

Kapitel 2

Klassische Schätz- und Testtheorie

Grundmodell:

Die Stichprobe $X = (X_1, \dots, X_n)$ besitzt die Verteilung $\mathbb{P} \in \mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^k$, wobei

- θ : k -dimensionaler Parameter
- Θ : Parameterraum
- $k < n$, oft $k \ll n$, mit $\dim(\theta) = k$ fest für asymptotische ($n \rightarrow \infty$)-Betrachtungen.
- In der Regel vorausgesetzt: Es existiert Dichte

$$f(x|\theta) = f(x_1, \dots, x_n|\theta) \text{ zu } \mathbb{P}_\theta,$$

so dass man analog schreiben kann:

$$\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}.$$

- Klassische Schätz- und Testtheorie für finite (d.h. für festen Stichprobenumfang n) i.i.d.-Stichprobe von besonderer Relevanz; es gilt:

$$f(x|\theta) = f(x_1|\theta) \cdot \dots \cdot f(x_n|\theta).$$

- Viele Begriffe, insbesondere der Schätztheorie, jedoch von genereller Bedeutung.
- Literatur: Lehmann & Casella (1998), Lehmann & Romano (2005), Rüger (1999, 2002) Band I+II

Definition 2.1 (Statistik). *Eine Statistik ist eine messbare Funktion*

$$T : \begin{cases} \mathcal{X} & \longrightarrow \mathbb{R}^l \\ x & \longmapsto T(x). \end{cases}$$

Normalerweise ist $l < n$, da mit der Statistik T eine Dimensionsreduktion erzielt werden soll.

Beispiel 2.1.

→ $T(x)$ Schätzfunktion

→ $T(x)$ Teststatistik

2.1 Klassische Schätztheorie

Gesucht: Punkt- oder Bereichsschätzung für θ oder einen transformierten Parametervektor $\tau(\theta)$.

Beispiel 2.2. $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$ mit $\theta = (\mu, \sigma^2)^\top$. Hier könnte $\tau(\theta) = \mu$ sein (d.h. σ^2 ist Nuisance-Parameter) oder $\tau(\theta) = 1/\sigma^2$ (d.h. die Präzision ist von Interesse).

Definition 2.2 (Punktschätzung, Schätzer, Schätzfunktion). Sei

$$T : \begin{cases} \mathcal{X} & \longrightarrow \Theta \subseteq \mathbb{R}^k \\ x & \longmapsto T(x) \end{cases}$$

eine messbare Abbildung. Man bezeichnet mit $T(x)$ den Schätzwert oder die Punktschätzung (zu konkreter Realisation x) und mit $T(X)$ den Punktschätzer von θ , der eine Zufallsvariable ist (auch gebräuchlich: $\hat{\theta}(x)$ oder kurz $\hat{\theta}$, d.h. notationell wird nicht zwischen Schätzwert und Schätzfunktion unterschieden).

2.1.1 Suffizienz

Der Begriff der Suffizienz ist von grundlegender Bedeutung in der klassischen parametrischen Inferenz; darüber hinaus ist die Bedeutung (stark) abgeschwächt, vgl. auch Statistik IV.

Definition 2.3. Eine Statistik T heißt suffizient für θ (oder auch für \mathcal{P}) $\stackrel{\text{def}}{\iff}$ die bedingte Verteilung bzw. Dichte von X gegeben $T(x) = t$ ist für alle Werte von $T(x) = t$ von θ unabhängig, d.h.

$$f_{X|T}(x|T(x) = t, \theta) = f_{X|T}(x|T(x) = t)$$

hängt nicht von θ ab.

Idee: Zusätzliche Information in X , die nicht in T enthalten ist, ist durch $f_{X|T}$ gegeben. Falls $f_{X|T}$ von θ unabhängig ist, dann enthält die Stichprobe x nicht mehr Information über θ als $T(x)$.

Folgender Satz ist äquivalent und konstruktiv:

Satz 2.4 (Faktorisierungssatz, Neyman-Kriterium). Eine Statistik T ist suffizient für θ genau dann wenn

$$f(x|\theta) = h(x)g(T(x)|\theta)$$

für fast alle x , d.h. die Dichte lässt sich in zwei Teile faktorisieren, von denen ein Teil von x , aber nicht von θ , und der andere nur von θ und $T(x)$ abhängt.

Beweis.

„ \Rightarrow “: Falls T suffizient ist, gilt:

$$f_{X|T}(x|T(x) = t, \theta) = f_{X|T}(x|T(x) = t) = \frac{f_{X,T}(x, t|\theta)}{f_{T|\theta}(t|\theta)}.$$

Weiterhin ist

$$f_{X,T}(x, t|\theta) = \begin{cases} f_{X|\theta}(x|\theta) & \text{für } T(x) = t \\ 0 & \text{sonst,} \end{cases}$$

d.h.

$$\underbrace{f_{X|T}(x|t)}_{h(x)} \cdot \underbrace{f_{T|\theta}(t|\theta)}_{g(T(x)|\theta)} = f_{X|\theta}(x|\theta).$$

„ \Leftarrow “: Man erhält die Dichte von T , ausgewertet an t , indem man im obigen Faktorisierungskriterium über die x , für die $T(x) = t$ gilt, summiert (bzw. integriert). Im diskreten Fall also:

$$f_{T|\theta}(t|\theta) = \sum_{x:T(x)=t} h(x)g(T(x)|\theta) = g(t|\theta) \sum_{x:T(x)=t} h(x).$$

Damit ist die bedingte Dichte von X gegeben $T = t$,

$$\frac{f_{X|\theta}(x|\theta)}{f_{T|\theta}(t|\theta)} = \frac{h(x)g(T(x)|\theta)}{\sum_{x:T(x)=t} h(x)g(t|\theta)} = \frac{h(x)}{\sum_{x:T(x)=t} h(x)},$$

unabhängig von θ . Im stetigen Fall werden Summen durch Integrale ersetzt; im Detail werden Messbarkeitsbedingungen verwendet. \square

Beispiel 2.3 (Bernoulli-Experiment). Seien $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Bin}(1, \pi)$ und $Z = \sum_{i=1}^n X_i$ die Anzahl der Erfolge. Dann ist Z suffizient für π , denn

$$\begin{aligned} f_{X|Z}(x|z, \pi) &= \mathbb{P}_\pi(X = x|Z = z) \\ &= \frac{\prod_{i=1}^n \pi^{x_i} (1 - \pi)^{1-x_i}}{\binom{n}{z} \pi^z (1 - \pi)^{n-z}}, \quad \text{wobei } \sum_{i=1}^n x_i = z \\ &= \binom{n}{z}^{-1} \end{aligned}$$

ist unabhängig von π . Gemäß Faktorisierungssatz ist

$$f(x|\pi) = \underbrace{\frac{1}{\binom{n}{z}}}_{=h(x)} \underbrace{\binom{n}{z} \pi^z (1 - \pi)^{n-z}}_{=g(z|\pi)} = \underbrace{1}_{=h^*(x)} \underbrace{\pi^z (1 - \pi)^{n-z}}_{=g^*(z|\pi)=f(x|\pi)}.$$

Beispiel 2.4 (Normalverteilung). Sei $X = (X_1, \dots, X_n)$ mit $X_i \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$ und $\theta = (\mu, \sigma^2)^\top$.

$$\begin{aligned} f_{X|\theta}(x|\theta) &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right) \\ &= \underbrace{(2\pi)^{-n/2} (\sigma^2)^{-n/2}}_{h(x)} \underbrace{\exp \left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \right) \right)}_{g((\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)|\theta)}, \end{aligned}$$

d.h. $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ ist suffizient für $\theta = (\mu, \sigma^2)^\top$. Aber: Die bijektive Transformation $\tilde{T}(X) = (\bar{X}, S^2)$ ist auch suffizient für θ , wobei S^2 die Stichprobenvarianz bezeichnet.

Beispiel 2.5 (Exponentialverteilung). Sei $X = (X_1, \dots, X_n) \stackrel{i.i.d.}{\sim} \text{Exp}(\lambda)$, dann

$$f(x|\lambda) = \prod_{i=1}^n f(x_i|\lambda) = \underbrace{1}_{h(x)} \cdot \underbrace{\lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right)}_{g(T(x)|\lambda)}$$

mit $T(x) = \sum_{i=1}^n x_i$. Nach der ursprünglichen Definition ist

$$\frac{f_{X,T|\lambda}(x, t|\lambda)}{f_{T|\lambda}(t|\lambda)} = \frac{\lambda^n \exp(-\lambda \sum_{i=1}^n x_i)}{\frac{\lambda^n}{\Gamma(n)} (\sum_{i=1}^n x_i)^{n-1} \exp(-\lambda \sum_{i=1}^n x_i)} = \frac{\Gamma(n)}{(\sum_{i=1}^n x_i)^{n-1}},$$

unabhängig von λ . Dabei wird benutzt, dass die Summe von n unabhängigen und identisch exponentialverteilten Zufallsvariablen mit Parameter λ gammaverteilt ist mit Parametern n und λ .

