$ schreiben wir $I(\theta) := I_1(\theta)$ und nennen $I(\theta)$ die *Fisher-Information einer Zufallsvariable*.

Definition (Erwartete und beobachtete Fisher-Information)

$\upsilon := \upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells $\mathcal{M}$ mit eindimensionalem Parameter θ, ℓ_n sei die zugehörige Log-Likelihood-Funktion und $\hat{\theta}_n^{\text{ML}}$ sei ein ML-Schätzer von θ.

- Die *beobachtete Fisher-Information der Stichprobe* ist definiert als

$$I_n\left(\hat{\theta}_n^{\text{ML}}\right) := -\frac{d^2}{d\theta^2}\ell_n\left(\hat{\theta}_n^{\text{ML}}\right), \tag{46}$$

  d.h. die beobachtete Fisher-Information der Stichprobe ist die Fisher-Information an der Stelle des ML-Schätzers $\hat{\theta}_n^{\text{ML}}$.

- Die *erwartete Fisher-Information der Stichprobe* ist definiert als

$$J_n(\theta) := \mathbb{E}_\theta(I_n(\theta)). \tag{47}$$

  Für $n = 1$ schreiben wir $J(\theta) := J_1(\theta)$ und nennen $J(\theta)$ die *erwartete Fisher-Information einer Zufallsvariable.*

Theorem (Additivität der Fisher-Information)

$\upsilon := \upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells $\mathcal{M}$ mit eindimensionalem Parameter θ, ℓ_n sei die zugehörige Log-Likelihood-Funktion, und $I_n(\theta)$ und $J_n(\theta)$ seien die Fisher-Information und die erwartete Fisher-Information der Stichprobe, respektive. Dann gilt

$$I_n(\theta) = nI_1(\theta) \text{ und } J_n(\theta) = nJ_1(\theta). \tag{48}$$

Bemerkungen

- Um $I_n(\theta)$ oder $J_n(\theta)$ zu berechnen, genügt es also $I(\theta)$ oder $J(\theta)$ zu berechnen.
- Die Additivität der beobachteten Fisher-Information ist in der Additivität von $I_n(\theta)$ implizit.

Beweis

Wir zeigen das Resultat für die erwartete Fisher-Information, das Resultat für die Fisher-Information ist dann implizit. Per definitionem und mit der Linearität von Ableitungen und Erwartungswerte gilt

$$
\begin{aligned}
J_n(\theta) &= \mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\ell_n(\theta)\right) \\
&= \mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\ln\left(\prod_{i=1}^n p_\theta(\upsilon_i)\right)\right) \\
&= \mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\sum_{i=1}^n \ln p_\theta(\upsilon_i)\right) \\
&= \mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\sum_{i=1}^n \ln p_\theta(y_1)\right) \\
&= \mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\ell_1(\theta)n\right) \\
&= n\mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\ell_1(\theta))\right) \\
&= nJ_n(\theta).
\end{aligned}
\tag{49}
$$

Theorem (Erwartungswert und Varianz der Scorefunktion)

Der Erwartungswert der Scorefunktion einer Zufallsvariable ist

$$\mathbb{E}_\theta(S(\theta)) = 0 \tag{50}$$

und die Varianz der Scorefunktion einer Zufallsvariable ist

$$\mathbb{V}_\theta(S(\theta)) = J(\theta). \tag{51}$$

Bemerkungen

- Der Erwartungswert der Ableitung der Log-Likelihood-Funktion ist Null.
- Die erwartete Fisher-Information ist gleich der Varianz der Scorefunktion.

