week11a: Kapitel 7: Effizienz von Schätzern, Teil 1

Die Maximum-Likelihood-Schätzer für den Mittelwert und die Varianz von normalverteilten Zufallszahlen aus dem Beispiel 1 waren gegeben durch

$$\hat{\mu}(x_1, \cdots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\widehat{\sigma^2}(x_1, \cdots, x_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2}$$

mit $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}$. Das $\hat{\mu}$ war erwartungstreu,

$$\mathsf{E}\big[\,\hat{\mu}\,\big] = \mu \tag{3}$$

aber das $\widehat{\sigma^2}$ war nur im Limes $n \to \infty$ erwartungstreu. Deshalb hatten wir den erwartungstreuen Schätzer $\widehat{s^2}$ mit Vorfaktor $1/(n-1)$ definiert, das war also

$$\widehat{s^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{4}$$

und wir haben dann die Identität

$$\mathsf{E}\big[\,\widehat{s^2}\,\big] = \sigma^2 \tag{5}$$

Wir wollen jetzt zeigen, dass diese Schätzer effizient sind. Das heisst, dass sie innerhalb der Menge aller erwartungstreuen Schätzer eine minimale Varianz besitzen. Das ist jetzt schon eine etwas nichttrivialere Sache, weil das einzige, was man hier fordert, Erwartungstreue ist, ansonsten können die Schätzer, mit denen man das $\hat{\mu}$ und das $\widehat{s^2}$ vergleichen möchte, völlig beliebige Funktionen sein. Also definieren wir diese Menge von Schätzern, mit denen wir das $\hat{\mu}$ und das $\widehat{s^2}$ vergleichen wollen (etwa $\mathcal{M}$ für Mittelwert und $\mathcal{V}$ für Varianz):

$$\mathcal{M} := \Big\{\, \tilde{\mu} : \mathbb{R}^n \to \mathbb{R} \;\Big|\; \mathsf{E}[\,\tilde{\mu}(x_1, \cdots, x_n)\,] = \mu \,\Big\} \tag{6}$$

$$\mathcal{V} := \Big\{\, \widetilde{\sigma^2} : \mathbb{R}^n \to \mathbb{R} \;\Big|\; \mathsf{E}[\,\widetilde{\sigma^2}(x_1, \cdots, x_n)\,] = \sigma^2 \,\Big\} \tag{7}$$

wobei die Erwartungswerte gemäss

$$\mathsf{E}\big[\,F(x_1, \cdots, x_n)\,\big] = \int_{\mathbb{R}^n} F(x_1, \cdots, x_n) \prod_{i=1}^{n} \Big\{\, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \,\frac{dx_i}{\sqrt{2\pi\sigma^2}} \,\Big\}$$

zu berechnen sind. Dann gilt das folgende

Theorem 7.1.1: a) Der Maximum-Likelihood-Schätzer $\hat{\mu}$ aus Gleichung (1) ist effizient. Das heisst, für jeden Schätzer $\tilde{\mu}$ aus $\mathcal{M}$ gilt:

$$\mathsf{V}\big[\,\tilde{\mu}\,\big] \;\geq\; \frac{\sigma^2}{n} \;=\; \mathsf{V}\big[\,\hat{\mu}\,\big] \qquad\qquad \forall\,\tilde{\mu}\in\mathcal{M} \tag{8}$$

b) Die erwartungstreue Version (4) des Maximum-Likelihood-Schätzers für die Varianz ist asymptotisch effizient. Das heisst genauer, für jeden Schätzer $\widetilde{\sigma^2}$ aus $\mathcal{V}$ gilt:

$$\mathsf{V}\big[\,\widetilde{\sigma^2}\,\big] \;\geq\; \frac{2\sigma^4}{n} \;=\; \frac{n-1}{n}\,\mathsf{V}\big[\,\widehat{s^2}\,\big] \;\overset{n\to\infty}{\approx}\; \mathsf{V}\big[\,\widehat{s^2}\,\big] \qquad\qquad \forall\,\widetilde{\sigma^2}\in\mathcal{V} \tag{9}$$

Beweis: ..machen wir gleich mit der Cramer-Rao Abschätzung.