Beispiel 2.6 (Ordnungsstatistik). Sei $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} f(x|\theta)$ (wobei f stetige Dichte ist) und $T(X) = X_{(\cdot)} = (X_{(1)}, \dots, X_{(n)})$ die Ordnungsstatistik. Dann gilt

$$f_{X|T,\theta}(x|T = x_{(\cdot)}, \theta) = \frac{1}{n!}.$$

Die Gleichheit folgt aus der Stetigkeit, denn $x_i \neq x_j \forall i \neq j$ (mit Wahrscheinlichkeit 1). $X_{(\cdot)}$ ist suffizient für θ . Wir haben also bei i.i.d.-Beobachtungen keinen Informationsverlust durch Ordnen der Daten.

Bemerkung.

- Offensichtlich ist $T(X) = X$, d.h. die Stichprobe selbst, suffizient.
- Ebenso ist jede eindeutige Transformation von X oder von einer suffizienten Statistik $T(X)$ suffizient.
- Ist T suffizient, dann auch (T, T^*) , wobei T^* eine beliebige weitere Statistik darstellt.

Dies zeigt: Die Dimension einer suffizienten Statistik sollte soweit wie möglich reduziert werden.

Definition 2.5 (Minimalsuffizienz). Eine Statistik T heißt *minimalsuffizient* für $\theta \stackrel{\text{def}}{\Leftrightarrow} T$ ist suffizient, und zu jeder anderen suffizienten Statistik V existiert eine Funktion H mit

$$T(X) = H(V(X)) \quad \mathcal{P} - \text{fast überall.}$$

Frage: Existieren minimal-suffiziente Statistiken? Wenn ja, sind sie eindeutig?

Beispiel 2.7 (Normalverteilung).

Seien $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$.

1. $T(X) = \bar{X}$ ist minimal-suffizient für μ bei bekanntem σ^2 .
2. $T(X) = \sum_{i=1}^n (X_i - \mu)^2$ ist minimal-suffizient für σ^2 bei bekanntem μ .

3. $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ ist minimal suffizient für μ und σ^2 .

Lemma 2.6. Sind T und S minimal suffiziente Statistiken, dann existieren injektive Funktionen g_1, g_2 , so dass $T = g_1(S)$ und $S = g_2(T)$.

Satz 2.7 (Charakterisierung von Minimal suffizienz durch Likelihood-Quotienten). Definiere den Likelihood-Quotienten

$$\Lambda_x(\theta_1, \theta_2) = \frac{f(x|\theta_1)}{f(x|\theta_2)}.$$

Eine notwendige und hinreichende Bedingung für die Minimal suffizienz einer Statistik T für θ ist, dass für alle θ_1 und θ_2 mit $\theta_1 \neq \theta_2$ gilt:

$$T(x) = T(x') \Leftrightarrow \Lambda_x(\theta_1, \theta_2) = \Lambda_{x'}(\theta_1, \theta_2).$$

Beispiel 2.8 (Suffizienz in Exponentialfamilien). Die Dichte einer k -parametrischen Exponentialfamilie hat die Form

$$\begin{aligned} f(x|\theta) &= h(x) \cdot c(\theta) \cdot \exp(\gamma_1(\theta)T_1(x) + \dots + \gamma_k(\theta)T_k(x)) \\ &= h(x) \cdot \exp(b(\theta) + \gamma(\theta)^\top \mathbf{T}(x)), \end{aligned}$$

d.h. $\mathbf{T}(X) = (T_1(X), \dots, T_k(X))^\top$ ist suffizient für θ nach Faktorisierungssatz. Falls Θ ein offenes Rechteck in \mathbb{R}^k enthält, ist \mathbf{T} auch minimal suffizient.

Es folgt nun die Charakterisierung der Minimal suffizienz nach Lehmann-Scheffé. Dazu wird der Begriff der Vollständigkeit benötigt.

Definition 2.8. Eine Statistik T ist vollständig $\stackrel{\text{def}}{\Leftrightarrow}$ für jede reelle (messbare) Funktion g gilt:

$$\mathbb{E}_\theta[g(T)] = 0 \quad \forall \theta \Rightarrow \mathbb{P}_\theta(g(T) = 0) = 1 \quad \forall \theta.$$

Aus der Definition wird nicht unmittelbar klar, warum „Vollständigkeit“ eine wünschenswerte Eigenschaft eines Schätzers sein sollte. Einen möglichen Grund liefert der folgende Satz.

Satz 2.9 (Lehmann-Scheffé). Angenommen, X besitzt eine Dichte $f(x|\theta)$ und $T(X)$ ist suffizient und vollständig für θ . Dann ist $T(X)$ minimal suffizient für θ .

Beweis. Vorausgesetzt wird, dass eine minimal suffiziente Statistik existiert - bewiesen wurde dies von Lehmann und Scheffé (1950). Ist dies der Fall, so ist diese bis auf bijektive Transformationen eindeutig. Bezeichne $S = g_1(T)$ eine solche minimal suffiziente Statistik für eine Funktion g_1 . Definiere $g_2(S) = \mathbb{E}[T|S]$. Da S suffizient für θ ist, hängt $g_2(S)$ nicht von θ ab. Betrachte nun

$$g(T) = T - g_2(S) = T - g_2(g_1(T)).$$

Anwendung des Satzes von der iterierten Erwartung liefert:

$$\mathbb{E}_\theta[g(T)] = \mathbb{E}_\theta[T] - \mathbb{E}_\theta[\mathbb{E}[T|S]] = 0.$$

Da T vollständig ist, ist $g(T) = 0$ bzw. $g_2(S) = T$ mit Wahrscheinlichkeit 1, d.h. S ist eine Funktion von T . Wähle x und x' , so dass $g_2(S(x)) = g_2(S(x'))$, dann ist $T(x) = T(x')$ und $S(x) = g_1(T(x)) = g_1(T(x')) = S(x')$, d.h. g_1 und g_2 sind injektiv, so dass T und S äquivalent. Somit ist also auch T minimal suffizient. \square

Bemerkung (Ancillarity einer Statistik). Eine Statistik $V(X)$ heißt ancillary („Hilfsstatistik“) für \mathcal{P} , wenn ihre Verteilung nicht von θ abhängt (also bekannt ist).

Häufiger Sachverhalt: $T = (U, V)$ ist suffizient für θ , V ancillary, U nicht suffizient.

Beispiel 2.9. $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. Man kann dann zeigen (Davison, 2004), dass mit

$$\begin{aligned} U &= \frac{1}{2}(X_{(1)} + X_{(n)}) \\ V &= X_{(n)} - X_{(1)} \end{aligned}$$

$T = (U, V)$ suffizient, aber nicht vollständig für θ ist. Ferner ist U alleine nicht suffizient und V ancillary.

2.1.2 Erwartungstreue, Varianz und MSE

- Fehler eines Schätzers $\hat{\theta} = \hat{\theta}(X)$ ist $\hat{\theta} - \theta$.
- Messung des Fehlers durch Verlustfunktion, zum Beispiel

$$\begin{aligned} L(\hat{\theta}, \theta) &= |\hat{\theta} - \theta| && \text{Abstand } (\theta \text{ skalar}), \\ L(\hat{\theta}, \theta) &= \|\hat{\theta} - \theta\|^2 && \text{quadratischer Fehler,} \\ L(\hat{\theta}, \theta) &= \frac{\|\hat{\theta} - \theta\|^2}{\|\theta\|^2} && \text{relativer quadratischer Fehler,} \\ L(\hat{\theta}, \theta) &= (\hat{\theta} - \theta)^\top \mathbf{D}(\hat{\theta} - \theta) && \text{gewichteter quadratischer Fehler } (\mathbf{D} \text{ positiv definit}). \end{aligned}$$

- Risikofunktion $R(\hat{\theta}, \theta) = \mathbb{E}_\theta[L(\hat{\theta}, \theta)]$.
- Hier wird (hauptsächlich) quadratischer Verlust betrachtet.

Definition 2.10 (Erwartungstreue, Bias, Varianz eines Schätzers).

- $\hat{\theta}$ heißt erwartungstreu $\stackrel{def}{\iff} \mathbb{E}_\theta[\hat{\theta}] = \theta$.
- $\text{Bias}_\theta(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta$.
- $\text{Var}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2]$, θ skalar.

Definition 2.11 (MSE). Der mittlere quadratische Fehler (mean squared error) ist definiert als

$$\text{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = \text{Var}_\theta(\hat{\theta}) + (\text{Bias}_\theta(\hat{\theta}))^2.$$

Der Gesamtfehler lässt sich also aufteilen in einen zufälligen Fehler (Varianz) und einen systematischen (quadratischer Bias).

Vergleicht man zwei Schätzer bezüglich ihres MSE, kann für einen Teilbereich von Θ der MSE des einen, für andere Teilbereiche der MSE des zweiten Schätzers kleiner sein:

Beispiel 2.10. $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} B(1, \pi)$.