Beweis

Wir betrachten nur den Fall, dass p_θ eine WDF ist und zeigen zunächst, dass $\mathbb{E}_\theta(S(\theta)) = 0$ ist:

$$
\begin{aligned}
\mathbb{E}_\theta(S(\theta)) &= \int S(\theta) p_\theta(x)\, dx \\
&= \int \frac{d}{d\theta} \ell(\theta) p_\theta(x)\, dx \\
&= \int \frac{d}{d\theta} \ln L(\theta) p_\theta(x)\, dx \\
&= \int \frac{1}{L(\theta)} \frac{d}{d\theta} L(\theta) p_\theta(x)\, dx \\
&= \int \frac{1}{p_\theta(x)} \frac{d}{d\theta} L(\theta) p_\theta(x)\, dx \\
&= \int \frac{d}{d\theta} L(\theta)\, dx \\
&= \frac{d}{d\theta} \int p_\theta(x)\, dx = \frac{d}{d\theta} 1 = 0.
\end{aligned} \tag{52}
$$

Beweis (fortgeführt)

Mit der Definition der Varianz folgt dann sofort, dass $\mathbb{V}_\theta(S(\theta)) = \mathbb{E}_\theta(S(\theta)^2)$ ist. Als nächstes zeigen wir, dass $J(\theta) = \mathbb{E}_\theta(S(\theta)^2)$ und deshalb $\mathbb{V}_\theta(S(\theta)) = J(\theta)$ ist:

$$
\begin{aligned}
J(\theta) &= \mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\ln L(\theta)\right) \\
&= \mathbb{E}_\theta\left(-\frac{d}{d\theta}\frac{\frac{d}{d\theta}L(\theta)}{L(\theta)}\right) \\
&= \mathbb{E}_\theta\left(-\frac{\frac{d^2}{d\theta^2}L(\theta)L(\theta) - \frac{d}{d\theta}L(\theta)\frac{d}{d\theta}L(\theta)}{L(\theta)L(\theta)}\right) \\
&= -\mathbb{E}_\theta\left(\frac{\frac{d^2}{d\theta^2}L(\theta)}{L(\theta)}\right) + \mathbb{E}_\theta\left(\frac{\left(\frac{d}{d\theta}L(\theta)\right)^2}{(L(\theta))^2}\right) \\
&= -\int \frac{\frac{d^2}{d\theta^2}L(\theta)}{L(\theta)}p_\theta(x)\,dx + \int \frac{\left(\frac{d}{d\theta}L(\theta)\right)^2}{(L(\theta))^2}p_\theta(x)\,dx \\
&= -\frac{d^2}{d\theta^2}\int p_\theta(x)\,dx + \int\left(\frac{1}{L(\theta)}\frac{d}{d\theta}L(\theta)\right)^2 p_\theta(x)\,dx \\
&= -\frac{d^2}{d\theta^2}1 + \int\left(\frac{d}{d\theta}\ln L(\theta)\right)^2 p_\theta(x)\,dx = \mathbb{E}_\theta\left(S(\theta)^2\right).
\end{aligned}
\tag{53}
$$

Theorem (Fisher-Information bei Bernoulli-Verteilung)

Es sei $\upsilon := \upsilon_1, ..., \upsilon_n \sim \text{Bern}(\mu)$ mit $\mu \in]0,1[$. Dann gilt:

- Die Scorefunktion der Stichprobe ist

$$S_n :]0,1[\to \mathbb{R}, \mu \mapsto S_n(\mu) := \frac{1}{\mu}\sum_{i=1}^n y_i - \frac{1}{1-\mu}\left(n - \sum_{i=1}^n y_i\right). \tag{54}$$

- Die Fisher-Information der Stichprobe ist

$$I_n :]0,1[\to \mathbb{R}, \mu \mapsto I_n(\mu) := I_n(\mu) = \frac{ny}{\mu^2} + \frac{n(1-y)^2}{1-\mu}. \tag{55}$$

- Die beobachtete Fisher-Information der Stichprobe ist

$$I_n :]0,1[\to \mathbb{R}, \hat{\mu}_n^{\text{ML}} \mapsto I_n\left(\hat{\mu}_n^{\text{ML}}\right) := \frac{ny}{\hat{\mu}_n^{\text{ML}^2}} + \frac{n(1-y)}{1-\hat{\mu}_n^{\text{ML}}}. \tag{56}$$