Der Beweis ergibt sich als Folgerung aus der sogenannten Cramer-Rao Abschätzung. Das ist eine untere Schranke für die Varianz von erwartungstreuen Schätzern, und das bemerkenswerte an dieser Schranke ist, dass man sie für ein sehr allgemeines Setting hinschreiben kann. Wir müssen nicht unbedingt eine unabhängige Folge von Zufallszahlen haben wie in den Beispielen 1 und 2, sondern die Cramer-Rao Abschätzung funktioniert auch noch für das Beispiel 3. Das genaue Setting ist das folgende:

Wir haben Zufallszahlen oder zufällige Grössen $x_1, x_2, \cdots, x_n$, die von einer Wahrscheinlichkeitsverteilung

$$p_\theta(x) \;=\; p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n) \tag{10}$$

generiert worden sind. Es gelte also

$$\int_{\mathbb{R}^n} p_\theta(x)\; d^n x \;=\; \int_{\mathbb{R}^n} p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n)\; dx_1 dx_2\cdots dx_n \;=\; 1 \tag{11}$$

Die theta's $\theta_1,\cdots,\theta_m$ sind die Modellparameter. Wenn wir das $p_\theta(x)$ nur als Funktion von θ auffassen, weil wir für die $x_1,\cdots,x_n$ die uns gegebenen realisierten Daten einsetzen, dann ist das genau die Likelihood-Funktion,

$$L(\theta) \;=\; L(\theta_1,\cdots,\theta_m) \;=\; p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n) \tag{12}$$

Schreiben wir das $p_\theta(x)$ für unsere 3 Beispiele noch einmal hin:

Beispiel 1: Für Beispiel 1 haben wir

$$p_\theta(x) \;=\; p_{\mu,\sigma^2}(x_1,\cdots,x_n) \;=\; \prod_{i=1}^{n}\,\Big\{\,\tfrac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\,\Big\} \tag{13}$$

Beispiel 2: Für Beispiel 2 haben wir

$$p_\theta(x) \;=\; p_\lambda(x_1,\cdots,x_n) \;=\; \prod_{i=1}^{n} \left\{ \tfrac{\lambda^{x_i}}{x_i!}\, e^{-\lambda} \right\} \tag{14}$$

mit $x_i \in \{0,1,2,\cdots\}$, die x_i sind hier also diskret. In einem solchen Fall wollen wir in der Gleichung (11) das Integral dann immer als eine Summe interpretieren, also hier wäre das dann

$$\sum_{x_1,\cdots,x_n=0}^{\infty} p_\lambda(x_1,\cdots,x_n) \;=\; 1$$

mit $\theta = \theta_1 = \lambda$ als einzigen Modellparameter.

Beispiel 3: Und für das Beispiel 3, den zeitdiskreten Ornstein-Uhlenbeck Prozess, hatten wir (mit $x_k := x_{t_k}$)

$$p_\theta(x) \;=\; p_{\alpha,\beta,\eta}(x_1,\cdots,x_n) \;=\; \prod_{k=1}^{n} \left\{ \tfrac{1}{\sqrt{2\pi\eta^2}}\, e^{-\frac{(\,x_k-[\alpha x_{k-1}+\beta]\,)^2}{2\eta^2}} \right\} \tag{15}$$

mit den Modellparametern α, β, η oder den ursprünglichen Modellparametern

$$\begin{aligned} \kappa &= \tfrac{1-\alpha}{\Delta t} \\ \bar{x} &= \tfrac{\beta}{1-\alpha} \\ \sigma^2 &= \tfrac{\eta^2}{\Delta t} \end{aligned}$$

Auch in diesem Fall gilt die Gleichung (11).