1. MSE von $\hat{\pi} = \bar{X}$:

$$\mathbb{E}_\pi[(\bar{X} - \pi)^2] = \text{Var}_\pi(\bar{X}) = \frac{\pi(1 - \pi)}{n}.$$

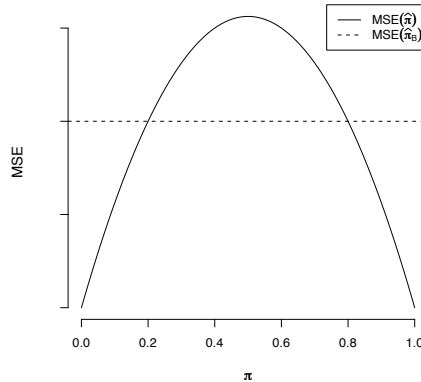
2. MSE des Bayes-Schätzers (Posteriori-Erwartungswert) bei einer Priori $p(\pi) \sim \text{Be}(\alpha, \beta)$:

$$\hat{\pi}_B = \frac{Y + \alpha}{\alpha + \beta + n}, \quad Y = \sum_{i=1}^n X_i,$$

$$\begin{aligned} \text{MSE}(\hat{\pi}_B) &= \text{Var}_\pi\left(\frac{Y + \alpha}{\alpha + \beta + n}\right) + \left(\mathbb{E}_\pi\left[\frac{Y + \alpha}{\alpha + \beta + n} - \pi\right]\right)^2 \\ &= \frac{n\pi(1 - \pi)}{(\alpha + \beta + n)^2} + \left(\frac{n\pi + \alpha}{\alpha + \beta + n} - \pi\right)^2. \end{aligned}$$

Für $\alpha = \beta = \sqrt{n/4}$ ergibt sich

$$\text{MSE}_\pi(\hat{\pi}_B) = \mathbb{E}_\pi[(\hat{\pi}_B - \pi)^2] = \frac{1}{4} \frac{n}{(n + \sqrt{n})^2} = \text{const bezüglich } \pi.$$



Fazit: In der Regel wird man keinen „MSE-optimalen“ Schätzer $\hat{\theta}^{\text{opt}}$ finden in dem Sinne, dass $\text{MSE}_\theta(\hat{\theta}^{\text{opt}}) \leq \text{MSE}_\theta(\hat{\theta})$ für alle θ und alle konkurrierenden $\hat{\theta}$. Bei Einschränkung auf erwartungstreue Schätzer ist dies öfter möglich. Deshalb die Forderung:

Definition 2.12 (zulässiger („admissible“) Schätzer). Ein Schätzer $\hat{\theta}$ heißt zulässig $\stackrel{\text{def}}{\iff}$ es gibt keinen Schätzer $\tilde{\theta}$ mit $\text{MSE}_\theta(\tilde{\theta}) \leq \text{MSE}_\theta(\hat{\theta})$ für alle θ und $\text{MSE}_\theta(\tilde{\theta}) < \text{MSE}_\theta(\hat{\theta})$ für mindestens ein θ , d.h. es gibt keinen Schätzer $\tilde{\theta}$, der $\hat{\theta}$ gleichmäßig/strikt „dominiert“.

Definition 2.13 (Verallgemeinerungen des MSE auf $\theta \in \mathbb{R}^p, p > 1$). *Üblich sind die folgenden zwei Alternativen:*

1. *MSE (skalar):*

$$\begin{aligned} \text{MSE}_\theta^{(1)}(\hat{\theta}) &= \mathbb{E}_\theta[\|\hat{\theta} - \theta\|^2] \\ &= \sum_{j=1}^p \mathbb{E}_\theta[(\hat{\theta}_j - \theta_j)^2] \\ &= \sum_{j=1}^p \text{MSE}_\theta(\hat{\theta}_j) \end{aligned}$$

2. *MSE-Matrix:*

$$\begin{aligned} \text{MSE}_\theta^{(2)}(\hat{\theta}) &= \mathbb{E}_\theta[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^\top] \\ &= \text{Cov}_\theta(\hat{\theta}) + (\mathbb{E}_\theta[\hat{\theta}] - \theta)(\mathbb{E}_\theta[\hat{\theta}] - \theta)^\top \end{aligned}$$

Diese Variante wird häufig bei linearen Modellen betrachtet.

Bemerkung. *Das j -te Diagonalelement der MSE-Matrix ist $\text{MSE}_\theta(\hat{\theta}_j)$. Vergleich von MSE-Matrizen gemäß „Löwner“-Ordnung:*

$$\text{MSE}_\theta(\tilde{\theta}) \stackrel{(\leq)}{<} \text{MSE}_\theta(\hat{\theta})$$

bedeutet, dass die Differenz $\text{MSE}_\theta(\hat{\theta}) - \text{MSE}_\theta(\tilde{\theta})$ positiv (semi-)definit ist. Man definiert allgemein für symmetrische $(p \times p)$ -Matrizen \mathbf{A}, \mathbf{B} :

$$\begin{aligned} \mathbf{A} \leq \mathbf{B} &\stackrel{\text{def}}{\Leftrightarrow} \mathbf{B} - \mathbf{A} \text{ ist positiv semidefinit,} \\ \mathbf{A} < \mathbf{B} &\stackrel{\text{def}}{\Leftrightarrow} \mathbf{B} - \mathbf{A} \text{ ist positiv definit.} \end{aligned}$$

Beispiel 2.11 (Gauß-Experiment). *Seien $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$.*

- σ^2 bekannt, μ unbekannt: *MSE-Vergleich von \bar{X} und $T = b\bar{X} + a$.*

- σ^2 unbekannt, μ bekannt:

- *Eine Möglichkeit:*

$$S_\mu^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2, \mathbb{E}_{\sigma^2}(S_\mu^2) = \sigma^2$$

- *Weitere Möglichkeit:*

$$V_\mu^2 = \frac{1}{n+2} \sum_{i=1}^n (X_i - \mu)^2, \mathbb{E}_{\sigma^2}(V_\mu^2) = \frac{n}{n+2} \sigma^2$$

Es stellt sich heraus, dass $\text{MSE}_{\sigma^2}(V_\mu^2) < \text{MSE}_{\sigma^2}(S_\mu^2)$ ist.

- μ und σ^2 unbekannt:

– Eine Möglichkeit:

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

$$\mathbb{E}_{\sigma^2}(S^2) = \sigma^2, \quad \text{MSE}_{\sigma^2}(S^2) = \text{Var}_{\sigma^2}(S^2) = \frac{2}{n-1} \sigma^4.$$

– Weitere Möglichkeit:

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

$$\mathbb{E}_{\sigma^2}(V^2) = \frac{n-1}{n+1} \sigma^2, \quad \text{MSE}_{\sigma^2}(V^2) = \frac{2}{n+1} \sigma^4,$$

d.h. V^2 dominiert S^2 .

– Der sogenannte Stein-Schätzer

$$T = \min \left\{ V^2, \frac{1}{n+2} \sum_{i=1}^n X_i^2 \right\}$$

dominiert V^2 (und damit S^2). Plausibilitätsbetrachtung: Ist $\mu = 0$, so ist $\sum_{i=1}^n X_i^2 / (n+2)$ besserer Schätzer als V^2 . Ist $\mu \neq 0$, so ist V^2 ein besserer Schätzer als $\sum_{i=1}^n X_i^2 / (n+2)$. Beim Stein-Schätzer wird fallweise mit hoher Wahrscheinlichkeit der jeweils bessere Schätzer benutzt.

Beispiel 2.12 (Steins Paradoxon). Seien $(X_1, \dots, X_m)^\top \sim N_m(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_m)$ multivariat normalverteilt mit $\boldsymbol{\mu} = (\mu_1, \dots, \mu_m)^\top$ und σ^2 bekannt. Es sollen simultan die Erwartungswerte μ_1, \dots, μ_m geschätzt werden. Man beachte dabei, dass die einzelnen Komponenten als unabhängig angenommen werden. Die Stichprobe hat die Form

$$X_{11}, \dots, X_{1n_1}, \dots, X_{m1}, \dots, X_{mn_m}$$

(i.i.d. Stichproben aus „Gruppen“ $1, \dots, m$). Übliche Schätzer:

$$T_j = \bar{X}_j, \quad j = 1, \dots, m, \quad \mathbf{T} = (T_1, \dots, T_m)^\top = (\bar{X}_1, \dots, \bar{X}_m)^\top.$$

Der (skalare) MSE ist:

$$\mathbb{E}_{\boldsymbol{\mu}}[\|\mathbf{T} - \boldsymbol{\mu}\|^2] = \sum_{j=1}^m \mathbb{E}_{\mu_j}[(\bar{X}_j - \mu_j)^2] = \sum_{j=1}^m \frac{\sigma_j^2}{n_j}.$$

Paradoxe Weise gilt:

1. Für $m \leq 2$ ist \mathbf{T} zulässig.
2. Für $m \geq 3$ ist \mathbf{T} nicht zulässig und wird dominiert durch den Stein-Schätzer

$$\mathbf{T}^* = \left(1 - \frac{(m-2)\sigma^2}{\sum_{j=1}^m n_j \bar{X}_j^2} \right) \mathbf{T}.$$

Es lässt sich zeigen: T^* ist selbst unzulässig. Der Stein-Schätzer ist ein sogenannter Shrinkage-Schätzer.