- Die erwartete Fisher-Information der Stichprobe ist

$$J_n :]0,1[\to \mathbb{R}, \mu \mapsto J_n(\mu) := \frac{n}{\mu(1-\mu)}. \tag{57}$$

Beweis

Die Scorefunktion wurde bereits im Kontext der Maximum-Likelihood-Schätzung von μ hergeleitet. Wir betrachten die Fisher-Information einer einzelnen Bernoulli Zufallsvariable υ:

$$
\begin{aligned}
I(\mu) &:= -\frac{d^2}{d\mu^2}\ell_1(\mu) \\
&= -\frac{d^2}{d\mu^2}\ln p_\mu(y) \\
&= -\frac{d^2}{d\mu^2}\left(y\ln\mu + (1-y)\ln(1-\mu)\right) \\
&= -\frac{d}{d\mu}\left(\frac{d}{d\mu}\left(y\ln\mu + (1-y)\ln(1-\mu)\right)\right) \\
&= -\frac{d}{d\mu}\left(\frac{y}{\mu} + \frac{(1-y)}{1-\mu}\right) \\
&= -\left(-\frac{y}{\mu^2} - \frac{(1-y)}{1-\mu}^{2}\right) \\
&= \frac{y}{\mu^2} + \frac{(1-y)}{1-\mu}^{2}.
\end{aligned}
\tag{58}
$$

Beweis (fortgeführt)

Damit ergibt sich die erwartete Fisher-Information der Zufallsvariable υ als

$$
\begin{aligned}
J(\mu) &= \mathbb{E}_\mu(I(\mu)) \\
&= \mathbb{E}_\mu\left(\frac{\upsilon}{\mu^2} + \frac{(1-\upsilon)^2}{1-\mu}\right) \\
&= \frac{\mathbb{E}_\mu(\upsilon)}{\mu^2} + \frac{(1-\mathbb{E}_\mu(\upsilon))^2}{1-\mu} \\
&= \frac{\mu}{\mu^2} + \frac{(1-\mu)^2}{1-\mu} \\
&= \frac{1}{\mu(1-\mu)}.
\end{aligned} \tag{59}
$$

Mit der Additivitätseigenschaft der Fisher-Information und der Definition der beobachteten Fisher-Information ergibt sich dann sofort

$$
I_n(\mu) = \frac{ny}{\mu^2} + \frac{n(1-y)^2}{1-\mu} \text{ und } J_n(\mu) = \frac{n}{\mu(1-\mu)}. \tag{60}
$$

Theorem (Fisher-Information bei Normalverteilung I)

Es sei $\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$ und σ^2 als bekannt vorausgesetzt. Dann gilt:

- Die Scorefunktion der Stichprobe ist

$$S_n : \mathbb{R} \to \mathbb{R}, \mu \mapsto S_n(\mu) := \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu). \tag{61}$$

- Die Fisher-Information der Stichprobe ist

$$I_n : \mathbb{R} \to \mathbb{R}, \mu \mapsto I_n(\mu) := \frac{n}{\sigma^2}. \tag{62}$$

- Die beobachtete Fisher-Information der Stichprobe ist

$$I_n(\hat{\mu}_n^{\text{ML}}) = \frac{n}{\sigma^2}. \tag{63}$$

- Die erwartete Fisher-Information der Stichprobe ist

$$J_n : \mathbb{R} \to \mathbb{R}, \mu \mapsto J_n(\mu) := \frac{n}{\sigma^2}. \tag{64}$$

Beweis

Wir erinnern uns, dass die Log-Likelihood-Funktion der Stichprobe $\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$ bei bekanntem Varianzparameter σ^2 durch

$$\ell_n : \mathbb{R} \to \mathbb{R}, \mu \mapsto \ell_n(\mu) := -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2. \tag{65}$$

gegeben ist. Damit ergibt sich die Scorefunktion als

$$s_n(\mu) = \frac{d}{d\mu}\ell_n(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \mu) \tag{66}$$