In allen 3 Beispielen können wir Ableitungen nach den Modellparametern mit den Integralen oder den Summen über die x_i vertauschen, es gilt also

$$\begin{aligned} 0 \;=\; \tfrac{\partial}{\partial\theta_k} \underbrace{\int_{\mathbb{R}^n} p_\theta(x)\, d^n x}_{=\,1} \;&=\; \tfrac{\partial}{\partial\theta_k} \int_{\mathbb{R}^n} p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n)\, dx_1 dx_2 \cdots dx_n \\ &=\; \int_{\mathbb{R}^n} \tfrac{\partial p_{\theta_1,\cdots,\theta_m}}{\partial\theta_k}(x_1,\cdots,x_n)\, dx_1 dx_2 \cdots dx_n \end{aligned} \tag{16}$$

Dass man die Ableitung mit dem Integral vertauschen kann, ist jetzt nicht immer ganz selbstverständlich, da wir etwa bei gleichverteilten Zufallszahlen ja eine Rechteck-Funktion als W'keitsdichte haben, und die ist im klassischen Sinne nicht differenzierbar (nur im Distributions-Sinn), und solche Fälle wollen wir hier mal ausschliessen. Schreiben wir jetzt die Cramer-Rao Abschätzung hin:

Theorem 7.1.2 (Cramer-Rao Lower Bound): Gegeben sei eine Wahrscheinlichkeitskeits-verteilung $p_\theta(x) = p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n)$ mit

$$\int_{\mathbb{R}^n} p_\theta(x)\, d^n x \;=\; \int_{\mathbb{R}^n} p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n)\, dx_1 dx_2 \cdots dx_n \;=\; 1$$

Für $k = 1, \cdots, m$ seien

$$\hat{\theta} \;=\; \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_m \end{pmatrix} \;:\; \mathbb{R}^n \;\to\; \mathbb{R}^m$$

erwartungstreue Schätzer, d.h. es gilt

$$\mathsf{E}[\,\hat{\theta}_k\,] \;:=\; \int_{\mathbb{R}^n} \hat{\theta}_k(x)\; p_\theta(x)\; d^n x \;=\; \theta_k$$

Es bezeichne

$$\mathsf{Cov}(\hat{\theta}) \;:=\; \Big(\; \mathsf{Cov}[\hat{\theta}_k, \hat{\theta}_\ell] \;\Big)_{k,\ell=1,\ldots,m} \;\in\; \mathbb{R}^{m\times m}$$

die Covarianz-Matrix von $\hat{\theta}$, insbesondere ist dann also

$$\mathsf{V}[\,\hat{\theta}_k\,] \;=\; \mathsf{Cov}(\hat{\theta})_{k,k}$$

die Varianz von $\hat{\theta}_k$. Wir definieren die sogenannte Fisher-Informations-Matrix $I(\theta)$ durch

$$I(\theta) \;:=\; \Big(\; -\mathsf{E}\Big[\; \tfrac{\partial^2 \log p_\theta}{\partial\theta_k \partial\theta_\ell} \;\Big] \;\Big)_{k,\ell=1,\ldots,m} \;\in\; \mathbb{R}^{m\times m}$$

mit

$$\mathsf{E}\Big[\; \tfrac{\partial^2 \log p_\theta}{\partial\theta_k \partial\theta_\ell} \;\Big] \;=\; \int_{\mathbb{R}^n} \tfrac{\partial^2 \log p_\theta}{\partial\theta_k \partial\theta_\ell}(x_1, \cdots, x_n)\; p_\theta(x_1, \cdots, x_n)\; d^n x$$

und $I^{-1}(\theta)$ sei das Inverse von $I(\theta)$. Dann gilt:

$$\big\langle\, v\,, \mathsf{Cov}(\hat{\theta})\, v \,\big\rangle \;\geq\; \big\langle\, v\,, I^{-1}(\theta)\, v \,\big\rangle \qquad \forall v \in \mathbb{R}^m \tag{17}$$

Insbesondere gilt also:

$$\mathsf{V}[\,\hat{\theta}_k\,] \;\geq\; \big[\, I^{-1}(\theta) \,\big]_{k,k} \tag{18}$$

für jeden erwartungstreuen Schätzer $\hat{\theta}_k : \mathbb{R}^n \to \mathbb{R}$.