Beispiel 2.13 (Lineares Modell).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim (N)(\mathbf{0}, \sigma^2 \mathbf{I})$$

$$\text{KQ-Schätzer:} \quad \hat{\boldsymbol{\beta}}_{KQ} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

$$\text{Ridge-Schätzer:} \quad \hat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{D})^{-1} \mathbf{X}^\top \mathbf{y},$$

wobei \mathbf{D} eine Diagonalmatrix mit positiven Diagonalelementen ist. Für einen MSE-Vergleich siehe Übung.

Fazit: Bereits im einfachen Beispiel der Schätzung von π in $B(1, \pi)$ (siehe Beispiel 2.10) zeigt sich, dass es im Allgemeinen keine MSE-optimalen Schätzer gibt.

Auswege:

1. Einschränkung auf Teilklasse von Schätzern, zum Beispiel erwartungstreu (und lineare) Schätzer, äquivalente Schätzer, ...
2. MSE-Kriterium verändern:
 - Ersetze $\text{MSE}_\theta(\hat{\theta})$ durch Minimierung von $\max_{\theta \in \Theta} \text{MSE}_\theta(\hat{\theta})$ (Minimax-Kriterium)
 - oder ersetze $\text{MSE}_\theta(\hat{\theta})$ durch $\mathbb{E}_{p(\theta)}[\text{MSE}_\theta(\hat{\theta})]$ bei einer Priori-Verteilung $p(\theta)$ (Bayes-Schätzer).

Hier: Strategie 1 mit erwartungstreuen Schätzern, vgl. 2.1.4.

2.1.3 Fisher-Information und Suffizienz

Definition 2.14 (Fisher-reguläre Verteilungsfamilien). Eine Familie von Verteilungen \mathcal{P}_θ mit Dichte $f(x|\theta) = f(x_1, \dots, x_n|\theta)$, $\theta \in \Theta$, heißt Fisher-regulär, wenn Folgendes gilt:

1. Der Träger $\{x \in \mathcal{X} : f(x|\theta) > 0\}$ ist unabhängig von θ (dies ist zum Beispiel bei $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} U[0; \theta]$ oder bei der Pareto-Verteilung verletzt).
2. Θ ist offen in \mathbb{R}^p (verletzt zum Beispiel bei $\sigma^2 \geq 0$).
3. Die ersten und zweiten Ableitungen von $f(x|\theta)$ bzgl. θ existieren und sind für jedes θ endliche Funktionen von x .
4. Vertauschbarkeit: Sowohl für $f(x|\theta)$ als auch für $\log(f(x|\theta))$ kann erstes und zweites Differenzieren nach θ und Integration über x vertauscht werden.

Definition 2.15 (Log-Likelihood, Scorefunktion und Information).

$$\ell(\theta; x) = \log f(x|\theta) \quad (\text{Log-Likelihood von } \theta \text{ bzgl. der Stichprobe } x)$$

$$s(\theta; x) = \frac{\partial}{\partial \theta} \ell(\theta; x) = \left(\frac{\partial}{\partial \theta_1} \ell(\theta; x), \dots, \frac{\partial}{\partial \theta_p} \ell(\theta; x) \right)^\top \quad (\text{Score-Funktion})$$

$$J(\theta; x) = -\frac{\partial^2 \ell(\theta; x)}{\partial \theta \partial \theta^\top} \quad (\text{beobachtete Informationsmatrix der Stichprobe mit Elementen})$$

$$(J(\theta; x))_{ij} = -\frac{\partial^2 \log f(x|\theta)}{\partial \theta_i \partial \theta_j}$$

$$I(\theta) = \mathbb{E}_\theta [J(\theta; X)] \quad (\text{erwartete oder Fisher-Informationsmatrix})$$

Satz 2.16. Ist \mathcal{P}_θ Fisher-regulär, so gilt:

1. $\mathbb{E}_\theta [s(\theta; X)] = 0$
2. $\mathbb{E}_\theta \left[-\frac{\partial^2 \ell(\theta; X)}{\partial \theta \partial \theta^\top} \right] = \text{Cov}_\theta(s(\theta; X))$

Beweis.

Zu 1.:

$$\begin{aligned} \mathbb{E}_\theta [s(\theta; X)] &= \int s(\theta; x) f(x|\theta) dx \\ &= \int \frac{\partial}{\partial \theta} \log(f(x|\theta)) f(x|\theta) dx \\ &= \int \frac{\frac{\partial}{\partial \theta} f(x|\theta)}{f(x|\theta)} f(x|\theta) dx \\ &= \frac{\partial}{\partial \theta} \int f(x|\theta) dx = 0 \end{aligned}$$

Zu 2.:

$$\begin{aligned} \mathbb{E}_\theta \left[-\frac{\partial^2 \ell(\theta; X)}{\partial \theta \partial \theta^\top} \right] &= -\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \left(\frac{\frac{\partial}{\partial \theta^\top} f(X|\theta)}{f(X|\theta)} \right) \right] \\ &= -\mathbb{E}_\theta \left[\frac{f(X|\theta) \frac{\partial^2}{\partial \theta \partial \theta^\top} f(X|\theta) - (\frac{\partial}{\partial \theta} f(X|\theta)) (\frac{\partial}{\partial \theta^\top} f(X|\theta))}{f(X|\theta)^2} \right] \end{aligned}$$

unter Verwendung der Quotientenregel der Differentiation. Dies ist gleich

$$\begin{aligned} &- \mathbb{E}_\theta \left[\frac{\frac{\partial^2}{\partial \theta \partial \theta^\top} f(X|\theta)}{f(X|\theta)} \right] + \mathbb{E}_\theta \left[\frac{\frac{\partial}{\partial \theta} f(X|\theta)}{f(X|\theta)} \cdot \frac{\frac{\partial f(X|\theta)}{\partial \theta^\top}}{f(X|\theta)} \right] \\ &= - \int \frac{\partial^2}{\partial \theta \partial \theta^\top} f(x|\theta) dx + \mathbb{E}_\theta [s(\theta; X) s(\theta; X)^\top] \end{aligned}$$

Der erste Summand ist unter Vertauschung von Differentiation und Integration gleich null. Für den zweiten Teil ergibt sich mit Teil 1.

$$\mathbb{E}[s(\theta; X) s(\theta; X)^\top] = \text{Cov}_\theta(s(\theta; X)).$$

□

Weitere Eigenschaften:

- Sind X_1, \dots, X_n unabhängig und gemäß $X_i \sim f_i(x|\theta)$, $i = 1, \dots, n$, verteilt, so gilt:

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^n \ell_i(\theta) \quad , \quad \ell_i(\theta) = \log f_i(x_i|\theta) \\ s(\theta) &= \sum_{i=1}^n s_i(\theta) \quad , \quad s_i(\theta) = \frac{\partial}{\partial \theta} \log f_i(x_i|\theta) \\ J(\theta) &= -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^\top} = \sum_{i=1}^n -\frac{\partial^2 \log f_i(x_i|\theta)}{\partial \theta \partial \theta^\top}\end{aligned}$$

- Für X_1, \dots, X_n i.i.d. wie $X_1 \sim f_1(x|\theta)$ folgt

$$I(\theta) = \mathbb{E}_\theta[J(\theta)] = n \cdot i(\theta),$$

wobei

$$i(\theta) = \mathbb{E}_\theta \left[-\frac{\partial^2 \ell_1(\theta; X)}{\partial \theta \partial \theta^\top} \right] = \text{Cov}_\theta \left(\frac{\partial \log f_1(X|\theta)}{\partial \theta} \right)$$

die erwartete Information einer Einzelbeobachtung ist, d.h. die erwartete Informationsmatrix der Stichprobe X_1, \dots, X_n ist die n -fache erwartete Information einer (typischen) Stichprobenvariable X_1 .

- Für eine Statistik $T = T(X)$, $X = (X_1, \dots, X_n)^\top$ mit $T \sim f_T(t|\theta)$ kann man die Begriffe Score-Funktion und Fisher-Information völlig analog definieren. Insbesondere ist

$$I_T(\theta) = \mathbb{E}_\theta \left[-\frac{\partial^2 \log f_T(t|\theta)}{\partial \theta \partial \theta^\top} \right].$$

Satz 2.17 (Suffizienz und Fisher-Information). *Sei $I(\theta)$ die Fisher-Information für X . Dann gilt unter Fisher-Regularität für jede Statistik T :*

1. $I_T(\theta) \leq I(\theta)$.
2. $I_T(\theta) = I(\theta) \Leftrightarrow T$ ist suffizient für θ .

Also: Bei einer suffizienten Statistik T wird keine (erwartete) Information „verschenkt“.