Die Fisher-Information der Stichprobe ergibt sich als

$$I_n(\mu) = -\frac{d^2}{d\mu^2}\ell_n(\mu) = -\frac{d}{d\mu}s_n(\mu) = -\frac{1}{\sigma^2}\frac{d}{d\mu}\left(\sum_{i=1}^n y_i - n\mu\right) = \frac{n}{\sigma^2}. \tag{67}$$

Die beobachtete Fisher-Information ist die Fisher-Information an der Stelle des Maximum-Likelihood Schätzes $\hat{\mu}_n^{\text{ML}}$. Die erwartete Fisher-Information schließlich ergibt sich als

$$J_n(\mu) = \mathbb{E}_\mu(I_n(\mu)) = \mathbb{E}_\mu\left(\frac{n}{\sigma^2}\right) = \frac{n}{\sigma^2}. \tag{68}$$

Theorem (Fisher-Information bei Normalverteilung II)

Es sei $\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$ und μ als bekannt vorausgesetzt. Dann gilt:

- die Scorefunktion ist

$$S_n : \mathbb{R}_{>0} \to \mathbb{R}, \sigma^2 \mapsto s_n(\sigma^2) := -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 \tag{69}$$

- die Fisher-Information der Stichprobe ist

$$I_n : \mathbb{R}_{>0} \to \mathbb{R}, \sigma^2 \mapsto I_n(\sigma^2) := \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - \mu)^2 - \frac{n}{2\sigma^4} \tag{70}$$

- die beobachtete Fisher-Information der Stichprobe ist

$$I_n(\hat{\sigma}_n^{2\,\text{ML}}) = \frac{n}{2\hat{\sigma}^4_{\text{ML}}} \tag{71}$$

- die erwartete Fisher-Information der Stichprobe ist

$$J_n : \mathbb{R}_{>0} \to \mathbb{R}, \sigma^2 \mapsto J_n(\sigma^2) := \frac{n}{2\sigma^4}. \tag{72}$$

Beweis

Wir erinnern uns, dass die Log-Likelihood-Funktion der Stichprobe $\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$ bei bekanntem Erwartungswert-Parameter μ durch

$$\ell_n : \mathbb{R}_{>0} \to \mathbb{R}, \sigma^2 \mapsto \ell_n(\sigma^2) := -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2. \tag{73}$$

gegeben ist. Die Scorefunktion ergibt sich also als

$$s_n(\sigma^2) = \frac{d}{d\sigma^2}\ell_n(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (y_i - \mu)^2. \tag{74}$$

Die Fisher-Information der Stichprobe $\upsilon := \upsilon_1, ..., \upsilon_n$ ergibt sich als

$$\begin{aligned} I_n(\sigma^2) = -\frac{d}{d\sigma^2}s_n(\sigma^2) &= -\left(\frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^n (y_i - \mu)^2\right) \\ &= \frac{1}{\sigma^6}\sum_{i=1}^n (y_i - \mu)^2 - \frac{n}{2\sigma^4}. \end{aligned} \tag{75}$$

Beweis (fortgeführt)

Die beobachtete Fisher-Information ist die Fisher-Information an der Stelle des Maximum-Likelihood Schätzes $\hat{\sigma}_n^{2\,\mathsf{ML}}$.