Bevor wir die Cramer-Rao Abschätzung in der nächsten Veranstaltung beweisen, wollen wir die Abschätzung (18) konkret für das Beispiel 1 hinschreiben und damit das Theorem 7.1.1 vom Anfang der Vorlesung beweisen:

Beweis Theorem 7.1.1: Wir haben mit $\nu := \sigma^2$

$$p_\theta(x) \;=\; p_{\mu,\nu}(x_1, \cdots, x_n) \;=\; \prod_{i=1}^{n} \Big\{\; \tfrac{1}{\sqrt{2\pi\nu}}\, e^{-\frac{(x_i-\mu)^2}{2\nu}} \;\Big\}$$

$$\log p_\theta(x) \;=\; \log p_{\mu,\nu}(x_1, \cdots, x_n) \;=\; -\tfrac{n}{2}\log(2\pi\nu) \;-\; \sum_{i=1}^{n} \tfrac{(x_i-\mu)^2}{2\nu}$$

und damit

$$\begin{aligned}\frac{\partial \log p_{\mu,\nu}}{\partial \mu} &= \sum_{i=1}^{n} \frac{x_i-\mu}{\nu} \\ \frac{\partial \log p_{\mu,\nu}}{\partial \nu} &= -\frac{n}{2\nu} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\nu^2}\end{aligned}$$

und

$$\begin{aligned}\frac{\partial^2 \log p_{\mu,\nu}}{\partial \mu^2} &= \sum_{i=1}^{n} \frac{-1}{\nu} = -\frac{n}{\nu} \\ \frac{\partial^2 \log p_{\mu,\nu}}{\partial \nu \partial \mu} &= -\sum_{i=1}^{n} \frac{x_i-\mu}{\nu^2} \\ \frac{\partial^2 \log p_{\mu,\nu}}{\partial \nu^2} &= \frac{n}{2\nu^2} - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\nu^3}\end{aligned}$$

Wegen

$$\begin{aligned}\mathsf{E}[\, x_i - \mu \,] &= 0 \\ \mathsf{E}[\, (x_i-\mu)^2 \,] &= \sigma^2 = \nu\end{aligned}$$

bekommen wir dann die folgenden Erwartungswerte:

$$\begin{aligned}\mathsf{E}\left[\, \frac{\partial^2 \log p_{\mu,\nu}}{\partial \mu^2} \,\right] &= -\frac{n}{\nu} \\ \mathsf{E}\left[\, \frac{\partial^2 \log p_{\mu,\nu}}{\partial \nu \partial \mu} \,\right] &= 0 \\ \mathsf{E}\left[\, \frac{\partial^2 \log p_{\mu,\nu}}{\partial \nu^2} \,\right] &= \frac{n}{2\nu^2} - \sum_{i=1}^{n} \frac{\nu}{\nu^3} = \frac{n}{2\nu^2} - \frac{n}{\nu^2} = -\frac{n}{2\nu^2}\end{aligned}$$

Also,

$$I(\theta) = I(\mu,\nu) = \begin{pmatrix} \frac{n}{\nu} & 0 \\ 0 & \frac{n}{2\nu^2} \end{pmatrix}$$

und damit

$$I^{-1}(\mu,\nu) = \begin{pmatrix} \frac{\nu}{n} & 0 \\ 0 & \frac{2\nu^2}{n} \end{pmatrix}$$

Die Cramer-Rao Abschätzung liefert also in diesem Fall

$$\begin{aligned}\mathsf{V}[\, \tilde{\mu} \,] &\geq I^{-1}(\mu,\nu)_{1,1} = \frac{\nu}{n} = \frac{\sigma^2}{n} \\ \mathsf{V}[\, \tilde{\nu} \,] &\geq I^{-1}(\mu,\nu)_{2,2} = \frac{2\nu^2}{n} = \frac{2\sigma^4}{n}\end{aligned}$$

für beliebige erwartungstreue Schätzer $\tilde{\mu}$ und $\tilde{\nu} = \widetilde{\sigma^2}$ und damit haben wir das Theorem 7.1.1 bewiesen. ■