2.1.4 Erwartungstreue Schätzer

- „Schöne“ Resultate für finites n , aber für vergleichsweise einfache statistische Modelle.
- Problem: Für komplexere Modelle existieren keine „vernünftigen“ erwartungstreuen Schätzer.
- Aber: Etliche Resultate besitzen allgemeine Eigenschaften für $n \rightarrow \infty$.

Informationsungleichungen

I. $\theta \in \mathbb{R}$ (skalar). Neben θ werden auch transformierte Parameter $\tau(\theta)$ betrachtet. Wenn Ableitungen benötigt werden, nehmen wir stillschweigend an, dass sie existieren.

Satz 2.18. Sei $f(x|\theta)$ Fisher-regulär.

1. Ist $\hat{\theta}$ erwartungstreu für θ , so gilt:

$$\text{Var}_\theta(\hat{\theta}) \geq \frac{1}{I(\theta)} \quad (\text{Cramer-Rao-Ungleichung}).$$

2. Ist $T = T(x)$ erwartungstreu für $\tau(\theta)$, so gilt:

$$\text{Var}_\theta(T) \geq \frac{(\tau'(\theta))^2}{I(\theta)}.$$

$\frac{(\tau'(\theta))^2}{I(\theta)}$ heißt Cramer-Rao-Schranke.

3. Besitzt $\hat{\theta}$ den Bias $B(\theta) = \mathbb{E}_\theta[\hat{\theta}] - \theta$, so gilt

$$\text{MSE}_\theta(\hat{\theta}) \geq B^2(\theta) + \frac{(1 + B'(\theta))^2}{I(\theta)}.$$

Beweis. Gezeigt wird 2. Daraus folgt 1. für $\tau(\theta) = \theta$ und 3. für $\tau(\theta) = \theta + B(\theta)$.
Differentiation von

$$\tau(\theta) = \mathbb{E}_\theta[T] = \int T(x)f(x|\theta) dx$$

bezüglich θ , und Verwendung der Fisher-Regularität liefert:

$$\begin{aligned} \tau'(\theta) &= \int T(x) \frac{d}{d\theta} f(x|\theta) dx \\ &= \int T(x) s(\theta; x) f(x|\theta) dx \\ &= \text{Cov}_\theta(T(X), s(\theta; X)). \end{aligned}$$

Unter Verwendung der Cauchy-Schwarz-Ungleichung

$$|\text{Cov}(U, V)| \leq \sqrt{\text{Var}(U)} \sqrt{\text{Var}(V)}$$

folgt

$$\begin{aligned} (\tau'(\theta))^2 &\leq \text{Var}_\theta(T(X))\text{Var}_\theta(s(\theta; X)) \\ &= \text{Var}_\theta(T(X))I(\theta). \end{aligned}$$

Also:

$$\text{Var}_\theta(T(X)) \geq \frac{(\tau'(\theta))^2}{I(\theta)}.$$

□

Bemerkung. Die Gleichheit wird genau dann angenommen, wenn eine einparametrische Exponentialfamilie $f(x|\theta) = h(x) \exp(b(\theta) + \gamma(\theta)T(x))$ vorliegt. In diesem Fall ist $T(x)$ ein effizienter Schätzer für seinen Erwartungswert $\tau(\theta) = -b'(\theta)/\gamma'(\theta)$. Also: eher eine kleine Modellklasse.

II. $\theta = (\theta_1, \dots, \theta_p)$ bzw. $\tau(\theta)$ mehrdimensional.

Satz 2.19. Sei $f(x|\theta)$ Fisher-regulär.

1. Ist $\hat{\theta}$ erwartungstreu für θ , so gilt:

$$\text{Cov}_\theta(\hat{\theta}) \geq \mathbf{I}^{-1}(\theta),$$

wobei sich das „ \geq “ auf die Löwner-Ordnung bezieht (vergleiche Seite 27). Daraus folgt insbesondere $\text{Var}_\theta(\hat{\theta}_j) \geq v_{jj}$, $j = 1, \dots, p$, wobei v_{jj} das j -te Diagonalelement von $\mathbf{I}^{-1}(\theta)$ bezeichnet.

2. Ist \mathbf{T} erwartungstreu für $\tau(\theta)$, so gilt

$$\text{Cov}_\theta(\mathbf{T}) \geq \mathbf{H}(\theta)\mathbf{I}^{-1}(\theta)\mathbf{H}(\theta)^\top$$

mit der Funktionalmatrix $(\mathbf{H}(\theta))_{ij} = \frac{\partial}{\partial \theta_j} \tau_i(\theta)$. Die Matrix $\mathbf{H}(\theta)\mathbf{I}^{-1}(\theta)\mathbf{H}(\theta)^\top$ ist die Cramer-Rao-Schranke.

Bemerkung. Obige Bemerkung für skalares θ gilt analog für

$$f(x|\theta) = h(x) \exp(b(\theta) + \gamma^\top(\theta)\mathbf{T}(x)),$$

d.h. für mehrparametrische Exponentialfamilien.

Beispiel 2.14 (Cramer-Rao-Schranke bei $X \sim N(\mu, \sigma^2)$). X_1, \dots, X_n i.i.d. wie $X \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$. Dann gilt für die Informationsmatrix

$$I(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \quad \text{bzw.} \quad I^{-1}(\theta) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}.$$

Beste erwartungstreue Schätzer

Erwartungstreue Schätzer minimaler Varianz innerhalb einer vorgegebenen Klasse nennt man *effizient*. Die Informationsungleichungen motivieren:

Definition 2.20 (Gleichmäßig bester erwartungstreuer (UMVU) Schätzer).

1. θ skalar:

Der Schätzer $\hat{\theta}_{\text{eff}}$ für θ heißt gleichmäßig bester erwartungstreuer oder UMVU („uniformly minimum variance unbiased“)-Schätzer $\stackrel{\text{def}}{\Leftrightarrow} \hat{\theta}_{\text{eff}}$ ist erwartungstreu, und es gilt $\text{Var}_{\theta}(\hat{\theta}_{\text{eff}}) \leq \text{Var}_{\theta}(\hat{\theta})$ für alle θ und jeden erwartungstreuen Schätzer $\hat{\theta}$.

2. θ mehrdimensional:

Ersetze in 1. $\text{Var}_{\theta}(\hat{\theta}_{\text{eff}}) \leq \text{Var}_{\theta}(\hat{\theta})$ durch $\text{Cov}_{\theta}(\hat{\theta}_{\text{eff}}) \leq \text{Cov}_{\theta}(\hat{\theta})$.

Satz 2.21 (Effizienz und Informationsungleichungen). Sei $f(x|\theta)$ Fisher-regulär und $\hat{\theta}$ erwartungstreu für θ . Falls $\text{Cov}_{\theta}(\hat{\theta}) = I^{-1}(\theta)$ für alle θ , so ist $\hat{\theta}$ ein UMVU-Schätzer.

Beweis. Die Aussage folgt direkt aus der Informationsungleichung und obiger Definition. \square

Beispiel 2.15 (Gauß-Experiment). Seien $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$ mit μ, σ^2 unbekannt. Aus [Beispiel 2.14](#) wissen wir, dass $I(\mu) = n/\sigma^2$ und somit $I^{-1}(\mu) = \sigma^2/n = \text{Var}(\bar{X})$. Dann ist \bar{X} UMVU für μ . Aber

$$\text{Var}(S^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n} = I^{-1}(\sigma^2).$$

Die Cramer-Rao-Schranke wird also nicht erreicht, somit kann nicht gefolgert werden, dass S^2 UMVU für σ^2 ist.

Beispiel 2.16 (Lineares Modell).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I) \quad \text{bzw.} \quad \mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$$

$$\hat{\boldsymbol{\beta}}_{KQ} = \hat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}^{\top} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbf{y} \text{ ist effizient für } \boldsymbol{\beta},$$

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \text{ ist UMVU-Schätzer für } \sigma^2.$$

Bemerkung. Zu unterscheiden sind folgende Situationen:

1. Es existiert ein UMVU-Schätzer, dessen Varianz gleich der Cramer-Rao-Schranke ist.
2. Es existiert ein UMVU-Schätzer, dessen Varianz größer als die Cramer-Rao-Schranke ist (findet man mit dem Satz von Lehmann-Scheffé, siehe [Satz 2.23](#)).
3. Der häufigste Fall: Es existiert (für finiten Stichprobenumfang) kein UMVU-Schätzer.

Fazit: Finite Theorie erwartungstreuer Schätzer ist von eingeschränkter Anwendungsrelevanz.

Aber: Es existiert eine analoge asymptotische Theorie mit breiter Anwendungsrelevanz, die sich an finiter Theorie orientiert (siehe [Abschnitt 2.1.5](#)).