$$
\begin{aligned}
I_n(\hat{\sigma}_n^{2\,\mathsf{ML}}) &= \frac{\sum_{i=1}^n (y_i - \mu)^2}{\left(\hat{\sigma}_n^{2\,\mathsf{ML}}\right)^3} - \frac{n}{2\left(\hat{\sigma}_n^{2\,\mathsf{ML}}\right)^2} \\
&= \frac{\sum_{i=1}^n (y_i - \mu)^2}{\frac{1}{n^3}\left(\sum_{i=1}^n (y_i - \mu)^2\right)^3} - \frac{n}{2\left(\hat{\sigma}_n^{2\,\mathsf{ML}}\right)^2} \\
&= \frac{1}{\frac{1}{n^3}\left(\sum_{i=1}^n (y_i - \mu)^2\right)^2} - \frac{n}{2\left(\hat{\sigma}_n^{2\,\mathsf{ML}}\right)^2} \\
&= \frac{n}{\left(\hat{\sigma}_n^{2\,\mathsf{ML}}\right)^2} - \frac{n}{2\left(\hat{\sigma}_n^{2\,\mathsf{ML}}\right)^2} \\
&= \frac{n}{2\left(\hat{\sigma}_n^{2\,\mathsf{ML}}\right)^2} \\
&= \frac{n}{2\hat{\sigma}_n^{4\,\mathsf{ML}}}.
\end{aligned}
\tag{76}
$$

Beweis (fortgeführt)

Die erwartete Fisher-Information ergibt sich als

$$
\begin{aligned}
J_n(\sigma^2) &= \mathbb{E}_{\sigma^2}(I_n(\sigma^2)) \\
&= \mathbb{E}_{\sigma^2}\left(\frac{1}{\sigma^6}\sum_{i=1}^n (y_i-\mu)^2 - \frac{n}{2\sigma^4}\right) \\
&= \frac{1}{\sigma^6}\sum_{i=1}^n \mathbb{E}_{\sigma^2}\left((y_i-\mu)^2\right) - \frac{n}{2\sigma^4} \\
&= \frac{1}{\sigma^6}\sum_{i=1}^n \sigma^2 - \frac{n}{2\sigma^4} \\
&= \frac{n\sigma^2}{\sigma^6} - \frac{n}{2\sigma^4} \\
&= \frac{n}{\sigma^4} - \frac{n}{2\sigma^4} \\
&= \frac{n}{2\sigma^4}.
\end{aligned}
\tag{77}
$$

Theorem (Cramér-Rao-Ungleichung)

$\mathcal{M}$ sei ein parametrisches statistisches Model mit WMF oder WDF p_θ und $\hat{\tau}_n$ sei ein erwartungstreuer Schätzer von $\tau(\theta)$. Dann gilt

$$\mathbb{V}_\theta(\hat{\tau}_n) \geq \frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{J(\theta)}. \tag{78}$$

Im Speziellen gilt für $\tau(\theta) := \theta$ und somit $\hat{\tau}_n = \hat{\theta}_n$ und $\left(\frac{d}{d\theta}\tau(\theta)\right)^2 = 1$, dass

$$\mathbb{V}_\theta(\hat{\theta}_n) \geq \frac{1}{J(\theta)}. \tag{79}$$

Die rechte Seite obiger Ungleichungen heißt *Cramér-Rao-Schranke*.

Bemerkungen

- Die Varianz eines erwartungstreuen Schätzers $\hat{\theta}$ von θ ist größer oder gleich der reziproken erwarteten Fisher-Information $J(\theta)$.
- Wenn $\mathbb{V}_\theta(\hat{\theta}_n) = \frac{1}{J(\theta)}$ ist, ist die Varianz des Schätzers minimal.

Beweis

Wir halten zunächst fest, dass für die Zufallsvariablen $S(\theta)$ und $\hat{\tau}_n$ mit der Korrelationsungleichung und $\mathbb{V}_\theta(S(\theta)) = J(\theta)$ gilt, dass

$$\begin{aligned}\frac{\mathbb{C}_\theta(S_n(\theta), \hat{\tau}_n)^2}{\mathbb{V}_\theta(S(\theta))\mathbb{V}_\theta(\hat{\tau}_n)} &\le 1 \\ \Leftrightarrow \mathbb{V}_\theta(\hat{\tau}_n) &\ge \frac{\mathbb{C}_\theta(S(\theta), \hat{\tau}_n)^2}{J(\theta)}.\end{aligned} \tag{80}$$