Zur Konstruktion von UMVU-Schätzern sind folgende zwei Aussagen nützlich:

Satz 2.22 (Rao-Blackwell). Sei $T = T(X)$ suffizient für θ bzw. \mathcal{P}_θ und $\hat{\theta}$ erwartungstreu für θ . Für den Schätzer

$$\hat{\theta}_{RB} = \mathbb{E}_\theta[\hat{\theta}|T] \quad (\text{„Rao-Blackwellization“})$$

gilt:

1. $\hat{\theta}_{RB}$ ist erwartungstreu für θ .
2. $\text{Var}_\theta(\hat{\theta}_{RB}) \leq \text{Var}_\theta(\hat{\theta})$.
3. In 2. gilt die Gleichheit, wenn $\hat{\theta}$ nur von T abhängt, d.h. $\hat{\theta}_{RB} = \hat{\theta}$ mit Wahrscheinlichkeit 1.

Satz 2.23 (Lehmann-Scheffé). Ist $T = T(X)$ suffizient und vollständig (also minimalsuffizient) und $\hat{\theta} = \hat{\theta}(x)$ ein erwartungstreuer Schätzer, so ist

$$\hat{\theta}^* = \mathbb{E}_\theta[\hat{\theta}|T]$$

der mit Wahrscheinlichkeit 1 eindeutig bestimmte UMVU-Schätzer für θ .

2.1.5 Asymptotische Eigenschaften und Kriterien

Wichtige Schätzer (Momentenschätzer, Shrinkage-Schätzer, ML- und Quasi-ML-Schätzer etc.) sind im Allgemeinen nicht erwartungstreu, besitzen aber günstige asymptotische ($n \rightarrow \infty$) Eigenschaften. Im Folgenden sei

$$\hat{\theta}_n = \hat{\theta}(X_1, \dots, X_n)$$

Schätzer für θ .

Definition 2.24 (Asymptotische Erwartungstreue). $\hat{\theta}_n$ heißt asymptotisch erwartungstreu $\stackrel{\text{def}}{\Leftrightarrow}$

$$\lim_{n \rightarrow \infty} \mathbb{E}_\theta[\hat{\theta}_n] = \theta \quad \text{für alle } \theta.$$

Definition 2.25 (Konsistenz).

1. $\hat{\theta}_n$ ist (schwach) konsistent für θ (in Zeichen: $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ (für alle θ)) $\stackrel{\text{def}}{\Leftrightarrow}$

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta(|\hat{\theta}_n - \theta| \leq \varepsilon) = 1 \quad \text{für alle } \varepsilon > 0 \text{ und alle } \theta.$$

2. $\hat{\theta}_n$ heißt MSE-konsistent für θ $\stackrel{\text{def}}{\Leftrightarrow}$

$$\lim_{n \rightarrow \infty} \text{MSE}_\theta(\hat{\theta}_n) = 0 \quad \text{für alle } \theta.$$

3. $\hat{\theta}_n$ ist stark konsistent für θ $\stackrel{\text{def}}{\Leftrightarrow}$

$$\mathbb{P}_\theta \left(\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta \right) = 1 \quad \text{für alle } \theta.$$

Bemerkung.

1. Aus der (verallgemeinerten) Tschebyscheff-Ungleichung folgt

$$\widehat{\theta}_n \text{ MSE-konsistent} \Rightarrow \widehat{\theta}_n \text{ schwach konsistent.}$$

2. Wegen $\text{MSE}_\theta(\widehat{\theta}_n) = \text{Var}_\theta(\widehat{\theta}_n) + (\text{Bias}_\theta(\widehat{\theta}_n))^2$ folgt:

$$\widehat{\theta}_n \text{ ist MSE-konsistent} \Leftrightarrow \text{Var}_\theta(\widehat{\theta}_n) \rightarrow 0 \text{ und } \text{Bias}_\theta(\widehat{\theta}_n) \rightarrow 0 \text{ f\u00fcr alle } \theta.$$

3. Ist $\widehat{\theta}_n$ konsistent f\u00fcr θ und g eine stetige Abbildung, so ist auch $g(\widehat{\theta}_n)$ konsistent f\u00fcr $g(\theta)$ (Continuous Mapping Theorem/Stetigkeitssatz).

4. Konsistenznachweise bestehen in der Regel in der Anwendung (schwacher) Gesetze gro\u00dfer Zahlen (f\u00fcr i.i.d. Variablen; i.n.i.d. Variablen; abh\u00e4ngige Variablen, z.B. Martingale, Markov-Prozesse, ...).

Beispiel 2.17.

1. $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ ist wegen $\mathbb{E}(\bar{X}_n) = \mu$ und $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \rightarrow 0$ f\u00fcr $n \rightarrow \infty$ konsistent.

2. $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ und $\widetilde{S}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ sind MSE-konsistent f\u00fcr σ^2 .

3. Mit $g(x) = \sqrt{x}$ folgt, dass

$$S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2} \quad \text{und} \quad \widetilde{S}_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

konsistent sind f\u00fcr σ .

4. S_n^2/\bar{X}_n ist konsistent f\u00fcr σ^2/μ f\u00fcr $\mu > 0$, da mit $\theta = (\mu, \sigma)$ und $g(\theta) = \sigma^2/\mu$ wieder der Stetigkeitssatz benutzt werden kann.

5. $\widehat{\pi}_n$ ist konsistent f\u00fcr π (im Bernoulli-Experiment).

6. $\widehat{\beta}_{KQ}, \widehat{\beta}_{Ridge}$ sind konsistent f\u00fcr β im linearen Modell unter gewissen schwachen Annahmen an \mathbf{X} , siehe Beispiel 2.19.

Asymptotische Normalit\u00e4t

Viele Sch\u00e4tzer (KQ-, Momenten-, ML-, Quasi-ML-, Bayes-Sch\u00e4tzer) sind unter Regularit\u00e4tsannahmen asymptotisch normalverteilt. Informell ausgedr\u00fcckt hei\u00dft das: F\u00fcr gro\u00dfe n ist $\widehat{\theta}_n$ nicht nur approximativ erwartungstreu, sondern zus\u00e4tzlich approximativ normalverteilt, kurz

$$\widehat{\theta}_n \stackrel{a}{\sim} N(\theta, V(\theta))$$

mit (approximativer) Kovarianzmatrix

$$\text{Cov}_\theta(\widehat{\theta}_n) \stackrel{a}{\sim} V(\theta),$$

die durch

$$\widehat{\text{Cov}}_{\theta}(\widehat{\theta}_n) := V(\widehat{\theta}_n)$$

geschätzt wird. In der Diagonalen von $V(\widehat{\theta}_n)$ stehen dann die (geschätzten) Varianzen

$$\widehat{\text{Var}}(\widehat{\theta}_j) = v_{jj}(\widehat{\theta}_n)$$

der Komponenten $\theta_j, j = 1, \dots, p$, von θ .

⇒ "Üblicher" Output statistischer Software ist

$$\underbrace{\widehat{\theta}_j}_{\text{Schätzer}} \quad \underbrace{\widehat{\sigma}_{\widehat{\theta}_j} = \sqrt{v_{jj}(\widehat{\theta})}}_{\text{Standardfehler}} \quad \underbrace{t}_{\text{t-Statistik}} \quad \underbrace{p}_{\text{p-Wert}}$$

Beispiel 2.18. Seien $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} F(x|\theta)$ mit $\mathbb{E}(X_i) = \mu$ und $\text{Var}(X_i) = \sigma^2$. Aber F sei nicht gleich Φ , sondern z.B. die Verteilungsfunktion von $B(\pi)$ oder $Po(\lambda)$. Für \bar{X}_n gilt

$$\mathbb{E}(\bar{X}_n) = \mu \text{ und } \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Aufgrund des zentralen Grenzwertsatzes folgt

$$\bar{X}_n \stackrel{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right),$$

zum Beispiel

$$\bar{X}_n \stackrel{a}{\sim} N\left(\pi, \frac{\pi(1-\pi)}{n}\right) \text{ bei } B(\pi).$$

Genauere Formulierung:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \text{ für } n \rightarrow \infty,$$

im Beispiel also

$$\sqrt{n}(\bar{X}_n - \pi) \xrightarrow{d} N(0, \pi(1-\pi)) \text{ für } n \rightarrow \infty$$

bzw.

$$\left. \begin{array}{l} \frac{\bar{X}_n - \mu}{\sigma} \sqrt{n} \xrightarrow{d} N(0, 1), \\ \frac{\bar{X}_n - \pi}{\sqrt{\pi(1-\pi)}} \sqrt{n} \xrightarrow{d} N(0, 1). \end{array} \right\} \text{ zentraler Grenzwertsatz}$$

Die \sqrt{n} -Normierung ist vor allem bei i.i.d. Stichprobenvariablen geeignet. Für nicht identisch verteilte Stichprobenvariablen wie zum Beispiel $y_1|x_1, \dots, y_n|x_n$ in Regressionssituationen benötigt man bei \sqrt{n} -Normierung Voraussetzungen, die (teilweise) unnötig restriktiv sind. Besser ist dann eine „Matrix-Normierung“ mit Hilfe einer „Wurzel“ $I^{\frac{1}{2}}(\theta)$ der Informationsmatrix.