Mit dem Translationstheorem für Kovarianzen, $\mathbb{E}_\theta(S(\theta)) = 0$ und der Erwartungstreue von $\hat{\tau}_n$ ergibt sich dann

$$\mathbb{C}_\theta(S(\theta), \hat{\tau}_n) = \frac{d}{d\theta}\tau(\theta) \tag{81}$$

wie unten gezeigt wird. Also gilt

$$\mathbb{V}_\theta(\hat{\tau}_n) \ge \frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{J(\theta)}. \tag{82}$$

Es bleibt also zu zeigen, dass $\mathbb{C}_\theta(S(\theta), \hat{\tau}_n) = \frac{d}{d\theta}\tau(\theta)$. Dies ergibt aber ergibt sich mit

$$
\begin{aligned}
\mathbb{C}_\theta(S(\theta), \hat{\tau}_n) = \mathbb{E}_\theta(S(\theta)\hat{\tau}_n) - \mathbb{E}_\theta(S(\theta))\mathbb{E}_\theta(\hat{\tau}_n) & = \mathbb{E}_\theta(S(\theta)\hat{\tau}_n) \\
& = \int S(\theta)\, \hat{\tau}_n\, p_\theta(x)\, dx \\
& = \int \frac{d}{d\theta} \ln L(\theta)\, \hat{\tau}_n\, p_\theta(x)\, dx \\
& = \int \frac{\frac{d}{d\theta} L(\theta)}{L(\theta)}\, \hat{\tau}_n\, p_\theta(x)\, dx \\
& = \int \frac{\frac{d}{d\theta} L(\theta)}{p_\theta(x)}\, \hat{\tau}_n\, p_\theta(x)\, dx \\
& = \int \frac{d}{d\theta} L(\theta)\, \hat{\tau}_n\, dx \\
& = \frac{d}{d\theta} \int L(\theta)\, \hat{\tau}_n\, dx \\
& = \frac{d}{d\theta} \int \hat{\tau}_n\, p_\theta(x)\, dx = \frac{d}{d\theta} \mathbb{E}_\theta(\hat{\tau}_n) = \frac{d}{d\theta} \tau(\theta).
\end{aligned}
\tag{83}
$$

Appendix | Asymptotische Effizienz

Definition (Asymptotische Effizienz)

$\upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells $\mathcal{M}$ und $\hat{\theta}_n$ sei ein Parameterschätzer für θ. Weiterhin sei $J_n(\theta)$ die erwartete Fisher-Information der Stichprobe $\upsilon := \upsilon_1, ..., \upsilon_n$. Wenn gilt, dass

$$\hat{\theta}_n \overset{a}{\sim} N\left(\theta, J_n(\theta)^{-1}\right), \tag{84}$$

dann heißt $\hat{\theta}_n$ *asymptotisch effizient*.

Bemerkungen

- Asymptotische Effizienz impliziert asymptotische Normalität.
- Asymptotische Effizienz impliziert asymptotische Erwartungstreue.
- Die Varianz der asymptotischen Verteilung heißt *asymptotische Varianz*.
- Die Varianz eines asymptotisch effizienten Schätzers ist gleich der Cramér-Rao-Schranke.
- Der Begriff der *Effizienz* wird in der Literatur nicht einheitlich verwendet.

Appendix | Konsistenz des Erwartungswertschätzers

Konsistenz des Erwartungswertschätzers bei Normalverteilung

Theorem (Mittlerer-Quadratischer-Fehler-Kriterium für Konsistenz)

$\upsilon := \upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells $\mathcal{M}$ und $\hat{\tau}_n$ sei ein Schätzer von τ. Wenn $\lim_{n\to\infty} \mathsf{MQF}(\hat{\tau}_n) = 0$ gilt, dann ist $\hat{\tau}_n$ ein konsistenter Schätzer.

Beweis

Mit der Chebychev-Ungleichung gilt, dass

$$\mathbb{P}_\theta \left(|\hat{\tau}_n(\upsilon) - \tau(\theta)| \geq \epsilon\right) \leq \frac{\mathbb{E}_\theta\left((\hat{\tau}_n(\upsilon) - \tau(\theta))^2\right)}{\epsilon^2} \tag{85}$$

Grenzwertbildung ergibt dann

$$\lim_{n\to\infty} \mathbb{P}_\theta \left(|\hat{\tau}_n(\upsilon) - \tau(\theta)| \geq \epsilon\right) \leq \frac{1}{\epsilon^2} \lim_{n\to\infty} \mathbb{E}_\theta\left((\hat{\tau}_n(\upsilon) - \tau(\theta))^2\right). \tag{86}$$

Wenn also $\lim_{n\to\infty} \mathbb{E}_\theta\left((\hat{\tau}_n(\upsilon) - \tau(\theta))^2\right) = 0$ gilt, dann gilt mit $\mathbb{P}_\theta(|\hat{\tau}_n(\upsilon) - \tau(\theta)| \geq \epsilon) \geq 0$, dass

$$\lim_{n\to\infty} \mathbb{P}_\theta \left(|\hat{\tau}_n(\upsilon) - \tau(\theta)| \geq \epsilon\right) = 0. \tag{87}$$

Also ist $\hat{\tau}_n$ ein konsistenter Schätzer.

Theorem (Bias-Varianz-Kriterium für Konsistenz)

$\upsilon_1, ..., \upsilon_n \sim p_\theta$ sei die Stichprobe eines parametrischen statistischen Produktmodells $\mathcal{M}$ und $\hat{\tau}_n$ sei ein Schätzer von τ. Wenn

$$\lim_{n\to\infty} \mathrm{B}(\hat{\tau}_n) = 0 \text{ und } \lim_{n\to\infty} \mathbb{V}_\theta(\hat{\tau}_n) = 0 \tag{88}$$

gilt, dann ist $\hat{\tau}_n$ ein konsistenter Schätzer

Beweis

Wenn $n \to \infty$, dann gilt $\mathrm{B}(\hat{\tau}_n) \to 0$, also auch $\mathrm{B}(\hat{\tau}_n)^2 \to 0$. Wenn für $n \to \infty$ sowohl $\mathrm{B}(\hat{\tau}_n)^2 \to 0$ als auch $\mathbb{V}_\theta(\hat{\tau}_n) \to 0$, dann gilt auch $\lim_{n\to\infty} \mathrm{MQF}(\hat{\theta}_n) = 0$. Also gilt mit dem MQF-Kriterium, dass $\hat{\tau}_n$ konsistent ist.

Theorem (Konsistenz des Erwartungswertschätzers bei Normalverteilung)

Es sei $\upsilon := \upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$. Dann ist $\bar{\upsilon}_n$ ein konsistenter Schätzer von $\mathbb{E}(\upsilon_1)$.

Beweis

Mit der Erwartungstreue des Stichprobenmittels als Schätzer für den Erwartungswert gilt zunächst

$$\lim_{n\to\infty} \mathrm{B}(\bar{\upsilon}_n) = 0 \tag{89}$$

Weiterhin gilt mit der Varianz des Stichprobenmittels

$$\lim_{n\to\infty} \mathbb{V}_\theta(\bar{\upsilon}_n) = \lim_{n\to\infty} \frac{1}{n}\mathbb{V}(\upsilon_1) = 0. \tag{90}$$

Mit dem Bias-Varianz-Kriterium folgt dann die Konsistenz von $\bar{\upsilon}_n$ als Schätzer von $\mathbb{E}(\upsilon_1)$

References

Held, Leonhard, and Daniel Sabanés Bové. 2014. *Applied Statistical Inference*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37887-4.

Vaart, A. W. van der. 1998. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK ; New York, NY, USA: Cambridge University Press.