Einschub: Wurzel einer positiv definiten Matrix

- \mathbf{A} ist positiv definit, wenn \mathbf{A} symmetrisch ist und $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$ für alle $\mathbf{x} \neq \mathbf{0}$ gilt.
- Dann heißt eine Matrix $\mathbf{A}^{\frac{1}{2}}$ (*linke*) Wurzel von $\mathbf{A} \stackrel{\text{def}}{\Leftrightarrow}$

$$\mathbf{A}^{\frac{1}{2}} \underbrace{(\mathbf{A}^{\frac{1}{2}})^\top}_{=\mathbf{A}^{\frac{1}{2}}, \text{ rechte Wurzel}} = \mathbf{A}.$$

Allerdings ist $\mathbf{A}^{\frac{1}{2}}$ nicht eindeutig, da für eine beliebige orthogonale Matrix \mathbf{Q} auch $\mathbf{A}^{\frac{1}{2}} \mathbf{Q}$ eine linke Wurzel ist:

$$\mathbf{A}^{\frac{1}{2}} \mathbf{Q} (\mathbf{A}^{\frac{1}{2}} \mathbf{Q})^\top = \mathbf{A}^{\frac{1}{2}} \underbrace{\mathbf{Q} \mathbf{Q}^\top}_{=\mathbf{I}} \mathbf{A}^{\frac{1}{2}} = \mathbf{A}.$$

- Zwei gebräuchliche Wurzeln sind:
 1. **Symmetrische Wurzel:** Betrachte die Spektralzerlegung von $\mathbf{A} \in \mathbb{R}^{p \times p}$. Mit der Matrix $\mathbf{P} \in \mathbb{R}^{p \times p}$ der orthonormalen Eigenvektoren als Spalten ist

$$\mathbf{P}^\top \mathbf{A} \mathbf{P} = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix},$$

wobei für alle i die $\lambda_i > 0$ die Eigenwerte von \mathbf{A} sind. (Diese Zerlegung ist numerisch aufwändig!) Dann gilt auch

$$\mathbf{A} = \mathbf{P} \mathbf{\Lambda} \mathbf{P}^\top = \underbrace{\mathbf{P} \mathbf{\Lambda}^{\frac{1}{2}}}_{=\mathbf{A}^{\frac{1}{2}}} \underbrace{(\mathbf{\Lambda}^{\frac{1}{2}})^\top \mathbf{P}^\top}_{=\mathbf{A}^{\frac{1}{2}}},$$

und $\mathbf{A}^{\frac{1}{2}}$ heißt *symmetrische Wurzel* von \mathbf{A} .

2. **Cholesky-Wurzel:** Sei $\mathbf{A}^{\frac{1}{2}} := \mathbf{C}$ untere Dreiecksmatrix mit positiven Diagonalelementen und $\mathbf{C} \mathbf{C}^\top = \mathbf{A}$. Dann ist \mathbf{C} die *eindeutig* bestimmte *Cholesky-Wurzel* von \mathbf{A} . (Diese ist numerisch vergleichsweise einfach zu erhalten!)

• Anwendungen in der Statistik

1. Erzeugen von $N_p(\mathbf{0}, \mathbf{\Sigma})$ -verteilten Zufallszahlen ($\mathbf{\Sigma}$ vorgegeben): Falls $\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{I})$, ist einfache Simulation möglich, indem p unabhängige $N(0, 1)$ -verteilte Zufallsvariablen Z_1, \dots, Z_p simuliert werden. Dann gilt auch

$$\mathbf{\Sigma}^{1/2} \mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{\Sigma}^{1/2} \mathbf{I} \mathbf{\Sigma}^{\top/2}) \doteq N(\mathbf{0}, \mathbf{\Sigma}).$$

Also: Berechne Cholesky-Wurzel von $\mathbf{\Sigma}$, ziehe p $N(0, 1)$ -verteilte Zufallsvariablen $\mathbf{Z} = (z_1, \dots, z_p)^\top$, berechne $\mathbf{Y} = \mathbf{\Sigma}^{1/2} \mathbf{Z}$. Dann ist $\mathbf{Y} = (Y_1, \dots, Y_p)^\top$ ein $N_p(\mathbf{0}, \mathbf{\Sigma})$ -verteilter Zufallsvektor.

2. Matrixnormierung bei asymptotischer Normalverteilung:

Beispiel 2.19 (Asymptotische Normalität des KQ-Schätzers im linearen Modell).
Seien $y_1|\mathbf{x}_1, \dots, y_n|\mathbf{x}_n$ unabhängig. Dann gilt

$$\begin{aligned}\mathbb{E}[y_i|\mathbf{x}_i] &= \mathbf{x}_i^\top \boldsymbol{\beta}, \quad \text{Var}(y_i|\mathbf{x}_i) = \sigma^2, \quad i = 1, \dots, n, \\ \Leftrightarrow \mathbf{y}_n &= \mathbf{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \quad \mathbb{E}[\boldsymbol{\varepsilon}_n] = \mathbf{0}, \quad \text{Cov}(\boldsymbol{\varepsilon}_n) = \sigma^2 \mathbf{I}_n.\end{aligned}$$

Der KQ-Schätzer ist

$$\hat{\boldsymbol{\beta}}_n = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{y}_n, \quad \mathbb{E}[\hat{\boldsymbol{\beta}}_n] = \boldsymbol{\beta}, \quad \text{Cov}(\hat{\boldsymbol{\beta}}_n) = \sigma^2 (\mathbf{X}_n^\top \mathbf{X}_n)^{-1}.$$

Die Informationsmatrix unter der Normalverteilungsannahme ist

$$\mathbf{I}(\boldsymbol{\beta}) = \frac{\mathbf{X}_n^\top \mathbf{X}_n}{\sigma^2} = \text{Cov}(\hat{\boldsymbol{\beta}}_n)^{-1}.$$

Zentrale Grenzwertsätze (für unabhängige, nicht identisch verteilte Zufallsvariablen, kurz: i.n.i.d.) liefern unter geeigneten Voraussetzungen (informell):

$$\hat{\boldsymbol{\beta}}_n \stackrel{a}{\sim} N_p(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}_n^\top \mathbf{X}_n)^{-1}).$$

Genauere Formulierungen nehmen an, dass

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n =: \mathbf{A} > 0 \tag{2.1}$$

existiert (also: $\mathbf{X}_n^\top \mathbf{X}_n \approx n\mathbf{A} \Leftrightarrow (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \approx \mathbf{A}^{-1}/n$ für große n). Anwendung des (multivariaten) zentralen Grenzwertsatzes liefert dann:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N_p(0, \sigma^2 \mathbf{A}^{-1})$$

bzw.

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n &\stackrel{a}{\sim} N_p(\boldsymbol{\beta}, \sigma^2 \mathbf{A}^{-1}/n) \\ \hat{\boldsymbol{\beta}}_n &\stackrel{a}{\sim} N_p(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}_n^\top \mathbf{X}_n)^{-1}).\end{aligned}$$

Die Annahme (2.1) ist zum Beispiel erfüllt, wenn \mathbf{x}_i , $i = 1, \dots, n$, i.i.d. Realisierungen stochastischer Kovariablen $\mathbf{X} = (X_1, \dots, X_p)^\top$ sind. Dann gilt nach dem Gesetz der großen Zahlen:

$$\frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \xrightarrow{n \rightarrow \infty} \mathbb{E}[\mathbf{X} \mathbf{X}^\top] =: \mathbf{A}.$$

Typischerweise ist die Annahme (2.1) nicht erfüllt bei deterministischen Regressoren mit Trend. Das einfachste Beispiel hierfür ist ein linearer Trend: $x_i = i$ für $i = 1, \dots, n$ und $y_i = \beta_1 i + \varepsilon_i$. Dann ist

$$\mathbf{X}_n^\top \mathbf{X}_n = \sum_{i=1}^n i^2$$

und daher

$$\frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n = \frac{\sum_{i=1}^n i^2}{n} \geq n \xrightarrow{n \rightarrow \infty} \infty.$$

In diesem Fall ist eine andere Normierung nötig, zum Beispiel eine Matrixnormierung mit

$$\mathbf{C}_n = (\mathbf{X}_n^\top \mathbf{X}_n).$$

Dann lässt sich die asymptotische Normalität des KQ-Schätzers

$$\mathbf{C}_n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N_p(\mathbf{0}, \sigma^2 \mathbf{I})$$

bzw.

$$\tilde{\mathbf{C}}_n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) := \frac{\mathbf{C}_n^{1/2}}{\sigma}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I})$$

unter folgenden, sehr schwachen Bedingungen zeigen:

(D) Divergenzbedingung: Für $n \rightarrow \infty$ gilt:

$$(\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \rightarrow \mathbf{0}.$$

Eine äquivalente Forderung ist:

$$\lambda_{\min}(\mathbf{X}_n^\top \mathbf{X}_n) \rightarrow \infty,$$

wobei λ_{\min} den kleinsten Eigenwert von $\mathbf{X}_n^\top \mathbf{X}_n$ bezeichnet. Die Divergenzbedingung sichert, dass die „Informationsmatrix“

$$\mathbf{X}_n^\top \mathbf{X}_n = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$$

für $n \rightarrow \infty$ gegen ∞ divergiert, die Information mit $n \rightarrow \infty$ also laufend wächst.

Es gilt: (D) ist hinreichend und notwendig für die (schwache und starke) Konsistenz des KQ-Schätzers $\hat{\boldsymbol{\beta}}_n$.

(N) Normalitätsbedingung:

$$\max_{i=1, \dots, n} \mathbf{x}_i^\top (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{x}_i \rightarrow 0 \quad \text{für } n \rightarrow \infty$$

(N) sichert, dass die Information jeder Beobachtung i asymptotisch gegenüber der Gesamtinformation $\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$ vernachlässigbar ist.

Unter (D) und (N) gilt

$$(\mathbf{X}_n^\top \mathbf{X}_n)^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N_p(\mathbf{0}, \sigma^2 \mathbf{I})$$

(Beweis mit Grenzwertsätzen für unabhängige, nicht identisch verteilte Zufallsvariablen), d.h. für praktische Zwecke:

$$\hat{\boldsymbol{\beta}}_n \stackrel{a}{\sim} N_p(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}_n^\top \mathbf{X}_n)^{-1})$$

für genügend großen Stichprobenumfang n . Dabei darf zusätzlich σ^2 durch einen konsistenten Schätzer $\hat{\sigma}^2$ ersetzt werden.

Definition 2.26 (Asymptotische Normalität).

1. Mit \sqrt{n} -Normierung: $\hat{\theta}_n$ heißt asymptotisch normalverteilt für $\theta \stackrel{\text{def}}{\Leftrightarrow}$

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta)) \quad \text{für } n \rightarrow \infty$$

mit nicht-negativ definiten (in der Regel positiv definiten) asymptotischer Kovarianzmatrix $V(\theta)$.

2. Mit Matrix-Normierung: $\hat{\theta}_n$ heißt asymptotisch normalverteilt für $\theta \stackrel{\text{def}}{\Leftrightarrow}$ es existiert eine Folge von Matrizen \mathbf{A}_n mit $\lambda_{\min}(\mathbf{A}_n) \rightarrow \infty$, so dass

$$\mathbf{A}_n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta)).$$

Bemerkung.

1. Praxisformulierung:

$$\hat{\theta}_n \stackrel{a}{\sim} N(\theta, V(\theta)/n)$$

bzw.

$$\hat{\theta}_n \stackrel{a}{\sim} N(\theta, (\mathbf{A}_n^{1/2})^{-1}V(\theta)(\mathbf{A}_n^{1/2})^{-\top}).$$

Dabei darf θ in $V(\theta)$ durch $\hat{\theta}_n$ ersetzt werden.

2. Oft: $V(\theta) = \mathbf{I}$ möglich, wenn geeignet normiert wird, zum Beispiel bei ML-Schätzung.

Beispiel 2.20. Seien X_1, \dots, X_n i.i.d. Zufallsvariablen mit (bekanntem) Erwartungswert μ und Varianz σ^2 .

$$S_\mu^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

ist asymptotisch normal für σ^2 mit $V(\theta) = \mu_4 - \sigma^4$, $\mu_4 = \mathbb{E}[(X_i - \mu)^4] < \infty$. S_μ^2 ist erwartungstreu. Für die Varianz erhält man:

$$\begin{aligned} \text{Var}(S_\mu^2) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right) \\ &= \frac{1}{n^2} \cdot n \cdot \text{Var}[(X_1 - \mu)^2] \\ &= \frac{1}{n} \left(\mathbb{E}[(X_1 - \mu)^4] - (\mathbb{E}[(X_1 - \mu)^2])^2 \right) \\ &= \frac{1}{n} (\mu_4 - \sigma^4). \end{aligned}$$

Es liegen die Voraussetzungen zur Anwendung des zentralen Grenzwertsatzes vor. Aus ihm folgt:

$$S_\mu^2 \stackrel{a}{\sim} N(\sigma^2, (\mu_4 - \sigma^4)/n) \quad \text{bzw.} \quad \sqrt{n}(S_\mu^2 - \sigma^2) \xrightarrow{d} N(0, \mu_4 - \sigma^4).$$

Die Delta-Methode

$\hat{\theta}_n$ sei asymptotisch normalverteilter Schätzer für θ .

Frage: Wie ist für eine gegebene Abbildung

$$h : \mathbb{R}^p \rightarrow \mathbb{R}^k, k \leq p$$

der Schätzer $h(\hat{\theta})$ für $h(\theta)$ verteilt?

Satz 2.27 (Delta-Methode). *Sei h wie oben.*

1. θ skalar: Für alle θ , für die h stetig differenzierbar ist mit $h'(\theta) \neq 0$, gilt:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta)) \Rightarrow \sqrt{n}(h(\hat{\theta}_n) - h(\theta)) \xrightarrow{d} N(0, [h'(\theta)]^2 V(\theta))$$

2. θ vektoriell: Sei

$$\theta = (\theta_1, \dots, \theta_p)^\top \mapsto h(\theta) = (h_1(\theta), \dots, h_k(\theta))^\top$$

mit Funktionalmatrix

$$(H(\theta))_{ij} = \frac{\partial h_i(\theta)}{\partial \theta_j}$$

mit vollem Rang. Für alle θ , für die $h(\theta)$ komponentenweise stetig partiell differenzierbar ist und jede Zeile von $H(\theta)$ ungleich dem Nullvektor ist, gilt:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta)) \Rightarrow \sqrt{n}(h(\hat{\theta}_n) - h(\theta)) \xrightarrow{d} N(0, H(\theta)V(\theta)H(\theta)^\top).$$

Beweisskizze für skalares θ . Taylorentwicklung von $h(\hat{\theta}_n)$ um θ liefert:

$$h(\hat{\theta}_n) = h(\theta) + (\hat{\theta}_n - \theta)h'(\theta) + o(\hat{\theta}_n - \theta)^2.$$

Dabei ist für eine Folge von Zufallsvariablen X_n

$$X_n = o(a_n) \quad \text{falls } X_n/a_n \xrightarrow{P} 0 \text{ für } n \rightarrow \infty.$$

Also:

$$h(\hat{\theta}_n) \approx h(\theta) + (\hat{\theta}_n - \theta)h'(\theta)$$

bzw.

$$\sqrt{n}(h(\hat{\theta}_n) - h(\theta)) \approx \sqrt{n}(\hat{\theta}_n - \theta)h'(\theta)$$

Aus $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta))$ folgt dann, dass $\sqrt{n}(h(\hat{\theta}_n) - h(\theta)) \xrightarrow{d} N(0, h'(\theta)^2 V(\theta))$. \square

Asymptotische Cramer-Rao Schranke und asymptotische Effizienz

Seien $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f(x|\theta)$ und

$$i(\theta) = -\mathbb{E} \left[\frac{\partial^2 \log f(x|\theta)}{\partial \theta \partial \theta^\top} \right]$$

die erwartete Fisher-Information einer Beobachtung X_i . Die Information der gesamten Stichprobe X_1, \dots, X_n ist dann

$$\mathbf{I}(\theta) = n \cdot i(\theta).$$

Satz 2.28 (Asymptotische Cramer-Rao Ungleichung). *Unter Fisher-Regularität sowie leichten Zusatzannahmen gilt:*

1. Aus $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta))$ folgt $V(\theta) \geq i^{-1}(\theta)$.
2. Aus $\sqrt{n}(h(\hat{\theta}_n) - h(\theta)) \xrightarrow{d} N(0, D(\theta))$ folgt

$$D(\theta) \geq H(\theta) i^{-1}(\theta) H(\theta)^\top$$

mit "≥" die Löwner-Ordnung (und den Bezeichnungen aus der Delta-Regel, Satz 2.27).

Definition 2.29 (Beste asymptotisch normaler (BAN)-Schätzer). $\hat{\theta}_n$ heißt BAN-Schätzer, falls in 1. oben gilt:

$$V(\theta) = i^{-1}(\theta).$$

Mit der Delta-Regel folgt unmittelbar:

Satz 2.30 (Transformation von BAN-Schätzern). *Ist $\hat{\theta}_n$ BAN-Schätzer für θ , so ist $h(\hat{\theta}_n)$ BAN-Schätzer für $h(\theta)$.*

Bemerkung. *Das Konzept der asymptotischen Effizienz lässt sich auf die Matrix-Normierung übertragen: $\hat{\theta}$ ist BAN-Schätzer für θ genau dann, wenn*

$$\mathbf{I}^{1/2}(\theta)(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I)$$

bzw. $\hat{\theta}_n \stackrel{a}{\sim} N(\theta, \mathbf{I}^{-1}(\hat{\theta}_n))$, mit $\mathbf{I}^{1/2}(\theta)$ Wurzel der Fisher-Information $\mathbf{I}(\theta)$ der Stichprobe X_1, \dots, X_n . Anstelle der erwarteten kann auch die beobachtete Fisher-Information $\mathbf{J}(\theta)$ verwendet werden.